MULTIMODAL DEEP LEARNING FOR DECODING VISUAL AND NEURAL REPRESENTATION

xianglin zheng, b.sc. (computer science)



University of
South Australia

Research Thesis for the Degree of Doctor of Philosophy

STEM
University of South Australia

December 18, 2024

Xianglin Zheng: *Multimodal Deep Learning for Decoding Visual and Neural Representation,* Research Thesis for the Degree of Doctor of Philosophy, © September 2024

supervisors:
A/Prof. Mark Billinghurst, Dr. Zehong (Jimmy) Cao and A/Prof. Lin Liu

institute:
Australian Research Centre for Interactive and Virtual Environments
University of South Australia
Mawson Lakes, SA, Australia

DECLARATION

I declare that this thesis presents work carried out by myself and that it has not been submitted, in whole or in part, in any previous application for a degree unless stated otherwise. This work does not contain any material previously published or written by another person except when reference is made in the text. All substantive contributions by others to the work presented, including jointly authored publications, are clearly declared and acknowledged.

Mawson Lakes, SA, Australia, September 2024



Xianglin Zheng

ABSTRACT

Multimodal Deep Learning for Decoding Visual and Neural Representation
Xianglin Zheng

Multimodal deep learning aims to model and integrate diverse data types, such as images and neural signals, to decode complex representations in tasks like image captioning and object classification. Traditional models often rely on large labelled datasets and focus on single modalities, which limits their ability to handle real-world complexities. This thesis explores deep learning and multimodal learning techniques to bridge the gap in visual and neural data processing, with a focus on novel approaches to image captioning, modality mapping, and visual neural decoding, as demonstrated through three distinct works.

First, a new framework for unsupervised image captioning is introduced, addressing the challenge of visual-text misalignment. By leveraging adversarial word-level modelling, this method captures object relationships within images to improve caption accuracy, achieving both global and local alignment between modalities.The second work focuses on mapping between images and EEG signals for object classification. By decoding EEG data associated with visual stimuli, this approach enhances object classification through cognitive signals, incorporating temporal features and effective connectivity to improve classification performance.Lastly, a trimodal contrastive learning model is proposed to decode visual neural representations by integrating visual, linguistic, and neural data. This approach provides a deeper understanding of how the brain processes visual information, enabling robust decoding of neural signals for tasks like object recognition.

Together, these contributions advance the field of multimodal deep learning by providing innovative methods for deep learning and modality integration, pushing the boundaries of image captioning, EEG-based classification, and brain-inspired computing.

PUBLICATIONS

0.1 thesis publications

The following publications have been generated from this thesis work:

Zheng, X., Cao, Z., & Bai, Q. (2020). An evoked potential-guided deep learning brain representation for visual classification. In Neural Information Processing: 27th International Conference, ICONIP 2020, Bangkok, Thailand, November 18–22, 2020, Proceedings, Part V 27 (pp. 54-61). Springer International Publishing.

The following publications are in progress:

Xianglin Zheng, Zehong Cao, Quan Bai, and Shiming Chen. 'Visual Relation Alignment by Adversarial Word-Level Modelling for Unsupervised Image Captioning'. To be submitted.

Xianglin Zheng, Zehong Cao, Lin Liu, and Mark Billinghurst. 'CCN-IG: Causal Convolutional Network-Inference Graph for Effective Brain Networks Modelling from EEG Data'. To be submitted.

Xianglin Zheng and Mark Billinghurst. 'Tri-model learning for decoding visual neural representation'. To be submitted.

I propose to consider the question,
'Can machines think?'

— Turing, A. M. [183]

ACKNOWLEDGMENTS

First and foremost, I would like to express my deepest gratitude to my supervisors, especially Professor Mark Billinghurst, for their invaluable guidance, support, and encouragement throughout my academic journey. Their expertise and patience have been instrumental in helping me overcome challenges and achieve my research goals.

I would also like to express my heartfelt thanks to my family, especially my parents and my brother, whose unconditional love and support have been my greatest source of strength.

To my friends, I owe a sincere thank you-to those in China, who supported me through the difficult times during the pandemic when I couldn't enter Australia, and to my friends in Australia, who have provided me with care and assistance since my arrival. Your kindness has made this experience much more manageable and memorable.

Lastly, I want to thank myself for not giving up. Despite the countless obstacles along the way-especially the loneliness and uncertainty during the difficult times of the pandemic-I kept going. This thesis is a testament to my perseverance, resilience, and unwavering determination.

CONTENTS

0.1 Thesis Publications vii
1 introduction 1
1.1 Motivation 1
1.1.1 Multimodal Learning and Visual Representation 1
1.1.2 Decode Visual Neural Representation 3
1.2 The Research Gap 6
1.3 Research Questions 8
1.4 Research Objectives 10
1.5 Chapter Overview 11
2 literature review 13
2.1 Visual Representation 13
2.2 Deep Learning and Multimodal learning Methods 18
2.2.1 Deep Learning 18
2.2.2 Reinforcement Learning (RL) 29
2.2.3 Multimodal learning 31
2.3 Image Captioning 33
2.3.1 Supervised Image Captioning 35
2.3.2 Unsupervised Image Captioning 36
2.4 Brain-inspired Visual Representations 37
2.4.1 Brain and Artificial Intelligence 38
2.4.2 Physiological Signals-EEG 40
2.4.3 Visual Neural Representations 41
2.5 Decode Visual Neural Representation with EEGs 42
2.5.1 Preprocessing 44
2.5.2 Feature Extraction 45
2.5.3 EEG-based Visual Classification Models 47
3 adversarial modality mapping between image and text 51
3.1 Introduction 51
3.2 Method 53
3.2.1 The adversarial network 54
3.2.2 The adversarial reward 55
3.2.3 The visual attention reward 55
3.2.4 Optimisation 58
3.3 Experiments and results 58
3.3.1 Experimental setup 59
3.3.2 Main results 61
3.3.3 Ablative analysis 62
3.3.4 Analysis with different visual relation detectors 63
3.3.5 Qualitative analysis 63
3.4 Conclusions 65
4 modality mapping between image and eeg 67

4.1 Evoked Potential-Guided Deep Learning for Visual Classification 67
4.1.1 ERP-LSTM Framework 67
4.1.2 The Experiment 69
4.1.3 Results 71
4.1.4 Conclusion 72
4.2 Causal Inference Graph for EEG-based Brain Networks 73
4.2.1 Introduction 73
4.2.2 Method 75
4.2.3 Experiments 80
4.2.4 Results 82
4.2.5 Conclusion 85
5 trimodal learning of image, text and eeg 89
5.1 Introduction 89
5.2 Method 92
5.2.1 Single-Modal Encoders 92
5.2.2 Trimodal Encoder 93
5.2.3 Multi-objective Modelling 93
5.2.4 Training 95
5.3 Experiments 95
5.3.1 **Datasets** 95
5.3.2 Implementation Details 96
5.4 Results 97
5.4.1 Visual Classification performance 97
5.5 Conclusion 98
6 discussion 101
6.1 Study 1 101
6.2 Study 2 103
6.3 Study 3 105
6.4 Study 4 107
7 conclusions and directions for future research 111

i appendix 113
a appendix 115
a.1 Statement of Authorship Form 115

bibliography 117

LIST OF FIGURES

Figure 1.1 Encoder-Decoder framework for image captioning. The process begins with a CNN acting as the encoder, which processes the input image to extract a feature vector representing the key visual features. This feature vector is then fed into an RNN decoder. The decoder starts with a special <start> token and generates a sequence of words to describe the image. Each RNN cell predicts the next word in the sequence, eventually leading to a complete caption, such as "A person riding a motorcycle on a dirt road." 3

Figure 1.2 Illustrate the task of visual encoding and decoding of neural representations in the brain. When the human brain perceives a stimulus, it encodes the attributes of the stimulus, activating specific neurons and creating distinct patterns of brain activity. Once a sufficient amount of brain activity data in response to stimuli is collected, it is possible to train a mapping function between brain activation and visual stimuli. Since each image corresponds to a specific category, the trained decoder can then be used to translate physiological signals into image categories. 5

Figure 1.3 The relationship of two objects is crucial for image captioning 7

Figure 2.1 The illustration of CNNs process for visual representation. 14

Figure 2.2 Illustration of AlexNet architecture (adapted from [97]). 15

Figure 2.3 Illustration of ResNet architecture (adapted from [63]). 16

Figure 2.4 Illustration of Fast R-CNN and Faster R-CNN architecture (adapted from [115]). 17

Figure 2.5 A general framework of a basic RNN cell (left) and an LSTM (right) memory cell [36]. 20

Figure 2.6 A general framework of GAN [51] 22

Figure 2.7 A Multi-layer Graph Convolutional Network. Source: https://tkipf.github.io/graph-convolutional-networks 23

Figure 2.8 An overview of the Transformer [187]. 26

Figure 2.9 The illustration of self-attention mechanism in the Transformer [187]. 27
Figure 2.10 A general framework of RL [175]. 29
Figure 2.11 Image Captioning Examples 34
Figure 2.12 The overall encoder and decoder framework propsed in [189]. 35
Figure 2.13 An image captioning model in detail, where a CNN encodes the image into features, and LSTM networks serve as the decoder to generate a descriptive caption for the image word by word. 36
Figure 2.14 Hierarchical visual processing in the brainis is hierarchical–one layer feeds into the next, computing progressively more complex features. This is the inspiration for the “layered” design of feed-forward neural networks. Source: https://doi.org/10.6084/m9.figshare.106794.v3 39
Figure 2.15 Examples of brain signals evoked by visual stimuli of two different ImageNet object classes [166]. This study utilised 72 images, with twelve selected from each of six categories: human body, human face, animal body, animal face, fruits and vegetables, and inanimate objects. 43
Figure 2.16 Examples of visual stimulus used in [85]. 43
Figure 2.17 The averaging of several EEG segments into the ERP waveform.. 46
Figure 3.1 An example of the four possible visual relations (A) "ride" (B) "hold" (C) "behind" and (D) "next to" between the same pair of objects: a person and bike. 52
Figure 3.2 The developed adversarial and word-level visual relation alignment model for the unsupervised image captioning task. It consists of a sentence generator, a discriminator, and a visual relation detector. The GAN is formalised as an RL process that can jointly update the parameters by leveraging the visual attention reward estimated from the detected relations. 53
Figure 3.3 Illustration of the visual attention model. Each component of the detected visual relation from the visual relation detector (D_r) is subsequently fed into the model. We can extract the cumulative maximum visual attention reward by mapping the component word to the reference sentence words that is outputted from the generator (G). 56

Figure 3.4 Three example images from MSCOCO with corresponding captions. 60
Figure 3.5 Qualitative captions generated from unsupervised image captioning models. 64
Figure 3.6 Illustrations of the error analysis generated from unsupervised image captioning models. 64
Figure 3.7 The average visual attention reward of each sentence gained during the training process. 65
Figure 4.1 Overview of the ERP-LSTM framework 68
Figure 4.2 Sample photographs from the six categories 70
Figure 4.3 Illustrations of statistical association, statistical causation, and CCN-IG causation. (a) Instantaneous coherence based on a statistical association. (b) Time lagged statistical causation (e.g., Granger causality). (c) Infer causation between EEG channels via CCN-IG with instantaneous and time-lagged (Granger) causal effects. 74
Figure 4.4 Illustration of CCN-IG. (a) EEG Causal Inference: EEG trials are inputted to the causal convolutional networks (CCNs) to identify the potential causes for each channel, which are then verified in the attentive causal validation module to generate a global causal inference map (GCIM). (b) Attentive Graph Learning: the obtained GCIM is fed into the spatial attentive layer followed by spatial-temporal graph convolution neural network to learn the effective brain networks for EEG-based BCI classification tasks. 76
Figure 4.5 Confusion matrices of CCN-IG in four classification tasks with spectral EEG inputs. 83
Figure 4.6 (a) AUROC scores of Coh-AGL, PDC-AGL and CCN-IG in four classification tasks. (b) AUROC scores of CCN-IG w/o SAL and CCN-IG in four classification tasks. 85
Figure 4.7 The significantly discriminative brain networks across multiple classes identified based on Coh (in the upper panel), PDC (in the middle panel) and GCIM (in the bottom panel) as graph topologies with spectral (frequency-domain) EEG inputs for four EEG-based classification tasks (a)-(d). 86
Figure 5.1 (a) Block design and randomised design. (b) Dual-coding of knowledge in the human brain. 90

Figure 5.2 Overview of the trimodal contrastive learning frammwork. 91

LIST OF TABLES

Table 1 Performance comparisons with existing unsupervised methods under the experimental setting of [41]. 61
Table 2 Performance comparisons with existing unsupervised methods under the experimental setting of [100]. 62
Table 3 Ablation study on MSCOCO across the VG dataset test split. The adversarial network, object and relation detector are trained on the images and ground-truth relationships of the VG dataset. 62
Table 4 Performance comparisons of different visual relation detectors for our unsupervised image captioning model on the test split of MSCOCO. 63
Table 5 Performance of six-category visual classification 71
Table 6 Performance of category- and exemplar-level visual classification 72
Table 7 Classification performance (ACC, F1, and AUROC) of CCN-IG (w/o SAL) and six baseline methods using temporal and spectral EEG inputs. 80
Table 8 The data summary of datasets for emotion recognition, fatigue assessment, motor imagery control and visual object detection. 81
Table 9 Statistics of the datasets 96
Table 10 Visual classification accuracy (%) for all subjects of Purdue-1.0 dataset trained using EEG, visual (V), textual (T) and combined (V&T) features, respectively. 97
Table 11 Visual classification accuracy (%) on Purdue-2.0 dataset 98
Table 12 Transposed Data with Bold Maximums 98

ACRONYMS

Acronym	Definition
AI	Artificial Intelligence
CNN	Convolutional Neural Network
RNN	Recurrent Neural Network
CV	Computer Vision
DNN	Deep Neural Network
EEG	Electroencephalogram
ERP	Event-Related Potential
fMRI	Functional Magnetic Resonance Imaging
GAN	Generative Adversarial Network
GNN	Graph Neural Network
GRU	Gated Recurrent Unit
LSTM	Long Short-Term Memory
MEG	Magnetoencephalography

1 INTRODUCTION

1.1 motivation

This thesis explores the problem of developing multimodal deep learning models to decode both visual and neural representations, drawing inspiration from the mechanisms of the human visual system. By shifting from a traditional "top-down" approach to a "bottom-up" framework, this research investigates how the physiological signal recordings of human visual processing-such as electroencephalography (EEG)-can enhance automated visual tasks. Through this integration, the study aims to leverage insights from human visual expertise to improve the synergy between human and machine vision, ultimately advancing computational models that better emulate human perception.

1.1.1 *Multimodal Learning and Visual Representation*

Vision plays a central role in human information processing, with approximately 80% of external information received through the visual system [184]. Consequently, equipping machines with the ability to "see" and interpret visual information as humans do has long been a significant challenge in Artificial Intelligence (AI).

In recent years, computer vision (CV) technologies have advanced rapidly, enabling machines to autonomously extract meaningful information from digital images and videos. Techniques such as image processing, pattern recognition, machine learning, and deep learning now drive this process. CV is a key focus within AI, with applications ranging from autonomous driving and medical imaging to facial recognition, video surveillance, and augmented reality.

A critical aspect of CV is visual representation, which typically involves using deep neural networks to extract and encode visual information into distinct features for tasks such as classification, detection, and recognition [64, 87]. Convolutional Neural Networks (CNNs) are commonly used for this purpose. In CCNs, images are passed through successive convolutional layers, which apply filters to detect features of varying complexity, starting from basic shapes like edges to more sophisticated elements such as object parts and textures as the network deepens [97]. Pooling layers help reduce the spatial dimensions of feature maps, improving computational efficiency while preserving the most relevant information. The final layers of the network interpret these high-level features to classify the image, for exam-

ple, distinguishing between different objects like animals or vehicles. High-quality representations can clearly differentiate between image categories, whereas low-quality ones may lead to confusion. CNNs have achieved remarkable success in various vision applications. For instance, face recognition models based on CNNs, such as DeepFace and FaceNet, have achieved over 99% accuracy on standardised datasets like LFW (Labelled Faces in the Wild), making them suitable for security systems and smartphone authentication [156, 178]. In autonomous driving, CNNs such as YOLO (You Only Look Once) and Faster R-CNN reliably detect and classify objects, including pedestrians, vehicles, and road signs, providing critical support for real-time decision-making in complex environments [149, 151]. In object detection tasks, for instance, feature pyramid networks (FPNs) enhance detection accuracy by integrating visual coding at multiple scales [211].

However, in today's deep learning landscape, relying on a single modality to process information has proven insufficient for addressing the complexities of real-world environments. Unimodal approaches typically excel in handling specific types of data but struggle to integrate contextual information from other sources [27]. For example, a model that relies only on visual data for image classification may perform poorly in tasks involving text or audio information. In contrast, the human brain integrates multiple sensory inputs, including visual, auditory, and tactile data, and draws upon context and experience to interpret complex information. This multisensory fusion enables humans to make accurate perceptions and decisions, even in noisy or incomplete environments. Inspired by this, multimodal learning has emerged as a growing area of interest within machine learning. By combining visual data with other modalities, such as text or audio, multimodal models leverage the strengths of each modality to improve image comprehension, thus enhancing both the robustness and expressiveness of the models [130]. For instance, in sentiment analysis, combining text and audio has significantly improved emotion recognition accuracy, with multimodal models achieving up to a 10% increase in accuracy over text-only approaches [142, 182]. Similarly, in visual question answering, models integrating visual and textual data outperform unimodal models by more than 15% in accuracy on standardised datasets such as VQA [7, 43]. In image captioning tasks, multimodal models that incorporate visual and linguistic information generate more semantically rich outputs, producing captions that are not only descriptive but also contextually appropriate [198, 203]. These examples underscore how multimodal learning enables more accurate and context-sensitive interpretations, particularly in complex environments.

Image captioning goes beyond object recognition by also identifying attributes and relationships between objects within an image. Based on extracted visual features, the model generates descriptive

text (see Figure 1.1), making it applicable to areas like image search, content generation, and assistance for the visually impaired. Researchers have made significant advances in this field [6, 28, 90, 120, 198, 200], largely following the encoder-decoder framework first proposed by Vinyals et al. [189], which draws inspiration from machine translation techniques. Generally, this framework consists of an encoder that employs a CNN to read the input image and a recurring neural network (RNN) decoder to generate text descriptions by translating the encoded representations (as shown in Figure 1.1). To automatically generate descriptive sentences from input images, these studies use paired image-sentence annotations to train the end-to-end neural network in a fully supervised manner. However, these models rely on large datasets of paired image-sentence annotations for supervised training, which can be expensive and time-consuming to collect, despite the availability of public datasets like MSCOCO [113] and Visual Genome (VG) [96]. MSCOCO, for example, only includes around 100 common object categories, limiting the scope of models trained on it to handle real-world problems. Therefore, developing unsupervised methods or innovative training strategies that enable models to learn the relationships between images and text without explicit annotations remains a major challenge in this area.

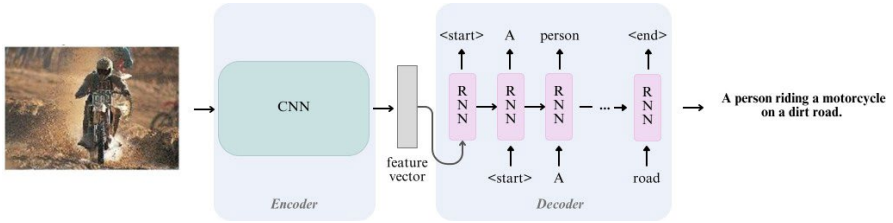


Figure 1.1: Encoder-Decoder framework for image captioning. The process begins with a CNN acting as the encoder, which processes the input image to extract a feature vector representing the key visual features. This feature vector is then fed into an RNN decoder. The decoder starts with a special <start> token and generates a sequence of words to describe the image. Each RNN cell predicts the next word in the sequence, eventually leading to a complete caption, such as "A person riding a motorcycle on a dirt road."

1.1.2 *Decode Visual Neural Representation*

Despite significant progress, it remains challenging to enable machines to interpret images at a higher semantic level, as humans do. Humans benefit from extravisual perception, relying on a lifetime of experience. For instance, we naturally focus on objects using our central vision, an intuitive process rooted in years of evolution. Unlike machines, which require examples and instructions to learn, humans

understand the world through an innate, unsupervised learning process.

Through years of research, neuroscience has unveiled many of the neural mechanisms underlying brain functions, particularly in visual processing. For example, the studies by David Hubel and Torsten Wiesel, Nobel laureates in 1981 [75, 76], revealed that visual information is processed hierarchically in the brain. The brain works through a continuous and iterative process. After the retina receives raw visual stimulus, the primary visual cortex (V1) extracts basic features like edges and orientations. Subsequently, higher areas, such as V2, abstract these features further into contours and shapes. This process iterates in stages until the brain forms a perception of complete images, objects, and scenes [40].

Early deep neural networks were inspired by this hierarchical structure, with their layers mimicking the brain's feedforward visual pathways. However, significant differences remain between how humans and CNNs process visual information. Studies have shown that CNNs still lack the robustness of human perception, particularly in generalising to unseen distortions and processing negative images [46, 73]. To achieve AI systems that can rival human intelligence, deeper exploration into the brain's information processing mechanisms are needed, especially to understand how the brain activity patterns are in relation to perception, cognition, emotion, and memory. This raises the question of whether the human brain, as an 'advanced' image encoder, can provide more modal information for the computer vision models.

Recent advancements in cognitive psychology, neuroscience, and brain imaging technologies, such as EEG, functional magnetic resonance imaging (fMRI), and magnetoencephalography (MEG), has made it possible to explore visual neural representations more deeply. These technologies allow the development of computational models that decode neural representations associated with visual input. When the brain processes visual information, it quickly encodes images into neural responses, a process known as "encoding" [67, 88]. The reverse process, predicting visual stimuli from known brain activity, is referred "decoding" [127]. By capturing and decoding these brain signals, we can gain insights into the brain's focus on specific image features and its overall response to visual stimuli. This not only helps to uncover the mechanisms of perception but also enhances image processing in multimodal learning by integrating brain signals with computer vision models.

Figure 1.2 illustrates the mapping between external stimuli, visual encoding in the brain, and the decoding of visual neural representations. Numerous studies have focus on the field to decode visual neural representations by interpreting the neural patterns activated during visual tasks [2, 60, 70, 71, 77, 106, 157, 159]. Specifically, when

the brain perceives a stimulus, it encodes the attributes of the stimulus, activating specific neurons and forming distinct brain activation patterns. Once sufficient brain activity data is collected, a mapping function can be developed to predict visual stimuli based on neural signals. Each image category corresponds to specific neural activation patterns, allowing the decoder to identify visual stimuli from brain activity and convert these physiological signals into image categories. When the brain responds strongly to specific parts of an image, those areas likely contain features critical to visual perception. This can can enhance the accuracy of tasks like image classification and object detection. For instance, research has shown that certain categories, such as faces or objects, are represented in specialised regions of the brain's cortex, as indicated by studies using hemodynamic responses [26, 59, 78].

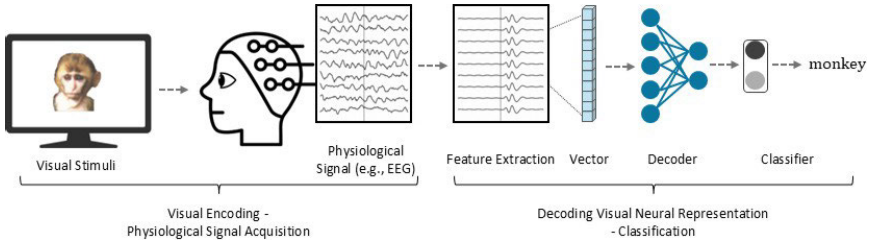


Figure 1.2: Illustrate the task of visual encoding and decoding of neural representations in the brain. When the human brain perceives a stimulus, it encodes the attributes of the stimulus, activating specific neurons and creating distinct patterns of brain activity. Once a sufficient amount of brain activity data in response to stimuli is collected, it is possible to train a mapping function between brain activation and visual stimuli. Since each image corresponds to a specific category, the trained decoder can then be used to translate physiological signals into image categories.

This concept of interactive approaches involving the "human-in-the-loop" paradigm is not entirely new. Various experiments have examined visual neural representation and decoding across different levels, from low-level features such as orientation and colour, to mid-level features like depth, and high-level semantic categories [37, 71, 77, 84, 157, 159, 166]. Most research in this area has used fMRI due to its high spatial resolution, which provides detailed spatial information about brain activity. However, fMRI is expensive, requires stringent experimental conditions, and has low temporal resolution, limiting its real-time applications. An electroencephalogram (EEG), by contrast, offers very high temporal resolution, capturing brain activity at the millisecond level, which is vital for real-time image processing. Moreover, EEG equipment is portable, affordable, and less restrictive, making it more suitable for large-scale studies. Despite these advantages, EEG-based research on visual neural representation is still limited. This thesis aims to address this gap by using EEG to decode vi-

sual neural representations. By classifying EEG signals, this research seeks to develop brain-inspired computational models that can classify or detect visual stimuli from brain activity. This approach not only contributes to the development of AI models but also advances our understanding of the cognitive processes underlying human visual perception.

1.2 the research gap

There are several key research gaps in the development of multimodal models for decoding visual and neural representations, particularly in two tasks: unsupervised image captioning and EEG-based decoding of visual neural representations.

1) Unsupervised image captioning has gained attention due to the limitations of traditional models, which rely on large, annotated datasets that are costly and time-consuming to produce. In contrast, unsupervised approaches, which use unpaired images and text, offer a more flexible and cost-effective alternative, making them highly scalable and applicable to new data. The main challenge is to minimise the misalignment between visual content and the generated language representation. This arises from the disjoint feature space of image sets and sentence corpora, which weakens the direct correspondence between the two modalities. Existing methods attempt to create a global mapping between visual and textual data, but this often leads to poor local alignment [55, 56], where models struggle to generate captions that accurately reflect the details within images. Additionally, corpus bias further complicates this, causing models to rely on generic patterns in the text data rather than the actual visual input [41, 100].

   To add more local alignment, several approaches have used weak supervision with object detection to improve the semantic accuracy of captions. While this has enhanced object recognition, it has been less successful in capturing the relationships between objects. For example, Figure 1.3 shows four images of the same subject and object, a person and bike. Models may correctly identify a person and a bicycle from the object detector, but fail to distinguish the relationships as it may generate different descriptions, like Figure 1.3(A) [person riding a bike], Figure 1.3(B) [person holding a bike], Figure 1.3(C) [person behind a bike], Figure 1.3(D) [person next to a bike], etc. This inability to capture the nuances of object relationship results in captions that lack semantic accuracy. Addressing these challenges requires improving both global and local alignment through techniques such as adversarial training and visual at-

tention mechanisms. These advancements would enable models to better capture the visual relationship knowledge in images, improving the overall quality and consistency of generated captions.



Figure 1.3: The relationship of two objects is crucial for image captioning

2) In the context of decoding visual neural representations based on EEG classification, significant gaps persist. While research on EEG-based object recognition has progressed [2, 11, 84, 140, 141, 166], most machine learning approaches primarily rely on unprocessed EEG waveforms, which represent the brain's spontaneous electrical activity. Generally, EEG data is often contaminated by noise and artifacts, making it crucial to pre-process the raw data before classification. Moreover, effective feature extractions are also essential to classification, as they enable the identification of brain patterns associated with visual stimuli. Classification based on EEG feature extraction is more likely to achieve higher accuracy compared to classification using raw EEG data [117]. Without robust feature extraction, the classification accuracy of these models remains limited.

2.1) Temporal features are commonly used in traditional EEG analysis, such as event-related potentials (ERP), which reflect the brain's response to specific visual stimuli [45]. ERPs are important for understanding the timing and sequence of neural processing. However, the current implementations do not fully exploit the use of ERP features in visual tasks, which limits the accuracy of EEG classification. Incorporating deep learning models, particularly RNNs, that can capture temporal dynamics may improve the performance of EEG classification.

2.2) Another important feature involves brain network connectivity patterns. Traditional deep learning models, such as CNNs and RNNs, process EEG data in one-dimensional sequences or two-dimensional grids, failing to capture the complex interactions between different brain regions [165, 180]. These models are more commonly utilised to extract temporal and spatial domain feature representations, overlooking non-Euclidean relationships that crucial to brain function. While several Graph

Neural Networks (GNNs)-based methods attempt to address this by modeling functional connectivity, they often focus on statistical associations rather than directed causal relationships [65, 74]. A few studies have explored effective connectivity, such as Granger causality; however these methods require extensive manual tuning, and primarily focus on time-lagged causality, neglecting instantaneous interactions. Therefore, there is a clear need for models that can automatically learn both time-lagged and instantaneous causal relationships, offering a more accurate and efficient representation of brain connectivity for EEG-based classification tasks.

2.3) Additionally, there is an under-exploitation of the multimodal semantic knowledge embedded in EEG data. Most research has focused on unimodal supervised learning frameworks, neglecting the potential benefits of of combining brain activity with other modalities, such as text, to jointly learn visual, brain, and linguistic features. Cognitive neuroscience suggests that the brain processes visual stimuli through a dual coding mechanism, combining visual and textual elements [13, 133]. However, few studies have explored the capability of EEG visual representation across multiple modalities.

In conclusion, significant research gaps remain in unsupervised image captioning, the effective use of EEG features for visual recognition, and the integration of multimodal semantic knowledge in EEG-based decoding. Addressing these gaps could lead to major advancements in developing models that more closely resemble human cognitive processes in visual understanding.

1.3 research questions

The overarching research aim of this thesis is to investigate how multimodality can be effectively modelled in tasks involving image captioning and the decoding of visual neural representations. This research seeks to understand the relationships between images, text, and EEG data, focusing on how these modalities can be mapped and integrated through unsupervised and multimodal learning techniques.

This work is guided by three primary research questions (RQs) that build on each other:

RQ1 (How can modality mapping between image and text be modelled in image captioning tasks?)

This question involves the following sub-quetions:

- RQ1.1: How to model a pure unsupervised modality mapping framework without relying on annotated image-sentence datasets while achieving both local and global alignment in image captioning?
- RQ1.2: How does incorporating adversarial learning, such as GANs, enhance the performance of global alignment in unsupervised image captioning?
- RQ1.3: How can visual relations contribute to enhancing local alignment of the generated caption with specific visual content in an image?

***RQ2** (How can modality mapping between image and EEG be modelled in visual-related classification tasks?)*

This question involves the following sub-quetions:

- RQ2.1: How does the feature extraction process from EEG signals and image data affect the multimodal object classification?
- RQ2.2: What is the impact of utilising temporal features of EEG (such as visual evoked potentials (ERPs)) on the feature representation and classification accuracy?
- RQ2.3: What is the impact of using graph-based neural network models in representing effective brain networks on the classification performance of multimodal data?
- RQ2.4: How does the modelling of time-lagged and instantaneous causal relationships in EEG signals improve the representation of brain network for visual classification?

***RQ3** (How to design multimodal ;earning of image, text and EEG in decoding visual neural representation?)*

This question involves the following sub-quetions:

- RQ3.1: How does the integration of brain, visual, and linguistic features through trimodal learning enhance the decoding of visual neural representations from EEG signals?
- RQ3.2: What is the effect of contrastive learning on aligning EEG, image, and text data for enhancing visual object classification from neural representations?
- RQ3.3: How does randomised experimental design influence the robustness of trimodal models in decoding neural representations across different visual categories?

As can be seen, these research questions are related. First, RQ1 explores the modelling of modality mapping between images and text in image captioning tasks using unsupervised learning techniques. This foundational investigation is essential for understanding how visual and linguistic information interact without explicit supervision. The insights from this question provide a framework for RQ2, which extends modality mapping to images and EEG signals in object classification tasks. By linking brain activity with visual data, this question seeks to uncover how the brain processes visual stimuli and how this can be translated into classification models. Finally, RQ3 integrates the findings from the previous questions to address the more complex task of multimodal learning. Here, the focus is on modelling the combined relationships between images, text, and EEG data, to decode visual neural representations more effectively.

Together, these research questions are critical for advancing our understanding of modality mapping and its role in multimodal learning, particularly in tasks that bridge visual and neural data. The ability to model these relationships has significant implications for fields such as brain-computer interfaces, image recognition, and cognitive computing.

1.4 research objectives

The work presented in this thesis aims to fill the important research gap in the current study of decoding visual and neural representation, as mentioned in Section 1.2. The main research objectives (OBJs) that will be discussed throughout this thesis are:

- **OBJ1: Formulating an Unsupervised Image Captioning Framework**

  To answer **RQ1.1 - 1.3**, OBJ1 aims to explore a well-performed unsupervised image captioning framework by identifying the gap between information misalignment in the image and the generated sentence, especially the visual relation between detected objects.

- **OBJ2: Formulating an EEG-based Object Classfication Model with Extracted Temporal Feature**

  To answer **RQ2.1 - 2.2**, OBJ2 aims to explore an evoked signal-guided object classification model by extracting more distinct EEG features from the distilled event-related potential (ERP) and utilising deep neural network (i.e., RNN) to merge the feature extraction and classifier.

- **OBJ3: Formulating an visual-related EEG Classfication Model with Effective Connectivity Feature**

To answer **RQ2.1 and 2.3 - 4**, OBJ3 aims to explore an GNNs-based EEG classification model by capturing effective connectivity to decode discriminate brain networks and enhance graph representation for four visual related tasks.

- **OBJ4: Formulating a Trimodal Learning Framework for Decoding Visual Neural Representations from EEG**

  To answer **RQ3.1 - 3.3**, OBJ3 aims to explore a novel trimodal contrast learning model by the joint learning of brain-visual-text features, with specially designed instance-, modality-level modelling to achieve more consistent brain-visual-linguistic joint representations.

1.5 chapter overview

This section provides an overall structure of this thesis, and the approach to each RQ is explained in detail in the corresponding chapters. The thesis is organised into distinct sections as described here:

chapter 1: introduction Introduces the motivation behind the research for enabling the multimodal learning to facilitate visual and neural representation, where this research is placed among the literature.

chapter 2: literature review Review of key literature to form a solid foundation for the experimental study being conducted in this thesis. Some key research areas are covered in this chapter including visual representation, image captioning, brain-inspired visual representation and decoding visual neural representation with EEGs.

chapter 3: adversarial word-level modelling for modality mapping between image and text in unsupervised image captioning An adversarial word-level visual relation alignment model for unsupervised image captioning to explore rich semantic captions is presented here to answer to answer **RQ1.1 - 1.3**, together with the experiments and results.

chapter 4: modality mapping between image and eeg in brain-inspired object classification Two brain-inspired visual classification models are presented here to answer **RQ2.1 - 2.4**, together with the experiments and results.

chapter 5: trimodal contrastive learning for decoding visual neural representation A novel trimodal contrast learning model for decoding visual neural representations from

EEG is presented here to answer **RQ3.1 - 3.3**, together with the experiments and results.

chapter 6: discussions Insights collected from this research are discussed here, focusing on key components that assess aspects of multimodal learning to improve visual and neural representations. Further, some fundamental limitations of the study are addressed.

chapter 7: conclusions and directions for future research Thesis conclusions and directions for future research are discussed here, summarising the advancements in multimodal deep learning for visual and neural representation. This chapter also highlights implications of this research and outlines future opportunities for advancing multimodal learning.

2

LITERATURE REVIEW

2.1 visual representation

Visual representation, particularly in the context of image processing, refers to how machines interpret and understand visual data, typically by transforming raw pixel data into high-dimensional features. Generating high-quality visual representations is fundamental to image classification tasks. Different images possess distinct visual features, which allow classifiers to predict the category of an image based on these features. An effective representation can better distinguish between different image classes, while a low-quality representation may lead to category confusion. This has emerged as a central research focus within computer vision and is extensively applied in various tasks, including object localisation, detection, and segmentation [64, 87].

The concept of visual representation has its roots in traditional image processing techniques, where handcrafted features, such as edges, corners, and textures, were designed manually by experts. Historically, before the advent of deep learning, visual representation was achieved using handcrafted feature extraction methods. Techniques such as Scale-Invariant Feature Transform (SIFT) and Histogram of Oriented Gradients (HOG) were some of the most popular approaches used to describe local structures in an image. SIFT, for instance, identified key points and computed descriptors that were invariant to scale and orientation, making it a powerful tool for tasks such as object recognition and image matching [118]. Similarly, HOG represented an image by capturing the gradient direction or intensity of edges, which was effective in detecting objects in images [29]. However, these handcrafted features were often task-specific and lacked the flexibility to generalise across different domains.

With the rise of CNNs, image representation began to be learned automatically through data-driven approaches rather than relying on handcrafted features. CNNs extract hierarchical feature representations by learning from the raw pixel values of an image, moving from simple features (like edges) in the lower layers to more abstract and complex features (such as faces or objects) in the higher layers. This hierarchical structure allows CNNs to capture a wide range of patterns and generalise across various visual tasks, thus significantly improved the performance.

A typical CNN starts with an input image that is passed through a series of convolutional layers, which apply filters to detect various fea-

tures. For instance, in the early stages of a CNN, convolutional layers apply filters to the input image to detect simple features like edges and corners. As the network deepens, these layers start to recognise more intricate features, such as object parts or textures. Finally, fully connected layers at the top of the network interpret the high-level features to classify the image into specific categories, such as distinguishing between animals, vehicles, or other objects [97]. Pooling layers are employed in between to reduce the spatial dimensions of the feature maps, improving computational efficiency while preserving the most relevant information. Followed fully connected layers at the end are responsible for making predictions based on the extracted features.

The following Figure 2.1 illustrates how CNNs process an image through multiple layers, progressively extracting more complex features: For example, in a task like classifying images of animals, the first few layers of a CNN might detect simple edges or textures like fur patterns. Deeper layers would identify parts of the animal, such as ears, eyes, or tails, and finally, the top layers would combine this information to recognize the full object, whether it's a cat, dog, or bird.

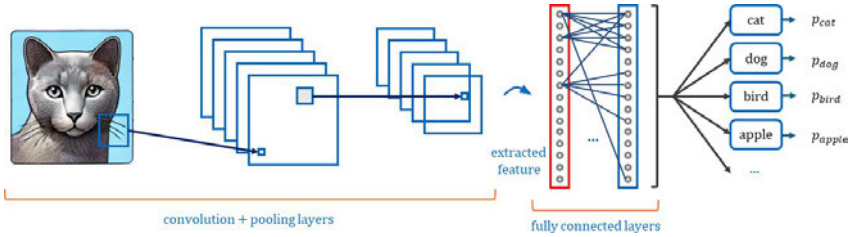


Figure 2.1: The illustration of CNNs process for visual representation.

The evolution of visual representation can be clearly observed when comparing traditional methods like HOG and SIFT to CNNs. The former required domain expertise to design features specific to certain tasks, while the latter automatically learns feature hierarchies from large datasets. This ability to learn representations end-to-end has allowed CNNs to outperform traditional methods on a wide range of vision tasks. In particular, AlexNet's breakthrough performance on the ImageNet dataset in 2012 was a clear demonstration of CNNs' ability to extract rich and meaningful visual representations [97].

AlexNet, consisting of five convolutional layers and three fully connected layers (as shown in Figure 2.2), achieved unprecedented success in the ImageNet Large Scale Visual Recognition Challenge (ILSVRC), demonstrating the strong feature learning capability of CNNs on large-scale datasets [97]. CNNs automatically learn hierarchical feature representations from raw image data, which eliminates the need for handcrafted feature extraction methods. This ability to learn both low-level features such as edges and textures and high-level semantic features such as shapes and objects across different layers has been

fundamental to their success. Following the success of AlexNet, researchers have focused on further optimizing CNNs to enhance their performance.

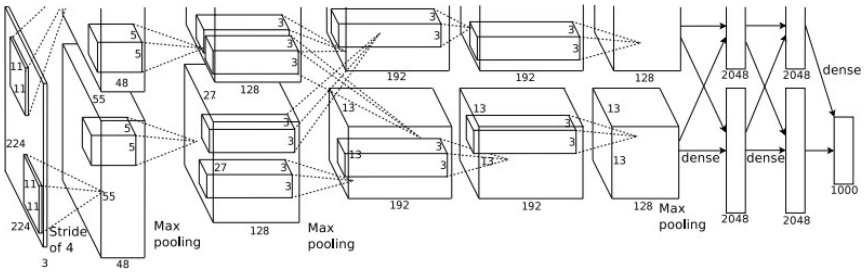


Figure 2.2: Illustration of AlexNet architecture (adapted from [97]).

VGG made a significant improvement by increasing network depth (16 and 19 layers), showing that deeper networks can learn more complex hierarchical features, achieving excellent classification performance on ImageNet [161]. One of VGG's key contributions was the use of smaller convolution filters, allowing the model to capture finer details of image features. Subsequently, ResNet introduced skip connections to address the problem of vanishing gradients in deep networks, enabling deeper networks without performance degradation. ResNet's architecture (as shown in Figure 2.3), ranging from 50 to 152 layers, significantly enhanced performance in both image classification and object detection tasks [63]. Inception networks, on the other hand, introduced multi-scale convolution filters to capture both fine and broad features within the same module, further improving CNNs' effectiveness in visual tasks [177].

In the domain of object detection, CNNs have also made significant contributions. The R-CNN model was one of the first to successfully apply CNNs to object detection tasks by generating region proposals and classifying each region [49]. Despite its success in performance, R-CNN had computational inefficiencies. Fast R-CNN and Faster R-CNN improved upon this by integrating region proposal generation with the classification network (see Figure 2.4), significantly reducing computational costs and improving detection speed and accuracy [47, 151]. Additionally, real-time object detection models such as YOLO (You Only Look Once) and SSD (Single Shot MultiBox Detector) further advanced the field. These models treat object detection as a single regression problem, directly predicting bounding boxes and classifications from feature maps, pioneering end-to-end object detection methods [116, 149].

Despite the significant achievements of CNNs in single-modal information processing, traditional single-modal approaches still have limitations in learning general features. These methods are typically specialized for handling specific data types. For example, a model trained solely on visual data may struggle with tasks that require

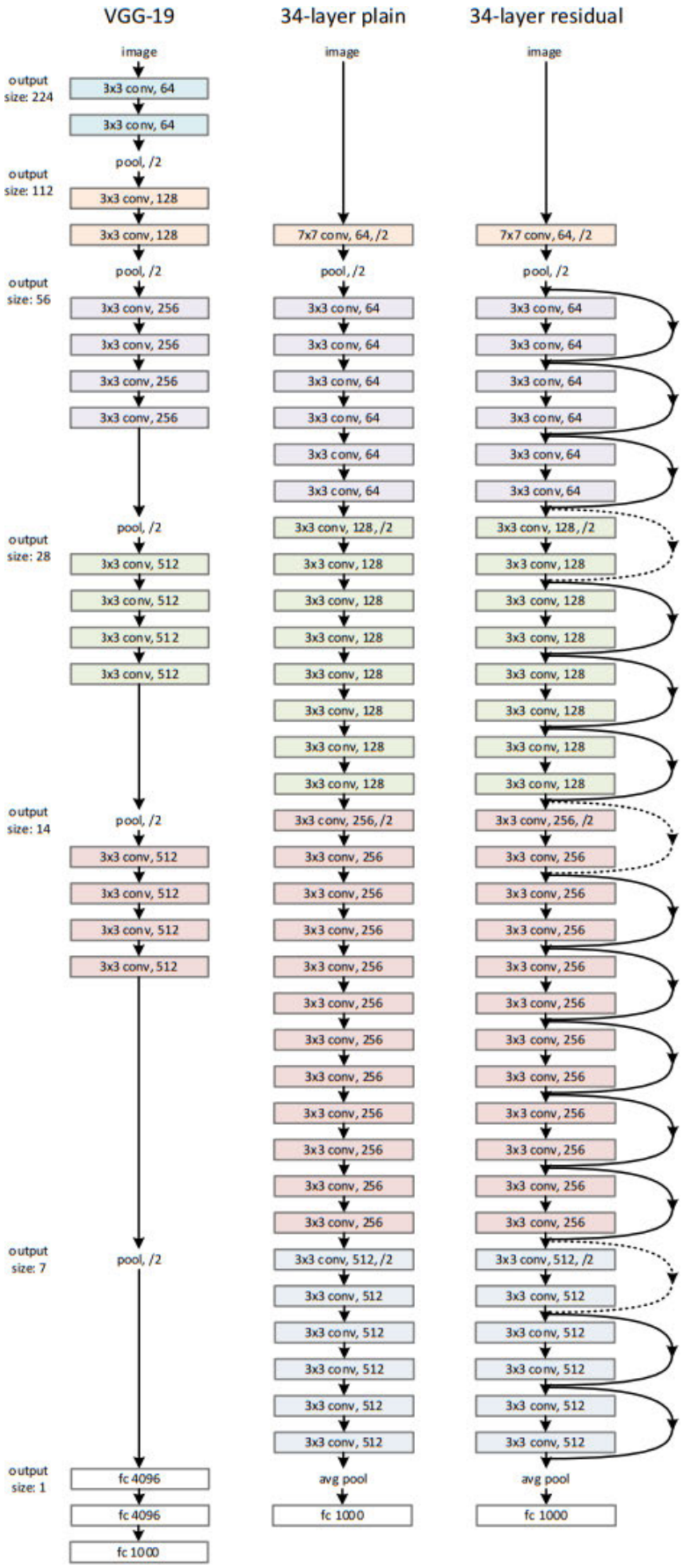


Figure 2.3: Illustration of ResNet architecture (adapted from [63]).

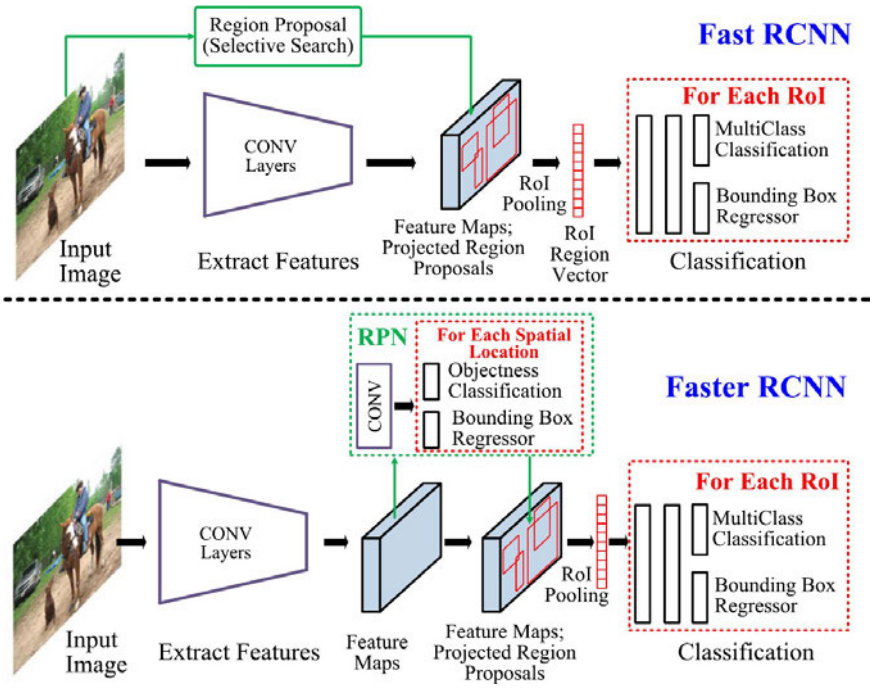


Figure 2.4: Illustration of Fast R-CNN and Faster R-CNN architecture (adapted from [115]).

understanding textual or audio data. Single-modal models often lack the ability to capture the rich context provided by integrating different types of sensory information, which is crucial for many real-world tasks.

To overcome the limitations of single-modal learning, researchers introduced multimodal learning, which integrates data from multiple modalities such as visual, textual, and auditory inputs to learn richer feature representations. Early multimodal CNN models focused on combining visual and textual data for tasks like image captioning, visual question answering (VQA), and video understanding [8]. These models typically use separate CNNs to extract visual features and RNNs or transformers to process textual data, with a final fusion step to combine the learned features.

Recently, transformer architectures, which have become dominant in natural language processing (NLP), have been extended to multimodal learning. Models like CLIP and Vision Transformers (ViTs) have been trained on both visual and textual data, enabling them to learn general feature representations that can be applied across various tasks. For example, CLIP uses contrastive learning to align images and text in a shared feature space, enabling zero-shot learning for image classification and object detection tasks [146]. In object detection, multimodal learning improves performance, especially when visual data alone is insufficient. Multimodal models can combine RGB and depth data (from LiDAR or stereo cameras) to detect objects

in 3D space, which is critical for applications like autonomous driving. Additionally, multimodal methods have been applied to tasks such as activity recognition, where visual data is combined with audio and motion data to accurately classify actions [131].

In conclusion, CNNs have significantly advanced the field of computer vision over the past decade, particularly in image classification and object detection. By learning hierarchical feature representations, CNNs have outperformed traditional machine learning models. However, the limitations of single-modal information processing have led to the rise of multimodal learning, which integrates data from multiple sources to build more generalizable and robust models. As multimodal learning continues to develop, it holds the potential for further breakthroughs in visual representation and understanding.

2.2 deep learning and multimodal learning methods

2.2.1 *Deep Learning*

Building on the previous section, which covered convolutional neural networks (CNNs) and their role in visual representation, this section delves deeper into four specific types of deep neural networks. In addition, it explores the principles of reinforcement learning and the emerging field of multimodal learning, highlighting their relevance and applications in decoding visual and neural representation. Together, these concepts offer a comprehensive understanding of advanced machine learning techniques and their interdisciplinary potential.

2.2.1.1 *RNNs and Long Short-Term Memory (LSTM)*

In many NLP tasks, the network need to predict the next word in a sentence conditioned on the words came before it, which is formalised as a sequence or time-series task. RNNs are proposed to handle such tasks because they perform the same task for every element of a sequence, with the output being depended on the previous computations [38]. RNNs acquire the ability to remember long-term information about what has been calculated so far in time-series tasks.

RNNs are a class of neural networks specifically designed to process sequential data by introducing loops within the architecture, allowing information to persist over time. Unlike feedforward neural networks, RNNs possess memory, making them suitable for tasks such as time-series prediction, speech recognition, and natural language processing. The architecture of a simple RNN can be described as:

$$h_t = f(W_{xh}x_t + W_{hh}h_{t-1} + b_h)$$

where h_t is the hidden state at time step t, x_t is the input, W_{xh} and W_{hh} are the input-to-hidden and hidden-to-hidden weight matrices respectively, and b_h is the bias term. The function f is usually a non-linear activation function, such as the hyperbolic tangent (tanh) or rectified linear unit (ReLU).

Despite the success of RNNs in processing sequential data, they encounter several limitations, notably the problem of vanishing and exploding gradients during training. This issue arises when backpropagating through time (BPTT), which involves calculating gradients over many time steps. The gradients tend to either vanish (become too small to influence learning) or explode (grow too large, causing unstable learning). The result is that vanilla RNNs struggle to capture long-range dependencies effectively. Formally, the update of the hidden state depends on the repeated application of the same weight matrix:

$$\frac{\partial L}{\partial W} = \sum_{t=1}^{T} \frac{\partial L}{\partial h_t} \frac{\partial h_t}{\partial W}$$

As the term $\frac{\partial h_t}{\partial W}$ involves the repeated multiplication of weight matrices, it leads to either exponential decay or growth, causing vanishing or exploding gradients.

long short-term memory (lstm) To address these limitations, the LSTM network was introduced by Hochreiter and Schmidhuber [66] in 1997. LSTMs are a specialised form of RNNs designed to overcome the vanishing gradient problem, enabling the network to capture long-term dependencies. The key innovation lies in the introduction of memory cells and gated mechanisms that control the flow of information. The difference between the RNN and LSTM is shown in Figure 2.5. A special connection called cell state was added between time steps to help information flowing through them as far as needed. In each connection, there is also a gate which controls whether to let a piece of information go through it or not-this allows LSTM to either remember or forget information in the previous time steps. While there are many variants of classic LSTM, e.g., gated recurrent unit (GRU) [24], we found that LSTM performed acceptably both at the image captioning framework and EEG feature extaction in our work.

An LSTM cell consists of three gates: the input gate, the forget gate, and the output gate. These gates regulate what information to store, discard, or output at each time step. Mathematically, the LSTM equations are as follows:

Forget Gate:

$$f_t = \sigma(W_f \cdot [h_{t-1}, x_t] + b_f)$$

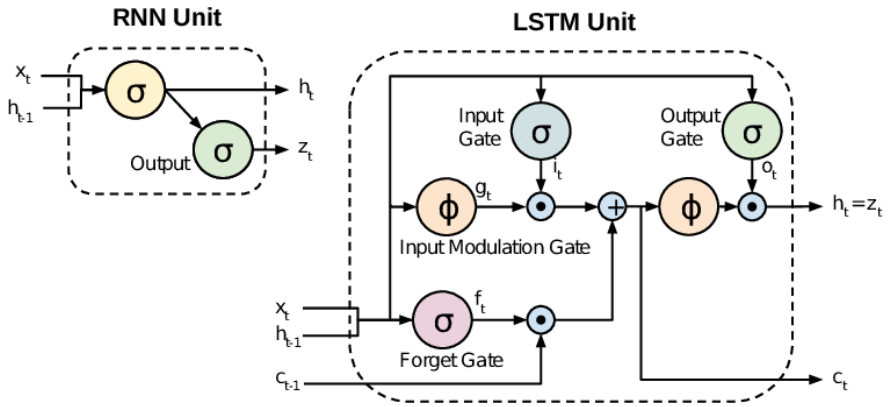


Figure 2.5: A general framework of a basic RNN cell (left) and an LSTM (right) memory cell [36].

The forget gate determines what portion of the previous cell state C_{t-1} should be carried forward. The sigmoid activation function σ squashes the values between 0 and 1, enabling the network to decide whether to keep or discard information.

Input Gate:

$$i_t = \sigma(W_i \cdot [h_{t-1}, x_t] + b_i)$$

$$\tilde{C}_t = \tanh(W_C \cdot [h_{t-1}, x_t] + b_C)$$

The input gate decides which values to update in the memory cell. The candidate cell state $\tilde{C}_t$ is computed and modulated by the input gate i_t.

Cell State Update:

$$C_t = f_t * C_{t-1} + i_t * \tilde{C}_t$$

The cell state C_t is updated based on the outputs of the forget gate and input gate. This allows the LSTM to carry forward relevant information over time, ensuring long-term dependencies are preserved.

Output Gate:

$$o_t = \sigma(W_o \cdot [h_{t-1}, x_t] + b_o)$$

$$h_t = o_t * \tanh(C_t)$$

The output gate determines the final output h_t at each time step, based on the updated cell state C_t and the hidden state from the previous time step. The memory cell effectively enables the LSTM to maintain a continuous flow of information across time steps without suffering from the vanishing gradient issue.

lstm training and optimisation The training process for LSTM networks follows the standard backpropagation through time

(BPTT) approach, though with the added advantage that the gated structure mitigates gradient degradation over long sequences. The optimisation is usually performed using stochastic gradient descent (SGD) or its variants, such as Adam or RMSprop, with a loss function typically defined as:

$$L = \frac{1}{N} \sum_{i=1}^{N} \ell(y_i, \hat{y}_i)$$

where y_i is the ground truth, $\hat{y}_i$ is the predicted output, and ℓ represents a loss function, often the cross-entropy loss for classification tasks or mean squared error (MSE) for regression tasks.

2.2.1.2 *Generative Adversarial Network (GAN)*

The GAN model is a generative model first proposed in [50], which was commonly applied in image and natural language generation. This architecture involves two sub-models: a generator model for generating new examples and a discriminator model for classifying whether generated examples are real, from the domain, or fake, generated by the generator model.

1) **Generator** G: Model that is used to captures the data distribution and generate new plausible examples from the problem domain.

2) **Discriminator** D: Model estimates the probability that a sample came from the training data rather than the generator (real or fake).

GANs are based on a game theoretic scenario in which the generator network must compete against an adversary. The generator network directly produces samples. Its adversary, the discriminator network, attempts to distinguish between samples drawn from the training data and samples drawn from the generator [51].

In the training process, the training procedure for G is to maximise the probability of D making a mistake. For example in an image generation task, the goal of the generator G is to generate real pictures as much as possible to deceive the discriminator D. The goal of D is to distinguish the pictures generated by G from the real pictures as much as possible. In this way, G and D constitute a dynamic "game process." What is the result of the final game? In the most ideal state, G can generate a picture G(z) that is "realistic" enough, where z is input noise variables.

In other words, D and G play the following two-player minimax game with value function V(G; D):

$$\min_G \max_D V(D, G) = E_{x \sim P_{data}(x)}[\log D(x)] + E_{z \sim P_z(z)}[\log(1 - D(z))]$$

(2.1)

where x denotes the real image from the training data. Figure 2.6 describes this training procedure.

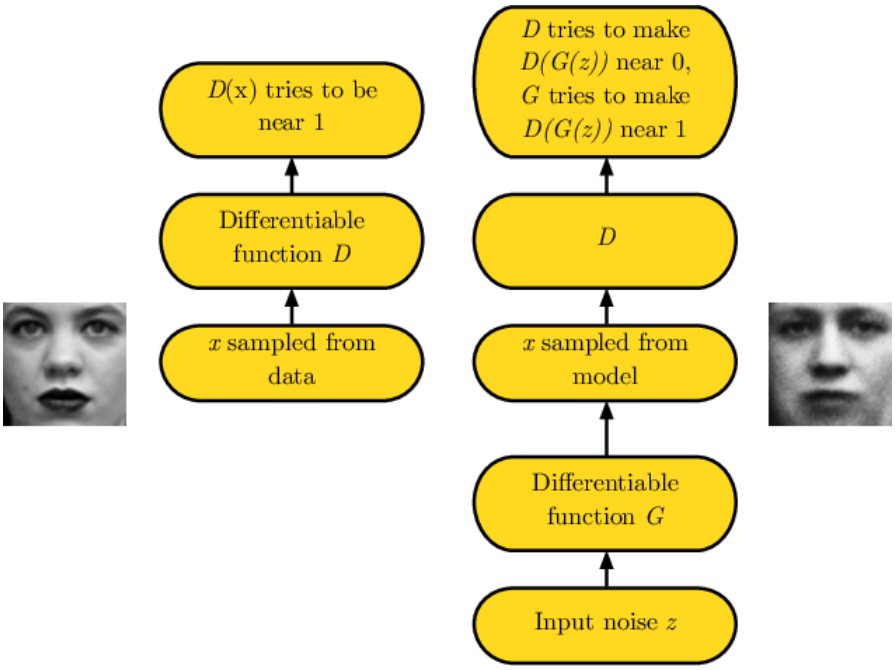


Figure 2.6: A general framework of GAN [51]

In the case where G and D are defined by multilayer perceptrons, the entire system can be trained with backpropagation and the parameters can be updated using Stochastic gradient descent. The whole training process alternates between G and D.

2.2.1.3 *Graph Neural Networks*

Graph Neural Networks (GNNs) are a class of neural networks that are inherently designed to work with graph-structured data [197]. Unlike traditional neural networks that operate on Euclidean data such as images or sequences, GNNs are capable of directly processing and learning from non-Euclidean data, such as social networks, molecular structures, and knowledge graphs. In the context of multimodal deep learning, GNNs play a critical role in understanding complex relational structures between data modalities. This section reviews the original methodology underlying GNNs, highlighting their core components, operational mechanisms, and mathematical foundation.

graph representation At the heart of GNNs is the graph, which is a data structure represented as $G = (V, E)$, where V denotes a set of nodes (or vertices) and E represents a set of edges connecting the nodes. Each node $v \in V$ can have its own feature vector $\mathbf{x}_v \in \mathbb{R}^F$,

where F is the dimension of the feature space, and edges $(u, v) \in E$ can also have associated weights $\mathbf{w}_{u,v} \in \mathbb{R}^W$. The adjacency matrix $A \in \mathbb{R}^{|V|\times|V|}$ encodes the connectivity between nodes, where A_{ij} denotes the presence or absence of an edge between nodes i and j.

The graph structure allows GNNs to capture both the features of individual nodes and the relationships between them, making them particularly effective in problems that involve relational data.

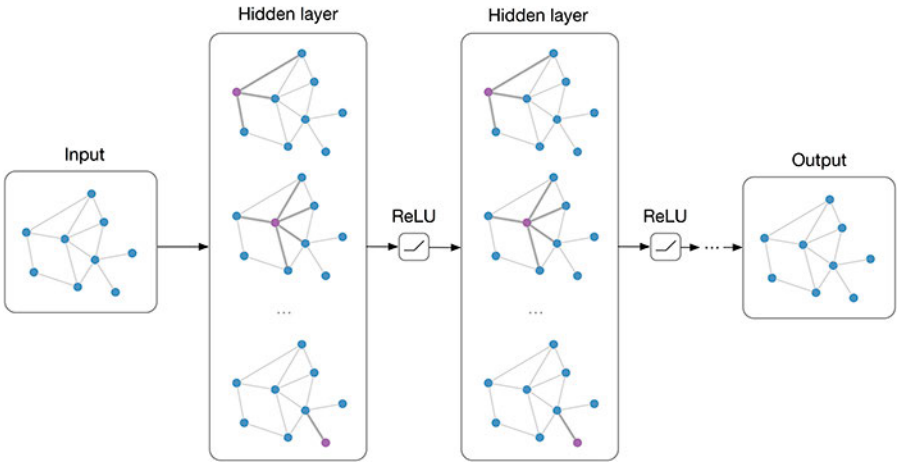


Figure 2.7: A Multi-layer Graph Convolutional Network. Source: https://tkipf.github.io/graph-convolutional-networks

gnn architecture overview The core idea behind GNNs is the iterative message-passing mechanism, where nodes aggregate information from their neighbors to update their own representations. The general formulation of a GNN layer is:

$$\mathbf{h}_v^{(k+1)} = \sigma\left(\mathbf{W}^{(k)} \cdot \text{AGGREGATE}\left(\{\mathbf{h}_u^{(k)} : u \in \mathcal{N}(v)\}\right)\right)$$

Here, $\mathbf{h}_v^{(k)}$ represents the hidden state of node v at layer k, $\mathbf{W}^{(k)}$ is a learnable weight matrix, σ is an activation function (e.g., ReLU), and $\mathcal{N}(v)$ denotes the neighbors of node v. The function AGGREGATE is key to the GNN architecture, determining how information from neighboring nodes is aggregated, which can be a summation, mean, or more complex functions depending on the specific GNN variant.

The message-passing process in GNNs can be described as the propagation of information across the graph through node features and edge relationships. At each layer, a node aggregates messages from its neighbors and updates its representation. Formally, for each node v, the message $m_v^{(k)}$ received from its neighbors at the k-th layer is given by:

$$m_v^{(k)} = \bigsqcup_{u\in\mathcal{N}(v)} W_e^{(k)} \cdot h_u^{(k)}$$

where $\mathbf{W}_e^{(k)}$ is a matrix of edge-specific weights. The updated node representation $\mathbf{h}_v^{(k+1)}$ is then computed by applying a non-linear transformation, typically using a neural network layer such as a multi-layer perceptron (MLP):

$$\mathbf{h}_v^{(k+1)} = \sigma\left(\mathbf{W}_n^{(k)} \cdot \mathbf{h}_v^{(k)} + m_v^{(k)}\right)$$

This process is repeated for a fixed number of layers, allowing the GNN to capture both local and global structures in the graph.

graph convolutional networks (gcn) One of the foundational GNN models is the Graph Convolutional Network (GCN), which extends the concept of convolution from Euclidean domains to graph domains (as shown in Figure 2.7). The GCN layer can be viewed as a localized convolution operation over the graph. Mathematically, the GCN layer updates the node representations as follows:

$$\mathbf{H}^{(k+1)} = \sigma\left(\tilde{D}^{-\frac{1}{2}}\tilde{A}\tilde{D}^{-\frac{1}{2}}\mathbf{H}^{(k)}\mathbf{W}^{(k)}\right)$$

where $\tilde{A} = A + I$ is the adjacency matrix with added self-loops, $\tilde{D}$ is the diagonal degree matrix of $\tilde{A}$, and $\mathbf{H}^{(k)}$ represents the matrix of node embeddings at layer k. This operation ensures that node features are smoothed over their local neighborhoods, allowing the GCN to capture more abstract representations of the graph structure.

node and graph-level prediction GNNs are versatile in their applicability, supporting both node-level and graph-level tasks. In node-level tasks, such as node classification, the objective is to predict the label of individual nodes. This is achieved by applying several GNN layers followed by a softmax function over the final node embeddings:

$$\hat{y}_v = \text{softmax}\left(\mathbf{h}_v^{(K)}\right)$$

where K is the number of GNN layers.For graph-level tasks, such as graph classification, the node embeddings are typically pooled to produce a fixed-size graph representation:

$$\mathbf{h}_G = \text{READOUT}\left(\{\mathbf{h}_v^{(K)} : v \in V\}\right)$$

The readout function can be a simple summation or a more sophisticated mechanism like attention pooling.

training gnns Training GNNs typically involves supervised or semi-supervised learning. In the supervised setting, the loss function is computed based on the prediction of labels for either nodes or graphs. A common choice for the loss function in node classification is the cross-entropy loss:

$$L = -\sum_{v \in V_L} y_v \log \hat{y}_v$$

where V_L represents the set of labeled nodes, y_v is the true label of node v, and $\hat{y}_v$ is the predicted label. Gradient-based optimization techniques, such as stochastic gradient descent (SGD), are used to update the GNN parameters during training.

2.2.1.4 *Transformer*

The Transformer architecture, first introduced by [187], has revolutionised the field of deep learning, especially in natural language processing (NLP). An overview of the transformer architecher is shown in Figure 2.8. Unlike earlier models that relied heavily on recurrence and convolutions, the Transformer dispenses with both, leveraging self-attention mechanisms to process sequences in parallel. This shift not only improved computational efficiency but also set the stage for further advancements in multimodal tasks, including the decoding of visual and neural representations. This review provides an in-depth exploration of the original Transformer architecture, focusing on the mechanisms of self-attention, positional encoding, and multi-head attention, with a clear explanation of the associated equations.

self-attention mechanism At the heart of the Transformer is the *self-attention* mechanism, which allows the model to focus on different parts of an input sequence (see Figure 2.9). Self-attention computes the relevance of each element in a sequence with respect to the others. For each token, queries Q, keys K, and values V are generated through linear transformations:

$$Q = XW_Q, \quad K = XW_K, \quad V = XW_V$$

where X is the input sequence, and W_Q, W_K, and W_V are learnable weight matrices. The attention score between two tokens is calculated by taking the dot product of their respective queries and keys, scaled by the square root of the dimensionality of the keys:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V$$

This scaling factor, $\frac{1}{\sqrt{d_k}}$, prevents the dot products from becoming too large, thereby ensuring stable gradients during training. The softmax function normalises these scores into a probability distribution, enabling the model to focus more on the relevant parts of the sequence.

positional encoding While self-attention allows for parallel processing of tokens, it does not inherently capture the order of tokens in a sequence. To address this limitation, *positional encoding* is

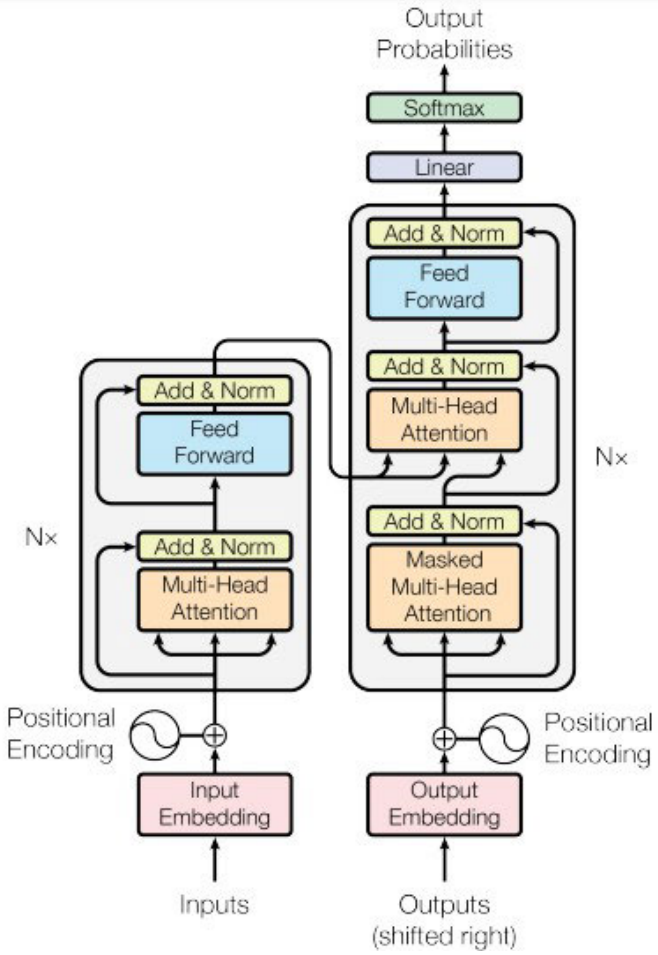


Figure 2.8: An overview of the Transformer [187].

added to the input embeddings, ensuring the model retains positional information. Positional encodings are computed using sine and cosine functions of different frequencies, with each position pos in the sequence represented as:

$$PE_{(pos,2i)} = \sin \frac{pos}{10000^{2i/d_{model}}}, \quad PE_{(pos,2i+1)} = \cos \frac{pos}{10000^{2i/d_{model}}}$$

Here, i is the dimension index, and d_{model} is the dimension of the model. This formulation ensures that the positional encodings are unique for each position and that they can generalise to sequences of varying lengths. By incorporating this information into the input, the Transformer can distinguish between different positions in the sequence, which is crucial for tasks requiring sequential order.

multi-head attention One of the major innovations of the Transformer is the *multi-head attention* mechanism. Instead of relying on a single self-attention operation, the Transformer computes attention in parallel across multiple "heads," each of which captures different aspects of the input relationships. The attention mechanism

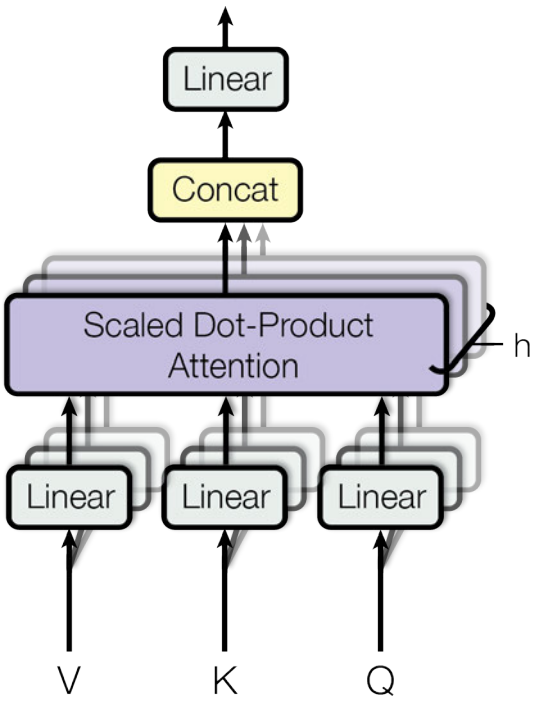


Figure 2.9: The illustration of self-attention mechanism in the Transformer [187].

described earlier is replicated across h heads, with each head having its own set of learnable parameters:

$$\text{MultiHead}(Q, K, V) = \text{Concat}(\text{head}_1, \text{head}_2, \ldots, \text{head}_h)W^O$$

where each head is computed as:

$$\text{head}_i = \text{Attention}(QW_{Q_i}, KW_{K_i}, VW_{V_i})$$

The results from each head are concatenated and passed through a final linear transformation using W^O. This approach allows the model to attend to different parts of the sequence simultaneously, capturing a wider range of dependencies and relationships between tokens. Importantly, multi-head attention improves the model's ability to focus on different subspaces of the input space, leading to richer representations.

feed-forward networks Following the multi-head attention mechanism, the Transformer employs *feed-forward neural networks* (FFNs) to further process the attention outputs. These are applied independently to each position in the sequence and consist of two linear transformations with a ReLU activation in between:

$$\mathrm{FFN}(x) = \max(0, xW_1 + b_1)W_2 + b_2$$

The FFN is applied identically to each token, enabling non-linear transformations of the attention output. Notably, the dimensionality of the intermediate layer (determined by W_1) is often much larger than the input and output dimensions, enhancing the model's capacity to learn complex patterns.

layer normalisation and residual connections To facilitate stable and efficient training, the Transformer incorporates *residual connections* and *layer normalisation*. Residual connections allow each sub-layer (multi-head attention or FFN) to learn corrections rather than completely new transformations. For an input x and a sub-layer $F(x)$, the residual connection is defined as:

$$\text{output} = \mathrm{LayerNorm}(x + F(x))$$

By adding the input back to the output of each sub-layer, residual connections help mitigate the vanishing gradient problem and allow for deeper networks. Layer normalisation, in turn, normalises the outputs of each sub-layer, ensuring consistent magnitudes of activations and improving convergence during training.

transformer architecture The Transformer consists of *encoder* and *decoder* stacks, each composed of multiple layers of the previously described components. The encoder processes the input sequence, while the decoder generates the output sequence, often for tasks such as translation or text generation.

Each encoder layer consists of two main sub-layers: multi-head self-attention and FFN. In contrast, each decoder layer includes an additional multi-head attention sub-layer, where the decoder attends to the encoder's output. The overall architecture of the encoder and decoder can be summarised as:

$$\text{EncoderLayer} = \text{MultiHeadSelfAttention} + \text{FFN}$$

$$\text{DecoderLayer} = \text{MultiHeadSelfAttention} + \text{MultiHeadEncoderAttention} + \text{FFN}$$

These layers are stacked N-times (usually 6 or 12) to form the full Transformer model, where each layer builds progressively more abstract representations of the input sequence. The decoder's ability to attend to the encoder's outputs enables the model to generate sequences in an autoregressive fashion.

training and optimisation The Transformer is trained using a variation of the *cross-entropy loss*, where the output is compared to the ground truth sequence. To stabilise training, the authors of the original paper employ *label smoothing*, a technique that prevents the model from becoming overconfident in its predictions.

For optimisation, the *Adam* optimiser is used with a warm-up learning rate schedule, where the learning rate increases linearly for the first 4000 steps and then decays proportionally to the inverse square root of the step number. This strategy ensures that the model trains smoothly without significant oscillations in the loss function.

2.2.2 *Reinforcement Learning (RL)*

Reinforcement Learning (RL) is a machine learning paradigm where an agent learns optimal actions through interactions with an environment by trial and error (as shown in Figure 2.10). Unlike supervised learning, RL operates without labelled data, making it suitable for dynamic and uncertain environments. The goal of the agent is to maximise cumulative rewards over time, which is essential for decision-making tasks such as game playing, robotics, and control systems. This review outlines the foundational concepts, methodologies, and mathematical formulations behind RL, with an emphasis on Markov Decision Processes (MDPs), value functions, and key algorithms. [175].

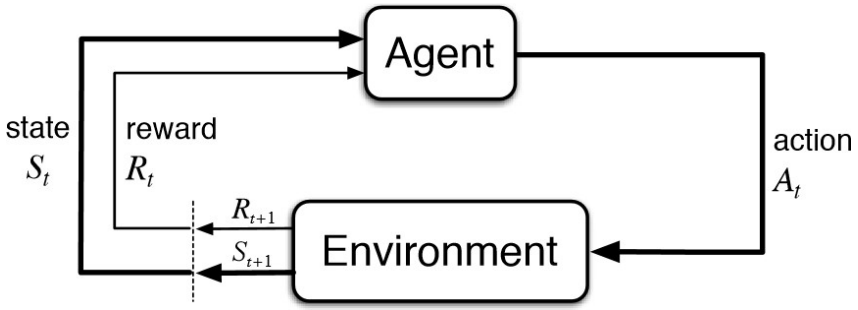


Figure 2.10: A general framework of RL [175].

Markov Decision Process At the foundation of RL is the Markov Decision Process (MDP), a mathematical model for decision-making problems where outcomes are partly random and partly under the control of the decision-maker. An MDP is defined by the tuple (S, A, P, R, γ), where S is the set of states, A is the set of actions, $P(s'|s, a)$ denotes the probability of transitioning from state s to state s' given action a, $R(s, a)$ is the reward function, and $\gamma \in [0, 1]$ is the discount factor that determines the importance of future rewards.

The agent's objective is to find a policy $\pi(a|s)$ that maximises the expected cumulative reward over time, also known as the return:

$$G_t = \sum_{k=0}^{\infty} \gamma^k r_{t+k+1}$$

Here, r_{t+k+1} represents the reward at time step $t + k + 1$, and γ^k discounts future rewards to ensure that immediate rewards are more significant.

Value Functions and Bellman Equations In RL, value functions are critical for evaluating the quality of states and actions. The state-value function $V^\pi(s)$ gives the expected return from state s following policy π:

$$V^\pi(s) = \mathbb{E}_\pi [G_t \mid s_t = s] = \mathbb{E}_\pi \left[\sum_{k=0}^{\infty} \gamma^k r_{t+k+1} \mid s_t = s \right]$$

Similarly, the action-value function, $Q^\pi(s, a)$, estimates the expected return for taking action a in state s:

$$Q^\pi(s, a) = \mathbb{E}_\pi [G_t \mid s_t = s, a_t = a]$$

The Bellman equations provide a recursive relationship between the value of a state and the value of subsequent states. For the state-value function:

$$V^\pi(s) = \sum_a \pi(a|s) \sum_{s'} P(s'|s, a) \left[R(s, a) + \gamma V^\pi(s') \right]$$

For the action-value function:

$$Q^\pi(s, a) = R(s, a) + \gamma \sum_{s'} P(s'|s, a) \sum_{a'} \pi(a'|s') Q^\pi(s', a')$$

These equations are essential for policy evaluation and optimisation.

Reinforcement Learning Algorithms One of the most prominent algorithms in RL is Q-learning, which is a model-free approach aimed at learning the optimal action-value function, $Q^*(s, a)$. The update rule for Q-learning is as follows:

$$Q(s, a) \leftarrow Q(s, a) + \alpha \left(r + \gamma \max_{a'} Q(s', a') - Q(s, a) \right)$$

Where α is the learning rate, r is the reward, and s' is the next state. This algorithm is off-policy, meaning it can learn the optimal policy regardless of the current policy being followed.

Another key algorithm is SARSA (State-Action-Reward-State-Action), which is an on-policy method. SARSA updates the action-value function based on the actual action taken by the agent:

$$Q(s, a) \leftarrow Q(s, a) + \alpha \left[r + \gamma Q(s', a') - Q(s, a) \right]$$

In contrast to Q-learning, SARSA follows the current policy when updating the action-value function, making it more sensitive to exploration strategies.

Actor-Critic Methods Actor-Critic methods combine the advantages of value-based and policy-based approaches. The actor selects actions based on a policy, while the critic evaluates the actions using a value function. The policy is updated using the gradient of the advantage function A(s, a), which measures how much better an action is compared to the expected action:

$$A(s, a) = Q(s, a) - V(s)$$

The actor's policy is updated by:

$$\theta \leftarrow \theta + \alpha \nabla_\theta \log \pi_\theta(a|s) A(s, a)$$

Here, θ represents the parameters of the policy, and α is the learning rate for the actor.

2.2.3 *Multimodal learning*

Multimodal learning aims to integrate data from different modalities (such as images, text, audio, video, etc.) to enhance a model's understanding of visual representation [130, 162, 169]. Traditional methods of visual representation typically rely on single-modal information, such as extracting features solely from images. However, this approach can result in missing information or modality-specific issues, limiting the model's ability to fully capture the multidimensional aspects of complex scenes. By introducing multimodal learning, information can be obtained from various perspectives, enabling the model to develop a more comprehensive understanding and representation of input data. By merging complementary information from different modalities, the limitations of single-modality methods are mitigated. In real-world tasks, many problems are inherently multimodal. For example, images and text together provide both visual and semantic information, while combining images and audio offers a synergy between visual and auditory perception. By integrating information from these modalities, models can learn richer and more distinctive representations. As a field of research, multimodal learning has been developing for a long time, well before 2012. However, significant advancements in multimodal learning have indeed accelerated since around 2012, primarily due to the rapid progress in deep learning and the availability of large-scale datasets across various modalities.

Before 2012, early work on multimodal learning involved statistical methods for integrating information from different data types, such as combining audio and video for tasks like speech recognition. However, with the rise of deep learning in 2012, starting with breakthroughs like the AlexNet model for image recognition, there has been a marked shift toward using neural networks to learn representations across multiple modalities in an end-to-end manner. This

allowed for the effective integration of different types of data, and has achieved great success in downstream tasks in vision and language [54, 81, 145], including visual recognition, caption generation, and more. Such models can not only make use of the visual features of images, but also enhance the understanding of images with the help of semantic information in text. Therefore, the field of multimodal learning for visual representation has undergone significant development, with various methodologies and applications explored. In particular, the core areas of interest include multimodal representation, modality mapping, alignment, multimodal fusion, and co-learning, all of which have been increasingly integrated with deep learning techniques.

Multimodal representation learning focuses on translating multiple modalities, such as images, text, and audio, into a unified feature space. Early approaches in this area, like the work by Zhang and Wang [208], propose a curriculum learning method for semi-supervised image classification using multimodal data to strengthen the robustness of visual recognition systems. This method leverages the complementary strengths of different modalities to improve the accuracy of image classification. Moreover, models such as CLIP, introduced by [146], present a major leap in integrating vision and language models, allowing image and text pairs to be represented in a shared latent space, thus greatly enhancing tasks like zero-shot classification and image retrieval.

Modality mapping has also seen progress, with studies focusing on mapping information from one modality to another. This is critical for tasks like image captioning, where textual descriptions are generated from visual inputs. Stefanini, Cornia, and Baraldi [172] conducted a comprehensive survey on image captioning techniques, highlighting the role of pre-training on large-scale multimodal datasets to improve the semantic richness of generated descriptions. Similarly, Zhu et al. [214] explore prompt-based learning for unpaired image captioning, where visual inputs are mapped to textual outputs without requiring paired datasets, showing the flexibility of modality mapping techniques in different application scenarios.

Alignment, the task of identifying correspondences between elements in different modalities, remains a crucial area of research. In [207], the authors explore how visual, semantic, and view consistencies across different modalities contribute to enhanced classification performance. This form of alignment ensures that semantically related elements across modalities are correctly linked, thereby improving model interpretability and accuracy in cross-modal tasks. Additionally, approaches like the work of [164] on multimodal GANs for unpaired image captioning incorporate joint semantic optimization, aligning visual and textual features to generate high-quality descriptions even without paired datasets.

Multimodal fusion, the process of integrating information from different modalities to create a comprehensive representation, has been extensively studied. For instance, the study by Wang, Li, and Zhu [191] on few-labeled visual recognition for self-driving applications uses a combination of visual and semantic representations to improve the robustness of recognition systems. This fusion of modalities ensures that the learned representations are more informative, leading to improved performance in downstream tasks like object detection and scene understanding. Furthermore, the integration of multimodal features has been shown to improve medical image analysis, as demonstrated by Dhanalakshmi, Maanasaa, and Maalikaa [33], where multimodal MRI data are fused to detect and classify brain tumors more accurately.

Co-learning, which involves transferring knowledge between modalities, has gained prominence in applications where one modality lacks sufficient data. A notable example is the use of multimodal learning in zero-shot learning, where models trained on one modality can generalize to tasks involving another modality. The work of Zhang, Cheng, and Tian [206] on multimodal distance metric learning showcases how the combination of visual and semantic information can improve performance in zero-shot recognition tasks. Co-learning has also been applied in self-driving systems, where visual and semantic features are jointly learned to enhance decision-making in autonomous vehicles.

2.3 image captioning

In today's life, images are everywhere, we can find images in almost every webpage like news sites, in mobile apps typical like social media platforms, and in other important areas such as the medical field which uses images to record and analyse disease. Images are regarded as a common way to convey information because they are more intuitive and more attractive to readers compared with other ways, and a more important reason is that humans are good at decoding image information. However, the scale of image data on the Internet is incredibly increasing in recent decades, and it could have great impacts if machines can understand the meaning of different images and interpret some forms of image captions if humans need automatic image captions from it. For instance, image captioning can be applied for automatic image indexing in a large image dataset, and also can help impaired people better understand the content of images on the Internet.

Image captioning is a research hotspot of machine learning, which is a combined task of computer vision and natural language processing, where the caption not only needs to recognise both the objects as well as their attributes but also expresses the relationships between

each object in an image, and then outputs a description in a natural language. See Figure 2.11 for some image captioning examples. The in-depth exploration of image caption shows a great potential to facilitate the research on the visual understanding of machine, and therefore, this task has been widely investigated in recent years.

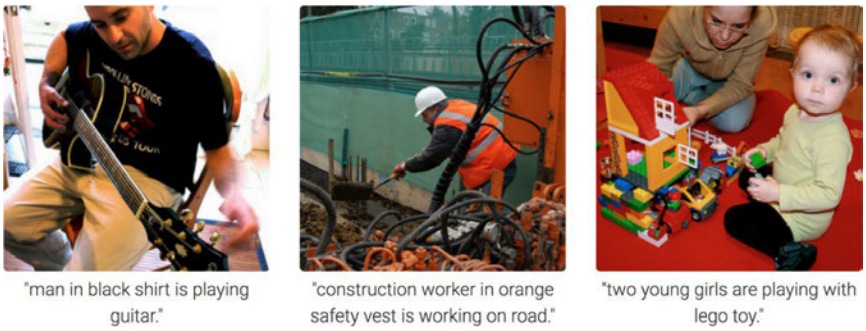


Figure 2.11: Image Captioning Examples

In the early stage, the image captioning methods were developed based on traditional machine learning. For example, Farhadi et al. [39] extracted image features by some handcrafted features like Histogram of Oriented Gradients (HOG) and used Support Vector Machines (SVM) to classify the objects, in the end, they attempted to mapping the image features to the sentence scene. Kulkarni et al. [99] employed a Conditional Random Field (CRF) to infer objects, attributes, and prepositions, which are then used to fill in the gaps. In another side, these methods have in common with rules/template-based approaches. After extracting the image features, the descriptions are generated by fixed language rules or filled to the template slots which are artificially defined for generating image captions. However, both the fixed rules and templates for generation are task-specific. In addition, extracting features from a large and diverse set of data is not feasible, which makes these methods hard to be applied to complex problems in the real world.

State-of-the-art captioning models predominantly rely on deep machine learning techniques, enabling automatic feature learning from training data and the ability to process large, diverse image datasets. For instance, CNNs are commonly employed for image feature extraction, while classifiers using SoftMax measurements are applied for categorisation. RNN are commonly used to generate captions. The following two sections will review deep learning-based methods for image captioning, which are divided into two categories: (1) supervised image captioning and (2) unsupervised image captioning.

2.3.1 *Supervised Image Captioning*

In the past few years, there has been a growing number of publications focus on image captioning. Since the "Show and Tell" study in [189] proposed the first study to attempt an end-to-end neural network on image captioning (see Figure 2.12 for details on the model structure), which consists of an encoder that uses CNN to read the input image and an RNN decoder (Fig Figure 2.13) to generate sentence descriptions by translating the representations learned from the encoder, enormous pieces of researches followed this encoder-decoder method to improve the captioning performance. Then, the introduction of the attention-based model has taken this progress a big step forward. It is a fact that human visual systems watch images rather than catch the entire visual information at first glance, but naturally, focus on the salient regions to effectively scratch the information needed. Incorporate with such an attention mechanism, Xu et al. [198] proposed two attention-based image caption generators and also visually showed "where" and "what" the attention model focused on when reading an image. Similarly, Anderson et al. [6] proposed a combined bottom-up and top-down attention captioning model that enables the model to consider not only the desired objects but also other unexpected or salient stimuli. More recently, some notable work applied reinforcement learning (RL) methods for image captioning. For example, Rennie et al. [153] proposed a self-critical sequence training algorithm to optimize the test captioning metrics. Xu et al. [199] proposed a multi-level policy and reward within the RL framework conditioned on the captioning model, language metrics, and visual-semantic function for optimisation.

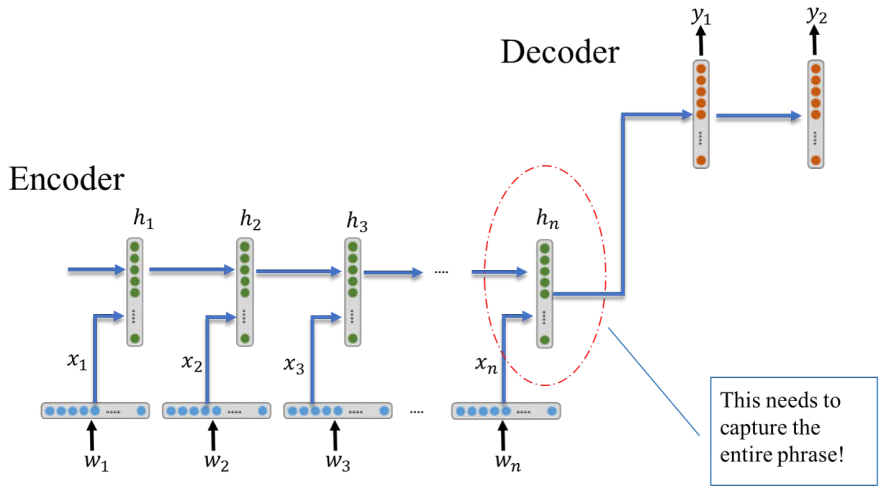


Figure 2.12: The overall encoder and decoder framework propsed in [189].

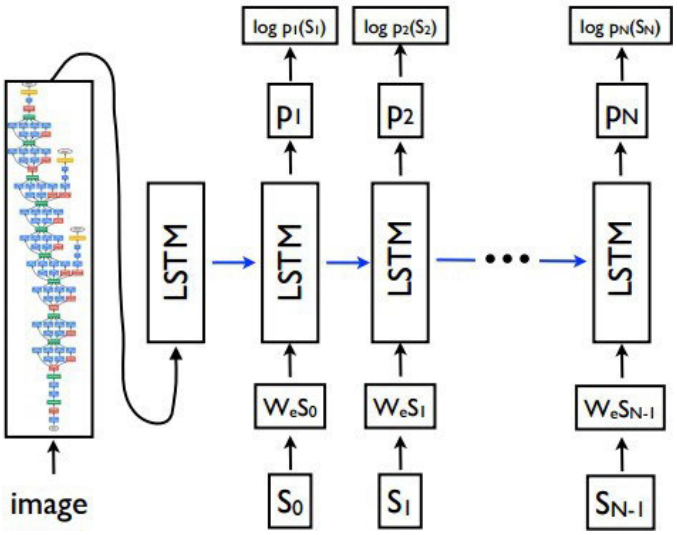


Figure 2.13: An image captioning model in detail, where a CNN encodes the image into features, and LSTM networks serve as the decoder to generate a descriptive caption for the image word by word.

2.3.2 *Unsupervised Image Captioning*

In terms of unsupervised image captioning in a step-by-step development, from 2019, Yang et al. [200] has introduced the inductive language bias from external dictionary corpus into the scene graph auto-encoder to help the model less focus on visual perception bias of the dataset. Similarly, Chen, Lucchi, and Hofmann [22] has put an approach that also used an external text corpus to establish a semi-supervised caption framework. However, they are still trained in an end-to-end supervised manner.

Indeed, the achievement for a good outcome of unsupervised image captioning is very challenging, but it is worth developing and train a purely but reliable unsupervised caption model with an independent image set and sentence corpus. The main superiority of unsupervised captioning includes reducing the reliance on labelled image-sentence datasets, which indicates the potential to dig out more information from the original input itself, compared with the supervised learning method. To date, several unsupervised models have been proposed for image captioning training. Similar to [200], Gu et al. [55] proposed their scene-graph-based learning method but for unpaired image captioning, in which the features that parsed from the image scene graph and the sentence scene graph are mapped to bridge the gap between unpaired images and sentences.

Thus, rather than leverage with a scene graph, some unsupervised approaches attacking the image captioning task with advanced stud-

ies of adversarial neural network. For example, Feng et al. [41] proposed a pure unsupervised image captioning model based on the GAN architecture and integrated it with object concept to generate captions that are more natural and difficult to be distinguished from real sentences in the external corpus. Then, Laina, Rupprecht, and Navab [100] proposed a domain alignment model to translate image representations into the embedding space learned from the language model in the image by leveraging visual concepts, and eventually decode these embeddings into descriptions. To move further, Guo et al. [56] adopted Recurrent Relational Memory Network (R2M) to learn the potential temporal correlation between objects on certain text corpus, and then fine-tuned the network to make much more appropriate sentences about the visual context in the image. Although the network has considered the temporal information about relationships between objects that are dug most from corpus sentences, it might miss the intrinsic spatial sides of images crucial for relationship prediction. However, in our work, we built from [41] and used the adversarial text generation method to achieve unsupervised training, and the critical improvement of our study is that we integrated the object and relationship information for relation reward function design, instead of involving in object concept knowledge only.

2.4 brain-inspired visual representations

In the human brain, the processing of visual information does not solely rely on individual visual inputs. The brain often integrates multiple sensory inputs, such as visual, auditory, and tactile information, combining them with environmental context and past experiences to achieve complex information integration and understanding. This multi-sensory information fusion allows the brain to make accurate perceptions and decisions, even when faced with ambiguous, noisy, or incomplete information. This multimodal characteristic of information processing has inspired research in machine learning. Brain-inspired multimodal learning of visual representation integrates information from various sensory modalities, such as brain signals and visual data, to develop systems that simulate how the brain processes visual stimuli, enhancing the ability of machines to understand and reason in complex environments. This approach utilizes techniques such as electroencephalography (EEG), which records brain activity with high temporal resolution, and applies machine learning methods to decode visual neural information from these signals. By combining EEG data with visual inputs, there is potential to create models that represent how the brain interprets and responds to images, thereby improving the accuracy of machine learning models in tasks like object recognition or visual classification. In this chapter, we discuss the inspiration provided by the brain's visual system for artifi-

cial intelligence and the application of EEG signals in decoding visual neural representations.

2.4.1 *Brain and Artificial Intelligence*

For humans, the benefit of visual exteroceptive perception is evident. For instance, individuals typically focus directly on objects they wish to identify, effectively utilising their foveal vision. Through years of research, neuroscience has revealed many neural foundations of brain functions, especially in understanding how the visual system processes information. For example, the studies by David H. Hubel and Torsten Wiesel, Nobel laureates in Physiology or Medicine in [75, 76] (1981), showed that information processing in the brain's visual system is hierarchical. The brain works through a continuous, iterative, and abstract process. After the retina receives the raw visual stimulus, the primary visual cortex (V1) first processes the information, extracting edge and orientation features. Subsequently, the V2 area further abstracts the information to identify contours and shapes. Through this iterative abstraction process, higher levels of the brain finally convert the information into visual perception of images, objects, and scenes [40], as shown in Figure 2.14. The cognitive behavior of the brain involves the collaboration of multiple neural circuits. Local neural circuits are responsible for processing specific functions, while long-range neural circuits coordinate between different brain regions. For instance, the interaction between the prefrontal cortex and the hippocampus is crucial for memory formation and decision-making processes. The collaboration of neural circuits across different brain regions allows humans to handle complex cognitive tasks such as language comprehension and problem-solving. Research into the higher functions of the human brain is not only significant for diagnosing and treating brain diseases but also serves as an inspiration for developing brain-inspired artificial intelligence technologies.

Since its inception, the field of artificial intelligence (AI) has aimed to simulate, extend, and enhance human intelligence. In recent years, many significant advancements in AI reflect a growing trend: even partial inspirations from neuroscience can significantly improve the intelligence of existing AI models and systems. Deep neural networks (DNNs), for example, are designed based on analogies to the brain's hierarchical processing. Particularly, CNNs mimic the hierarchical structure of the visual cortex, processing visual inputs layer by layer. Each network layer extracts increasingly abstract features, from edges to complex shapes, leading to the recognition of high-level objects or scenes. This characteristic of layer-by-layer processing has enabled CNNs to excel in visual information processing, particularly in tasks such as image classification and object detection [48, 61, 62, 97]. This neuroscience-inspired design provides a solid theoretical foundation

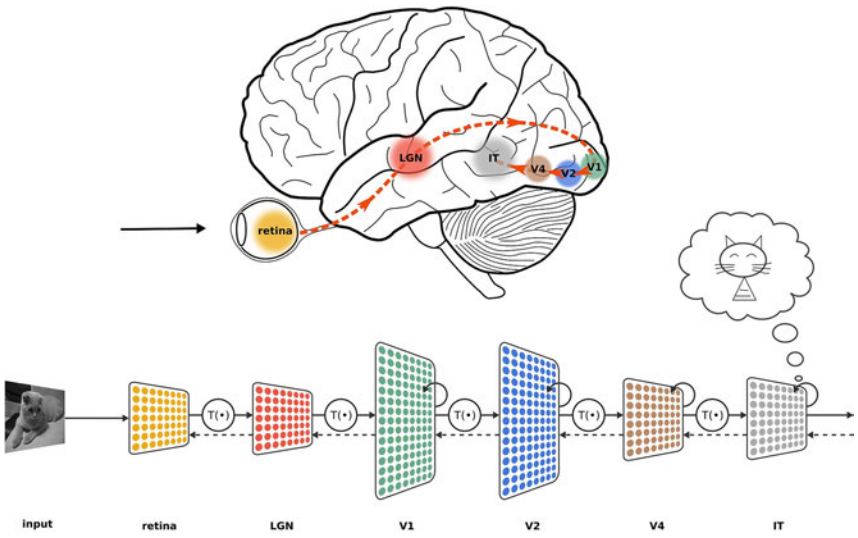


Figure 2.14: Hierarchical visual processing in the brainis is hierarchi- cal– one layer feeds into the next, computing progressively more complex features. This is the inspiration for the "layered" design of feed-forward neural networks. Source: https://doi.org/10.6084/m9.figshare.106794.v3

for AI models. Hassabis et al. [58] discussed how the brain's learning mechanisms, memory systems, and hierarchical processing models have inspired the design of modern deep neural networks. For instance, the backpropagation algorithm in DNNs corresponds to synaptic plasticity in the brain. Synaptic plasticity refers to the adjustment of the strength of connections between neurons based on experience and learning, which is analogous to how the backpropagation algorithm in DNNs adjusts network weights to gradually optimize the processing of input data [111]. However, some researchers have recognised notable differences between how humans and current CNNs process visual object information [46]. Specifically, studies evaluating CNNs on negative images [73] and comparing their generalisation to previously unseen distortions with human performance [46] have demonstrated that CNNs still fall short of human-level robustness in object recognition. Therefore, achieving AI that truly approaches or surpasses human-level intelligence requires deeper exploration and learning from the brain's information processing mechanisms. A crucial aspect of this research is understanding how brain activity patterns are linked to complex human brain functions such as perception, cognition, emotion, and memory. To establish these links, we need to leverage physiological signals.

2.4.2 *Physiological Signals-EEG*

In the field of neuroscience, the collection of physiological signals refers to the process of recording and measuring brain activity using various devices (see figure below). These signals reflect the brain's physiological state and neural activity while performing different tasks. Techniques for acquiring neural signals are generally divided into invasive and non-invasive methods. Invasive methods involve directly implanting electrodes into the brain to capture detailed neural data, while non-invasive methods record brain activity and metabolic changes through devices like EEG or functional magnetic resonance imaging (fMRI). EEG, as a non-invasive method, records electrical signals from the brain's surface using electrode arrays and is one of the most widely used techniques for neural signal collection. This paper focuses on using EEG data to study the relationship between brain activity patterns and complex brain functions, such as visual perception and affective computing.

EEG works by placing multiple electrodes on the scalp to detect cortical electrical activity from the brain's neurons, primarily capturing voltage fluctuations with high temporal resolution, has been widely used in many brain-computer interface (BCI) studies in recent years [9, 147]. Electrodes are typically arranged in a standardized manner (such as the international 10-20 system) to record electrical signals from various brain regions. When neurons are active, the ionic currents produce electrical signals, which are conducted through the electrodes on the scalp and recorded. EEG headgear can collect signals in multiple non-overlapping frequency bands (e.g. Delta, Theta, Alpha, Beta and Gamma). This is based on strong in-band connectivity with different behavioural states [210] and different frequency bands present diverse features and patterns that reflect different neural functions and cognitive states. EEG is widely applied in studying various neural and cognitive tasks due to its millisecond-level temporal resolution, enabling it to capture rapid neural dynamics. The risk to the subject is very low compared to invasive and other non-invasive techniques that require high intensity magnetic field exposure. The figure below shows typical EEG data, including examples of EEG waveforms recorded during different cognitive tasks (e.g., visual tasks).

Due to its non-invasive nature, convenience, and low cost, EEG holds an irreplaceable role in many brain-computer interface (BCI) applications and is widely used in classifying cognitive tasks such as emotion recognition, motor imagery, and fatigue detection. For instance, in emotion recognition tasks, EEG is used as part of affective computing to analyze brain responses associated with emotional stimuli, effectively categorizing individuals' emotional states. Studies have shown that different emotional states, such as happiness,

sadness, or stress, trigger distinct EEG patterns. Murugappan, Murugappan, and Ibrahim [126] explored the relationship between EEG signals and emotions, discovering that EEG signals from the frontal and parietal regions show significant differences between emotional states. These signals can be further classified into discrete emotional categories using machine learning techniques, providing a powerful tool for emotion recognition. Motor imagery refers to the process of mentally simulating movement without actual physical execution. EEG is widely applied in BCI systems for motor imagery tasks, allowing users to control external devices, such as robotic arms or computer cursors, using brain signals. Pfurtscheller and Neuper [139] conducted early research on EEG motor imagery classification, identifying characteristic brainwave patterns, particularly in the alpha and beta rhythms, associated with imagined movements. Fatigue detection is another important application of EEG, particularly in occupational and transportation settings. Prolonged cognitive or physical activity can lead to brain fatigue, which can be detected through changes in alpha, beta, and theta wave activities in EEG signals. [80] demonstrated that EEG can accurately detect fatigue states, which has applications in real-time monitoring of driver alertness and similar contexts.

2.4.3 *Visual Neural Representations*

Understanding how humans effortlessly achieve object recognition in daily life and modelling the mechanisms or discriminative feature spaces used for categorisation has long been a challenging topic in cognitive neuroscience [30, 53, 137]. Researchers have found that during visual cognitive processes, the brain performs functions such as feature extraction, shape description, and memory matching [45]. Further studies [57, 132, 160] have shown that analysing brain activity recordings in conjunction with the human visual system offers insights into how visual objects are represented in the brain's cortex. For example, hemodynamic responses have revealed that specific image categories (e.g., human faces and objects) are predominantly represented in specialised cortical regions [26, 59, 78]. Building on these findings, recent research has explored processing visual classification tasks by analysing neurophysiology and neuroimaging data recorded during human visual cognitive activities [2, 60, 84, 106, 166]. These studies have demonstrated the potential to identify the feature spaces humans utilise for visual representation. This makes us wonder if we can treat the human brain as a "high-level" image encoder and introduce human visual representation into multimodal computer vision models.

When the human brain processes visual information, it rapidly and efficiently encodes images, converting visual stimuli into neural ac-

tivity. This process is known as neural information encoding [68, 89, 128, 171, 185]. When a new image stimulus is provided, it is possible to predict the brain's response based on this encoding relationship. Figure 1.3 illustrates the mapping between external stimulus space, brain response space, and semantic space. Early computational neuroscience research focused primarily on exploring how the brain processes neural information from an encoding perspective, but more recent studies have shifted towards decoding neural information by interpreting brain activity [21, 128]. Specifically, decoding visual neural representation involves interpreting and mapping the feature space used by the brain for visual classification, achieved by analyzing physiological signals (e.g., EEG) that record brain activity. Researchers apply machine learning algorithms and neural networks to extract specific neural patterns related to visual stimuli from these signals, thereby decoding the neural representations formed during visual perception. Current research on neural information decoding can be divided into three categories, depending on the complexity and level of decoding: classification, retrieval, and reconstruction. Classification involves determining which category a presented image belongs to among known categories. Retrieval focuses on identifying which specific image from a known set is being viewed. Reconstruction aims to accurately reconstruct the presented image based on brain signals. This paper will focus on decoding visual neural representation through classification tasks.

Compared to other techniques like fMRI and magnetoencephalography (MEG), EEG provides a significant advantage in temporal resolution, allowing real-time tracking of the brain's responses to visual stimuli. This makes EEG particularly effective for tasks requiring high temporal precision in visual stimulus processing. While fMRI offers high spatial resolution, its lower temporal resolution limits its ability to capture rapid brain responses to transient visual events. In contrast, EEG can trace brain activity on a millisecond scale, capturing the early neural processes involved in visual stimulus processing. A qualitative difference between EEG signals evoked, on one subject, by visual stimuli of two different object classes is shown in Figure 2.15. Since EEG can record the brain's neural activity in real-time during visual processing, using this data to study how the brain encodes external stimuli and decode visual information is valuable for developing brain-inspired computational models that enhance machine perception. In this study, we will focus on decoding visual neural representation by classifying EEG signals.

2.5 decode visual neural representation with eegs

In the experiments to decode visual neural representations with EEG, researchers present visual stimuli (e.g., natural images or videos) to

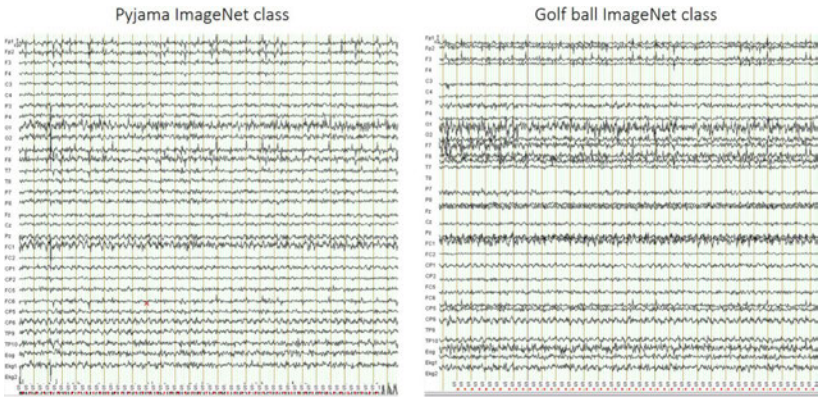


Figure 2.15: Examples of brain signals evoked by visual stimuli of two different ImageNet object classes [166]. This study utilised 72 images, with twelve selected from each of six categories: human body, human face, animal body, animal face, fruits and vegetables, and inanimate objects.

elicit brain responses, and simultaneously record the corresponding EEG signals. These visual stimuli are typically designed so that each stimulus corresponds to a specific category of visual information. For example, in [85], participants were shown a set of 72 images from six object categories (see Figure 2.16), while their brain responses were recorded. Participants viewed a series of images in a controlled environment, with the images ranging from simple shapes to more complex objects. During each image presentation, EEG devices recorded the participants' brain signals, and the authors analysed the signals using classification algorithms to decode the brain activity and identify the category of images the participants were viewing.

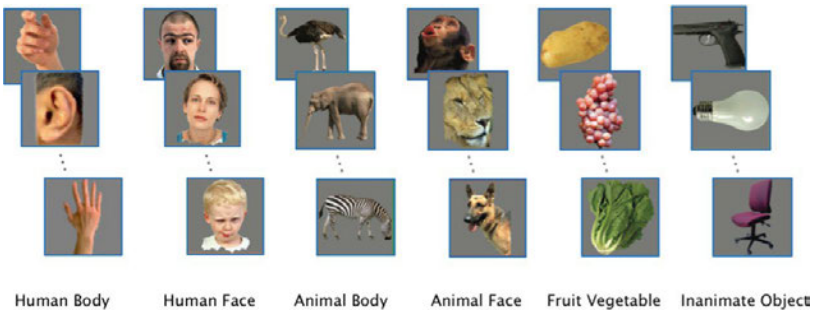


Figure 2.16: Examples of visual stimulus used in [85].

Subsequently, Spampinato et al. [166] used a deep learning approach to decode EEG signals. In their study, participants were shown 2,000 object images from 40 categories of the ImageNet dataset while their EEG signals were recorded. The researchers employed a CNN to process these EEG signals and trained the model to classify the visual objects viewed by participants. Their study demonstrated the

potential of deep learning in visual decoding tasks, especially for classifying complex neural signals.

These experiments highlight several key steps in using EEG for visual stimulus decoding: first, a carefully designed experimental environment ensures that each visual stimulus elicits a specific brain response; second, EEG devices record these responses in real-time; and finally, various machine learning models are employed to decode visual representation from EEG signals. These models typically rely on the extraction of relevant features from EEG data and the application of classification algorithms to distinguish between different visual stimuli.

Generally, EEG data is often contaminated by noise and artifacts, making it crucial to preprocess the raw data before feature extraction. Moreover, appropriate feature extraction and classification methods are essential to effectively decode visual information from the EEG signals. In this section, we discuss the standard preprocessing steps used in EEG analysis and four major feature extraction methods: time-domain analysis, frequency-domain analysis, time-frequency analysis, and brain network analysis. A variety of deep learning models for EEG classification are also discussed here.

2.5.1 *Preprocessing*

EEG data preprocessing is a vital step to remove artifacts, improve signal quality, and prepare the data for analysis. Preprocessing typically involves several stages, as described below:

Filtering Filtering is one of the most fundamental preprocessing steps in EEG analysis. Band-pass filters, typically between 0.1 Hz and 40 Hz, are used to remove low-frequency drifts and high-frequency noise, such as power line interference (50/60 Hz). High-pass filtering removes drifts, while low-pass filtering eliminates muscle artifacts and other high-frequency noise [194]. FIR (Finite Impulse Response) and IIR (Infinite Impulse Response) filters are commonly used to ensure that the EEG data focuses on frequencies that reflect neural activity relevant to decoding visual stimuli.

Electrode Localization and Re-referencing Proper electrode placement is based on international standards like the 10-20 system to ensure accurate and consistent data across subjects. Re-referencing is a common step to reduce noise, such as that coming from the reference electrode itself. The most common approaches include average reference (using the average of all electrodes) and linked mastoid reference.

Independent Component Analysis (ICA) ICA is a powerful tool used to separate independent sources from EEG data, such as eye blinks and muscle activity, that can contaminate brain signals [31]. This method decomposes the EEG signal into statistically indepen-

dent components, allowing the identification and removal of artifacts without losing valuable neural information.

Down-sampling EEG is often recorded at a high sampling rate (e.g., 500-1000 Hz), but down-sampling to a lower rate (e.g., 128-256 Hz) can reduce the computational load without compromising signal integrity. After artifact removal, down-sampling reduces data size while preserving critical information [15].

Bad Channel Interpolation Certain electrodes may record poor-quality signals due to faulty connections or hardware issues. Interpolating the signals from neighboring electrodes can replace bad channels, which improves the overall quality of the dataset. Bad channels are identified based on high variance or noise level.

Segmentation EEG data is often divided into time-locked epochs corresponding to specific events, such as visual stimuli. This allows researchers to focus on event-related potentials (ERPs) or other temporally related phenomena [121]. Segmentation ensures that only relevant portions of the signal are analysed.

Baseline Correction Baseline correction removes slow signal drifts by subtracting the mean signal in a pre-stimulus period from each epoch. This adjustment ensures that variations in baseline activity do not interfere with subsequent analyses, providing a cleaner signal for feature extraction.

Artifact Rejection Despite ICA and other preprocessing steps, some artifacts may remain. Methods such as thresholding or visual inspection can help remove remaining noise from the data [82].

2.5.2 *Feature Extraction*

Once the EEG data has been preprocessed, feature extraction techniques are applied to decode the relevant neural information. Different methods focus on various aspects of the signal, such as time, frequency, or brain connectivity. In the context of decoding visual information, four main approaches are commonly used: time-domain analysis, frequency-domain analysis, time-frequency analysis, and brain network analysis.

Time-Domain Analysis Time-domain analysis focuses on the direct characteristics of the EEG signal over time, such as amplitude and latency.

Event-related potentials (ERPs) are a common example of time-domain features that can be extracted by averaging EEG signals in response to visual stimuli (Luck, 2014). The peak amplitude and latency of ERP components (such as P300 or N170) can reflect different stages of visual processing (Woodman, 2010). Interestingly, when humans are exposed to different visual or auditory stimuli, EEG signals can capture varied responses in the form of evoked potentials [45]. However, the raw EEG waveforms reflect the brain's sponta-

neous activity in its natural state, making it challenging to discern the event-related information embedded within visual cognitive processes [45]. To address this, the event-related potential (ERP) was introduced to isolate real-time evoked response waveforms triggered by stimuli (e.g., specific visual or motor activities). These responses typically exhibit lower amplitudes compared to spontaneous EEG signals [45] and are extracted by averaging multiple trials to enhance the signal related to the event. The ERP involves a signal amplification that adds up and averages specifically time-locked epochs that may present a lower signal-to-noise ratio (SN)[2] than those of the original waveforms (see Figure 2.17).

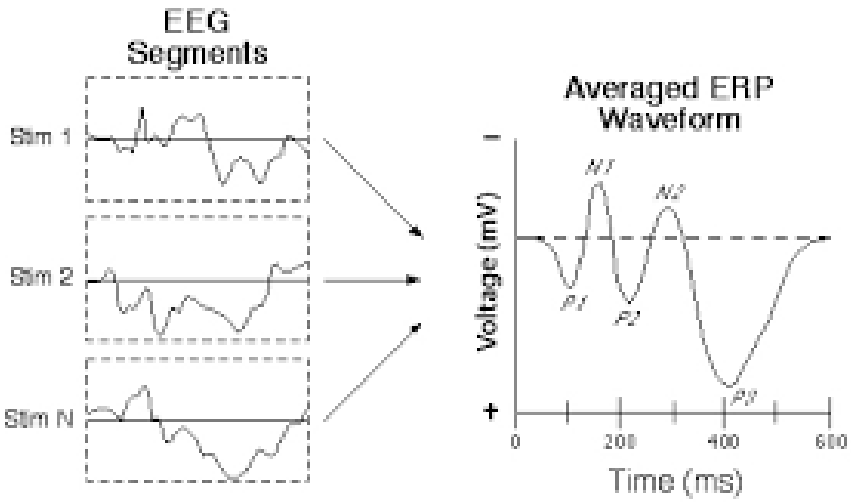


Figure 2.17: The averaging of several EEG segments into the ERP waveform..

Most current machine learning approaches for visual classification have largely overlooked the exploration of EEG-evoked potentials generated spontaneously. Even with advancements in deep learning, it remains challenging to identify distinctive patterns of evoked potentials directly from raw EEG waveforms associated with visual stimuli. This limitation suggests that neglecting visual-related evoked potentials may contribute to an uncertain feature representation space for visual classification, thereby restricting improvements in classification accuracy. Only two early-stage studies [144, 190] have preliminarily investigated evoked-guided models by extracting ERPs from EEG signals for visual classification. These studies utilised four ERP components (P1, N1, P2a, and P2b) as inputs for a linear classifier. However, they relied on extensive preprocessing involving multiple spatial and temporal components and did not incorporate deep learning to enhance evoke-guided feature representations.

Time-domain analysis also includes the use of statistical methods, such as principal component analysis (PCA), to identify relevant time-domain features [168].

Frequency-Domain Analysis Frequency-domain analysis examines the distribution of signal power across different frequency bands. EEG signals are typically divided into five primary frequency bands: delta (0.1-4 Hz), theta (4-8 Hz), alpha (8-13 Hz), beta (13-30 Hz), and gamma (30-100 Hz). The power spectral density (PSD) of each frequency band is computed to study brain oscillatory activity. Visual processing is particularly associated with changes in the alpha and gamma bands. For example, alpha suppression is often observed during tasks involving attention and visual processing [93]. Gamma-band activity, on the other hand, has been linked to visual object recognition and conscious perception [179]. The PSD can be computed using methods such as the Fourier Transform, Welch's method, or wavelet-based approaches. These methods provide insights into how neural oscillations in different frequency bands contribute to visual information processing.

Time-Frequency Analysis Time-frequency analysis bridges the gap between time and frequency domains by examining how frequency components evolve over time. Short-Time Fourier Transform (STFT) and wavelet transform are popular methods for time-frequency analysis, providing insights into transient changes in brain activity [181]. STFT uses a sliding window to compute the Fourier Transform across short segments of the signal, while wavelet analysis offers better resolution for higher frequencies, making it particularly useful for analysing rapid, transient events in visual perception.

Brain Network Analysis Brain network analysis explores the functional and effective connectivity between different brain regions. Functional connectivity measures such as coherence, phase-locking value (PLV), and weighted phase-lag index (wPLI) assess synchronization between EEG signals recorded from different electrodes, providing insights into how different regions of the brain coordinate their activity during visual tasks [42, 170].

Effective connectivity measures, such as Granger causality and dynamic causal modeling (DCM), go beyond correlation by assessing directed influences between brain regions, revealing how information flows during visual processing [173].

2.5.3 *EEG-based Visual Classification Models*

EEG-based visual classification models have attracted significant attention in recent years, particularly in brain-computer interface (BCI) systems used for tasks such as image, music, and speech recognition [11, 19, 174]. EEG signals, recorded through scalp electrodes, provide a non-invasive way to capture brain activity, making them valuable for decoding visual neural representations.

Early research sought to decode image-related EEG signals, aiming to classify brain responses to visual stimuli. For instance, Philiastides,

Ratcliff, and Sajda [140] and Philiastides and Sajda [141] worked on classifying single-trial EEG responses to photographs of faces and cars. Kaneshiro et al. [84] extended this with a linear classifier that processed EEG data related to six different object categories, achieving around 40% classification accuracy.

The integration of deep learning techniques, such as CNNs, significantly advanced this field. For instance, Bashivan et al. [11] transformed EEG signals into topographical maps and applied deep neural networks, improving classification performance. Khaleghi, Rezaii, and Beheshti [91] further enhanced this approach by developing a functional connectivity-based geometric deep network that uses Chebyshev graph convolution for EEG-based visual decoding, showing the effectiveness of multimodal representations in improving classification accuracy [91].

RNN and LSTM networks were also introduced to further enhance the accuracy of EEG-based visual classification. Spampinato et al. [166] developed an automated classification framework using RNNs to compute EEG features and a CNN-based regressor to map images onto these features. Despite its innovation, this framework faced criticism due to its block design, where presenting stimuli of the same class sequentially could inflate classification accuracies through temporal correlations rather than genuine brain activity [2, 106]. Subsequent research, such as [2], addressed this issue by adopting a rapid-event design with randomised stimulus presentations, resulting in more reliable classification outcomes. The application of LSTM models in these studies further demonstrated their ability to improve decoding accuracy of visual activities from EEG data. Zheng et al. [212] proposed an ensemble deep learning framework using LSTM-guided EEG visual classification, showing significant improvement by leveraging multiple networks for enhanced classification accuracy [212].

graph neural networks for eeg representation Traditional approaches to EEG analysis primarily relied on extracting temporal and spectral features using methods such as power spectral density [98, 143], common spatial patterns [136], fuzzy entropy [18], and wavelet transforms [12]. However, with the rise of deep learning, CNN and RNN-based models, such as EEGNet [102] and DepConvNet [155], have been increasingly employed to automatically extract high-level features from EEG data. These models capture both spatial and temporal characteristics, enhancing the performance of classifiers in a range of BCI tasks.

While CNNs and RNNs perform well on data with Euclidean structures, they are less suited for EEG data, which often involves non-Euclidean relationships between channels. Graph Neural Networks (GNNs) have proven to be more effective for this purpose, as they can model the complex inter-channel connections present in EEG

data. By constructing an adjacency matrix based on adaptive topologies or coherence measures, GNNs can represent the brain's network structure more accurately. These models have demonstrated superior performance in tasks like emotion recognition and motor imagery [74, 103, 165]. Ding et al. [34] introduced LGGNet, which uses local-global graph representations for EEG data to capture both local and global topological information, further enhancing EEG classification tasks [34]. Similarly, Jang, Moon, and Lee [79] developed a deep learning model that extracts graph structures from raw EEG data, improving classification by leveraging graph connectivity [79].

Granger causality (GC) is widely used to assess effective brain connectivity between EEG channels. For example, Varsehi and Firoozabadi [186] introduced a channel selection model based on GC correlations in a motor imagery task, while Manomaisaowapak, Nartkulpat, and Songsiri [122] applied GC to visual-evoked EEG potentials, characterising brain connectivity using state-space representation. However, these methods often require extensive parameter validation to extract causal connections. Recently, deep learning methods have shown promise in causal inference by automatically learning these connections, reducing the need for manual tuning [129].

multimodal learning with eeg Multimodal learning has emerged as a powerful technique to improve EEG-based classification by integrating data from various sensory inputs. This approach creates a shared representation of different modalities, such as visual and textual information, leading to improved performance in downstream tasks across fields like vision and language [54, 81, 130, 145, 162, 169]. One prominent method in multimodal learning is cross-modal contrastive learning, which aligns different modalities through contrastive loss, such as CLIP loss [145]. For example, Li et al. [104] applied contrast loss to align image and textual representations, thereby enhancing cross-modal attention and improving learning across both domains.

In EEG research, multimodal approaches have been used to decode visual stimuli from brain activity. Palazzo et al. [134] proposed a framework combining deep EEG and image encoders to optimise compatibility between visual stimuli and brain representations. Similarly, Ye et al. [201] developed a self-supervised cross-modal learning model to retrieve visual stimuli from EEG data. Peng et al. [138] advanced this field by developing the Multimodal Brain Graph Transformer (mBGT), which integrates EEG and fNIRS data using graph-based encoding to handle multimodal tasks effectively, highlighting the potential of graph learning in multimodal settings [138]. Furthermore, Chen and Wei [20] introduced the Mind's Eye framework, utilising a graph attention module for multimodal EEG-image contrastive learning, showing significant improvements in zero-shot EEG-based

image classification [20]. These models underscore the potential of multimodal learning to enhance the decoding of visual neural representations, leading to more accurate and robust classification systems.

3

ADVERSARIAL WORD-LEVEL MODELLING FOR MODALITY MAPPING BETWEEN IMAGE AND TEXT IN UNSUPERVISED IMAGE CAPTIONING

Our exploration into multimodal deep learning begins with a study that bridges the understanding of images and text, specifically through image captioning. This initial work aims to capture the intricate relationships between visual and language representations, drawing insights from both modalities. By focusing on the alignment of visual features and textual descriptions, this work lays the foundational framework for understanding broader multimodal interactions, crucial to the overall vision of decoding and interpreting neural and visual information.

3.1 introduction

In recent years, use of convolutional neural networks (CNNs) have led to significant achievements in pure computer vision tasks, yet it is still difficult for machines to fully understand images on a semantic level. Image captioning, as a combined task of computer vision and natural language processing, has been shown to be promising for translating complex visual perceptions into natural sentences [6, 72, 198, 200] and has demonstrated critical applications in semantic image retrieval [125] as well as assisting visually impaired people [196].

The objective of the image captioning task is to generate natural language descriptions for a given image. Traditionally, this task has been performed with supervised learning from paired image-sentence datasets [6, 28, 90, 120, 198, 200]. To reduce the dependence on extensive annotation datasets, several studies have attempted to train an unsupervised model without relying on any ground-truth caption samples. For example, Gu et al. [55] and Guo et al. [56] focused on beneficial global captioning at the sentence level by mapping the image and corpus text into a joint semantic space. However, the undesirable quantitative results they obtained implied a major challenge in bridging the two independent modalities purely through the shared global features under an unsupervised setting in a captioning task.

To address the issue raised above, there has been a trend towards weak supervision with the object concepts detected in the given image [41, 100]. However, these methods suffer from weak local alignment and corpus bias. Therefore, they fail to describe exact visual



Figure 3.1: An example of the four possible visual relations (A) "ride" (B) "hold" (C) "behind" and (D) "next to" between the same pair of objects: a person and bike.

relationships by referring only to the object features, and they may often yield words in the generated sentences that are visually inconsistent with the image. For example, in Figure 3.1, it is obvious that the same object categories appear in all instances, i.e., `person` and `bike`, but the "predicate" between these pairwise objects is different in each image. The captioner can generate various possible descriptions, such as are shown in (a) `the person is riding a bike`, (b) `the person is holding a bike`, (c) `the person is behind a bike`, and (d) `the person is next to a bike`. As the "captioner" chooses words by imitating the corpus, it is challenging to learn which is the correct sentence that contains the corresponding relationship to the image. This might be difficult when there is bias in the corpus; e.g., if `ride` is observed more often between `person` and `bike` than any other relations, the generator will tend to output mismatched captions, as by applying the relation in Figure 3.1(a) to the other three instances.

To overcome the aforementioned mismatch problem in unsupervised image captioning, we aim to promote complementary global and local alignment by focusing more on the visual relations between pairwise objects. Furthermore, a knowledge distillation approach is applied to capture the apparent relation features in the form of <*subject, predicate, object*>. This enables both the object and relation conception to be gathered from pairwise object interactions, which are crucial for building an image's contextual semantics [110, 204]. Subsequently, the recognised visual relation triplet form can be aligned between the visual relation and textual words without requiring grammatical rules to be followed as in image-text alignment, thereby facilitating more fine-grained and interpretable local matching to assist the global captioning training.

To enhance local alignment while retaining the global captioning capability, in this work, we proposed an adversarial word-level visual relation alignment model for unsupervised image captioning. Specifically, we adopted a generative adversarial network (GAN) as a global matching model trained on a reference sentence corpus. Then, to further guide the adversarial training and encourage its discriminator to distinguish between semantic and visually consistent captions, we

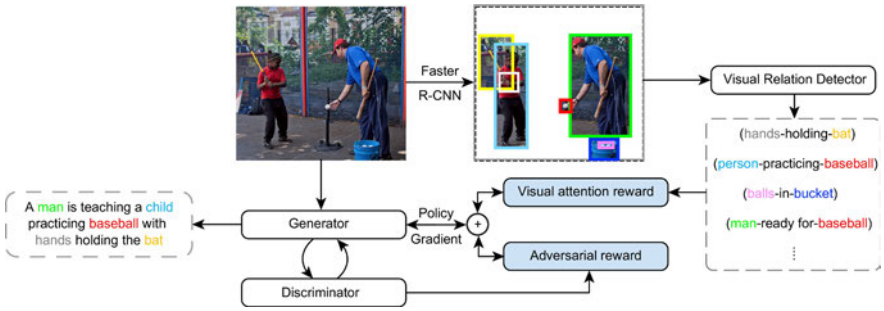


Figure 3.2: The developed adversarial and word-level visual relation alignment model for the unsupervised image captioning task. It consists of a sentence generator, a discriminator, and a visual relation detector. The GAN is formalised as an RL process that can jointly update the parameters by leveraging the visual attention reward estimated from the detected relations.

considered the whole unsupervised image captioning process as a reinforcement learning (RL) optimisation problem. Furthermore, we incorporated adversarial and visual attention rewards via a policy-gradient method to align the adversarial network between the distilled visual relation information and generated sentences. In this unsupervised setting, we implemented the proposed model to learn without the ground-truth captions of the input image and continually follow adversarial training to generate accurate captions until the reference sentences are indistinguishable. The model was extensively evaluated on two large-scale image captioning datasets, and the performance was compared with baselines and state-of-the-art methods.

The **main contributions** of this study can be summarised as follows:

- We developed an unsupervised image captioning model to explicitly learn from the complementary global and local word-level alignment between the image and text modalities.
- We developed adversarial and visual attention rewards that explore the matching degree between the detected visual relations and the generated sentence for each image to help the mapping process focus on the salient visual relation features.
- creating a prototype model which achieved competitive performance compared to existing unsupervised methods on two different prevailing corpus settings.

3.2 method

For the unsupervised image captioning task, the image dataset is defined as $I = \{I_1, ..., I_{N_m}\}$, and the extended sentence corpus is defined

as $E = \{E_1, ..., E_{N_s}\}$, where N_m and N_s are the total numbers of images and sentences, respectively. Note that I and E are not matched one-to-one. This task aims to build a model that generates natural descriptions of an image by learning from the sampled language corpus. As shown in Figure 3.2, our proposed model is trained based on GAN to bridge the connections between the different parsing modalities of I and E and integrated reward functions to support training and optimisation using an RL process.

3.2.1 *The adversarial network*

In the initial step, each input image I is encoded as a feature representation vector $\hat{I}$:

$$\hat{I} = E(I) \tag{3.1}$$

where E is a CNN image encoder: Inception-V4 [176].

Generator The generator **G** in the GAN decodes the obtained image feature $\hat{I}$ to generate descriptive words and construct them into a coherent captioning sentence. This is developed based on sequential long short-term memory (LSTM) cells [189], which applies the memory mechanism to share the same parameters at each step t among LSTM cells to reduce the training complexity. In particular, at the first step $t = -1$, the extracted image feature $\hat{I}$ is sent to the sequential LSTM as the initial input x_{-1}. Later, at each time step t, the LSTM accepts the predicted word from the previous LSTM cell as input with a one-hot word embedding z_t. The whole network is trained to predict the next generated word x_{t+1} by outputting a probability distribution p_{t+1} at time t. Then, x_{t+1} is sampled to a one-hot vector z_{t+1} from the counted vocabulary. The complete procedure will generate a captioning sentence $C = \{z_1, ..., z_{N_c}\}$ with word length N_c, which is defined as follows:

$$x_{-1} = \hat{I} \tag{3.2}$$

$$x_t = W_e z_t, \ t \in \{0...N_c - 1\} \tag{3.3}$$

$$p_{t+1} = LSTM^G(x_t), \ t \in \{0...N_c - 1\} \tag{3.4}$$

$$z_{t+1} \sim p_{t+1}, \ t \in \{0...N_c - 1\}$$

where x_{-1} is the initial input image feature $\hat{I}$ of the generator, W_e is the word embedding matrix, and $LSTM^G$ is the sequential LSTM cells of the generator. $\sim$ denotes the sampling process from the counted vocabulary. Note that we denote z_0 and z_{N_c} as the start word token <SOS>and the stop token <EOS>, respectively. These two words instruct the generator to start generating a sentence and to end a completed sentence.

Discriminator The discriminator **D** is developed to distinguish whether the generated sentence C is a real sentence from the corpus, which has the same structure as the generator with sequential long short-term memory (LSTM) cells. At each time step t, the LSTM accepts the t_{th} word from generated sentence C or corpus sentence E as input x_t with word embedding and predicts a high probability for real data and a low probability for fake data. The complete procedure can be defined as:

$$q_t = \mathrm{LSTM}^D(z_t),\ t \in \{0...N_c\} \tag{3.5}$$

where q_t is the output probability indicating that the word z_t sampled from x_t is judged as real by the discriminator and LSTM^D is the LSTM discriminator.

3.2.2 *The adversarial reward*

The sentences generated by the image captioning model must be plausible to human readers, which can be achieved through the adversarial network. The generator needs to take an image feature as the input and generate one sentence conditioned on the image feature. The discriminator determines whether a sentence was generated by the model or is a real sentence from the corpus. The generator tries to fool the discriminator by generating sentences that are as natural as possible. To achieve this goal, a reward is given to the generator at each time step that is defined as an adversarial reward. The reward value for the t_{th} generated word is the logarithm of the probability estimated by the discriminator:

$$r^t_{adv} = \log(q_t). \tag{3.6}$$

By maximising the adversarial reward, the generator gradually learns to generate plausible sentences. For the discriminator, the corresponding adversarial loss is formalised as:

$$L_{adv} = -\left[\frac{1}{N_s}\sum_{t=1}^{N_s}\log(\hat{q}_t) + \frac{1}{N_c}\sum_{t=1}^{N_c}\log(1-\hat{q}_t)\right] \tag{3.7}$$

3.2.3 *The visual attention reward*

To generate visually consistent image captions, the method must recognise the visual relationships in the image and involve this information in the generated sentence. Here, we distill the knowledge from a pre-trained visual relation detector to guide the image captioning model. The detector is adapted to detect and localise all objects presented in the input image I and to predict the relationships between any two

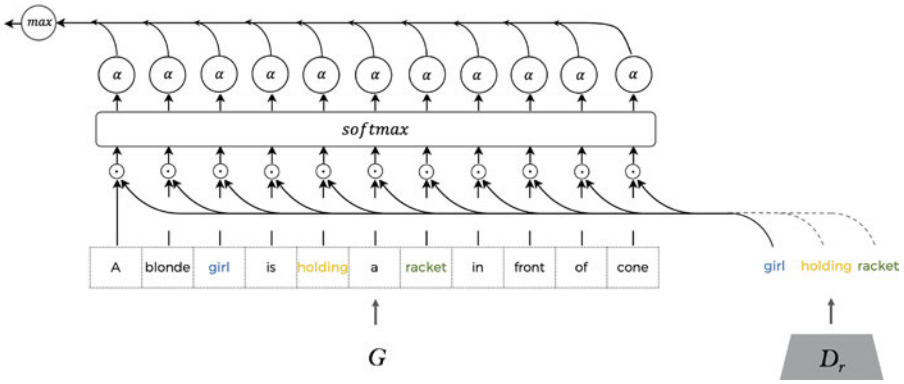


Figure 3.3: Illustration of the visual attention model. Each component of the detected visual relation from the visual relation detector (D_r) is subsequently fed into the model. We can extract the cumulative maximum visual attention reward by mapping the component word to the reference sentence words that is outputted from the generator (G).

of the observed objects. Generally, the predicted relationship is characterised in triplet form including the subject, predicate and object. Concretely, given an image I, the detection process starts with object recognition based on Faster R-CNN [152] to decode the objects in the image, which is formalised as:

$$[O_I, B_I] = \text{Faster R} - \text{CNN(I)} \tag{3.8}$$

where O_I and B_I are the observed object class distribution and the corresponding bounding box for the spatial location, respectively.

For each input image, the visual relation detector outputs a set of relationships $L = \{l_1, ..., l_i, ..., l_{N_r}\}$ to represent the relationships between every paired object. l_i denotes the i_{th} relation triplet of L in the form of $(s_i - p_i - o_i, c_i)$ and N_r is the total number of relationships. This complete process can be defined given as:

$$\mathbf{L} = D_r(O_I, B_I) \tag{3.9}$$

where D_r denotes the visual relation detector, Please note that "no relation" between two objects in L is considered "background" in the existing visual relation detector, but we excluded "background" to reduce the unnecessary calculation steps when evaluating the reward.

The visual attention reward is designed to guide the adversarial network in forming close local alignments. Specifically, it refers to a simple yet powerful function used to evaluate how the generated sentence C matches the detected visual relationships $\mathbf{L}$ at the word level. We characterise the mapping process as a visual attention model, and the illustration is shown in Figure 3.3. For each relation triplet l_i in L, We first arrange the discrete words from subject s_i, predicate p_i and object o_i, into a sequential fragment sentence $F = \{u_1, \ldots, u_{N_{l_i}}\}$,

where N_{l_i} denotes the total word count of l_i. Then the fragment sentence F will be input with the generated sentence C to the visual attention model, which can obtain a normalised attention weight α that maps to the generated sentence words. We encode this process as the scaled dot-product attention [187] to assign weights to specific input sequence as follows:

$$(\alpha_1^k, \alpha_2^k, \ldots, \alpha_{N_c}^k) = Softmax_{t=1}^{N_c}(\frac{1}{\sqrt{\sigma}} z_t \odot W_e u_k), u_k \in \{u_1, \ldots, u_{N_{l_i}}\} \tag{3.10}$$

where W_e is the word embedding matrix used in the generator, indicating that the attention weights will be calculated from the mapping distances of word embeddings to determine the similarity measure. $\odot$ is the elementwise product, and σ denotes the dimension of W_e as a constant scaling factor. Next, the key idea is to compute the maximum attention weights for each word in F and to give the reward as:

$$r_{va}^i = \sum_{k=1}^{N_{l_i}} \max(\alpha_1^k, \alpha_2^k, \ldots, \alpha_{N_c}^k) \tag{3.11}$$

We take the maximum weights of the different relationship components in order so that the "subject", "predicate" and "object" words will be presented sequentially in the generated sentence. After obtaining the first maximum weight of the "subject" and recording the corresponding word that maps to it, the next maximum weight of the "predicate" will be obtained by mapping to the rest of the sentence words, and the same process will yield the next computation for the "object". This rule is reasonable because the word order of "subject" and "object" is crucial for encoding a relationship from an intuitive point of view; e.g., < person − ride − bike > is completely different from < bike − ride − person >, and we should avoid matching these three detected concepts with a generated sentence repeatedly. Sometimes the generated sentence may present more than one relationship, which indicates that if the related subject or the object of two different relationships changes, the understanding of the image may not be accurate.

Finally, this attention process will be iterated from the first to the last relation triplet for an image, and the overall visual attention reward r_{va} for the given image will be:

$$r_{va} = \sum_{i=1}^{N_r} r_{va}^i \times c_i. \tag{3.12}$$

When the relation set is derived only from the ground-truth dataset, we define the reward $\hat{r}_{va}$ with $c_i = 1$ by default as:

$$\hat{r}_{va} = \sum_{i=1}^{N_r} i_{va}. \tag{3.13}$$

3.2.4 *Optimisation*

The adversarial reward and visual relation reward are jointly considered in training our proposed image captioning model. The training process takes the whole model as an RL policy and trains the generator using the policy gradient [175], and the gradients are estimated with respect to the trainable parameters given the combined rewards. The parameters of the generator and discriminator are updated alternately. Let θ be the trainable parameters of generator G; the objective function is given as:

$$\begin{aligned}\nabla_\theta L(\theta) = -\sum_{t=1}^{n} E[\sum_{s=t}^{n} \gamma^s(r^s_{adv} + \lambda_r r^s_{va}) + \lambda_{im} r^s_i - b_t) \nabla_\theta \log(s_t^T p_t)] \\ + \lambda_{sen} \nabla_\theta L_{sen}(\theta)\end{aligned} \tag{3.14}$$

where γ is the decay factor and b_t is the baseline reward computed using the self-critic method. r^s_i and $L_{sen}(\theta)$ are estimated according to the bidirectional image-sentence reconstruction, and λ_r, λ_{im}, and γ_{sen} are the hyperparametric weights of different terms.

In the training of the GAN network, the generator and discriminator are trained simultaneously. The objective of the discriminator D is to maximise the probability of correctly labelling both real sentences drawn from the reference corpus, and fake data generated by the generator G. In contrast, the generator aims to maximise the probability of fooling the discriminator into misclassifying the generated sentences as real data. Through this iterative process of dual optimisation, the generator and discriminator learn to improve their respective abilities to generate and distinguish between real and fake captioning sentences. The pseudo-code of this model is defined as:

3.3 experiments and results

In this section, we will first introduce the experimental settings and then show our evaluation details and the performance comparisons with existing unsupervised image captioning methods.

We ran the experiments as follows. First, we performed a quantitative analysis in two kinds of *unsupervised* settings with independent images and text corpora for a fair comparison with current methods.

Algorithm 1 Visual relation alignment by adversarial word-level modelling for unsupervised image captioning

Require: image I; corpus $E = E_{\{1:E_{N_s}\}}$; generator G; discriminator D; an image encoder E; visual relation detector D_r
Initialise G and D with random weights θ
Pre-train D_r
for number of training iterations **do**
  Encode image feature $\hat{I} \leftarrow \mathsf{E}(I)$;
  Generate a sentence $C = \{z_1, ..., z_{N_c}\}$ by Eq. 3.2-3.4;
  Predict probability $q_{\{1:N_c\}} \leftarrow D(z_{\{1:N_c\}})$
  Update D by minimising the loss:
  $\mathsf{L}_D \leftarrow -[\frac{1}{N_s} \sum_{t=1}^{N_s} \log(\hat{q}_t) + \frac{1}{N} \sum_{t=1}^{N_c} \log(1 - \hat{q}_t)]$ (refer to Eq. 3.7)
  Compute the adversarial reward r_{adv} by Eq. 3.6
  Compute the visual attention reward r_{va} by Eq. 3.8 - 3.13
  Update G by minimising the loss:
  $\nabla_\theta \mathsf{L}(\theta) \leftarrow -\mathsf{E}[\sum_{t=1}^{n} (\sum_{s=t}^{n} \gamma^s \, {}^{s}_{adv} + \lambda_r r^s_{va}) + \lambda_{im} r^s_i - b_t) \nabla_\theta \log(s_T^T p_t)] + \lambda_{sen} \nabla_\theta \mathsf{L}_{sen}(\theta)$ (refer to Eq. 3.14)
end for

Second, we evaluated the impact of different visual relation detectors in our unsupervised image captioning model. Then, we performed ablative analysis on the MSCOCO cross-Visual Genome (VG) dataset Krishna et al. [96] in an *unpaired* setting to evaluate the effectiveness of the components in the proposed model. Finally, the qualitative results are shown in the different settings.

3.3.1 *Experimental setup*

Dataset

Training image sets: MSCOCO [113] is a widely used image dataset for image captioning tasks that contains over 330,000 images across 91 object categories, with more than 200,000 of these images annotated with 5 captions, generated by human annotators. This large-scale dataset is specifically designed to reflect a wide variety of everyday scenes, with captions that describe not only individual objects but also the relationships between them within diverse contexts. Figure 3.4 displays three example images from MSCOCO along with their corresponding captions. We applied the split defined in Karpathy and Fei-Fei [86], with 113,287 images for adversarial training, 5,000 images for validation, and the remaining 5,000 images for testing. Additionally, we trained the visual relation detector on a VG dataset that provides ground-truth relationship triplets < object, predicate, subject > for each image and then filtered the objects and relationships by retaining the 150 most frequent objects and 50 most frequent predicates.

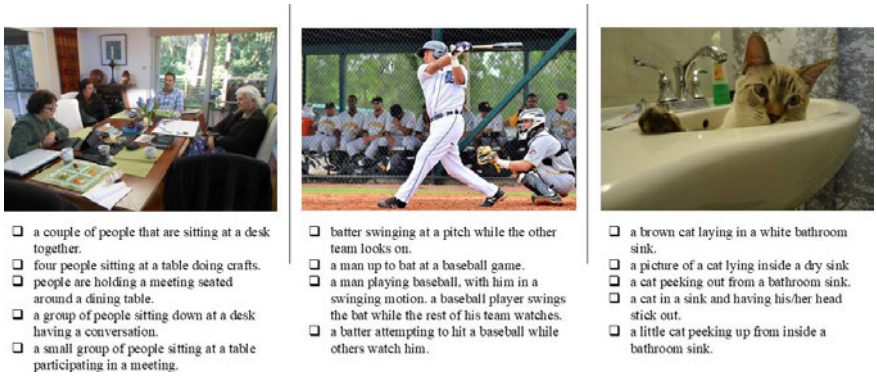


Figure 3.4: Three example images from MSCOCO with corresponding captions.

Training corpus sets: The current unsupervised approaches generally use two training corpora, that of Feng et al. [41] with MSCOCO captions and that of Laina, Rupprecht, and Navab [100] with Google's Conceptual Captions dataset (GCC); we evaluated our model on both corpora to compare each kind of method. This step guarantees that the corpus is related to the descriptive sentences of the images.

Evaluation metrics A single numerical score can evaluate how "good" image captioning methods are by comparing the provided reference sentences and the generated outputs. In this study, the widely used metrics BLEU (B1, B2, B3, B4) [135], METEOR (M) [32], ROUGE (R) [112], CIDEr-D (C) [188] and SPICE (S) [5] are used to evaluate the quality of the generated captions.

Implementation details In the preprocessing stage, the selected images were processed by filtering the objects and relationships and retaining the 150 most frequent object categories and 50 most frequent predicate categories, which followed the initialization guidelines proposed by Feng et al. [41] to tokenise the caption sentences and set the hyperparameters. Specifically, the weighting hyperparameters λ_c, λ_{im}, and λ_{sen} were set to 10, 0.2, and 1, respectively. γ was set to 0.9, and the Adam optimiser [92] was used with a learning rate of 0.0001. These steps could increase the visual relation's impact on the optimisation process.

Visual relation detector To assess the performance of our proposed model with visual relations, we used a basis detector, MOTIFS [204], to represent statistical visual relation detection methods. It is a scene graph parsing method with global context detection motivated by the statistical analysis of VG. Specifically, it uses a Faster R-CNN to detect and localise the objects in the input image, and then predicts the relationships between any two objects using a global context detection module. The predicted relationships are characterised in triplet form as <subject, predicate, object>, where subject and object are the object categories and predicate is the type of relation between them. The relation detector outputs a set of relationships with a confidence

score for each image and filters out the background relations that are trivial or less informative.

The visual triplet relationships were selected based on the maximum attention weights for each component word of the triplet in the generated sentence. The attention weights were calculated from the mapping distances of word embeddings between the triplet and the sentence, using a scaled dot-product attention function. The attention process iterates from the first to the last relation triplet for an image, and the overall visual attention reward was computed as the sum of the individual rewards multiplied by their confidence scores. By taking advantage of pretrained MOTIFS on VG, we employed the visual attention reward for optimisation, denoted as **MOTIFS RS (r_{va})**.

Table 1: Performance comparisons with existing unsupervised methods under the experimental setting of [41].

Method \ Metrics	B1	B2	B3	B4	M	R	C	S
Gu et al. [55]	46.2	24.0	11.2	5.4	13.2	-	17.7	-
Feng et al. [41]	58.9	40.3	27.0	18.6	17.9	43.1	54.9	11.1
Laina, Rupprecht, and Navab [100]	61.7	42.8	28.6	19.3	20.1	45.4	**63.6**	12.8
Shambharkar et al. [158]	60.3	41.1	27.8	18.9	18.2	43.3	55.2	-
Song et al. [164]	62.6	43.2	29.2	20.0	19.9	45.1	**63.6**	**12.9**
Honda et al. [69]	49.5	27.3	13.1	6.3	14.0	34.5	31.9	8.6
Gao et al. [44]	63.4	43.2	29.5	17.9	24.5	45.1	53.5	-
Ours:								
MOTIFS RS (r_{va})	**66.3**	**45.9**	**30.4**	**21.0**	**21.2**	**47.5**	60.9	**12.9**

3.3.2 *Main results*

To compare with the current state-of-the-art approaches, we tested the proposed model with two different corpora, as mentioned in Section 3.3.1, in an unsupervised setting. In the testing stage, for consistency with other models, the evaluation performance was tested on the split of the MSCOCO dataset.

Table 1 shows the performance across four existing unsupervised image captioning methods following the corpus setting of Feng et al. [41], and Table 2 lists the results of our model compared with three methods under the setting of Laina, Rupprecht, and Navab [100]. The results demonstrated that our proposed model outperformed the existing methods in most of the metrics, indicating the effectiveness and robustness of the proposed reward functions for visual relation alignment.

Table 2: Performance comparisons with existing unsupervised methods under the experimental setting of [100].

Method \ Metrics	B1	B2	B3	B4	M	R	C	S
Laina, Rupprecht, and Navab [100]	-	-	-	6.5	12.9	35.1	22.7	-
Guo et al. [56]	51.2	29.5	15.4	8.3	14.0	35.0	29.3	9.6
Honda et al. [69]	50.4	29.5	14.4	7.6	13.5	**37.3**	**31.8**	8.4
Ours:								
MOTIFS RS (r_{va})	**52.1**	**30.3**	**15.8**	**8.5**	**14.4**	36.0	30.3	**9.8**

Table 3: Ablation study on MSCOCO across the VG dataset test split. The adversarial network, object and relation detector are trained on the images and ground-truth relationships of the VG dataset.

Method \ Metrics	B1	B2	B3	B4	M	R	C	S
Init-GAN	45.1	24.5	11.1	5.7	12.4	30.6	34.8	8.2
Init-GAN + Adv	52.5	34.1	21.2	15.2	15.6	39.1	44.7	9.8
Init-GAN + Adv + Obj	56.4	38.9	25.1	17.2	16.8	42.7	54.5	11.0
Full	**64.9**	**44.1**	**27.8**	**19.0**	**19.5**	**43.4**	**58.8**	**11.7**

3.3.3 *Ablative analysis*

We applied four models and evaluated their performance based on the VG dataset with the corpus of Feng et al. [41] excluded using a visual relation detector. To enable the two image datasets to contain the ground-truth captions as well as relationships, we extracted a set of images that exist in both VG and MSCOCO for the adversarial network training, resulting in 46,000 images for training and 10,000 images for testing.

The four models are listed as follows: model (1) is initialized based on GAN without a reward, referred to as **(Init-GAN)**, which is used to verify the feasibility of the proposed initialization pipeline of the adversarial model [41]. Model (2) is based on GAN with an adversarial reward, referred to as r_{adv} **(GAN + Adv)**. Model (3) uses an object concept reward, referred to as **(GAN + Adv + Obj)**. It is a pretrained unsupervised image captioning model [41] on the VG dataset for the object concept reward. Model (4) uses the full model with relationship ground truth with $r\hat{}_{va}$, referred to as **(Full)**.

The results of the above four models are presented in Table 3. We observed that these four models generated reasonable results with high CIDEr scores, but "GAN" and "GAN + Adv" led to worse results. Their poor performance might be due to that the pure GAN sturucture or adversarial objective only enforces global sentence generation and does not ensure semantic correlation with the image con-

tent. Furthermore, we found that the full model with visual attention reward improved the average metric scores more than the object-concept reward model (GAN + Adv + Obj), which verified that our local relation-word alignment is beneficial for the connection between the image and sentence feature space, thereby proving to be promising in boosting the optimisation performance for unsupervised image captioning tasks.

3.3.4 *Analysis with different visual relation detectors*

Our prototype model for unsupervised image captioning aims to minimise the misalignment between visual content and generated sentences through the visual relation alignment, which is also designed to be compatible with commonly used detectors. Thus the effectiveness of the visual relation alignment is significantly influenced by the performance of different visual relation detectors. Some of the commonly used methods are based on statistical analysis that predicts the most frequent relationships in the dataset such as MOTIFS [204]. On the other hand, other research for relationship detection such as VRD [119] and NLG [109] relies on language priors derived from semantic word embeddings to predict the semantic connection between objects. Hence we want to know the impact of visual relation detectors in unsupervised image captioning tasks under Feng et al. [41] corpus setting. The results of the experiments are presented in Table 4. The findings reveal that our model with the MOTIFS detector outperforms the VRD and NLG detectors in all metrics, although the variation in performance among the detectors is relatively small. This suggests that the proposed visual relation alignment with visual attention reward is adaptable to various detectors.

Table 4: Performance comparisons of different visual relation detectors for our unsupervised image captioning model on the test split of MSCOCO.

Method	B1	B2	B3	B4	M	R	C	S
VRD RS (r_{va})	64.7	44.5	29.0	18.4	19.2	43.3	57.8	11.1
NLG RS (r_{va})	65.6	45.3	**30.8**	18.5	20.4	46.0	60.3	11.8
MOTIFS RS (r_{va})	**66.3**	**45.9**	30.4	**21.0**	**21.2**	**47.5**	**60.9**	**12.9**

3.3.5 *Qualitative analysis*

Figure 3.5 shows some qualitative captioning examples of captioning models. We present the results generated by four image captioning models (Only Obj, RS GT($\hat{r}_{va}$), MOTIFS RS(r_{va}) w/ MSCOCO, MOTIFS RS(r_{va}) w/ GCC) for three sampled images. For example, in

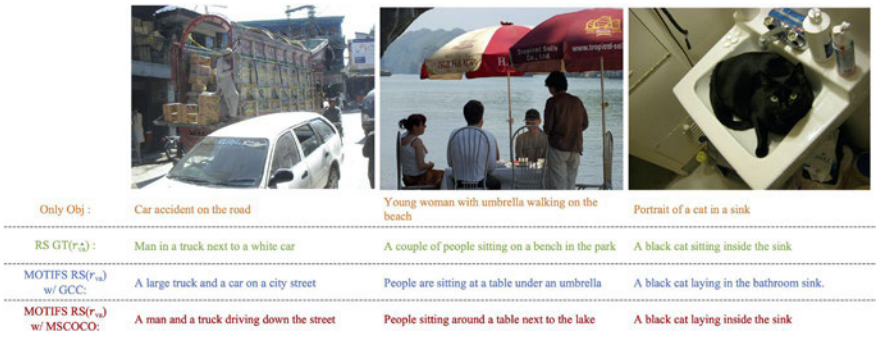


Figure 3.5: Qualitative captions generated from unsupervised image captioning models.



Figure 3.6: Illustrations of the error analysis generated from unsupervised image captioning models.

the first image on the left of Figure 3.5, the "Only Obj" models represented the correct object "car" but failed to depict the relationships with other objects in the image. In contrast, the remaining three models successfully recognised some relationships in the captions, and the last two models were more likely to predict precise relations instead of general geometric relations, such as "driving". The two "MOTIFS RS" models could choose more clear and meaningful relations to generate captions. For instance, in the third image Figure 3.5, it is intuitive to predict the relationships between the key objects "cat" and "sink", and we found that all models generated a plausible caption. The last two models could also outline a semantic relation, such as "laying".

Furthermore, we conducted an error analysis using our model under two experimental settings to provide insights on how the model could be improved, as shown in Figure 3.6. The relatively low performance in ROUGE and CIDEr-D compared with that of [69] might be caused by the biased visual relation detection trained from VG. The "bias" indicates that the detected relationships are trivial and less informative, e.g., because "`on`" is used rather than "`sitting at/standing on`", which implies that the captioning model is more likely to contain such relations in a shorter sentence.

Analysis of the visual attention reward We captured the cumulative average visual attention reward that was calculated from the

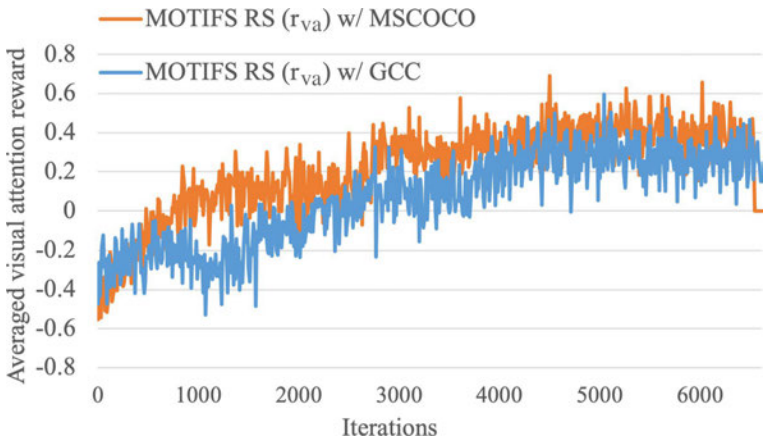


Figure 3.7: The average visual attention reward of each sentence gained during the training process.

output of each training iteration in the two **MOTIFS RS** experiments. As shown in Figure 3.7, both curves presented a stable increase in the reward values as the training progressed. This increase shows that the global and local gaps between the image and text modalities narrowed. After approximately 500 iterations, the "MOTIFS RS(r_{va}) w/ MSCOCO" value was consistently higher than that of "MOTIFS RS(r_{va}) w/ GCC". This is reasonable because the MSCOCO caption corpus contains the ground-truth references of the training image, but the image-caption assignments of GCC are from a different dataset, which brings much more unpaired noise for the adversarial training process.

3.4 conclusions

Unsupervised image captioning is one of the most challenging tasks in visual representation, as it requires minimising the misalignment between visual content and the generated language representation. Existing approaches primarily focus on global mappings between the independent visual and textual modalities, but they often struggle with weak local alignment and corpus bias, leading to less visually consistent captions.

In this part of work, we addressed these challenges by proposing an adversarial word-level visual relation alignment model. By formalising the image captioning process as a reinforcement learning (RL) optimisation problem, we introduced visual attention and adversarial rewards through a policy-gradient method to improve both global and local alignment. The visual attention reward, driven by distilled knowledge from the visual relation detector, served as a critical intermediary, aligning generated captions with visual features and enabling the discriminator to better distinguish plausible descriptions.

Extensive experiments on the MSCOCO and Visual Genome datasets demonstrated the effectiveness of this approach, with our model outperforming existing unsupervised methods across various evaluation metrics. While the model showed strong performance, further research is needed to explore its generalisation capabilities on more diverse or specialised datasets with varying visual styles or thematic content.

4 MODALITY MAPPING BETWEEN IMAGE AND EEG IN BRAIN-INSPIRED OBJECT CLASSIFICATION

In the previous chapter, we introduced the problem of multimodal learning of image and text in unsupervised image captioning and illustrated an algorithm to address it. Now, we shift our focus to explore how humans can be directly involved in computer vision tasks. Specifically, our journey into brain-inspired object classification explores how neural signals can enrich visual understanding through multimodal learning.

This chapter presents two EEG-based visual classification models that harness temporal and effective connectivity features. By integrating EEG temporal dynamics and graph-based neural representations, these models aim to enhance the interpretability and performance of object recognition tasks and other visual-related tasks, demonstrating the potential of human cognitive signals to enrich visual representation and improve the interpretability of deep learning methods.

4.1 an evoked potential-guided deep learning brain representation for visual classification

A novel approach in visual classification seeks to decode the feature representations of visual objects based on human brain activity. Electroencephalography (EEG) recording, collected from the brain's cortex, has become a popular method for examining cognitive processes during image classification tasks. In this study, we introduced a deep learning framework, guided by visual evoked potentials (ERPs) from EEG data, termed the ERP-Long short-term memory (LSTM) framework. This model leverages EEG signal-based ERPs for visual object classification. More specifically, we initially extracted ERP sequences from multiple EEG channels in response to visual stimuli. Subsequently, we trained an LSTM network to learn the feature representation space for object classification. In our experiment, we used an image dataset and recorded over 50,000 EEG trials from 10 subjects who viewed images from six different categories, totaling 72 unique exemplars [84].

4.1.1 *ERP-LSTM Framework*

The overview of the ERP-LSTM framework is presented in Figure 4.1. It is organized into two stages for visual classification. In Stage 1, we processed the raw EEG data from a visual experiment and extracted

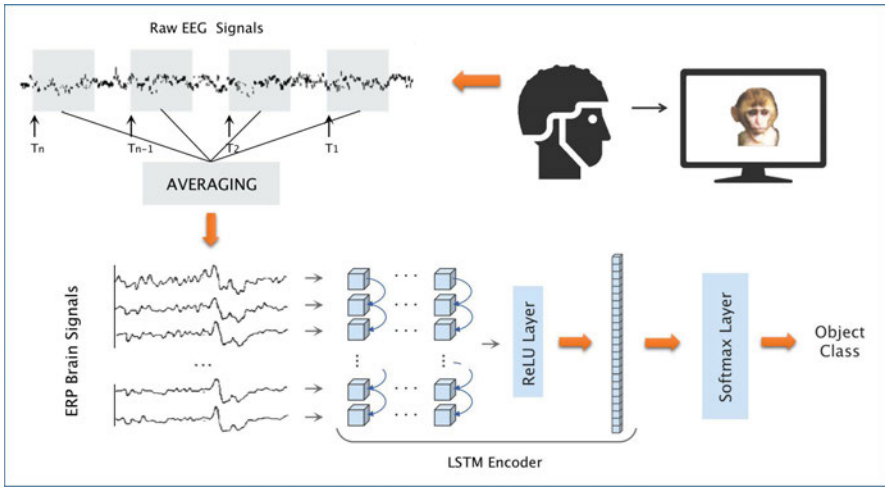


Figure 4.1: Overview of the ERP-LSTM framework

ERPs to isolate signals associated with visual stimuli. In Stage 2, the LSTM network was trained to learn the feature space of these ERP sequences, followed by a Softmax layer that classified the visual objects into categories.

4.1.1.1 *Stage 1: ERPs Extractions from EEG*

The key to EEG-based visual object classification is the extraction of representative signal features. In the first stage of our framework, we obtained representative visual-related ERP features by improving the signal-to-noise ratio (SNR) of raw EEG signals through smooth-averaging. A series of EEG segments from identical trials were averaged to create a single waveform. Specifically, this averaging process retained consistent segment features (the ERPs) while diminishing those that varied across segments (refer to the top left of Figure 4.1).

In formal terms, let $d_i^j = \{T_1^j, T_2^j, \ldots, T_n^j\}$, $i \times n = N$, d_i^j be the i_{th} subset of temporal EEG signals across multiple channels when a subject observes the j_{th} exemplar image. N is the number of EEG trials to be averaged, which contains n EEG trials, where trial $T_n^j \in R^c$ (c is the number of channels).

The averaging process follows this formula:

$$e_i^j = \sum T_n^j \ / n, \quad T_n^j \in d_i^j \tag{4.1}$$

where e_i^j represents the ERP sequence averaged from d_i^j.

Let E be the sum of extracted multi-channel ERPs, $E_i\{e_1^j, e_2^j, \ldots, e_i^j\}$, which serves as input to the LSTM encoder module described in the subsequent section, enabling it to learn feature representations for visual classification.

4.1.1.2 *Stage 2: Feature Representations and Classification*

To capture the spatial and temporal characteristics from the ERP data, we applied an LSTM encoder module, illustrated in the lower portion of Figure 4.1, following Spampinato's architecture of a "common LSTM + output layer" [166]. The encoder takes as input the multi-channel ERP signals extracted in the previous stage.

At each time step t, the first layer takes the input $s(\cdot, t)$ (representing the values across all channels at time t), allowing all ERP data from multiple channels to feed into a shared LSTM layer. After several LSTM layers, a ReLU layer is added to facilitate mapping of the feature space. The output of the LSTM encoder is a one-dimensional feature representation for each ERP. After obtaining these feature vectors, a Softmax layer is employed to classify the LSTM-derived features across different visual categories.

The LSTM encoder module's performance is evaluated using cross-entropy loss, measuring the discrepancy between the predicted class distributions and the actual class labels. The cross-entropy loss $H(p, q)$ is defined as follows:

$$H(p, q) = -\sum_{x}^{L} (p(x) \log q(x) + (1 - p(x)) \log(1 - q(x))) \qquad (4.2)$$

where $p(x)$ represents the probability distribution of the ground-truth class label, and $q(x)$ is the output probability distribution of the Softmax classification layer. denotes the probability output from the Softmax classification layer. This loss is back-propagated through the network to optimise the model parameters via gradient descent.

In summary, the ERP-LSTM framework utilizes an LSTM encoder to transform ERP sequences into feature representations, with a Softmax layer then classifying these features into the respective visual categories.

4.1.2 *The Experiment*

4.1.2.1 *Dataset*

In this study, we used an image datasetalong with EEG data collected from 10 subjects viewing images from various categories [84]. The dataset includes 72 images, classified into six categories: (1) Human Body (HB), (2) Human Face (HF), (3) Animal Body (AB), (4) Animal Face (AF), (5) Fruit and Vegetable (FV), and (6) Inanimate Object (IO). In Figure 4.2, we display example images from each of the six categories (HB, HF, AB, AF, FV, IO) used for visual classification.

The experiment involved each subject completing two sessions, each consisting of three blocks. Within each block, all 72 images were presented 12 times in a random order. During each trial, a single image appeared on the screen for 500 ms, followed by a 750 ms blank grey

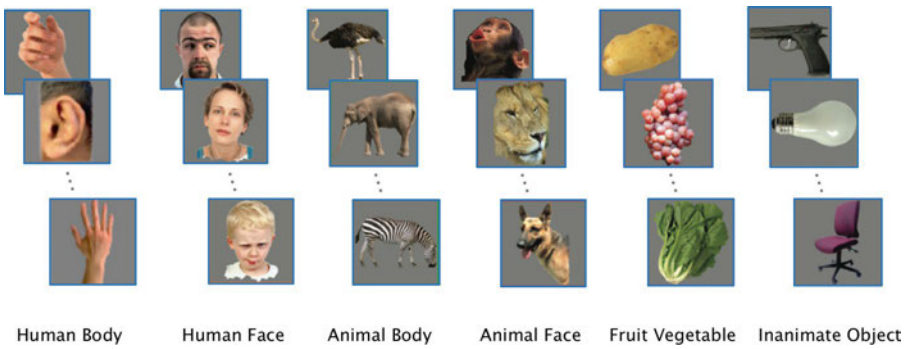


Figure 4.2: Sample photographs from the six categories

screen as an interval. Consequently, each subject underwent 72 trials per image, totaling 5,184 trials per participant, leading to a total of 51,840 trials across all subjects. Each trial was labeled according to the visual event category (the relevant one of the 6 categories) and image exemplar (one of the 72 specific images).

The EEG recording was collected using unshielded 128-channel EGI HCGSN 110 nets, preprocessed with a Butterworth filter set to a frequency range of 1 Hz to 25 Hz, and then temporally downsampled to a final rate of 62.5 Hz (496 ms). For artifact reduction, the four frontal cortex channels were removed, leaving 124 channels for ERP analysis.

4.1.2.2 *Experimental Settings*

In this experiment, we divided the 72 EEG trials into 6 sets, with each set containing 12 trials. The trials within each set were averaged to extract an ERP sequence, labeled according to the corresponding image and category. This process yielded six ERP sequences for each image, producing the ERP space E for all extracted 124-channel ERP sequences. The ERP space E was divided into a training set and a testing set in an 80:20 ratio, with 80% of ERP sequences used for training and 20% for testing. To further evaluate classification performance, we implemented two approaches: cross-subject and within-subject classification. Cross-subject classification involved training the model on data from certain subjects and testing it on data from different subjects, assessing generalizability across individuals. Within-subject classification, by contrast, trained and tested the model on data from the same subject, allowing us to evaluate the model's ability to capture subject-specific patterns.

Table 5: Performance of six-category visual classification

Accuracy	EEG-LSTM[28]	ERP-LSTM(our)	Improvement
Cross-Subject	36.72%	66.81%	**30.09**%
Within-Subject			
Subject 1	35.07%	89.06%	**53.99**%
Subject 2	35.30%	60.94%	**25.64**%
Subject 3	45.25%	71.88%	**26.63**%
Subject 4	35.88%	50.00%	**14.12**%
Subject 5	48.03%	65.62%	**17.59**%
Subject 6	47.80%	75.00%	**27.20**%
Subject 7	40.74%	62.50%	**21.76**%
Subject 8	31.37%	45.31%	**13.94**%
Subject 9	39.12%	60.94%	**21.82**%
Subject 10	47.45%	59.38%	**11.93**%

4.1.3 *Results*

4.1.3.1 *Performance of Six-category Visual Classification*

Table 5 presents the classification performance of both the standard EEG-LSTM framework (using raw EEG) [166] and our proposed ERP-LSTM framework. It also compares performance across the two classification types (cross-subject and within-subject). Results indicate that our ERP-LSTM framework achieved a cross-subject classification accuracy of approximately 66.8% and the highest within-subject accuracy of 89.1% for a single subject (subject 1). Both results surpassed the EEG-LSTM framework, where classification accuracy increased by 30.1% across all 10 subjects, reaching 54.0% for subject 1 and 23.5% for the average within-subject results.

These findings suggest that the feature representation space generated from ERPs is more effective for image object classification than raw EEG data. Furthermore, we infer that critical brain signal information for object cognition was retained during averaging, as the extracted ERPs preserved spatial and temporal characteristics related to visual evoked potentials.

4.1.3.2 *Performance of Exemplar-Level Visual Classification*

We extended our analysis to assess the performance of existing frameworks and our ERP-LSTM model at the level of individual exemplar

Table 6: Performance of category- and exemplar-level visual classification

Accuracy	Kaneshiro[18]	EEG-LSTM[28]	ERP-LSTM(our)
Categories (6 classes)	40.68%	36.72%	**66.81**%
Exemplars (72 classes)	14.46%	7.97%	**27.08**%

images. In this analysis, the category labels were removed, and instead, the task was to classify each specific image as a distinct exemplar. Table 6 presents the performance of the two baseline frameworks, Kaneshiro [84] and EEG-LSTM [166], as well as our ERP-LSTM framework in identifying the 72 exemplars across all 10 subjects. Our ERP-LSTM framework achieved an exemplar-level classification accuracy of 27.1%, outperforming Kaneshiro's 14.5% and EEG-LSTM's 8.0%. For comparative insights, we also included the results for classification at the category level, highlighting differences between the simpler category classification task and the more complex exemplar identification.

Our results indicate that, compared to prior models, the ERP-LSTM framework offers a representation space with lower confusion, yielding a more distinctive feature representation for visual classification tasks. Additionally, our ERP-LSTM framework outperformed a recent model from 2020 [2], which reported an accuracy of 17.1% for six-category classification, despite differences in data sources. This suggests that the LSTM network in our framework effectively encodes ERPs, leveraging its strength in handling the temporal dynamics of time-series EEG data to create a more representative feature space.

4.1.4 *Conclusion*

This section introduces a novel method for visual classification by decoding brain activity patterns related to visual objects. Using electroencephalogram (EEG) data, our ERP-LSTM framework utilizes Event-Related Potentials (ERPs) to guide the learning of neural representations for visual processing. The framework operates in two stages: initially extracting ERP sequences from multi-channel EEG signals triggered by visual stimuli, and subsequently using an LSTM network to encode these ERP-based features for visual object classification.

In our experiments, we analyzed EEG data from 10 subjects, comprising over 50,000 trials across six object categories and 72 exemplar images. In cross-subject evaluations, the ERP-LSTM framework achieved classification accuracies of 66.81% for object categories and 27.08% for individual exemplars. These outcomes represent substantial improvements over previous frameworks, with accuracy gains ranging from 12.62% to 53.99%.

Our findings highlight the effectiveness of using ERPs derived from EEG data to facilitate learning of discriminative neural representations for visual classification. This underscores the ERP-LSTM framework's potential to capture distinctive neural features, making it a valuable contribution to brain-inspired approaches in visual classification research.

4.2 causal convolutional network-inference graph for effective brain networks modelling from eeg data

4.2.1 *Introduction*

The emerging application of EEG-based brain-computer interface (BCI) has gained particular attention, especially in the area related to visual conception, which plays a crucial role in interpreting visual stimuli in EEG data. Visual conception tasks, such as emotion recognition [4, 209], fatigue assessment [23, 205], motor imagery control [1, 102, 195], and visual object detection [3, 84], are pivotal for decoding human visual cognitive patterns in EEG-based BCI systems. These tasks highlight how visual classification intersects with various domains of BCI, as it aids in interpreting visual cues in object recognition, cognitive state assessment, and motor imagery classification. Human cognitive patterns can be fundamentally revealed by brain network representation, where information integration and propagation correlate to multiple brain areas as a connectivity topology rather than working independently [14, 150].

Recently there has been a growing trend towards using graph neural networks (GNNs) to address the critical nature of brain network representation in classifying cognitive patterns from EEG data [74, 103, 165, 180]. GNN models can decode graph structure from non-Euclidean EEG data, and mimic the information propagation process in the brain network by aggregating the messages passing between connected channels [197]. However, one primary challenge that GNN models confront is defining the effective brain connectivity topology between two regions to present a brain network for multi-channel EEGs.

The conventional method to address this challenge is to employ statistical modelling of brain networks to capture the inter-channel relations in EEG before transforming it into a graph structure. The brain network modelling can be divided into functional connectivity and effective connectivity. For functional connectivity analysis, most previous studies attempted to investigate statistical associations between channels, such as instantaneous coherence, to assemble an undirected graph (see **Figure 4.3(a)**) [74, 180]. However, these approaches failed to identify direct causation between channels, potentially contaminat-

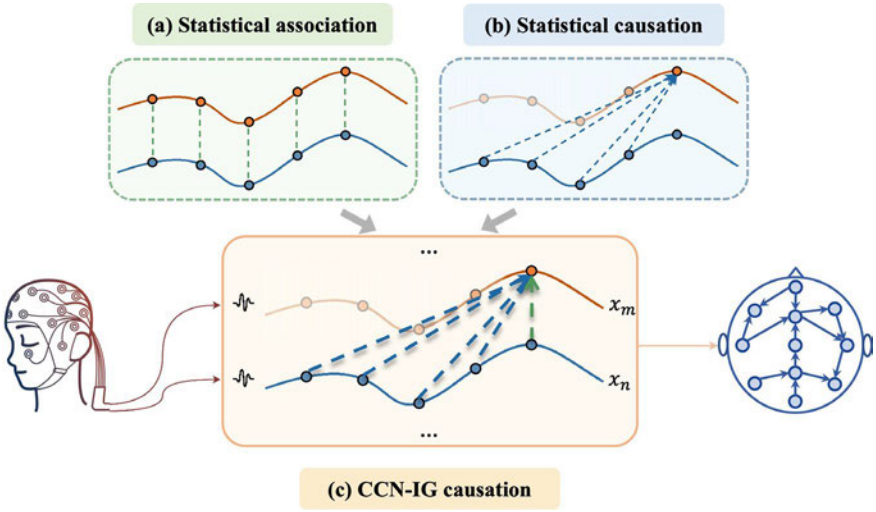


Figure 4.3: Illustrations of statistical association, statistical causation, and CCN-IG causation. (a) Instantaneous coherence based on a statistical association. (b) Time lagged statistical causation (e.g., Granger causality). (c) Infer causation between EEG channels via CCN-IG with instantaneous and time-lagged (Granger) causal effects.

ing the learning of directional information integration and propagation patterns within the brain networks [65].

Since EEG signals are generated by multiple electrodes on the scalp, statistical associations can be influenced by external factors, such as [129]. The observed results could be misleading due to indirect causation from intermediate channels, or influenced by common factors shared with other channels or unobserved variables [150].

For effective connectivity analysis, time-lagged statistical causation [123] is commonly measured in GNN models [95, 193], with classic methods like multivariate Granger-causality (GC) analysis [52]. GC focuses on inter-channel causation with directional connections and deals with confounders by fitting third-party channels within sensitivity tests [105]. However, it is time-consuming to adjust hyperparameter tuning for the time-lag validation. Moreover, the instantaneous causal effects at the same time step are often overlooked during the process (see **Figure 4.3(b)**), which, however, is essential for capturing crucial functional relationships that manifest immediately within brain networks and making full use of the high temporal resolution of EEG.

To address the limitations of statistical modelling of brain networks, in this part of work, we created a **causal convolutional network-inference graph (CCN-IG)** modelling method, which can concurrently learn both the functional and effective connectivity for the causal brain network representation. First, we built an EEG causal inference model to capture the inter-channel causation and generate the causal brain network. Specifically, a series of parallel causal convolutional

networks (CCNs) were developed to train a supervised multivariate autoregression model for each EEG channel. The autoregression process identifies the causation between EEG channels by performing time-series intervention and causal validation. In this process, both instantaneous and time-lagged causal effects were included as shown in **Figure 4.3(c)**, integrating the advantages of both functional and effective connectivity. Second, a global causal inference map (GCIM) wa generated to merge the extracted causation into a graph structure. Finally, we developed an attentive graph learning network based on the GCIM to represent the spatiotemporal patterns for EEG-based BCI classification tasks.

The main **contributions** of this work are as follows:

- we developed CCN-IG, the first approach to constructing a causal brain network in a graph structure based on deep neural networks, rather than relying on statistical association or causation, specifically for multi-channel EEGs and visual classification applications in BCI.
- the developed GCIM and attentive graph learning network in CCN-IG represent direct and effective causation between EEG channels, concurrently learning both functional and effective connectivity while decoding corresponding spatial and temporal patterns critical for visual classification tasks.
- running experiments that demonstrate that CCN-IG outperforms six baselines using deep neural networks (including CNN-, RNN-, and GNN-based methods) and a baseline using statistical association and statistical causation, with visualisation of discriminative graphs in four EEG-based classification tasks, underscoring its effectiveness for visual stimuli interpretation and object detection, cognitive state assessment, and motor imagery task classification.

4.2.2 *Method*

In this section, we introduce CCN-IG. As shown in **Figure 4.4**, CCN-IG consists of an EEG causal inference model and an attentive graph learning network for multi-channel EEG-based BCI classification tasks.

4.2.2.1 *EEG causal inference*

The EEG causal inference has two components: causal convolutional networks (CCNs) and attentive causal validation. CCNs learn to regress the input EEG channels and enable them to represent the inter-channel causation during the multivariate autoregression process. We employ attentive causal validation to evaluate the causation effectiveness and

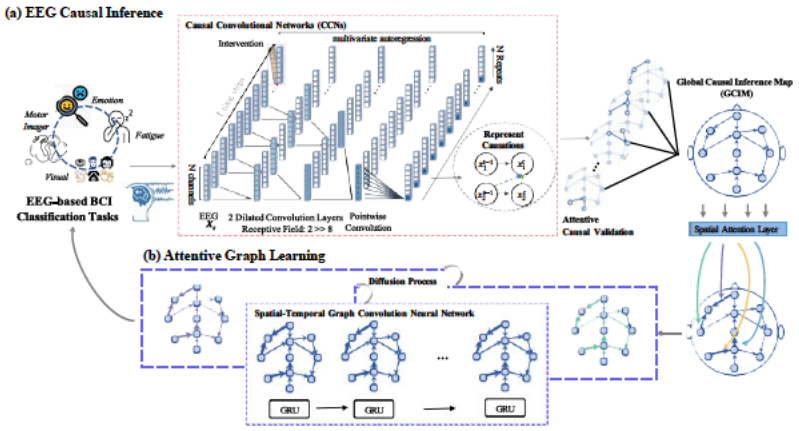


Figure 4.4: Illustration of CCN-IG. (a) EEG Causal Inference: EEG trials are inputted to the causal convolutional networks (CCNs) to identify the potential causes for each channel, which are then verified in the attentive causal validation module to generate a global causal inference map (GCIM). (b) Attentive Graph Learning: the obtained GCIM is fed into the spatial attentive layer followed by spatial-temporal graph convolution neural network to learn the effective brain networks for EEG-based BCI classification tasks.

identify valid causation for each channel, ultimately generating a global causal inference map (GCIM), as shown in Figure 4.4**(a)**.

causal convolutional networks (ccns) We use CCNs to capture the causation between EEG channels by training a multivariate autoregression model for each channel. The autoregression aims to formalise each channel sequence in EEG as a function of the current and past values of causally related channels. Given an EEG trial $\mathbf{X} = \{X_1, ..., X_j, ..., X_N\}$ with N channels, N numbers of independent CCNs are developed parallelly with the same architecture in the autoregression process of $\mathbf{X}$. We formulate this as follows:

$$X_i = \Theta(a_1 \odot F_{i1}(X_1), ..., a_j \odot F_{ij}(X_j), ..., a_N \odot F_{iN}(X_N)) \tag{4.3}$$

where $[a_1, ..., a_j, ..., a_N]$ is an N-dimensional attention vector, denoted by Z_i, F_{ij} is a CCN to regress the values of a target channel sequence X_i with the input value of X_j. $\odot$ denotes that each attention score in Z_i is element-wise multiplied with the according input channel. Θ is a pointwise convolution layer to merge the convolutional outputs from all CCNs. In the time dimensions, the CCNs also incorporate the instantaneous causal effects at the same time step by using both the past and current values of the input channels to predict the value of X_i at the moment t, noted as X_i^t, which can be defined as:

$$\begin{aligned} X_i^t = \Theta(&a_1 \odot F_{i1}(X_1^1, ..., X_1^t), ..., a_j \odot F_{ij}(X_j^1, ..., X_j^t), ..., \\ &a_N \odot F_{iN}(X_N^1 ..., X_N^t)) \end{aligned} \tag{4.4}$$

While multiple CCNs collaboratively perform the multivariate autoregression, here we introduce the structure of a single CCN. It captures the temporal causation by expanding the receptive field of convolutional layers beyond the causal time delay between EEG channels, which is built with a 1D convolutional layer and L dilated convolution layers. Given the EEG channel X_j, we first input it to a 1D convolutional layer with a convolutional kernel sliding on the sequence. To expand the perceptual domain of the convolutional kernel, we use dilated 1D convolutions with a dilation factor c that doubles at each subsequent hidden layer $l \in [1, L]$. The receptive field can grow exponentially by doubling the dilation factor at each layer as the convolutional kernel skips $f = c^l$ number of time steps at the l_{th} layer. The receptive field of a kernel in CCNs is defined as:

$$\text{ReceptiveField}(K) = 1 + \sum_{l=0}^{\lfloor L} (K-1) \cdot c^l \tag{4.5}$$

After dilated convolutions, a residual connection is used to facilitate gradient flow, and ReLU activation is applied for outputs. Thus, our proposed CCN can be described as follows:

$$H_{-1} = X_j, j \in [1 - N] \tag{4.6}$$

$$\hat{H}_l = \text{ReLU}\,(W_1 * H_{l-1} + b_1) \tag{4.7}$$

$$H_l = H_{l-1} + W_2 * \hat{H}_l + b_2 \tag{4.8}$$

where H_{-1} is the input of a CCN, H_l is the hidden output of layer l, $*$ denotes the convolution operator, $W_1 \in R^{K \times D}$ are the weights of the dilated convolution with kernel size K and D is the number of convolutional filters, $W_2 \in R^{1 \times D}$ are the weights of a 1×1 convolution for residual connection, and b1, b2 $\in R^D$ are bias vectors.

Additionally, we develop a pointwise convolution layer Θ to merge the convolutional outputs from N numbers of CCNs, into a regression result. This layer performs pointwise convolution using a kernel with dimensions $1 \times 1 \times N$ and takes the dot product with the hidden output of the last layer, H_L. The multiple CCNs are optimised for multivariate autoregression purposes, which learn to regress the input EEG channels. We use the attention weight of each single CCN to represent the causation, interpreting which channel attends to regress the target channel. Using multiple channels in the autoregression can effectively control the confounding factors and thus remove their confounding effects when exploiting the effective causal connections for multi-channel EEGs.

attentive causal validation After training the autoregression models and fixing the learned CCNs for the target channel sequence X_i, we obtain the attention vector Z_i with each attention score

representing the potential causation to other channels. The attention scores are further transformed as:

$$h_j = \text{HardSoftmax}(a_j) = \begin{cases} \sigma(a_j) & \text{if } a_j \geqslant \tau_i \\ 0 & \text{if } a_j < \tau_i \end{cases} \tag{4.9}$$

where h_j is the attention score transformed from a_j and τ_i is a pre-defined threshhold for X_i. A set of potential causes P_i from X_i is generated by selecting the causally related channels whose attention score $h_{i,j} \in h_i > 0$.

Here, a few potential causes of a channel can be observed in P_i, which are required to be evaluated as a valid causation by testing the change of autoregression performance after the intervention on the causes. For example, if the intervention on a fundamental cause of X_i (the target channel) does lead to a change of the autoregression result on X_i, we consider it a cause from P_i of the target. To verify this, we developed an attentive causal validation method based on Permutation Importance (PI), the same as that in [129].

To validate a potential cause $X_j \in P_i$, we generated an intervened EEG trial $\hat{\mathbf{X}}$ by randomly permuting the values of X_j while keeping the other channels constant in $\mathbf{X}$. Then, we deployed the fixed CCNs to regress the value of X_i based on $\hat{\mathbf{X}}$ and calculated the intervention loss $L_{j\to i}$ to test the change of autoregression performance for the target EEG channel, which can be defined as:

$$L_{j\to i} = \hat{L}_i - L_i \tag{4.10}$$

where L_i is the training loss of the CCNs for X_i and $\hat{L}_i$ is the test loss using the intervened EEG trial $\hat{\mathbf{X}}$. We identified the valid cause X_j if the intervention loss $L_{j\to i}$ is larger than a pre-defined condition significance threshold.

By iteratively training the CCNs for autoregression on EEG channels $[X_1 : X_N]$, we have to find the potential cause sets $[P_1 : P_N]$ for each channel and identify them to valid causation after the validation. Then we can create a global causal inference map (GCIM) $\mathbf{G}$ for EEG trial $\mathbf{X}$ with EEG channels as nodes and the verified causation connections as edges in the graph structure.

4.2.2.2 *Attentive graph learning (AGL)*

Given a GCIM $\mathbf{G}$, we consider a spatial attentive layer (SAL) with a trainable attentive mask to capture the most significant causal connections in brain networks for further graph learning, as shown in **Figure 4.4(b)**.

First, we define the trainable attentive mask, $M \in N^{N\times N}$ as:

$$\mathbf{M} = \begin{bmatrix} \mathbf{w}_{1,1} & \cdots & \mathbf{w}_{1,N} \\ \vdots & \ddots & \vdots \\ \mathbf{w}_{N,1} & \cdots & \mathbf{w}_{N,N} \end{bmatrix} \tag{4.11}$$

The trainable mask takes a dot product with the adjacent matrix $\mathbf{A}_{adj}$ of G, which is followed by the ReLU activation function to make the adjacent matrix non-negative. To maintain self-connections, we add self-loops to the final global adjacent matrix $\hat{\mathbf{A}}_{adj}$, which can be calculated as:

$$\hat{\mathbf{A}}_{adj} = \Phi_{ReLU}\left(\mathbf{A}_{adj} \circ \mathbf{M} + \mathbf{I}\right) \tag{4.12}$$

Then, we created a spatial-temporal graph convolution neural network to learn the spatial-temporal EEG patterns from the GCIM based on a recurrent neural network with graphical diffusion convolution [108]. The advantage of the graphical diffusion process includes extracting of the spatial patterns from the global causal connections by aggregating the message passing between causally connected channels. Our graphical diffusion is characterised by a bidirectional random walk on the directed graph G, which leads to the following diffusion convolution:

$$\mathbf{X}_{:,m\star\mathsf{G}} f_\theta = \sum_{s=0}^{s-1} \theta_{s,1} \left(\mathbf{D}_O^{-1}\mathbf{W}\right)^s + \theta_{s,2} \left(\mathbf{D}_I^{-1}\mathbf{W}^\top\right)^s \mathbf{X}_{:,m} \text{ for } m \in \{1,\ldots,M\} \tag{4.13}$$

where $\mathbf{X} \in \mathsf{R}^{N\times M}$ is the preprocessed EEG trial at time step $t \in \{1,\ldots,T\}$ with N nodes and M features, $\star\mathsf{G}$ denotes the diffusion convolution, f_θ is the convolution filter with parameters $\theta \in \mathsf{R}^{S\times 2}$, $\mathbf{D}_O$ and $\mathbf{D}_I$ are the out-degree and in-degree diagonal matrices of G, respectively. s is the number of maximum diffusion steps, and $D_O^{-1}W$ and $D_I^{-1}W^\top$ are the state transition matrices of the outward and inward diffusion processes, respectively.

Afterwards, we deployed Gated Recurrent Units (GRUs) to model the temporal dependency in EEG signals, a variant of RNNs with a gating mechanism, where matrix multiplications are replaced with

Method	Input	Emotion Recognition			Fatigue Assessment			Motor Imagery Control			Visual Object Detection		
		ACC	F1	AUROC	ACC	F1	AUROC	ACC	F1	AUROC	ACC	F1	AUROC
LSTM	Temporal	0.660	0.611	0.692	0.667	0.615	0.711	0.459	0.447	0.484	0.367	0.375	0.343
	Spectral	0.731	0.710	0.681	0.673	0.738	0.622	0.454	0.469	0.533	0.401	0.392	0.389
EEGNet	Temporal	0.846	0.842	0.906	0.744	0.736	0.759	0.523	0.517	0.612	0.407	0.385	0.386
	Spectral	0.885	0.884	0.874	0.776	0.774	0.810	0.560	0.539	0.568	0.404	0.398	0.399
Dense-CNN	Temporal	0.887	0.849	0.817	0.758	0.747	0.804	0.628	0.634	0.699	0.427	0.371	0.380
	Spectral	0.904	0.863	0.802	0.760	0.776	0.696	0.592	0.447	0.584	0.375	0.321	0.349
DGCNN	Temporal	0.873	0.842	0.823	0.769	0.739	0.793	0.565	0.499	0.561	0.441	0.397	0.400
	Spectral	0.867	0.860	0.791	0.786	0.782	**0.818**	0.554	0.546	0.552	0.497	0.485	0.513
RGNN	Temporal	0.816	0.818	0.796	0.773	0.805	0.714	0.712	0.664	0.679	0.485	0.482	0.502
	Spectral	0.879	0.873	0.923	0.681	0.653	0.775	0.729	0.734	**0.779**	0.511	0.513	0.569
SS-DCRNN	Temporal	0.778	0.791	0.773	0.729	0.724	0.757	0.643	0.616	0.641	0.479	0.421	0.485
	Spectral	0.864	0.866	0.892	0.742	0.732	0.770	0.687	0.624	0.625	**0.531**	**0.529**	0.572
CCN-IG w/o SAL	Temporal	0.828	0.782	0.856	0.675	0.650	0.700	0.671	0.639	0.672	0.447	0.415	0.441
	Spectral	0.919	0.872	0.887	0.788	0.805	0.804	0.699	0.657	0.685	0.473	0.467	0.498
CCN-IG	Temporal	0.853	0.850	0.895	0.732	0.707	0.761	0.707	0.692	0.746	0.488	0.426	0.509
	Spectral	**0.921**	**0.893**	**0.934**	**0.805**	**0.815**	0.808	**0.741**	**0.761**	0.766	0.516	0.520	**0.583**

Table 7: Classification performance (ACC, F1, and AUROC) of CCN-IG (w/o SAL) and six baseline methods using temporal and spectral EEG inputs.

diffusion convolutions, allowing for spatial-temporal modelling of EEG patterns as:

$$\mathbf{r}^{(t)} = \sigma\left(\Theta_{r\star G}\left[\mathbf{X}^{(t)}, \mathbf{H}^{(t-1)}\right] + \mathbf{b}_r\right) \tag{4.14}$$

$$\mathbf{u}^{(t)} = \sigma\left(\Theta_{t\star\Leftrightarrow}\left[\mathbf{X}^{(t)}, \mathbf{H}^{(t-1)}\right] + \mathbf{b}_{ru}\right) \tag{4.15}$$

$$\mathbf{C}^{(t)} = \tanh\left(\Theta_{C\star G}\left[\mathbf{X}^{(t)}, \mathbf{r}^{(t)} \odot \mathbf{H}^{(t-1)}\right] + \mathbf{b}_C\right) \tag{4.16}$$

$$\mathbf{H}^{(t)} = \mathbf{u}^{(t)} \odot \mathbf{H}^{(t-1)} + \left(1 - \mathbf{u}^{(t)}\right) \odot \mathbf{C}^{(t)} \tag{4.17}$$

where the input and output of GRUs at a given time step t are denoted by $\mathbf{X}^{(t)}$ and $\mathbf{H}^{(t)}$, respectively. The reset gate, update gate, and candidate at time step t are represented by $\mathbf{r}^{(t)}$, $\mathbf{u}^{(t)}$, and $\mathbf{C}^{(t)}$, respectively. The Sigmoid function is denoted by σ, and the Hadamard product is denoted by $\odot$. The weights and biases for the corresponding convolutional filters are represented by Θ_r, $\mathbf{b}_r$, Θ_u, $\mathbf{b}_u$, Θ_C and $\mathbf{b}_C$, respectively. They are used to connect with a fully-connected layer to perform classification output $\mathbf{H}^{(T)}$.

4.2.3 *Experiments*

datasets In this section we introduce more details about the four public datasets used to evaluate the proposed CCN-IG: (1) the DEAP dataset [94] for emotion recognition, (2) the fatigue dataset [17] for driving fatigue assessment, (3) the BCICIV2a [16] for motor imagery control and (4) the visual dataset [83] for object detection, respectively. A summary of the data format of these datasets is shown in Table 8.

DEAP [94] is a 32-channel EEG emotion recognition dataset collected from 32 participants while watching 40 one-minute-long videos.

Task	Emotion	Fatigue	Motor Imagery	Visual
Channel	32	32	22	128
Trial length (s)	4	3	4	0.5
Sample rate (Hz)	128	128	250	62.5
Subject	32	11	9	10
Classes	2	2	4	6

Table 8: The data summary of datasets for emotion recognition, fatigue assessment, motor imagery control and visual object detection.

They sampled EEG at 512 Hz and downsampled it to 128 Hz. We divided each trial lasting 60 seconds into 15 new trials (e.g., every 4 seconds) and marked negative (1∼5 points) and positive (6∼9 points) emotions based on the valence dimension as the classification labels.

Fatigue [17] is a 32-channel EEG dataset collected based on a sustainable driving task from 11 participants in a Virtual Reality environment. EEG data were sampled at 512 Hz and downsampled to 128 Hz. There were a total of 2952 experimental trials, and in each trial, the reaction time of each lane-departure event can correspond to driving fatigue levels (alertness vs. drowsiness) as the classification labels.

BCICIV2a [16] is a 22-channel EEG dataset from 9 participants who performed four motor imagery tasks (labelled as left hand, right hand, foot and tongue movements for classification task). Each participant generated a total of 576 trials from two sessions, and thus a total of 4608 trials of EEG data were obtained from all participants. The EEG signals were sampled at 250Hz.

Visual object detection [83] is a 128-channel EEG dataset collected from 10 participants when they were observing 72 images of six different types of object images (labelled as the human body (HB), human face (HF), animal body (AB), animal face (AF), fruit vegetable (FV), and inanimate object (IO) for classification task). EEG signals were sampled at 62.5 Hz and we considered the first 500ms of the onscreen time for each EEG trial to avoid noise.

data preprocessing We considered temporal and spectral features from each EEG trial, denoted as $X \in \mathbb{R}^{N \times M}$ for the four listed BCI classification tasks. As mentioned above, N was the number of EEG channels, and M was the time or frequency dimensions. We then normalised these dimensions based on the mean and standard deviation values. Of note, we used the fast Fourier transform (FFT) for frequency dimensions to retain the log amplitudes of spectral components.

baselines We considered the existing six methods based on RNNs, CNNs and GNNs for EEG classification purposes. They include LSTM [166], EEGNet (implemented following [102]), Dense-CNN [35], DGCNN [165], RGNN [213] and SS-DCRNN [180].

model training Training was accomplished in PyTorch on a single NVIDIA Volta GPU. We performed a grid search to determine the optimal hyperparameters for all four classification tasks. For training in CCNs, we selected CCN hidden layers among $L \in \{0, 1, 2\}$, kernel size among $K \in \{0, 2, 4, 6\}$, the dilation coefficient c is set same as kernel size and training epochs within range {500, 1000, 1500}. The CCN hidden layers were 1, the kernel size and dilation coefficient were 2 and we trained 1000 epochs for all datasets in the CCN training process. And the threshold for attentive causal validation is set as 0.8. For AGL training, we selected the hidden unit for SAL among {128, 256, 512}, the number of GRU layers within range {2, 3, 4}, hidden units within range {32, 64, 128} and the maximum diffusion step $K \in \{2, 3, 4\}$. Furthermore, the CCN-IG was trained for the maximum training epoch of 300 with the initial learning rate being 3e-4, the maximum number of diffusion steps and GRU layers being 2. The hidden unit for SAL and the hidden units of GRU were set to 128 and 32 for all tasks, respectively, excepting motor imagery controls (set to 256 and 64). Five-fold cross-validation for the subject-dependent setting was adopted to split training (4/5) and testing (1/5) datasets, and cross-entropy was selected as the loss function to optimise the training process.

evaluation metrics Three metrics were used to evaluate the classification performance: (1) accuracy (ACC), (2) F1-score, and (3) the area under the receiver operating characteristic curve (AUROC). For the generated EEG GCIM, we conducted a statistical analysis on each edge (nodes pair) in the learnable adjacent matrix for all subjects. To avoid the less significant EEG channel interactions in brain networks, we only considered the top 20 high correlations. Here, we use the Kruskal-Wallis test, a nonparametric method that does not assume a normal distribution of the residuals, and we set the significance level (p-value) as 0.05.

4.2.4 *Results*

4.2.4.1 *Classification performance*

We considered EEG temporal and spectral inputs to evaluate the performance using ACC, F1 score, and AUROC for all classification tasks, including emotion recognition, fatigue assessment, motor imagery control, and visual object detection. As shown in **Table 7**, we ob-

served a consistent trend where CCN-IG and most baselines had superior performance on spectral EEG inputs over temporal EEG inputs. This finding aligns with prior research indicating that transforming time-domain EEG signals into the frequency domain can offer more informative and less noise-contaminated EEG patterns [124, 180]. Appendix B.1 shows more detailed classification results of temporal and spectral EEG inputs on different sub-bands.

Highlighting CCN-IG's performance, for the emotion recognition task, it achieved the highest accuracy (92.1%), F1 scoreed (89.3%) and AUROC (93.4%) on the spectral EEG inputs over the baseline methods. For the fatigue assessment task, CCN-IG achieved the best accuracy of 80.5%, which was 13.2%, 2.9%, 4.5%, 1.9%, 12.4% and 6.3% higher than the baselines of LSTM, EEGNet, Dense-CNN, DGCNN, RGNN and SS-DCRNN on spectral EEG inputs, respectively. CCN-IG also achieved the best F1 score (81.5%) and second AUROC (80.8%), while DGCNN attained the best AUROC (81.8%).

For the motor imagery control task, we can observed that CCN-IG, RGNN and SS-DCRNN achieved better performance on spectral EEG inputs than on temporal EEG inputs, but this was not the case for LSTM, EEGNet, Dense-CNN, and DGCNN. Of note, CCN-IG achieved the best accuracy (74.1%), F1 score (76.1%) and AUROC (76.6%) on the spectral EEG inputs.

For the visual object detection task, CCN-IG outperformed most of the baseline methods on the spectral EEG inputs, achieving the second accuracy (51.6%) and F1 score (52.0%), and the best AUROC (58.3%) while SS-DCRNN attained the best accuracy and F1 score. CCN-IG also achieved the best accuracy of (48.8%) on the temporal EEG inputs, which is 12.1%, 8.1%, 6.1%, 4.7%, 0.3% and 0.9% higher than the accuracy of LSTM, EEGNet, Dense-CNN, DGCNN, RGNN and SS-DCRNN, respectively.

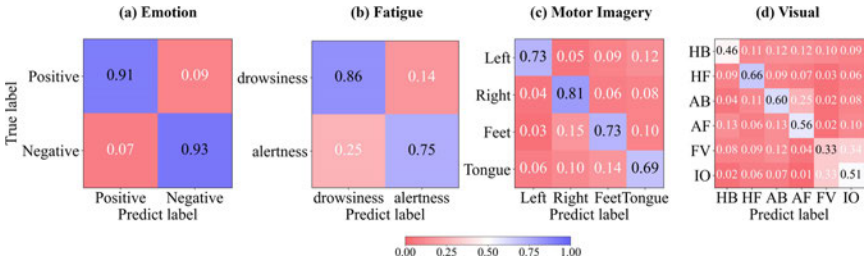


Figure 4.5: Confusion matrices of CCN-IG in four classification tasks with spectral EEG inputs.

In addition, for our proposed CCN-IG, we further analysed and visualised the detailed classification performance based on confusion matrices (see **Figure 4.5**) with spectral EEG input. For example, CCN-IG performed well on emotion recognition tasks, with relatively high true positive and negative rates. However, it also revealed substantial

confusion between the fruit vegetable (FV) and inanimate object (IO) classes for the visual object detection task.

4.2.4.2 *Analysis of graph structures*

To compare our EEG causal inference model to statistical modelling of brain networks, we performed two studies on spectral EEG inputs: 1) Coh-AGL, which uses the coherence correlation [202] - a classic statistical coherence measure for functional connectivity, to replace the global causal inference map (GCIM) in CCN-IG while remaining the same attentive graph learning network; 2) PDC-AGL, which uses the partial directed coherence [10] - a common GC measure for effective connectivity, to replace the GCIM in CCN-IG.

Figure 4.6 **(a)** reports the AUROC results and shows that CCN-IG consistently outperformed Coh-AGL and PDC-AGL. The observed performance advantages were 9.6%, 7.5%, 7.7% and 12.6% over Coh-AGL, and 6.5%, 4.7%, 9.8% and 2% over PDC-AGL. The improved classification performance demonstrated the generated global causal inference map's effectiveness for brain network representation compared to the statistical modelling of brain networks.

4.2.4.3 *Analysis of graph learning*

To look deeply into the effectiveness of the different components of CCN-IG, we performed an ablation study on all four classification tasks with spectral EEG inputs: CCN-IG w/o SAL, which removes the spatial attention layer from attentive graph learning in CCN-IG. **Figure 4.6** **(b)** reports the AUROC results. Without adding the spatial attention layer of attentive graph learning in CCN-IG, we observed that it reduced up to 14.1% AUROC score on motor imagery control task. This suggests that the spatial attention layer can extract the latent causation in the complex brain networks to assist the learning of spatial-temporal dependence from the GCIM.

4.2.4.4 *Analysis of discriminative brain networks*

To discriminate brain networks associated with different cognitive activities, such as from positive to negative sentiment in the emotion recognition task, we calculated and visualised significant differences (p-value < 0.05) of EEG graph topologies across multiple classes with spectral (frequency-domain) EEG inputs using the coherence correlation (Coh), partial directed coherence (PDC) and the generated GCIM from CCN-IG, respectively. See Appendix B.2 for additional analysis with temporal (time-domain) EEG inputs.

Figure 4.7 illustrates the top 20 significantly discriminative connections for the four classification tasks. As shown in **Figure 4.7(a)**, for the emotion recognition task, Coh and PDC tend to connect the left pre-frontal area to the frontocentral area, but GCIM can leverage the

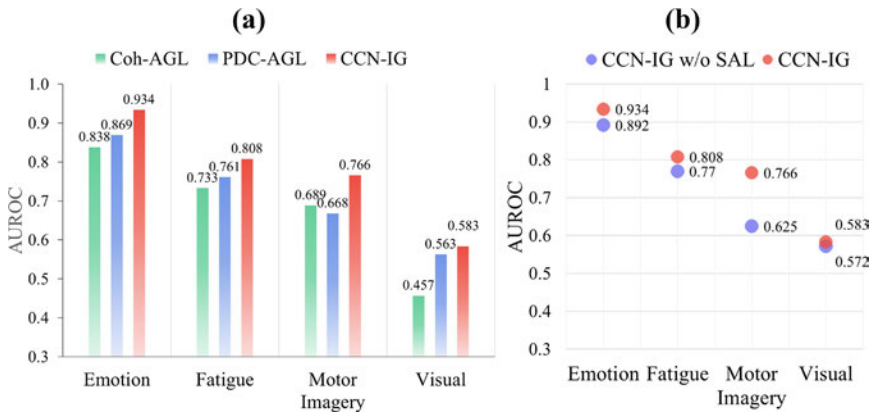


Figure 4.6: (a) AUROC scores of Coh-AGL, PDC-AGL and CCN-IG in four classification tasks. (b) AUROC scores of CCN-IG w/o SAL and CCN-IG in four classification tasks.

pre-frontal and parietal area (AF3 → P4). This finding is consistent with the existing knowledge that emotion involves a large information propagation among frontal and parietal channels [4]. For the fatigue assessment task (**Figure 4.7(b)**), Coh, PDC and GCIM present a few connections between pre-frontal and frontal areas. Still, GCIM forms a critical causal connection from the frontal to the occipital area (F7 → Oz).

For the motor imagery control task (**Figure 4.7(c)**), Coh primarily showed inter-connections of the frontocentral area, but PDC and GCIM focused more on the causal connections from the frontocentral to the central area to express effective control patterns in brain networks [192]. For the visual object detection task, as shown **Figure 4.7(d)**, Coh showed functional connections between the pre-frontal area and the frontocentral area, while GCIM highlighted a significant causal connection from the pre-frontal to the occipital area (AFF2h → Oz). This is consistent with the finding that visual processing and mapping are related to the occipital area [101]. In conclusion, visualisation of the discriminative brain networks showed that Coh, PDC and GCIM can identify the essential connections associated with different cognitive activities. However, GCIM discovered significant causation through directed long-distance connections across different brain areas or hemispheres, demonstrating its effectiveness in excluding confounding factors in EEG causal inference.

4.2.5 *Conclusion*

Identifying effective brain connectivity across multi-channel EEG signals is essential for capturing direct causal relationships within graph neural networks, especially in tasks related to visual conception, where interpreting visual stimuli in EEG data is crucial. In this study, we in-

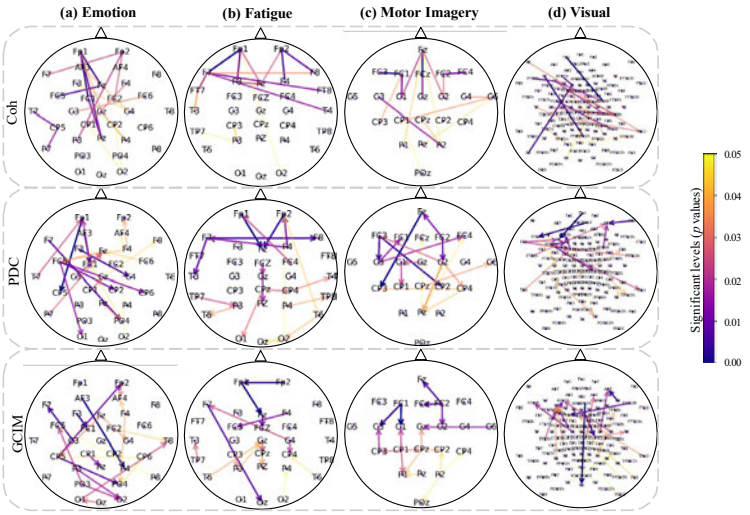


Figure 4.7: The significantly discriminative brain networks across multiple classes identified based on Coh (in the upper panel), PDC (in the middle panel) and GCIM (in the bottom panel) as graph topologies with spectral (frequency-domain) EEG inputs for four EEG-based classification tasks (a)-(d).

troduced the Causal Convolutional Network-Inference Graph (CCN-IG) as a novel modelling method to construct causal brain networks specifically designed for EEG-based classification tasks. By capturing both time-lagged and instantaneous causal relationships, CCN-IG overcomes limitations of previous statistical approaches, enabling a more dynamic and realistic representation of brain connectivity.

The CCN-IG framework was evaluated across four visual conception tasks-emotion recognition, driving fatigue assessment, motor imagery control, and visual object detection—where visual classification plays a pivotal role in decoding human cognitive patterns. Through parallel causal convolutional networks, CCN-IG generated a Global Causal Inference Map (GCIM) that underpins the construction of effective brain networks, allowing an attentive graph learning network to capture both temporal and spatial patterns. Extensive experiments demonstrated that CCN-IG significantly outperformed six deep learning baselines, underscoring its strengths in visual-related classification tasks.

Our findings highlight the centrality of visual classification in EEG-based brain-computer interface (BCI) systems. By establishing a comprehensive and directed model of visual brain connectivity, CCN-IG enhances the decoding of visual stimuli within EEG data, which is essential for applications in BCI domains that rely on visual conception tasks. This approach sets a new standard for interpreting visual information in EEG signals and opens promising avenues for further development of BCI applications, particularly in visual stimulus classification. Future research can explore optimising CCN-IG for com-

plex visual tasks, enhancing the system's potential in cognitive monitoring, neural feedback, and interactive visual environments.

5 TRIMODAL CONTRASTIVE LEARNING FOR DECODING VISUAL NEURAL REPRESENTATION

Building upon the previous chapter's focus on human involvement in multimodal learning via EEG-based object classification, we now broaden our exploration to include text as an additional modality. In this chapter, we introduce a trimodal learning framework that jointly uses image, text, and EEG data to decode visual neural representations. By using contrastive learning, this work aims to bridge cognitive, visual, and linguistic domains, ultimately achieving a more unified and consistent understanding of brain-visual-linguistic relationships.

5.1 introduction

In recent years, visual representation aims to convert visual inputs such as images or videos into feature representations via deep neural networks, which has driven the progress of a variety of downstream tasks, such as image classification, object detection, image segmentation, and target tracking [148].

Analogous to neural networks, the human brain uses visual stimuli as input to encode visual information using neuron activity to form so-called visual neural representations during visual cognition. Currently, using human brain activities recorded by EEG or functional magnetic resonance imaging (fMRI), researchers are increasingly interested in decoding visual neural representations with the aim of extracting information from neural activities induced by visual stimuli to infer visual categories [25, 163]. This not only enables researchers to reveal the visual cognitive mechanism of human brain, but also facilitates the development of brain-inspired computational models to classify or detect visual stimuli from brain signals [37, 201].

For decoding EEG-based visual representation tasks, several methods have attempted to classify human brain activity evoked from visual stimuli [167], with an accuracy of over 82.9%. However, these promising results may have benefited from the block design employed in EEG data acquisition experiments, whereby all images of the same class are presented to subject in a centralised manner. The block design introduces long-term temporal correlation in the training and testing of EEG trials [114] and tends to produce discriminative patterns associated with high-frequency bands (over 71 Hz) [107] that contain a lot of noise information, rather than patterns that truly reflect cognitive brain activity. Consequently, Li et al. [107] and Ahmed

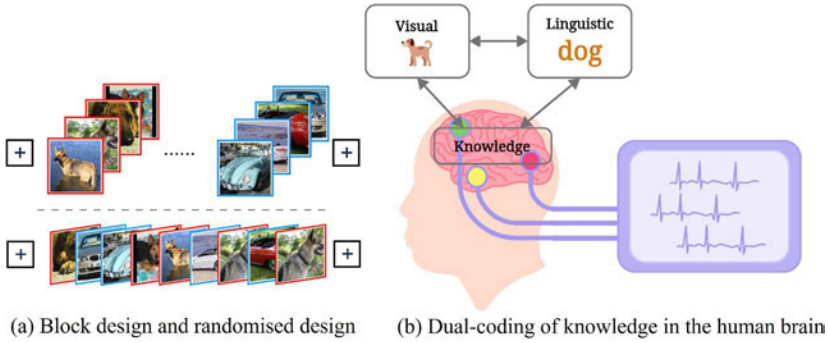


Figure 5.1: (a) Block design and randomised design. (b) Dual-coding of knowledge in the human brain.

et al. [3] introduced new datasets based on completely randomised design experiments (as shown in Figure 5.1 (a)), aiming to decode real visual representation from EEG. However, the experimental results showed a significant decrease in classification accuracy for the same visual stimulus across sessions, demonstrating the increased difficulty in recognising patterns from EEG signals in the face of significant knowledge gaps between cognitive and visual domains when temporal correlation bias is removed.

Existing research on visual neural representation mostly follows a unimodal supervised learning framework, whereas the encoding of EEG data is highly correlated with specific semantic categories. Although multimodal learning approaches have proven their effectiveness in combining EEG and visual modalities [134], their success is also mostly based on block design experiments. Research in cognitive neuroscience has revealed that the brain uses a dual coding process when processing visual stimuli [13, 133], integrating visual and text coding functions when detecting, discriminating, and recognising perceptual visual stimuli (as shown in Figure 5.1 (b)). Therefore, we argue that a promising solution to tackle this challenge could be trimodal learning of brain-visual-linguistic features to infer visual categories, by analysing recordings of human brain activity and decoding cognitive knowledge shaped by visual features and their linguistic context. So far, few works have explored the capability for EEG visual representation across all three modalities.

To fulfil this gap, we propose a trimodal contrastive learning method to learn visual representations from human brain activity by the joint learning of brain-visual-linguistic features. Our model focuses on modeling the relationships between brain activity and multimodal semantic knowledge, i.e., visual semantic knowledge extracted from images and textual semantic knowledge from class label. Specifically, we first encode the EEG, image, and text separately by three single modal encoders. We then fuse these three modalities to learn joint contextualised representations through a Transformer based tri-

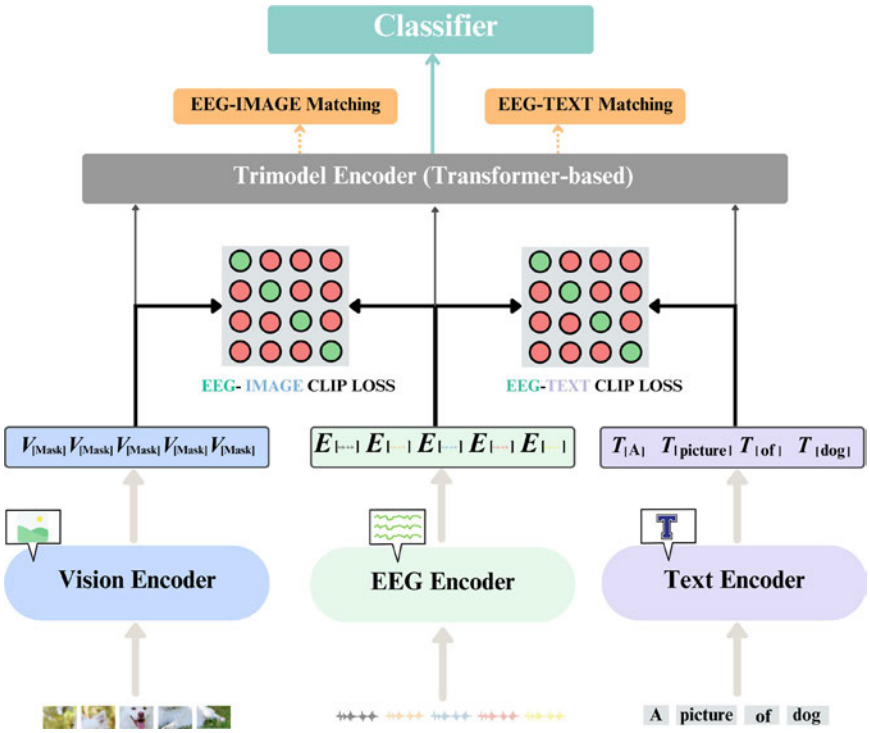


Figure 5.2: Overview of the trimodal contrastive learning frammwork.

modal encoder. To enhance the consistency of this joint representational space, we pretrained our model on a randomised EEG visual dataset with multi-objective modelling: instance- and modality-level optimisation. Instance-level modelling introduces EEG-image and EEG-text contrastive loss and matching to align the three modalities, making it easier for the trimodal encoder to perform cross-modal learning. Modality-level modelling includes masked image classification to randomly mask out the embedding from any one or two of the three modalities, allowing the trimodal encoder to adapt to different numbers of modalities with singole-, two-, or three-modal inputs. Finally, after pretraining, the outputs of the trimodal encoder were taken for the visual classification.

In summary, our main contributions are listed as follows:

- To the best of our knowledge, we are the first to propose a trimodal contrastive learning approach that combines visual and linguistic knowledge for neural decoding of visual objects with the randomised design experiments.
- We developed a new trimodal learning model with specially designed instance-, modality-level modelling to achieve more consistent brain-visual-linguistic joint representations.

5.2 method

The model architecture of our trimodal contrastive learning method is illustrated in Figure 5.2. First, the model utilises single-modal encoders to extract representations from the input EEG, image, and text. These representations are then integrated using a Transformer-based trimodal encoder, which facilitates fusion across the three modalities. Finally, the output from the trimodal encoder is used for the subsequent classification task.

5.2.1 *Single-Modal Encoders*

Since EEG signals, visual images and text label are different sources, we first preprocess the raw inputs using single-Modal encoders to extract modality-specific features.

To extract the feature embedding of EEG data, we adopt the 6-layer transformer as the EEG encoder, noted as E_b, and train from scratch to processes the input EEG trail E along the channel dimension to extract sequenced features. The final EEG embedding $B = \{b_1, ..., b_N\}$ is obtained by passing the EEG features through an layer normalisation layer, where N is the number of EEG channels. The calculation is

$$B = \{b_1, ..., b_N\} = LN(E_b(E)) \tag{5.1}$$

For image inputs, we used a powerful vision encoder, referred to as ViT-B/16, to extract hierarchical visual features from the images, which had been pretrained to achieve 80.2% top-1 accuracy on ImageNet-1k. An input image I was encoded into a sequence of embeddings: $\{v_{cls}, v_1, ..., v_K\}$, where v_{cls} is the embedding of the [CLS] token and K denotes the number of embedded image regions. The final vision embedding was obtained by passing through a fully-connected (FC) and a layer normalisation layers as follows:

$$V = \{v_{cls}, v_1, ..., v_K\} = LN(FC(E_v(I))) \tag{5.2}$$

For text inputs, we employed BERT as the text encoder. All words were first tokenised using WordPieces to produce the token sequence $T = \{t_{cls}, t_1, ..., t_L\}$, where t_i is the i-th token and L is the sequence length. To extract sentence embeddings from the BERT model, the most common method involves using the output of the first token, designated as the [CLS] token t_{cls}. The final embedding for each token was generated by summing its token embedding and position embedding, followed by applying a normalisation layer.

$$T = \{t_{cls}, t_1, ..., t_L\} = LN(E_t(C)) \tag{5.3}$$

5.2.2 *Trimodal Encoder*

Based on the initial EEG embedding B, vision embedding V, and text embedding T, we developed a Transformer-based trimodal encoder that jointly uses the visual and textual features to enhance the brain activity representation learning and downstream visual object classification performance. Specifically, we merged B, V and T into a unified feature token sequence, which was then passed through the trimodal encoder to obtain the contextualised representation U. The self-attention mechanism in the Transformer allowed the encoder to incorporate every element within the unified sequence to generate representations for a particular position, which not only fused the information at the current modality but also integrated information from its context and other modalitites. This can be represented as:

$$\mathrm{U} = \text{ Trimodal Encoder } (\mathrm{V} \oplus \mathrm{B} \oplus \mathrm{T}) \tag{5.4}$$

where $\oplus$ indicates the concatenate operation. It's crucial to highlight that this merging was conducted along the sequence dimension, not the hidden layer dimension, ensuring that the intrinsic temporal dynamics and patterns of each modality were preserved and utilised.

5.2.3 *Multi-objective Modelling*

We pretrained our model on the randomised EEG visual dataset with multi-objective modelling: instance-, modality-level optimisation, which are performed simultaneously.

5.2.3.1 *Instance-level Optimisation*

To enhance the alignment between each paired training instance from three modalities, i.e. EEG, vision, and text, we proposed four objectives: EEG-image contrastive learning (EIC) and EEG-text contrastive learning (ETC), EEG-image matching (EIM) and EEG-text matching (ETM) on the multimodal encoder.

EEG-image contrastive learning(EIC) Our primary objective was to harness the strengths of the CLIP (Contrastive Language-Image Pre-training) approach to establish a meaningful and robust similarity function between EEG sequences and images. Given that parallel EEG-image pairs are hypothesized to have inherent correlations, our method ensures these parallel pairs exhibit higher similarity scores when represented in the shared latent space.

We used a bilinear similarity function defined as:

$$\mathrm{s}(\mathrm{B}, \mathrm{V}) = \mathrm{g_b}(\mathrm{B})^{\mathrm{T}}\, \mathrm{g_v}(\mathrm{V}) \tag{5.5}$$

where g_b, g_v are transformation functions for EEG and image embeddings, respectively. These functions project the respective representations into a shared latent space. To ensure that the similarity score sis higher for parallel EEG-image pairs and lower for non-parallel pairs, we used a contrastive loss function. The contrastive loss was given by:

$$L^{eic} = -\log \left(\frac{\exp(s(B, V)/\tau)}{\exp(s(B, V)/\tau) + \sum_{n=1}^{N} \exp(s(B, V_n)/\tau)} \right) \quad (5.6)$$

where N represents the set of negative (non-parallel) EEG-image pairs and τ is a temperature parameter controlling the sharpness of the distribution. The objective was to minimise this loss during training, driving the model to increase the similarity score for genuine pairs while reducing it for non-matching pairs.

EEG-text contrastive learning(ETC) Same as EIC, we used a bilinear similarity function defined as:

$$s(B, T) = g_b(B)^T g_t(T) \quad (5.7)$$

where g_b, g_t were transformation functions for EEG and image embeddings, respectively. These functions project the respective representations into a shared latent space. To ensure that the similarity score sis higher for parallel EEG-image pairs and lower for non-parallel pairs, we used a contrastive loss function. The contrastive loss was given by:

$$L_{etc} = -\log \left(\frac{\exp(s(B, T)/\tau)}{\exp(s(B, T)/\tau) + \sum_{n=1}^{M} \exp(s(B, T_n)/\tau)} \right) \quad (5.8)$$

where N represents the set of negative (non-parallel) EEG-text pairs and τ is a temperature parameter controlling the sharpness of the distribution.

EEG-image matching (EIM) In EEG-Text Matching, the goal is to determine if a given EEG and image pairing is positive (aligned) or negative (misaligned). To achieve this, we used the output from the multimodal encoder as the combined representation of the image-EEG-text pair. Subsequently, we added a fully-connected (FC) layer, which was then followed by a softmax function, yielding a probability p^{eim} across three classes. The EIM loss was:

$$L_{eim} = E_{(B,V)\sim D} H\left(\mathbf{y}^{eim}, p^{eim}(B, V, T)\right) \quad (5.9)$$

where y^{eim} is a 2-dimensional one-hot vector representing the ground-truth label, where negative pairs have a probability of 0 and the positive pair has a probability of 1. H is the cross-entropy function.

EEG-text matching (ETM) Same as EIM, the ETM loss is:

$$L_{etm} = E_{(B,T)\sim D} H\left(\mathbf{y}^{etm}, p^{etm}(B, V, T)\right) \tag{5.10}$$

where y^{etm} is a 2-dimensional one-hot vector representing the ground-truth label.

5.2.3.2 *Modality-level Modelling*

To enhance the trimodal encoder's generative capability and improve representation learning, we incorporated a modality-level modelling task combined with a masking mechanism for classification. Instead of masking individual token embeddings, one or two entire modalities are masked from the input, with each modality having a 30% chance of being masked independently. To ensure functionality, scenarios where all modalities are masked are explicitly avoided. This approach allows the trimodal encoder to adapt flexibly to varying input configurations, accommodating single-, dual-, or tri-modal inputs effectively.

Specifically, We attached a linear layer after the average pooling output of trimodal encoder for classification. The loss function was,

$$L_{mlm} = E_{(B,V,T)\sim D} H\left(\mathbf{y}^{mlm}, p(FC(\mathbf{U}))\right) \tag{5.11}$$

where U is the representations generated by the trimodal encoder and $\mathbf{y}^{mlm}$ denotes a one-hot vector of ground-truth label.

5.2.4 *Training*

Based on the trimodal encoder, instance-level optimisation minimised the representation similarity between instances of paired modalities and the modality-level optimisation enhanced the generation capability to adapt to different numbers of modalities with single-, two-, or three-modal inputs. We combine different loss from the instance- and modality-level optimisation to perform the overall learning by minimising:

$$L = L_{eic} + L_{etc} + L_{eim} + L_{etm} + L_{mlm} \tag{5.12}$$

5.3 experiments

5.3.1 ***Datasets***

We validated our method using two publicly available EEG datasets that measured brain activity evoked by visual stimuli. As described in the Section 5.1, both of their data acquisition are based on the

Table 9: Statistics of the datasets

	Purdue-1.0	Purdue-2.0
Subjects	6	1
Classes	40	40
Images per class	50	1000
Total images	2000	40000
Total EEG trails	2000	40000
Trail length (s)	0.5	0.5
Channels	96	96
Downsampled (Hz)	1024	1024

randomized experimental design. Table 9 summarises the statistics of the datasets used in our experiments.

Li et al. [107] is a 10-subject EEG dataset, denoted as **Purdue-1.0**. The EEG trails were collected from each participant while observing 2000 images of visual objects. These 2000 images were selected from ImageNet [154] and belonged to 40 visual categories, each category containing 50 images. For pre-processing, EEG signals were down-sampled at 1024 Hz and we considered the first 500ms of the onscreen time for classification to avoid noise.

Ahmed et al. [3] is a single-subject EEG dataset, denoted as **Purdue-2.0**. It expands upon Purdue-1.0 by increasing the number of images for each visual category. EEG trials were collected by presenting a total of 40,000 images across 40 visual categories. The EEG apparatus used was the same as Purdue-1.0. We thus followed the same pre-processing steps as Purdue-1.0.

For training, each dataset was split into five folds. Two sets were randomly chosen for validation and testing, while the remaining sets were used to train the model.

5.3.2 *Implementation Details*

The training process was conducted using PyTorch on a single NVIDIA Volta GPU. A grid search was employed to optimise the hyperparameters across all four classification tasks. For the single-modal encoder, we utilised the pre-trained wave2vec 2.0 framework to tokenise the input audio and extract audio token features. For the cross-modal encoder, we adopted the BERT-base model [7], which consists of 12 layers of Transformer blocks, each with 12 attention heads and a hidden size of 768. The text decoder and the Transformer decoder within the vision decoder comprised 6 layers of blocks. The reconstructed image size was 64×64, while the image code size was 8×8, with an embedding dimensionality of 8192.

Table 10: Visual classification accuracy (%) for all subjects of Purdue-1.0 dataset trained using EEG, visual (V), textual (T) and combined (V&T) features, respectively.

Class	Modality	Subject 1		Subject 2		Subject 3		Subject 4		Subject 5		Subject 6		Average	
		top-1	top-5	top-1	top-5	top-1	top-5	top-1	top-5	top-1	top-5	top-1	top-5	top-1	top-5
2	EEG	0.435	-	0.493	-	0.508	-	0.518	-	0.523	-	0.503	-	0.497	-
	+ V	0.491	-	0.519	-	0.589	-	0.583	-	0.590	-	0.562	-	0.556	-
	+ T	0.298	-	0.233	-	0.298	-	0.273	-	0.269	-	0.215	-	0.264	-
	+ V&T	0.651	-	0.665	-	0.645	-	0.658	-	0.660	-	0.641	-	0.653	-
5	EEG	0.335	-	0.358	-	0.350	-	0.395	-	0.349	-	0.344	-	0.355	-
	+ V	0.379	-	0.409	-	0.395	-	0.354	-	0.404	-	0.387	-	0.388	-
	+ T	0.254	-	0.207	-	0.259	-	0.221	-	0.201	-	0.218	-	0.227	-
	+ V&T	0.470	-	0.458	-	0.450	-	0.454	-	0.461	-	0.441	-	0.456	-
10	EEG	0.178	0.445	0.207	0.405	0.174	0.392	0.186	0.389	0.171	0.415	0.186	0.405	0.184	0.409
	+ V	0.179	0.410	0.224	0.426	0.198	0.429	0.208	0.425	0.201	0.446	0.215	0.445	0.204	0.430
	+ T	0.089	0.277	0.158	0.273	0.085	0.299	0.090	0.310	0.072	0.304	0.084	0.299	0.096	0.294
	+ V&T	0.233	0.478	0.249	0.463	0.236	0.459	0.245	0.460	0.233	0.472	0.248	0.468	0.241	0.467
20	EEG	0.089	0.254	0.131	0.260	0.102	0.261	0.107	0.287	0.075	0.309	0.132	0.306	0.106	0.280
	+ V	0.102	0.273	0.135	0.281	0.094	0.283	0.096	0.308	0.066	0.318	0.133	0.318	0.104	0.297
	+ T	0.064	0.151	0.088	0.147	0.068	0.146	0.081	0.157	0.062	0.152	0.068	0.179	0.072	0.155
	+ V&T	0.126	0.302	0.142	0.296	0.113	0.302	0.114	0.316	0.101	0.326	0.140	0.320	0.123	0.310
30	EEG	0.072	0.237	0.095	0.241	0.062	0.253	0.088	0.247	0.068	0.252	0.054	0.287	0.073	0.253
	+ V	0.079	0.243	0.106	0.255	0.078	0.256	0.092	0.289	0.059	0.271	0.058	0.274	0.079	0.265
	+ T	0.061	0.138	0.085	0.126	0.056	0.137	0.077	0.142	0.050	0.131	0.062	0.168	0.065	0.140
	+ V&T	0.102	0.242	0.118	0.260	0.078	0.282	0.105	0.322	0.075	0.296	0.069	0.300	0.091	0.284
40	EEG	0.063	0.237	0.066	0.230	0.051	0.236	0.078	0.247	0.058	0.211	0.074	0.242	0.065	0.234
	+ V	0.058	0.249	0.065	0.218	0.054	0.240	0.077	0.250	0.064	0.207	0.076	0.228	0.066	0.232
	+ T	0.054	0.104	0.057	0.106	0.044	0.098	0.059	0.100	0.054	0.107	0.056	0.119	0.054	0.106
	+ V&T	0.075	0.254	0.082	0.251	0.055	0.247	0.092	0.286	0.072	0.239	0.082	0.279	0.076	0.259

5.4 results

5.4.1 *Visual Classification performance*

To achieve viusal classification, a linear layer was introduced following the average pooling output generated by the trimodal encoder. During this phase, the pre-trained model parameters were held constant, allowing only the newly added linear layer to be trained. We evaluated the classification performance using top-1 and top-5 accuracy for predicting the correct class out of 2/5/10/20/30/40 classes.

Table 10 presents the average decoding accuracy of various models across multiple subjects in the Purdue-1.0 dataset and under different modalities—EEG, +V, +T, and +V&T. Two significant observations emerged from the experimental outcomes. First, models using the combination of multiple modalities, notably +V&T, exhibited substantially higher performance than those relying on a single modality. Specifically, the model based on +V&T features leads to a remarkable improvement in the average accuracy, reaching up to 0.653. This strongly suggests that even though the stimuli presented to the subjects contain only visual information, it is plausible that the subjects subconsciously activate corresponding linguistic or other sensory representations, thereby influencing the visual decoding process. Hence, it is not surprising that multimodal models capable of learning joint representations of multiple features demonstrate superior decoding performance on 'pure' visual neural response data. Second,

the model trained on +V&T features attained an average accuracy of 0.653, markedly outperforming the baseline and well-known models. This underscores that the EEG signals triggered by natural visual stimuli contain sufficient information to significantly distinguish their categories under our specially designed multimodal decoding framework.

We evaluated our model on Purdue-2.0 dataset for predicting the correct class out of 2/5/10/20/30/40 classes. We compared our method with LSTM, k-NN, SVM, MLP, 1D CNN, EEGNet and SyncNet tested in [3], as shown in Table 11. First of all, the SVM model had the highest accuracy in class 2 level, reaching 0.708 points. However, our method outperformed all other models in higher classes, especially in class 10 and class 20 where it achieved an accuracy of 0.259 and 0.121, respectively. This suggests that our model is particularly effective for more complex classification tasks. Second, traditional models such as LSTM and k-NN consistently performed at a lower level, with accuracy rarely exceeding 0.513. These results highlight the efficacy of our model, which not only performs well on simpler tasks, but also shows strong performance in more subtle classification scenarios.

Table 11: Visual classification accuracy (%) on Purdue-2.0 dataset

Table 12: Transposed Data with Bold Maximums

Method	2	5	10	20	30	40
LSTM	0.5	0.208	0.101	0.048	0.034	0.023
k-NN	0.513	0.207	0.097	0.047	0.032	0.021
SVM	**0.708**	0.375	0.206	0.108	0.072	0.052
MLP	0.5	0.211	0.105	0.053	0.033	0.024
1D CNN	0.664	0.384	0.22	0.11	0.073	0.054
EEGNet	0.5	0.248	0.173	**0.127**	0.099	0.072
SyncNet	0.5	0.241	0.127	0.06	0.034	0.025
Ours	0.691	**0.462**	**0.259**	0.121	**0.103**	**0.09**

5.5 conclusion

In this work, we advanced our exploration into decoding human visual neural representations through multimodal learning, building on previous EEG-based and brain-inspired object classification models with the addition of a novel trimodal framework. Specifically, we introduced a trimodal contrastive learning model designed to decode visual neural representations from EEG, addressing key challenges in existing approaches, such as reliance on block-designed EEG datasets and the limited use of multimodal semantic knowledge. Our model

integrates brain, visual, and textual features through a Transformer-based trimodal encoder, effectively aligning these modalities using both instance-level and modality-level optimization techniques.

Experimental results on two randomized EEG visual datasets revealed that our model significantly outperformed existing methods, offering a more consistent and interpretable joint representation of brain-visual-linguistic features. Moreover, the trimodal framework exhibited superior classification performance compared to baseline models, emphasizing the potential of multimodal semantic integration to enhance neural decoding of visual information.

This method not only achieved enriched and cohesive brain-visual-linguistic representations but also underscores the role of human cognitive signals in advancing computer vision tasks. By bridging cognitive processes with artificial intelligence, our findings advance the understanding of visual cognition and contribute to the development of brain-inspired computational models for effective neural decoding and interpretation of visual data.

6 DISCUSSION

The significance of this thesis lies in its exploration of multimodal learning for effective integration of visual and neural representations. This research addresses an important gap in understanding about how different modalities, such as visual and textual information, can interact synergistically with brain signals to enhance representation learning. Multimodal learning remains a promising yet challenging domain, with much potential to influence fields ranging from artificial intelligence to cognitive neuroscience.

This discussion chapter synthesises the findings from the three main chapter of the thesis, which collectively contained four comprehensive studies. These studies were designed to evaluate different aspects of the proposed research questions (RQs) and to advance our understanding of multimodal integration and effective representation learning. The chapter begins by linking the findings of each study to the original research questions laid out in Chapter 1, highlighting their contributions to the overarching goals of the thesis. It also summarises the limitations of each study, providing a balanced view of the strengths and areas for improvement in the research approach. It then proceeds to discuss both the conceptual advancements and applied contributions made by this work, demonstrating the broader implications of this research. Finally, an outlook for future research and concluding remarks are provided, pointing towards new opportunities for advancing multimodal learning and its practical applications.

6.1 study 1

The study is presented in Chapter 3. This part of work focused on unsupervised image captioning, a task central to the integration of visual and linguistic modalities. It leveraged adversarial and reinforcement learning techniques to enhance the alignment between generated language and the visual relations present in images (**RQ1.1**). By addressing local misalignments, the proposed model improved both the semantic richness and accuracy of image captions. Through the use of adversarial word-level modeling, the model aligns visual relations between objects with corresponding textual representations, thereby overcoming challenges posed by corpus bias and weak local alignment.

The results show that the proposed model significantly improved caption generation by employing visual attention rewards that help

align textual captions with visual relationships (**RQ1.2**). This approach effectively mitigates common issues in unsupervised methods, such as the over-reliance on corpus-driven global features that often fail to capture local object relations. Notably, the model outperformed existing methods in several key metrics, particularly in visual consistency and the accuracy of local alignments between paired objects in an image. The successful application of reinforcement learning allows for continual improvement of captioning capabilities, which contributes to the model's competitive edge in performance.

In comparison with previous studies on image captioning, the current model sets itself apart by focusing on fine-grained local alignment rather than solely relying on global visual and textual matching. While earlier models by [55] and [41] predominantly focused on global sentence-level captioning, this study advances the field by proposing a method that incorporates adversarial training to achieve more locally coherent and semantically accurate captions. Furthermore, the comparison of different visual relation detectors shows the flexibility of the proposed model in adapting to various detection methods (**RQ1.3**), which enhances its robustness.

theoretical or practical significance of the study The study holds significant theoretical value in advancing the understanding of how multimodal learning can bridge visual and linguistic representations through unsupervised methods. By formalizing image captioning as a reinforcement learning problem, it sets a foundation for future explorations in vision-language tasks, particularly in environments where paired data is scarce. Practically, this research could lead to improvements in applications such as assistive technologies for visually impaired users, automated image description generation for social media platforms, and advanced semantic retrieval systems.

limitations Despite its promising results, the study acknowledges several limitations. The generalisability of the model across more diverse datasets with varying visual styles and thematic content remains untested. This limitation is particularly relevant given that the datasets used, such as MSCOCO and Visual Genome, primarily consist of common objects and scenes, which may not fully capture the variability of real-world visual data. To overcome this, future work could involve training and evaluating the model on more diverse datasets, including specialised or domain-specific collections that feature unique visual contexts, such as medical imaging, satellite imagery, or historical art collections.

Another limitation is the model's reliance on pre-trained object detectors and its sensitivity to corpus bias, which may constrain its applicability to niche domains where object relationships differ significantly from those found in standard datasets. One way to address this

would be to fine-tune or retrain object detectors on domain-specific data, thus improving their adaptability and accuracy in specialised contexts. Additionally, using datasets with fewer biases or incorporating data augmentation techniques could help mitigate the impact of existing biases on model performance.

Furthermore, the study is limited by its computational demands. The model's complexity, particularly in handling multimodal data, requires substantial processing power and memory, which may restrict its use in resource-limited environments or real-time applications. Future research could explore model optimisation techniques, such as pruning, quantisation, or the use of lighter architectures, to make the model more accessible without sacrificing performance.

Finally, although the model demonstrates effective multimodal integration, it does not yet incorporate adaptive learning mechanisms, which would allow it to update its knowledge in response to new data over time. Integrating continual or transfer learning approaches could enhance the model's adaptability and robustness, enabling it to refine its predictions and representations dynamically as it encounters new scenarios.

next step Integrating Physiological Signals (EEG) Future research could build upon this study by incorporating physiological signals, such as electroencephalography (EEG), into the multimodal learning framework. The combination of visual data and EEG signals could enable the decoding of visual and neural representations more holistically. By integrating EEG with visual features, researchers could develop models capable of classifying and decoding visual neural representations in a manner that reflects both external (visual) and internal (neural) responses. This approach could have substantial implications for fields such as brain-computer interfacing and cognitive neuroscience, where understanding the interaction between visual stimuli and neural activity is critical.

In conclusion, this study represents a notable contribution to the field of unsupervised multimodal learning, particularly in the domain of image captioning. While limitations exist, the proposed model offers a pathway for future exploration, particularly in multimodal learning scenarios that incorporate physiological data for more comprehensive decoding of visual and neural representations.

6.2 study 2

This study is presented in Chapter 4. The study successfully introduced an Event-Related Potential (ERP)-Long Short-Term Memory (LSTM) framework, which decoded visual evoked potentials for image classification. By extracting ERP sequences from EEG recordings and feeding them into an LSTM model, the framework achieves su-

perior classification accuracy for both category and exemplar levels (**RQ2.1**). The ERP-LSTM framework demonstrates a marked improvement over existing EEG-LSTM models, achieving a classification accuracy of 66.81% for categories and 27.08% for exemplars, substantially outperforming previous models. These results validate the efficacy of using ERP features for decoding neural representations of visual objects (**RQ2.2**), offering significant potential for future applications in brain-computer interfaces (BCIs).

The ERP-LSTM model demonstrated that extracting ERP sequences allows for a more focused representation of brain activity in response to visual stimuli. The enhanced classification performance, particularly in the cross-subject setting, suggests that ERP signals retain critical spatial and temporal features of visual processing. These results support the hypothesis that ERPs can offer a more discriminative feature space than raw EEG signals, reflecting the underlying neural mechanisms of visual recognition more effectively. The findings indicate that incorporating time-sensitive neural data through LSTM provides valuable insights into both short-term and long-term cognitive processes related to visual classification.

The findings build on previous studies that explored visual classification using EEG signals but often struggled with lower classification accuracy due to the use of raw EEG data. Notably, earlier studies such as [85] achieved lower exemplar classification accuracies (14.46%), which this study improved upon significantly. The results also advance the work of [166], whose block design had limitations in random EEG trials. By adopting a more refined ERP extraction process and an advanced LSTM model, the current study addresses these shortcomings and provides a more accurate framework for visual classification.

theoretical or practical significance The theoretical significance of this study lies in its contribution to understanding how neural representations of visual stimuli can be decoded using non-invasive EEG techniques. The ERP-LSTM framework adds value to the domain of brain-computer interfaces by offering an effective tool for decoding neural responses. Practically, the ability to classify visual objects from EEG data opens up applications in areas such as neuroprosthetics, cognitive neuroscience, and assistive technology, where understanding and translating brain activity into actionable outputs are essential. Furthermore, this study provides a foundation for integrating multimodal approaches, combining neural signals with visual inputs to create more robust AI models for cognitive tasks.

limitations Despite the promising results, the study has several limitations. Firstly, the dataset involves only a limited number of subjects (10), which may affect the generalisability of the results. To im-

prove robustness, future studies could incorporate larger and more diverse participant samples, potentially including subjects with varied demographic and cognitive profiles to capture a wider range of neural response patterns.

Additionally, while the ERP-LSTM framework demonstrates strong performance, it is constrained by the inherent limitations of EEG, such as low spatial resolution, which makes it difficult to precisely localise neural activity. To address this, future research could combine EEG with complementary neuroimaging techniques like fMRI or MEG, which offer higher spatial resolution, thereby enabling a more detailed mapping of brain activity. Such multimodal approaches could provide a richer and more accurate understanding of the neural processes involved in visual recognition.

Furthermore, the study's focus on visual classification tasks may limit the framework's applicability across other cognitive domains, such as language processing or memory recall. To enhance the generalisability of the ERP-LSTM framework, future studies could explore its application to a broader range of cognitive tasks, potentially enabling insights into different areas of cognitive function and expanding its relevance across neuroscience and psychology.

Finally, the framework heavily relies on preprocessed ERP data, which may not capture the full complexity of neural activity during visual recognition. Future work could experiment with raw EEG data to preserve more nuanced information or apply advanced preprocessing techniques that retain both temporal and spatial details. Additionally, exploring alternative neural network architectures, such as convolutional neural networks (CNNs), may allow for a more sophisticated handling of raw EEG signals, potentially capturing subtle neural dynamics that are lost in traditional ERP preprocessing.

next step For future research, it would be valuable to extend this framework towards building effective brain network models. Specifically, developing an EEG inference graph that models brain activity as a network could further improve visual EEG classification. This approach would explore how different brain regions interact during visual tasks and could enhance the interpretability of the ERP-LSTM model. Moreover, incorporating multimodal data, such as functional MRI, alongside EEG could provide a richer understanding of the neural correlates of visual recognition and further improve classification accuracy.

6.3 study 3

This study was presented in Chapter 4. The findings of this study presented a significant advance in decoding brain connectivity for visual related classification tasks, through a novel multimodal approach.

Specifically, the causal convolutional network-inference graph (CCN-IG) introduced in this work addresses both time-lagged and instantaneous causal effects, which are crucial in understanding effective brain connectivity. Multimodal deep learning techniques have been employed to accurately represent neural connections, capturing both temporal and spatial dependencies. This approach outperforms traditional methods, establishing a clear path forward in brain network modelling, particularly within the domain of visual-related EEG tasks, such as emotion recognition, fatigue assessment, motor imagery control, and visual object detection.

The results demonstrate that CCN-IG provides superior performance compared to existing methods, particularly in EEG classification tasks across various domains. The use of a multimodal deep learning framework enables the extraction of high-level feature representations, leading to more accurate models of brain connectivity (**RQ2.3**). By incorporating both time-lagged and instantaneous causality, CCN-IG creates a more dynamic and realistic model of how brain regions interact. This is particularly significant as previous studies have primarily relied on static connectivity measures, failing to capture the true complexity of brain networks. The global causal inference map (GCIM) generated by CCN-IG enhances the ability to decode neural patterns, enabling more precise classification across tasks.

When comparing these results to existing literature, it becomes clear that CCN-IG represents a substantial leap forward. Previous models, particularly those based on functional connectivity such as Pearson's Correlation and Phase Locking Value (PLV), often struggled to capture the directed causal relationships between brain regions. These approaches, while useful, relied heavily on static or average measures of connectivity and were unable to differentiate between causality and mere correlation. In contrast, CCN-IG's integration of time-lagged and instantaneous causal effects places it ahead of existing Graph Neural Networks (GNNs), which also faced limitations in accurately depicting brain connectivity (**RQ2.4**). This development provides a more complete picture of how brain regions communicate, significantly improving classification tasks as demonstrated through the experiments in this study.

theoretical and practical implications of the research

The theoretical implications of this research are far-reaching. Multimodal learning, when applied to decoding visual and neural representations, provides a framework through which we can better understand the intricacies of brain function. The introduction of CCN-IG, which is grounded in causal inference, represents a shift towards more robust models of brain connectivity. Practically, this can have a significant impact on fields such as brain-computer interfaces (BCIs), where accurate brain network models are crucial for improving the

effectiveness of control systems. The implications for emotion recognition, fatigue detection, and other EEG-based applications are profound, opening up possibilities for more responsive and adaptive systems in both clinical and everyday settings.

limitations Despite its contributions, this study is not without limitations. The reliance on EEG data, while useful, presents inherent challenges due to its susceptibility to noise and its relatively low spatial resolution compared to other neuroimaging techniques like fMRI. Additionally, the CCN-IG model, although effective, still requires extensive computational resources, which may limit its scalability for real-time applications. Furthermore, while the model was tested across several EEG classification tasks, its generalisability to other neuroimaging data, such as MEG or fNIRS, remains to be explored. Lastly, the study's focus on classification tasks, while important, does not fully exploit the potential of the model in other areas such as prediction or anomaly detection.

next step Looking forward, an exciting avenue for future research lies in the extension of CCN-IG to a trimodal learning framework. Trimodal contrastive learning, which integrates EEG, fMRI, and behavioural data, could provide a more comprehensive understanding of neural activity. By fusing these three data types, future models could overcome the limitations of single or dual-modal approaches, allowing for richer and more accurate representations of brain function. This would not only improve the decoding of visual and neural representations but could also lead to advances in areas such as cognitive state monitoring and neurorehabilitation. Future work could explore how such a model might be implemented in real-time applications, particularly in neurofeedback systems where rapid and accurate brain state classification is essential.

6.4 study 4

This study was presented in Chapter 5. This study introduced a novel trimodal contrastive learning model designed to decode visual neural representations from EEG signals. The model fused brain, visual, and linguistic modalities to create joint contextual representations using a Transformer-based trimodal encoder (**RQ3.1**). Experimental evaluations demonstrated the model's superiority in decoding EEG-based visual neural representations over existing unimodal and bimodal methods (**RQ3.2**), especially when using a randomised experimental design. This comprehensive exploration of trimodal learning not only advances neural decoding but also provides insights into the underlying cognitive mechanisms of visual neural processing in the human brain.

The results showed that the trimodal model significantly outperformed unimodal and bimodal models, with a notable improvement in classification accuracy. By incorporating visual, linguistic, and neural data, the model achieved better alignment and integration of cross-modal information, enhancing the fidelity of visual category decoding. This suggests that the fusion of semantic knowledge from different modalities contributes to a more nuanced understanding of visual neural representations. Notably, the introduction of multi-objective modelling, both at the instance and modality levels, allowed for a robust and flexible model capable of handling varying modality inputs.

Compared to traditional EEG decoding approaches, which often rely on block designs that introduce temporal biases, this study addressed the challenges of randomised designs (**RQ.3**), as proposed by Ahmed et al. (2021) and Li et al. (2021). Previous research in visual neural decoding often focused on unimodal data, particularly EEG or visual inputs, but this study expands upon that by integrating text, reflecting cognitive neuroscience findings on dual coding in the brain. This aligns with studies such as Bi (2021), which emphasise the importance of combining visual and linguistic information for cognitive processing. By moving beyond conventional unimodal frameworks, this research contributes to the growing field of multimodal learning, where cross-modal learning strategies, such as contrastive loss and matching, have shown significant promise (e.g., [166]).

theoretical or practical significance The theoretical significance of this work lies in its demonstration of how multimodal learning frameworks can enhance our understanding of the brain's visual processing capabilities. It illustrates that integrating diverse modalities (brain signals, visual stimuli, and textual representations) offers a more complete picture of neural representations. Practically, this research has potential implications for brain-computer interfaces (BCIs) and artificial intelligence (AI), where more accurate decoding of neural activity could lead to better human-computer interactions, cognitive analysis, and even medical applications in diagnosing or monitoring neurological disorders.

limitations Despite the promising results, this trimodal learning approach for image-text-EEG data has some limitations. First, the model was trained on specific datasets (Purdue-1.0 and Purdue-2.0) that may not generalise to all types of visual stimuli or broader datasets. To improve generalisability, future research could incorporate more diverse and expansive datasets, including those with a wider array of visual contexts and complexities. Additionally, training the model on open-source datasets with diverse visual and semantic features could help improve its robustness across various applications and visual domains.

The EEG data used in this study was restricted to a set of predefined visual categories, which may limit the model's applicability to more complex visual environments or real-world scenarios. Expanding the dataset to include EEG responses to a broader range of dynamic and naturalistic visual stimuli could address this limitation, enabling the model to generalise better to everyday visual experiences. Incorporating more complex visual scenes, such as natural landscapes or urban environments, could also help the model adapt to real-world contexts.

Additionally, while the use of transformer-based architectures enhanced the model's performance, it demands substantial computational resources, which may limit its applicability in resource-constrained settings, such as mobile devices or real-time applications. To mitigate this, future work could explore lightweight transformer variants, such as MobileViT or DistilBERT, or employ model optimisation techniques like pruning and quantisation to reduce computational demands while maintaining accuracy.

Furthermore, the lack of integration with other physiological signals, such as eye-tracking data, limits the scope of the multimodal framework. Adding eye-tracking or other relevant physiological data, such as galvanic skin response (GSR) or heart rate variability (HRV), could capture more comprehensive information about attention, arousal, and engagement, enriching the model's understanding of user responses. This expanded multimodal approach could provide deeper insights into cognitive states and further enhance the interpretive power of the model across different contexts.

next steps Future research should focus on expanding the scope of multimodal learning by incorporating additional physiological signals such as eye-tracking data to further enhance the accuracy of visual neural decoding. Eye movement patterns, often correlated with cognitive processes during visual tasks, could provide complementary information to EEG signals, thereby enriching the model's multimodal representation capabilities. Another avenue for future exploration would be to adapt this trimodal model to real-time applications, particularly in the field of BCIs, where it could be used to improve interaction with machines through more precise neural decoding.

7

CONCLUSIONS AND DIRECTIONS FOR FUTURE RESEARCH

This chapter presents a brief summary of our work which address the challenges and advances in multimodal deep learning for visual and neural representation. The focus of the research is on image captioning, EEG-based visual classification, and trimodal learning that integrates visual, neural, and linguistic data. The research significance lies in its potential to enhance machine learning models by using the strengths of multiple data modalities to more effectively decode complex representations. Multimodal learning, which seeks to bridge the information gaps between visual and neural data, was central to the investigations in this work, with a particular emphasis on the tasks of unsupervised image captioning and neural decoding through EEG signals.

The primary research questions examined the potential of modality mapping between visual, linguistic, and neural data, exploring how each modality contributes to decoding and understanding representations at various levels. The identified research gap highlighted limitations in existing unimodal approaches that restrict real-world applications due to their reliance on single-modality inputs and fully supervised data. This thesis aimed to advance multimodal learning by developing unsupervised frameworks and enhancing alignment between image, EEG, and text data in real-world scenarios, where diverse sensory inputs often need to be combined to achieve meaningful insights.

The four main studies conducted across the three chapters systematically addressed each research question and objective. Chapter 3 introduced an innovative framework for unsupervised image captioning, using adversarial modelling to enhance global and local alignment. This approach allowed for more accurate semantic representations by capturing relationships within visual data, significantly improving on existing models. In Chapter 4, the focus shifted to EEG-based object classification, where an ERP-LSTM model was employed to decode EEG signals in response to visual stimuli. By introducing methods for temporal feature extraction and effective connectivity through graph neural networks, this research presented a novel EEG-based classification framework that demonstrated improved classification performance through the use of brain-inspired connectivity patterns. Lastly, Chapter 5 developed a trimodal learning framework that integrated visual, neural, and linguistic data, employing contrastive learning to refine the alignment of these modali-

ties. This approach not only enhanced classification accuracy but also provided insights into brain activity patterns associated with visual processing.

Through these contributions, this thesis advances multimodal deep learning for visual and neural representation, pushing the boundaries of unsupervised learning, EEG-based decoding, and brain-inspired computing. The methods developed address gaps in data alignment, semantic representation, and multimodal integration, ultimately enabling models that better mirror human cognitive processes in understanding complex visual and neural data. This work thus opens pathways for further research on brain-computer interfaces, robust image recognition, and deeper explorations into the cognitive processes underlying human visual perception.

Future research could extend these approaches to other modalities, such as audio signals or eye-tracking data, further refining multimodal integration to enhance performance across a broader range of applications. Additionally, incorporating advanced transfer learning and continual learning techniques could enable models to adapt dynamically to new data and settings, promoting generalisability across diverse domains. Exploring real-time EEG-based decoding for interactive applications, such as adaptive learning environments or neurofeedback systems, could also be a valuable direction, enabling the development of more responsive brain-computer interfaces. Another promising area is the application of these techniques in healthcare, where multimodal models could support diagnostics by combining neuroimaging data with patient history and behavioural assessments. Lastly, future research might investigate the ethical implications of using neural data in AI systems, focusing on data privacy, interpretability, and fairness, to ensure that these innovations serve societal needs responsibly and inclusively. These avenues hold potential to further the impact of multimodal learning on both technological and societal levels.

Part I

APPENDIX

A

APPENDIX

a.1 statement of authorship form

University of South Australia

Statement of Authorship

Title of Paper	An Evoked Potential-Guided Deep Learning Brain Representation for Visual Classification
Publication Status	☑ Published ☐ Accepted for Publication ☐ Submitted for Publication ☐ Unpublished and Unsubmitted work written in manuscript style
Publication Details Relevant reference details	Zheng, X., Cao, Z. and Bai, Q., 2020. An evoked potential-guided deep learning brain representation for visual classification. In Neural Information Processing: 27th International Conference, ICONIP 2020, Bangkok, Thailand, November 18– 22, 2020, Proceedings, Part V 27 (pp. 54-61). Springer International Publishing.
PRINCIPAL AUTHOR	
Name of Principal Author (Candidate)	Xianglin Zheng
Contribution to the Paper Brief description of your work in this publication	As the first author, I was primarily responsible for conducting the research, developing the ERP-LSTM algorithm, and designing as well as implementing the experiments. I carried out extensive data analysis, including preprocessing EEG signals and evaluating the proposed framework's performance. Additionally, I prepared the initial draft of the manuscript and contributed significantly to its revisions.
Overall Percentage (%)	60%
Certification	This paper reports on original research I conducted during the period of my Higher Degree by Research candidature and is not subject to any obligations or contractual agreements with a third party that would constrain its inclusion in this thesis. I am the primary author of this paper.
Signature *Signatures will be redacted by the Library at the time of publication*	Xianglin Zheng — Date 11-Dec-24
CO-AUTHOR CONTRIBUTIONS	
By signing the Statement of Authorship, each author certifies that: I. the candidate's stated contribution to the publication is accurate (as detailed above); II. permission is granted for the candidate in include the publication in the thesis; and III. the sum of all co-author contribution is equal to 100% less the candidate's stated contribution.	
Name of Co-Author 1	Zehong Cao

The University of South Australia
www.unisa.edu.au
CRICOS Provider No. 00121B

Last updated: May 2023

University of South Australia

% Contribution to the Paper	25%		
Signature *Signatures will be redacted by the Library at the time of publication*		Date	
Name of Co-Author 2	Quan Bai		
% Contribution to the Paper	15%		
Signature *Signatures will be redacted by the Library at the time of publication*	Quan Bai Digitally signed by Quan Bai Date: 2024.12.12 09:20:01 +11'00'	Date	12-Dec-24
Name of Co-Author 3			
% Contribution to the Paper			
Signature *Signatures will be redacted by the Library at the time of publication*		Date	
Name of Co-Author 4			
% Contribution to the Paper			
Signature *Signatures will be redacted by the Library at the time of publication*		Date	
Name of Co-Author 5			
% Contribution to the Paper			
Signature *Signatures will be redacted by the Library at the time of publication*		Date	
Name of Co-Author 6			
% Contribution to the Paper			
Signature *Signatures will be redacted by the Library at the time of publication*		Date	
Name of Co-Author 7			
% Contribution to the Paper			

The University of South Australia
www.unisa.edu.au
CRICOS Provider No. 00121B

BIBLIOGRAPHY

[1] Reza Abiri, Soheil Borhani, Eric W. Sellers, Yang Jiang, and Xiaopeng Zhao. 'A Comprehensive Review of EEG-based Brain–Computer Interface Paradigms.' In: *Journal of Neural Engineering* 16.1 (2019), p. 011001.

[2] Hamad Ahmed, Ronnie B Wilbur, Hari M Bharadwaj, and Jeffrey Mark Siskind. 'Object classification from randomized EEG trials.' In: *arXiv preprint arXiv:2004.06046* (2020).

[3] Hamad Ahmed, Ronnie B. Wilbur, Hari M. Bharadwaj, and Jeffrey Mark Siskind. 'Object Classification from Randomized EEG Trials.' In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2021, pp. 3845–3854.

[4] Soraia M Alarcao and Manuel J Fonseca. 'Emotions Recognition Using EEG Signals: A survey.' In: *IEEE Transactions on Affective Computing* 10.3 (2017), pp. 374–393.

[5] Peter Anderson, Basura Fernando, Mark Johnson, and Stephen Gould. 'SPICE: Semantic Propositional Image Caption Evaluation.' In: *European Conference on Computer Vision (ECCV)*. Springer, 2016, pp. 382–398. doi: 10.1007/978-3-319-46454-1_24.

[6] Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. 'Bottom-Up and Top-Down Attention for Image Captioning and Visual Question Answering.' en. In: *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*. Salt Lake City, UT: IEEE, June 2018, pp. 6077–6086. isbn: 978-1-5386-6420-9. doi: 10.1109/CVPR.2018.00636.

[7] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh. 'VQA: Visual Question Answering.' In: *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*. 2015. doi: 10.1109/ICCV.2015.279.

[8] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. 'VQA: Visual question answering.' In: *Proceedings of the IEEE international conference on computer vision*. 2015, pp. 2425–2433.

[9] Pietro Aricò, Gianluca Borghini, Gianluca Di Flumeri, Nicolina Sciaraffa, and Fabio Babiloni. 'Passive BCI beyond the lab: current trends and future directions.' In: *Physiological measurement* 39.8 (2018), 08TR02.

[10] Luiz A Baccalá and Koichi Sameshima. 'Partial Directed Coherence: A New Concept in Neural Structure Determination.' In: *Biological Cybernetics* 84.6 (2001), pp. 463–474.

[11] Pouya Bashivan, Irina Rish, Mohammed Yeasin, and Noel Codella. 'Learning representations from EEG with deep recurrent-convolutional neural networks.' In: *arXiv preprint arXiv:1511.06448* (2015).

[12] Adnan Mehmood Bhatti, Muhammad Majid, Syed Muhammad Anwar, and Bilal Khan. 'Human Emotion Recognition and Analysis in Response to Audio Music Using Brain Signals.' In: *Computers in Human Behavior* 65 (2016), pp. 267–275.

[13] Yanchao Bi. 'Dual coding of knowledge in the human brain.' In: *Trends in Cognitive Sciences* 25.10 (2021), pp. 883–895.

[14] Andrea Brovelli, Mingzhou Ding, Anders Ledberg, Yonghong Chen, Richard Nakamura, and Steven L Bressler. 'Beta Oscillations in A Large-scale Sensorimotor Cortical Network: Directional Influences Revealed by Granger Causality.' In: *Proceedings of the National Academy of Sciences* 101.26 (2004), pp. 9849–9854.

[15] Clemens Brunner, Arnaud Delorme, and Scott Makeig. 'EEGLAB—An open source toolbox for analysis of single-trial EEG dynamics.' In: *Frontiers in Neuroscience* 7 (2013), p. 127. doi: 10.3389/fnins.2013.00127.

[16] Clemens Brunner, Robert Leeb, Gernot Müller-Putz, Alois Schlögl, and Gert Pfurtscheller. 'BCI Competition 2008–Graz data set A.' In: *Institute for Knowledge Discovery (Laboratory of Brain-Computer Interfaces), Graz University of Technology* 16 (2008), pp. 1–6.

[17] Zehong Cao, Chun-Hsiang Chuang, Jung-Kai King, and Chin-Teng Lin. 'Multi-Channel EEG Recordings during a Sustained-Attention Driving Task.' In: *Scientific Data* 6.1 (1 2019), p. 19.

[18] Zehong Cao and Chin-Teng Lin. 'Inherent Fuzzy Entropy for the Improvement of EEG Complexity Evaluation.' In: *IEEE Transactions on Fuzzy Systems* 26.2 (2017), pp. 1032–1035.

[19] Thomas A Carlson, Hinze Hogendoorn, Ryota Kanai, Juraj Mesik, and Jeremy Turret. 'High temporal resolution decoding of object position and category.' In: *Journal of Vision* 11.10 (2011), pp. 9–9.

[20] CS Chen and CS Wei. 'Mind's Eye: Image Recognition by EEG via Multimodal Similarity-Keeping Contrastive Learning.' In: *arXiv preprint arXiv:2406.16910* (2024). url: https://arxiv.org/abs/2406.16910.

[21] Ming Chen, Jun Han, Xintao Hu, et al. 'Survey of encoding and decoding of visual stimulus via fMRI: an image analysis perspective.' In: *Brain Imaging and Behavior* 8.1 (2014), pp. 7–23.

[22] Wenhu Chen, Aurelien Lucchi, and Thomas Hofmann. 'A Semi-Supervised Framework for Image Captioning.' In: *arXiv:1611.05321 [cs]* (June 2017). arXiv: 1611.05321 [cs].

[23] Chun-Hsiang Chuang, Zehong Cao, Jung-Tai King, Bing-Syun Wu, Yu-Kai Wang, and Chin-Teng Lin. 'Brain Electrodynamic and Hemodynamic Signatures Against Fatigue During Driving.' In: *Frontiers in neuroscience* 12 (2018), p. 181.

[24] Junyoung Chung, Caglar Gulcehre, KyungHyun Cho, and Yoshua Bengio. 'Empirical evaluation of gated recurrent neural networks on sequence modeling.' In: *arXiv preprint arXiv:1412.3555* (2014).

[25] Michael Cogswell, Faruk Ahmed, Ross Girshick, Larry Zitnick, and Dhruv Batra. 'Reducing overfitting in deep networks by decorrelating representations. arXiv 2015.' In: *arXiv preprint arXiv:1511.06068* ().

[26] David D Cox and Robert L Savoy. 'Functional magnetic resonance imaging (fMRI)"brain reading": detecting and classifying distributed patterns of fMRI activity in human visual cortex.' In: *Neuroimage* 19.2 (2003), pp. 261–270.

[27] Mutlu Cukurova, Michail Giannakos, and Roberto Martinez-Maldonado. 'The promise and challenges of multimodal learning analytics.' In: *British Journal of Educational Technology* 51.5 (2020), pp. 1441–1449.

[28] Bo Dai and Dahua Lin. 'Contrastive learning for image captioning.' In: *arXiv preprint arXiv:1710.02534* (2017).

[29] Navneet Dalal and Bill Triggs. 'Histograms of oriented gradients for human detection.' In: *IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*. Vol. 1. IEEE. 2005, pp. 886–893.

[30] Hans P Op De Beeck, Katrien Torfs, and Johan Wagemans. 'Perceived shape similarity among unfamiliar objects and the organization of the human object vision pathway.' In: *Journal of Neuroscience* 28.40 (2008), pp. 10111–10123.

[31] Arnaud Delorme, Terrence Sejnowski, and Scott Makeig. 'Enhanced detection of artifacts in EEG data using higher-order statistics and independent component analysis.' In: *NeuroImage* 34.4 (2007), pp. 1443–1449.

[32] Michael Denkowski and Alon Lavie. 'Meteor universal: Language specific translation evaluation for any target language.' In: *Proceedings of the ninth workshop on statistical machine translation*. 2014, pp. 376–380.

[33] S. Dhanalakshmi, R.S. Maanasaa, and R.S. Maalikaa. 'A review of emergent intelligent systems for the detection of Parkinson's disease.' In: *Biomedical Engineering Systems* (2023). url: https://link.springer.com/article/10.1007/s13534-023-00319-2.

[34] Y Ding, N Robinson, C Tong, and Q Zeng. 'LGGNet: Learning from local-global-graph representations for brain-computer interface.' In: *Neural Networks and Learning Systems* (2023). url: https://ieeexplore.ieee.org/abstract/document/10025569/.

[35] Yi Ding, Neethu Robinson, Qiuhao Zeng, Duo Chen, Aung Aung Phyo Wai, Tih-Shih Lee, and Cuntai Guan. 'Tsception: A Deep Learning Framework for Emotion Detection Using EEG.' In: *2020 International Joint Conference on Neural Networks (IJCNN)*. IEEE. 2020, pp. 1–7.

[36] Jeffrey Donahue, Lisa Anne Hendricks, Sergio Guadarrama, Marcus Rohrbach, Subhashini Venugopalan, Kate Saenko, and Trevor Darrell. 'Long-term recurrent convolutional networks for visual recognition and description.' In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2015, pp. 2625–2634.

[37] Changde Du, Kaicheng Fu, Jinpeng Li, and Huiguang He. 'Decoding Visual Neural Representations by Multimodal Learning of Brain-Visual-Linguistic Features.' In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2023), pp. 1–17. doi: 10.1109/TPAMI.2023.3263181. (Visited on 04/23/2023).

[38] Jeffrey L. Elman. 'Finding Structure in Time.' In: *Cognitive Science* 14.2 (1990), pp. 179–211. doi: 10.1207/s15516709cog1402_1.

[39] Ali Farhadi, Mohsen Hejrati, Mohammad Amin Sadeghi, Peter Young, Cyrus Rashtchian, Julia Hockenmaier, and David Forsyth. 'Every picture tells a story: Generating sentences from images.' In: *European conference on computer vision*. Springer. 2010, pp. 15–29.

[40] Daniel J. Felleman and David C. Van Essen. 'Distributed hierarchical processing in the primate cerebral cortex.' In: *Cerebral Cortex* 1.1 (1991), pp. 1–47.

[41] Yang Feng, Lin Ma, Wei Liu, and Jiebo Luo. 'Unsupervised Image Captioning.' en. In: *arXiv:1811.10787 [cs]* (Apr. 2019). arXiv: 1811.10787 [cs].

[42] Karl J Friston. 'Functional and effective connectivity: A review.' In: *Brain Connectivity* 1.1 (2011), pp. 13–36. doi: 10.1089/brain.2011.0008.

[43] Hao Gao, Junhua Mao, Jie Zhou, Zhiheng Huang, Lei Wang, and Wei Xu. 'Are You Talking to a Machine? Dataset and Methods for Multilingual Image Question.' In: *Proceedings of Advances in Neural Information Processing Systems (NeurIPS)*. 2015. url: https://proceedings.neurips.cc/paper/2015/file/8c52d7db9b7de81dcf60682d942e4d44-Paper.pdf.

[44] Jiahui Gao, Yi Zhou, Philip L. H. Yu, Shafiq Joty, and Jiuxiang Gu. 'UNISON: Unpaired Cross-Lingual Image Captioning.' In: *Proceedings of the 36th AAAI Conference on Artificial Intelligence (AAAI)*. 2022.

[45] Michael S Gazzaniga, Richard B Ivry, and GR Mangun. *Cognitive Neuroscience. The biology of the mind,(2014)*. 2006.

[46] Robert Geirhos, Carlos RM Temme, Jonas Rauber, Heiko H Schütt, Matthias Bethge, and Felix A Wichmann. 'Generalisation in humans and deep neural networks.' In: *Advances in Neural Information Processing Systems (NIPS)*. 2018, pp. 7538–7550.

[47] Ross Girshick. 'Fast R-CNN.' In: *Proceedings of the IEEE international conference on computer vision*. 2015, pp. 1440–1448.

[48] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. 'Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation.' In: *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2014.

[49] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. 'Rich feature hierarchies for accurate object detection and semantic segmentation.' In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2014, pp. 580–587.

[50] Ian J Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 'Generative adversarial networks.' In: *arXiv preprint arXiv:1406.2661* (2014).

[51] Ian Goodfellow. 'Nips 2016 tutorial: Generative adversarial networks.' In: *arXiv preprint arXiv:1701.00160* (2016).

[52] Clive WJ Granger. 'Investigating causal relations by econometric models and cross-spectral methods.' In: *Econometrica: journal of the Econometric Society* (1969), pp. 424–438.

[53] Kalanit Grill-Spector, Tammar Kushnir, Talma Hendler, Shimon Edelman, Yacov Itzchak, and Rafael Malach. 'A sequence of object-processing stages revealed by fMRI in the human occipital lobe.' In: *Human Brain Mapping* 6.4 (1998), pp. 316–328. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002.

[54] Jiaxi Gu, Xiaojun Meng, Guansong Lu, Lu Hou, Niu Minzhe, Xiaodan Liang, Lewei Yao, Runhui Huang, Wei Zhang, Xin Jiang, et al. 'Wukong: A 100 million large-scale chinese cross-modal pre-training benchmark.' In: *Advances in Neural Information Processing Systems* 35 (2022), pp. 26418–26431.

[55] Jiuxiang Gu, Shafiq Joty, Jianfei Cai, and Gang Wang. 'Unpaired Image Captioning by Language Pivoting.' In: *arXiv:1803.05526 [cs]* (July 2018). arXiv: 1803.05526 [cs].

[56] Dan Guo, Yang Wang, Peipei Song, and Meng Wang. 'Recurrent Relational Memory Network for Unsupervised Image Captioning.' en. In: *arXiv:2006.13611 [cs]* (June 2020). arXiv: 2006.13611 [cs].

[57] Stephen José Hanson, Toshihiko Matsuka, and James V Haxby. 'Combinatorial codes in ventral temporal lobe for object recognition: Haxby (2001) revisited: is there a "face" area?' In: *Neuroimage* 23.1 (2004), pp. 156–166.

[58] Demis Hassabis, Dharshan Kumaran, Christopher Summerfield, and Matthew Botvinick. 'Neuroscience-inspired artificial intelligence.' In: *Neuron* 95.2 (2017), pp. 245–258.

[59] John-Dylan Haynes and Geraint Rees. 'Predicting the orientation of invisible stimuli from activity in human primary visual cortex.' In: *Nature Neuroscience* 8.5 (2005), pp. 686–691. issn: 1546-1726.

[60] John Dylan Haynes and Geraint Rees. 'Decoding mental states from brain activity in humans.' In: *Nature Reviews Neuroscience* 7.7 (2006), pp. 523–534.

[61] Kaiming He, Georgia Gkioxari, Piotr Dollar, and Ross Girshick. 'Mask R-CNN.' In: *The IEEE International Conference on Computer Vision (ICCV)*. 2017.

[62] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 'Spatial pyramid pooling in deep convolutional networks for visual recognition.' In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 37.9 (2015), pp. 1904–1916.

[63] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 'Deep residual learning for image recognition.' In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016, pp. 770–778.

[64] Tong He, Zhi Zhang, Hang Zhang, Zhongyue Zhang, Junyuan Xie, and Mu Li. 'Bag of tricks for image classification with convolutional neural networks.' In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2019, pp. 558–567.

[65] Christoph S Herrmann, Daniel Strüber, Randolph F Helfrich, and Andreas K Engel. 'EEG Oscillations: From Correlation to Causality.' In: *International Journal of Psychophysiology* 103 (2016), pp. 12–21.

[66] Sepp Hochreiter and Jürgen Schmidhuber. 'Long short-term memory.' In: *Neural computation* 9.8 (1997), pp. 1735–1780.

[67] CR Holdgraf, JW Rieger, C Micheli, and S Martin. 'Encoding and decoding models in cognitive electrophysiology.' In: *Frontiers in Systems Neuroscience* 11 (2017), p. 61. url: https://www.frontiersin.org/articles/10.3389/fnsys.2017.00061/full.

[68] Crone R. Holdgraf, Jana W. Rieger, Claudio Micheli, et al. 'Encoding and decoding models in cognitive electrophysiology.' In: *Frontiers in Systems Neuroscience* 11 (2017), p. 61.

[69] Ukyo Honda, Yoshitaka Ushiku, Atsushi Hashimoto, Taro Watanabe, and Yuji Matsumoto. 'Removing Word-Level Spurious Alignment between Images and Pseudo-Captions in Unsupervised Image Captioning.' In: *EACL*. Online, Apr. 2021, pp. 3692–3702.

[70] T Horikawa and Y Kamitani. 'Generic decoding of seen and imagined objects using hierarchical visual features.' In: *Nature Communications* 8 (2017), p. 15037. url: https://www.nature.com/articles/ncomms15037.

[71] T Horikawa, M Tamaki, Y Miyawaki, and Y Kamitani. 'Neural decoding of visual imagery during sleep.' In: *Science* 340.6132 (2013), pp. 639–642. url: https://www.science.org/doi/abs/10.1126/science.1234330.

[72] Md Zakir Hossain, Ferdous Sohel, Mohd Fairuz Shiratuddin, and Hamid Laga. 'A Comprehensive Survey of Deep Learning for Image Captioning.' In: *arXiv:1810.04020 [cs, stat]* (Oct. 2018). arXiv: 1810.04020 [cs, stat].

[73] Hossein Hosseini, Baicen Xiao, Mayoore Jaiswal, and Radha Poovendran. 'On the limitation of convolutional neural networks in recognizing negative images.' In: *2017 16th IEEE International Conference on Machine Learning and Applications (ICMLA)*. IEEE. 2017, pp. 352–358.

[74] Yimin Hou, Shuyue Jia, Xiangmin Lun, Ziqian Hao, Yan Shi, Yang Li, Rui Zeng, and Jinglei Lv. 'GCNs-net: A Graph Convolutional Neural Network Approach for Decoding Time-resolved EEG Motor Imagery Signals.' In: *IEEE Transactions on Neural Networks and Learning Systems* (2022).

[75] David H. Hubel and Torsten N. Wiesel. 'Receptive fields, binocular interaction and functional architecture in the cat's visual cortex.' In: *The Journal of Physiology* 160.1 (1962), pp. 106–154.

[76] David H. Hubel and Torsten N. Wiesel. *Brain and visual perception: the story of a 25-year collaboration*. Oxford University Press, 2004.

[77] AG Huth, T Lee, S Nishimoto, and NY Bilenko. 'Decoding the semantic content of natural movies from human brain activity.' In: *Frontiers in Systems Neuroscience* 10 (2016), p. 81. url: https://www.frontiersin.org/articles/10.3389/fnsys.2016.00081/full.

[78] Alumit Ishai, Leslie G Ungerleider, Alex Martin, Jennifer L Schouten, and James V Haxby. 'Distributed representation of objects in the human ventral visual pathway.' In: *Proceedings of the National Academy of Sciences* 96.16 (1999), pp. 9379–9384.

[79] S Jang, SE Moon, and JS Lee. 'EEG-based emotional video classification via learning connectivity structure.' In: *IEEE Transactions on Affective Computing* (2021). url: https://ieeexplore.ieee.org/abstract/document/9609641/.

[80] Boon Ting Jap, Sundeep Lal, Paul Fischer, and Efthymios Bekiaris. 'Using EEG spectral components to assess algorithms for detecting fatigue.' In: *Expert Systems with Applications* 36.2 (2009).

[81] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. 'Scaling up visual and vision-language representation learning with noisy text supervision.' In: *International conference on machine learning*. PMLR. 2021, pp. 4904–4916.

[82] Tzyy-Ping Jung, Scott Makeig, Colin Humphries, Te-Won Lee, Martin J McKeown, Vicente Iragui, and Terrence J Sejnowski. 'Removing electroencephalographic artifacts by blind source separation.' In: *Psychophysiology* 37.2 (2000), pp. 163–178.

[83] Blair Kaneshiro, Marcos Perreau Guimaraes, Hyung-Suk Kim, Anthony M. Norcia, and Patrick Suppes. 'A Representational Similarity Analysis of the Dynamics of Object Processing Using Single-Trial EEG Classification.' In: *PLOS ONE* 10.8 (2015), e0135697.

[84] Blair Kaneshiro, Marcos Perreau Guimaraes, Hyung-Suk Kim, Anthony M Norcia, and Patrick Suppes. 'A representational similarity analysis of the dynamics of object processing using single-trial EEG classification.' In: *Plos One* 10.8 (2015), e0135697.

[85] Blair Kaneshiro, Marcos Perreau Guimaraes, Hyung-Suk Kim, Anthony M Norcia, and Patrick Suppes. 'A representational similarity analysis of the dynamics of object processing using single-trial EEG classification.' In: *Plos one* 10.8 (2015), e0135697.

[86] Andrej Karpathy and Li Fei-Fei. 'Deep Visual-semantic Alignments for Generating Image Descriptions.' In: *CVPR*. 2015, pp. 3128–3137.

[87] Andrej Karpathy et al. 'Cs231n convolutional neural networks for visual recognition.' In: *Neural Networks* 1 (2016), p. 1.

[88] KN Kay, T Naselaris, RJ Prenger, and JL Gallant. 'Identifying natural images from human brain activity.' In: *Nature* 452.7185 (2008), pp. 352–355. url: https://www.nature.com/articles/nature06713.

[89] Kendrick N. Kay, Thomas Naselaris, Ryan J. Prenger, and Jack L. Gallant. 'Identifying natural images from human brain activity.' In: *Nature* 452.7185 (2008), pp. 352–355.

[90] Lei Ke, Wenjie Pei, Ruiyu Li, Xiaoyong Shen, and Yu-Wing Tai. 'Reflective Decoding Network for Image Captioning.' en. In: *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*. Seoul, Korea (South): IEEE, Oct. 2019, pp. 8887–8896. isbn: 978-1-72814-803-8. doi: 10.1109/ICCV.2019.00898.

[91] N Khaleghi, TY Rezaii, and S Beheshti. 'Developing an efficient functional connectivity-based geometric deep network for automatic EEG-based visual decoding.' In: *Biomedical Signal Processing and Control* (2023). url: https://www.sciencedirect.com/science/article/pii/S1746809422006759.

[92] Diederik P Kingma and Jimmy Ba. 'Adam: A method for stochastic optimization.' In: *arXiv preprint arXiv:1412.6980* (2014).

[93] Wolfgang Klimesch. 'EEG alpha and theta oscillations reflect cognitive and memory performance: A review and analysis.' In: *Brain Research Reviews* 29.2–3 (1999), pp. 169–195.

[94] Sander Koelstra, Christian Mühl, Mohammad Soleymani, Jong-Seok Lee, Ashkan Yazdani, Touradj Ebrahimi, Thierry Pun, Anton Nijholt, and Ioannis Patras. 'DEAP: A Database for Emotion Analysis ;Using Physiological Signals.' In: *IEEE Transactions on Affective Computing* 3.1 (2012), pp. 18–31.

[95] Wanzeng Kong, Min Qiu, Menghang Li, Xuanyu Jin, and Li Zhu. 'Causal Graph Convolutional Neural Network For Emotion Recognition.' In: *IEEE Transactions on Cognitive and Developmental Systems* (2022), pp. 1–1. doi: 10.1109/TCDS.2022.3175538.

[96] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. 'Visual genome: Connecting language and vision using crowdsourced dense image annotations.' In: *International journal of computer vision* 123 (2017), pp. 32–73.

[97] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. 'ImageNet classification with deep convolutional neural networks.' In: *Advances in Neural Information Processing Systems*. Vol. 25. 2012, pp. 1097–1105.

[98] Eleni Kroupi, Jean-Marc Vesin, and Touradj Ebrahimi. 'Subject-independent Odor Pleasantness Classification Using Brain and Peripheral Signals.' In: *IEEE Transactions on Affective Computing* 7.4 (2015), pp. 422–434.

[99] Girish Kulkarni, Visruth Premraj, Vicente Ordonez, Sagnik Dhar, Siming Li, Yejin Choi, Alexander C Berg, and Tamara L Berg. 'Babytalk: Understanding and generating simple image descriptions.' In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 35.12 (2013), pp. 2891–2903.

[100] Iro Laina, Christian Rupprecht, and Nassir Navab. 'Towards Unsupervised Image Captioning With Shared Multimodal Embeddings.' en. In: *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*. Seoul, Korea (South): IEEE, Oct. 2019, pp. 7413–7423. isbn: 978-1-72814-803-8. doi: 10.1109/ICCV.2019.00751.

[101] Jonas Larsson and David J Heeger. 'Two Retinotopic Visual Areas in Human Lateral Occipital Cortex.' In: *Journal of neuroscience* 26.51 (2006), pp. 13128–13142.

[102] Vernon J. Lawhern, Amelia J. Solon, Nicholas R. Waytowich, Stephen M. Gordon, Chou P. Hung, and Brent J. Lance. 'EEGNet: A Compact Convolutional Neural Network for EEG-based Brain–Computer Interfaces.' In: *Journal of Neural Engineering* 15.5 (2018), p. 056013.

[103] Cunbo Li, Peiyang Li, Yangsong Zhang, Ning Li, Yajing Si, Fali Li, Dezhong Yao, and Peng Xu. 'Hierarchical Emotion-recognition Framework Based on Discriminative Brain Neural Network Topology and Ensemble Co-decision Strategy.' In: *arXiv preprint arXiv:2002.10804* (2020).

[104] Junnan Li, Ramprasaath Selvaraju, Akhilesh Gotmare, Shafiq Joty, Caiming Xiong, and Steven Chu Hong Hoi. 'Align before fuse: Vision and language representation learning with momentum distillation.' In: *Advances in neural information processing systems* 34 (2021), pp. 9694–9705.

[105] Peiyang Li, Xiaohui Gao, Cunbo Li, Chanlin Yi, Weijie Huang, Yajing Si, Fali Li, Zehong Cao, Yin Tian, and Peng Xu. 'Granger Causal Inference Based on Dual Laplacian Distribution and Its Application to MI-BCI Classification.' In: *IEEE Transactions on Neural Networks and Learning Systems* (2023).

[106] Ren Li, Jared S. Johansen, Hamad Ahmed, Thomas V. Ilyevsky, Ronnie B Wilbur, Hari M Bharadwaj, and Jeffrey Mark Siskind. 'Training on the test set? An analysis of Spampinato et al. [31].' In: *arXiv preprint arXiv:1812.07697* (2018).

[107] Ren Li, Jared S. Johansen, Hamad Ahmed, Thomas V. Ilyevsky, Ronnie B. Wilbur, Hari M. Bharadwaj, and Jeffrey Mark Siskind. 'The Perils and Pitfalls of Block Design for EEG Classification Experiments.' In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 43.1 (Jan. 2021), pp. 316–333. doi: 10.1109/TPAMI.2020.2973153.

[108] Yaguang Li, Rose Yu, Cyrus Shahabi, and Yan Liu. 'Diffusion Convolutional Recurrent Neural Network: Data-driven Traffic Forecasting.' In: *arXiv preprint arXiv:1707.01926* (2017).

[109] Renjie Liao, Xin Tao, Ruiyu Li, Ziyang Ma, and Jiaya Jia. 'Multi-Scale Structure-Aware Network for Human Pose Estimation.' In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2019.

[110] Wentong Liao, Bodo Rosenhahn, Ling Shuai, and Michael Ying Yang. 'Natural Language Guided Visual Relationship Detection.' In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*. CVPR, 2019, pp. 0–0.

[111] Timothy P. Lillicrap, Adam Santoro, Luke Marris, Colin J. Akerman, and Geoffrey Hinton. 'Backpropagation and the brain.' In: *Nature Reviews Neuroscience* 21.6 (2020), pp. 335–346.

[112] Chin-Yew Lin. 'Rouge: A package for automatic evaluation of summaries.' In: *Text summarization branches out*. 2004, pp. 74–81.

[113] Tsung-Yi Lin, Michael Maire, Serge Belongie, Lubomir Bourdev, Ross Girshick, James Hays, Pietro Perona, Deva Ramanan, C. Lawrence Zitnick, and Piotr Dollár. 'Microsoft COCO: Common Objects in Context.' In: *arXiv:1405.0312 [cs]* (Feb. 2015). arXiv: 1405.0312 [cs].

[114] Klaus Linkenkaer-Hansen, Vadim V Nikouline, J Matias Palva, and Risto J Ilmoniemi. 'Long-range temporal correlations and scaling behavior in human brain oscillations.' In: *Journal of Neuroscience* 21.4 (2001), pp. 1370–1377.

[115] Li Liu, Wanli Ouyang, Xiaogang Wang, Paul Fieguth, Jie Chen, Xinwang Liu, and Matti Pietikäinen. 'Deep learning for generic object detection: A survey.' In: *International journal of computer vision* 128 (2020), pp. 261–318.

[116] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C. Berg. 'SSD: Single shot multibox detector.' In: *European conference on computer vision*. Springer. 2016, pp. 21–37.

[117] F Lotte, L Bougrain, A Cichocki, M Clerc, M Congedo, A Rakotomamonjy, and F Yger. 'A Review of Classification Algorithms for EEG-based Brain–Computer Interfaces: A 10 Year Update.' In: *Journal of Neural Engineering* 15.3 (2018), p. 031005. doi: 10.1088/1741-2552/aab2f2.

[118] David G. Lowe. 'Distinctive image features from scale-invariant keypoints.' In: *International journal of computer vision* 60.2 (2004), pp. 91–110.

[119] Cewu Lu, Ranjay Krishna, Michael Bernstein, and Li Fei-Fei. 'Visual Relationship Detection with Language Priors.' In: *European Conference on Computer Vision (ECCV)*. 2016.

[120] Jiasen Lu, Caiming Xiong, Devi Parikh, and Richard Socher. 'Knowing When to Look: Adaptive Attention via a Visual Sentinel for Image Captioning.' en. In: *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. Honolulu, HI: IEEE, July 2017, pp. 3242–3250. isbn: 978-1-5386-0457-1. doi: 10.1109/CVPR.2017.345.

[121] Steven J Luck. *An Introduction to the Event-Related Potential Technique*. MIT Press, 2014.

[122] Parinthorn Manomaisaowapak, Anawat Nartkulpat, and Jitkomut Songsiri. 'Granger Causality Inference in EEG Source Connectivity Snalysis: A State-space Approach.' In: *IEEE Transactions on Neural Networks and Learning Systems* 33.7 (2021), pp. 3146–3156.

[123] Mariusz Maziarz. 'A Review of the Granger-causality Fallacy.' In: *The journal of philosophical economics: Reflections on economic and social issues* 8.2 (2015), pp. 86–105.

[124] Santiago Morales and Maureen E Bowers. 'Time-frequency Analysis Methods and Their Application in Developmental EEG Data.' In: *Developmental cognitive neuroscience* 54 (2022), p. 101067.

[125] Henning Müller and Devrim Unay. 'Retrieval from and understanding of large-scale multi-modal medical datasets: a review.' In: *IEEE transactions on multimedia* 19.9 (2017), pp. 2093–2104.

[126] Murugappan Murugappan, Siti Murugappan, and Norsharina Mohd Ibrahim. 'EEG signals based mental task classification using wavelet transform and support vector machine.' In: *Signal Processing: An International Journal* 4.5 (2010), pp. 32–42.

[127] Thomas Naselaris, Kendrick N Kay, Shinji Nishimoto, and Jack L Gallant. 'Encoding and decoding in fMRI.' In: *Neuroimage* 56.2 (2011), pp. 400–410.

[128] Thomas Naselaris, Kendrick N. Kay, Shinji Nishimoto, and Jack L. Gallant. 'Encoding and decoding in fMRI.' In: *NeuroImage* 56.2 (2011), pp. 400–410.

[129] Meike Nauta, Doina Bucur, and Christin Seifert. 'Causal Discovery with Attention-Based Convolutional Neural Networks.' In: *Machine Learning and Knowledge Extraction* 1.1 (Jan. 2019), pp. 312–340. issn: 2504-4990. doi: 10.3390/make1010019.

[130] Jiquan Ngiam, Aditya Khosla, Mingyu Kim, Juhan Nam, Honglak Lee, and Andrew Y Ng. 'Multimodal deep learning.' In: *Proceedings of the 28th international conference on machine learning (ICML-11)*. 2011, pp. 689–696.

[131] Jiquan Ngiam, Aditya Khosla, Mingyu Kim, Juhan Nam, Honglak Lee, and Andrew Y. Ng. 'Multimodal deep learning.' In: *Proceedings of the 28th International Conference on Machine Learning*. 2011, pp. 689–696.

[132] Kenneth A Norman, Sean M Polyn, Greg J Detre, and James V Haxby. 'Beyond mind-reading: multi-voxel pattern analysis of fMRI data.' In: *Trends in cognitive sciences* 10.9 (2006), pp. 424–430.

[133] Allan Paivio. *Imagery and verbal processes*. Psychology Press, 2013.

[134] Simone Palazzo, Concetto Spampinato, Isaak Kavasidis, Daniela Giordano, Joseph Schmidt, and Mubarak Shah. 'Decoding Brain Representations by Multimodal Learning of Neural Activity and Visual Features.' In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 43.11 (Nov. 2021), pp. 3833–3849. doi: 10.1109/TPAMI.2020.2995909.

[135] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 'Bleu: a method for automatic evaluation of machine translation.' In: *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*. 2002, pp. 311–318.

[136] Yongkoo Park and Wonzoo Chung. 'Frequency-Optimized Local Region Common Spatial Pattern Approach for Motor Imagery Classification.' In: *IEEE Transactions on Neural Systems and Rehabilitation Engineering* (2019), pp. 1378–1388.

[137] Marius V. Peelen and Paul E. Downing. 'The neural basis of visual body perception.' In: *Nature Reviews Neuroscience* 8.8 (2007), pp. 636–648.

[138] C Peng, T Guo, C Xie, X Bai, and J Zhou. 'mBGT: Encoding Brain Signals With Multimodal Brain Graph Transformer.' In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2024). url: https://ieeexplore.ieee.org/abstract/document/10589581/.

[139] Gert Pfurtscheller and Christa Neuper. 'Motor imagery activates primary sensorimotor area in humans.' In: *Neuroscience Letters* 239.2-3 (1997), pp. 65–68.

[140] Marios G Philiastides, Roger Ratcliff, and Paul Sajda. 'Neural representation of task difficulty and decision making during perceptual categorization: a timing diagram.' In: *Journal of Neuroscience* 26.35 (2006), pp. 8965–8975.

[141] Marios G. Philiastides and Paul Sajda. 'Temporal Characterization of the Neural Correlates of Perceptual Decision Making in the Human Brain.' In: *Cerebral Cortex* 16.4 (2005), pp. 509–518. issn: 1047-3211. eprint: https://academic.oup.com/cercor/article-pdf/16/4/509/1084096/bhi130.pdf.

[142] Soujanya Poria, Erik Cambria, Rajiv Bajpai, and Amir Hussain. 'A Review of Affective Computing: From Unimodal Analysis to Multimodal Fusion.' In: *Information Fusion* 37 (2017), pp. 98–125. doi: 10.1016/j.inffus.2017.02.003.

[143] A. F. Pérez-Zapata, A. F. Cardona-Escobar, J. A. Jaramillo-Garzón, and Gloria M. Díaz. 'Deep Convolutional Neural Networks and Power Spectral Density Features for Motor Imagery Classification of EEG Signals.' In: *Augmented Cognition: Intelligent Technologies*. Cham: Springer International Publishing, 2018, pp. 158–169. isbn: 978-3-319-91470-1. doi: 10.1007/978-3-319-91470-1_14.

[144] Yu Qin, Yu Zhan, Changming Wang, Jiacai Zhang, Li Yao, Xiaojuan Guo, Xia Wu, and Bin Hu. 'Classifying four-category visual objects using multiple ERP components in single-trial ERP.' In: *Cognitive Neurodynamics* 10.4 (2016), pp. 275–285.

[145] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 'Learning transferable visual models from natural language supervision.' In: *International conference on machine learning*. PMLR. 2021, pp. 8748–8763.

[146] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, et al. 'Learning transferable visual models from natural language supervision.' In: *Proceedings of the 38th International Conference on Machine Learning*. 2021, pp. 8748–8763.

[147] Ramadan A. Ramadan and Athanasios V. Vasilakos. 'Brain computer interface: Control signals review.' In: *Neurocomputing* 223 (2017), pp. 26–44.

[148] Waseem Rawat and Zenghui Wang. 'Deep convolutional neural networks for image classification: A comprehensive review.' In: *Neural computation* 29.9 (2017), pp. 2352–2449.

[149] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. 'You only look once: Unified, real-time object detection.' In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016, pp. 779–788.

[150] Andrew T. Reid et al. 'Advancing Functional Connectivity Research from Association to Causation.' In: *Nature Neuroscience* 22.11 (Nov. 2019), pp. 1751–1760. doi: 10.1038/s41593-019-0510-4.

[151] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. 'Faster R-CNN: Towards real-time object detection with region proposal networks.' In: *Advances in neural information processing systems*. Vol. 28. 2015, pp. 91–99.

[152] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. 'Faster R-CNN: towards real-time object detection with region proposal networks.' In: *IEEE transactions on pattern analysis and machine intelligence* 39.6 (2016), pp. 1137–1149.

[153] Steven J. Rennie, Etienne Marcheret, Youssef Mroueh, Jarret Ross, and Vaibhava Goel. 'Self-Critical Sequence Training for Image Captioning.' In: *arXiv:1612.00563 [cs]* (Nov. 2017). arXiv: 1612.00563 [cs].

[154] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. 'Imagenet large scale visual recognition challenge.' In: *International journal of computer vision* 115 (2015), pp. 211–252.

[155] Robin Tibor Schirrmeister, Jost Tobias Springenberg, Lukas Dominique Josef Fiederer, Martin Glasstetter, Katharina Eggensperger, Michael Tangermann, Frank Hutter, Wolfram Burgard, and Tonio Ball. 'Deep Learning with Convolutional Neural Networks for EEG Decoding and Visualization.' In: *Human brain mapping* 38.11 (2017), pp. 5391–5420.

[156] Florian Schroff, Dmitry Kalenichenko, and James Philbin. 'FaceNet: A Unified Embedding for Face Recognition and Clustering.' In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2015, pp. 815–823. doi: 10.1109/CVPR.2015.7298682.

[157] K Seeliger, U Güçlü, L Ambrogioni, Y Güçlütürk, et al. 'Generative adversarial networks for reconstructing natural images from brain activity.' In: *NeuroImage* 181 (2018), pp. 775–785. url: https://www.sciencedirect.com/science/article/pii/S105381191830658X.

[158] Prashant Giridhar Shambharkar, Priyanka Kumari, Pratik Yadav, and Rajat Kumar. 'Generating Caption for Image using Beam Search and Analyzation with Unsupervised Image Captioning Algorithm.' In: *ICICCS*. 2021, pp. 857–864.

[159] G Shen, T Horikawa, and K Majima. 'Deep image reconstruction from human brain activity.' In: *PLOS Computational Biology* 15.1 (2019), e1006633. url: https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1006633.

[160] Irina Simanova, Marcel Van Gerven, Robert Oostenveld, and Peter Hagoort. 'Identifying object categories from event-related EEG: toward decoding of conceptual representations.' In: *Plos One* 5.12 (2010), e14465.

[161] Karen Simonyan and Andrew Zisserman. 'Very deep convolutional networks for large-scale image recognition.' In: *arXiv preprint arXiv:1409.1556* (2014).

[162] Kihyuk Sohn, Wenling Shang, and Honglak Lee. 'Improved multimodal deep learning with variation of information.' In: *Advances in neural information processing systems* 27 (2014).

[163] Mohammad Soleymani, Sadjad Asghari-Esfeden, Yun Fu, and Maja Pantic. 'Analysis of EEG signals and facial expressions for continuous emotion detection.' In: *IEEE Transactions on Affective Computing* 7.1 (2015), pp. 17–28.

[164] P. Song, D. Guo, J. Zhou, and M. Xu. 'Memorial GAN with joint semantic optimization for unpaired image captioning.' In: *IEEE Transactions on Multimedia* (2022). url: https://ieeexplore.ieee.org/abstract/document/9784948/.

[165] Tengfei Song, Wenming Zheng, Peng Song, and Zhen Cui. 'EEG Emotion Recognition Using Dynamical Graph Convolutional Neural Networks.' In: *IEEE Transactions on Affective Computing* 11.3 (Mar. 2018), pp. 532–541. doi: 10.1109/TAFFC.2018.2817622.

[166] Concetto Spampinato, Simone Palazzo, Isaak Kavasidis, Daniela Giordano, Nasim Souly, and Mubarak Shah. 'Deep learning human mind for automated visual classification.' In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2017, pp. 6809–6817.

[167] Concetto Spampinato, Simone Palazzo, Isaak Kavasidis, Daniela Giordano, Nasim Souly, and Mubarak Shah. 'Deep_Learning_Human_CVPR_2017 In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2017, pp. 6809–6817.

[168] Kevin M Spencer, Joseph Dien, and Emanuel Donchin. 'A componential analysis of the ERP elicited by novel events using a dense electrode array.' In: *Psychophysiology* 38.2 (2001), pp. 238–247.

[169] Nitish Srivastava and Russ R Salakhutdinov. 'Multimodal learning with deep boltzmann machines.' In: *Advances in neural information processing systems* 25 (2012).

[170] Cornelis J Stam, Guido Nolte, and Andreas Daffertshofer. 'Phase lag index: Assessment of functional connectivity from multi channel EEG and MEG with diminished bias from common sources.' In: *Human Brain Mapping* 28.11 (2007), pp. 1178–1193.

[171] Dustin E. Stansbury, Thomas Naselaris, and Jack L. Gallant. 'Natural scene statistics account for the representation of scene categories in human visual cortex.' In: *Neuron* 79.5 (2013), pp. 1025–1034.

[172] M. Stefanini, M. Cornia, and L. Baraldi. 'From show to tell: A survey on deep learning-based image captioning.' In: *IEEE Transactions on Image Processing* (2022). url: https://ieeexplore.ieee.org/abstract/document/9706348/.

[173] Klaas E Stephan, Will D Penny, Rosalyn J Moran, Hanneke EM den Ouden, Jean Daunizeau, and Karl J Friston. 'Ten simple rules for dynamic causal modeling.' In: *NeuroImage* 49.4 (2010), pp. 3099–3109.

[174] Sebastian Stober, Avital Sternin, Adrian M Owen, and Jessica A Grahn. 'Deep feature learning for EEG recordings.' In: *arXiv preprint arXiv:1511.04306* (2015).

[175] Richard S Sutton, David A McAllester, Satinder P Singh, Yishay Mansour, et al. 'Policy gradient methods for reinforcement learning with function approximation.' In: *NIPs*. Vol. 99. Citeseer. 1999, pp. 1057–1063.

[176] Christian Szegedy, Sergey Ioffe, Vincent Vanhoucke, and Alexander Alemi. 'Inception-v4, Inception-ResNet and the Impact of Residual Connections on Learning.' In: *Proceedings of the AAAI Conference on Artificial Intelligence* 31.1 (2017). url: https://ojs.aaai.org/index.php/AAAI/article/view/11231.

[177] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, et al. 'Going deeper with convolutions.' In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2015, pp. 1–9.

[178] Yaniv Taigman, Ming Yang, Marc'Aurelio Ranzato, and Lior Wolf. 'DeepFace: Closing the Gap to Human-Level Performance in Face Verification.' In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2014, pp. 1701–1708. doi: 10.1109/CVPR.2014.220.

[179] Catherine Tallon-Baudry and Olivier Bertrand. 'Oscillatory gamma activity in humans and its role in object representation.' In: *Trends in Cognitive Sciences* 3.4 (1999), pp. 151–162.

[180] Siyi Tang, Jared Dunnmon, Khaled Kamal Saab, Xuan Zhang, Qianying Huang, Florian Dubost, Daniel Rubin, and Christopher Lee-Messer. 'Self-Supervised Graph Neural Networks for Improved Electroencephalographic Seizure Analysis.' In: *International Conference on Learning Representations*. 2022.

[181] Christopher Torrence and Gilbert P Compo. 'A practical guide to wavelet analysis.' In: *Bulletin of the American Meteorological Society* 79.1 (1998), pp. 61–78.

[182] Yao-Hung Hubert Tsai, Paul Pu Liang, Amir Zadeh, Louis-Philippe Morency, and Ruslan Salakhutdinov. 'Learning Factorized Multimodal Representations.' In: *Proceedings of the International Conference on Representation Learning (ICLR)*. 2019. url: https://openreview.net/forum?id=rJl0KP5cF7.

[183] Alan M Turing. *Computing machinery and intelligence*. Springer, 2009.

[184] DC Van Essen and CH Anderson. 'Information processing strategies and pathways in the primate visual system.' In: *An introduction to neural and electronic systems* (1995). url: https://www.rctn.org/bruno/public/papers/vanessen-anderson.pdf.

[185] Marcel A. J. Van Gerven. 'A primer on encoding models in sensory neuroscience.' In: *Journal of Mathematical Psychology* 76 (2017), pp. 172–183.

[186] Hesam Varsehi and S Mohammad P Firoozabadi. 'An EEG Channel Selection Method for Motor Imagery Based Brain–Computer Interface and Neurofeedback Using Granger Causality.' In: *Neural Networks* 133 (2021), pp. 193–206.

[187] A Vaswani. 'Attention is all you need.' In: *Advances in Neural Information Processing Systems* (2017).

[188] Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. 'Cider: Consensus-based image description evaluation.' In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2015, pp. 4566–4575.

[189] Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. 'Show and Tell: A Neural Image Caption Generator.' en. In: *arXiv:1411.4555 [cs]* (Apr. 2015). arXiv: 1411.4555 [cs].

[190] Changming Wang, Shi Xiong, Xiaoping Hu, Li Yao, and Jiacai Zhang. 'Combining features from ERP components in single-trial EEG for discriminating four-category visual objects.' In: *Journal of Neural Engineering* 9.5 (2012), p. 056013. issn: 1741-2552.

[191] D.H. Wang, J. Li, and S. Zhu. 'Few-labeled visual recognition for self-driving using multi-view visual-semantic representation.' In: *Neurocomputing* (2021). url: https://www.sciencedirect.com/science/article/pii/S0925231220311978.

[192] Tao Wang and Bin He. 'An Efficient Rhythmic Component Expression and Weighting Synthesis Strategy for Classifying Motor Imagery EEG in A Brain–Computer Interface.' In: *Journal of neural engineering* 1.1 (2004), p. 1.

[193] ZhongMin Wang, Yang Liu, Rong Zhang, Jie Zhang, and XiaoMin Guo. 'EEG-Based Emotion Recognition Using Partial Directed Coherence Dense Graph Propagation.' In: *2022 14th International Conference on Measuring Technology and Mechatronics Automation (ICMTMA)*. IEEE. 2022, pp. 610–617.

[194] Andreas Widmann, Erich Schröger, and Burkhard Maess. 'Digital filter design for electrophysiological data—A practical approach.' In: *Journal of Neuroscience Methods* 250 (2015), pp. 34–46.

[195] Dongrui Wu, Xue Jiang, and Ruimin Peng. 'Transfer Learning for Motor Imagery Based Brain-Computer Interfaces: A Tutorial.' In: *Neural Networks* (2022).

[196] Shaomei Wu, Jeffrey Wieland, Omid Farivar, and Julie Schiller. 'Automatic alt-text: Computer-generated image descriptions for blind users on a social network service.' In: *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*. 2017, pp. 1180–1192.

[197] Zonghan Wu, Shirui Pan, Fengwen Chen, Guodong Long, Chengqi Zhang, and S Yu Philip. 'A Comprehensive Survey on Graph Neural Networks.' In: *IEEE transactions on neural networks and learning systems* 32.1 (2020), pp. 4–24.

[198] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhutdinov, Richard Zemel, and Yoshua Bengio. 'Show, Attend and Tell: Neural Image Caption Generation with Visual Attention.' en. In: *arXiv:1502.03044 [cs]* (Apr. 2016). arXiv: 1502.03044 [cs].

[199] Ning Xu, Hanwang Zhang, An-An Liu, Weizhi Nie, Yuting Su, Jie Nie, and Yongdong Zhang. 'Multi-level policy and reward-based deep reinforcement learning framework for image captioning.' In: *IEEE Transactions on Multimedia* 22.5 (2019), pp. 1372–1383.

[200] Xu Yang, Kaihua Tang, Hanwang Zhang, and Jianfei Cai. 'Auto-Encoding Scene Graphs for Image Captioning.' en. In: *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. Long Beach, CA, USA: IEEE, June 2019, pp. 10677–10686. isbn: 978-1-72813-293-8. doi: 10.1109/CVPR.2019.01094.

[201] Zesheng Ye, Lina Yao, Yu Zhang, and Sylvia Gustin. *See What You See: Self-supervised Cross-modal Retrieval of Visual Stimuli from Brain Activity*. Aug. 18, 2022. arXiv: 2208.03666 [cs]. (Visited on 10/13/2022). preprint.

[202] Jae-Chern Yoo and Tae Hee Han. 'Fast Normalized Cross-correlation.' In: *Circuits, Systems and Signal Processing* 28 (2009), pp. 819–843.

[203] Quanzeng You, Hailin Jin, Zhaowen Wang, Chen Fang, and Jiebo Luo. 'Image Captioning with Semantic Attention.' In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2016. doi: 10.1109/CVPR.2016.374.

[204] Rowan Zellers, Mark Yatskar, Sam Thomson, and Yejin Choi. 'Neural Motifs: Scene Graph Parsing with Global Context.' In: *arXiv:1711.06640 [cs]* (Mar. 2018). arXiv: 1711.06640 [cs].

[205] Hong Zeng, Chen Yang, Guojun Dai, Feiwei Qin, Jianhai Zhang, and Wanzeng Kong. 'EEG Classification of Driver Mental States by Deep Learning.' In: *Cognitive neurodynamics* 12.6 (2018), pp. 597–606.

[206] C. Zhang, J. Cheng, and Q. Tian. 'Multiview semantic representation for visual recognition.' In: *IEEE Transactions on Cybernetics* (2018). url: https://ieeexplore.ieee.org/abstract/document/8525319/.

[207] C. Zhang, J. Cheng, and Q. Tian. 'Multi-view image classification with visual, semantic and view consistency.' In: *IEEE Transactions on Image Processing* (2019). url: https://ieeexplore.ieee.org/abstract/document/8804379/.

[208] C. Zhang and D.H. Wang. 'Exploring the prediction consistency of multiple views for transductive visual recognition.' In: *IEEE Signal Processing Letters* (2021). url: https://ieeexplore.ieee.org/abstract/document/9381669/.

[209] Jianhua Zhang, Zhong Yin, Peng Chen, and Stefano Nichele. 'Emotion Recognition Using Multi-modal Data and Machine Learning Techniques: A Tutorial and Review.' In: *Information Fusion* 59 (2020), pp. 103–126.

[210] Xiang Zhang, Lina Yao, Chaoran Huang, Tao Gu, Zheng Yang, and Yunhao Liu. 'DeepKey: an EEG and gait based dual-authentication system.' In: *arXiv preprint arXiv:1706.01606* (2017).

[211] Qijie Zhao, Tao Sheng, Ying Wang, Zhishen Tang, Ying Chen, and Liang Cai. 'M2Det: A single-shot object detector based on multi-level feature pyramid network.' In: *Proceedings of the AAAI Conference on Artificial Intelligence* 33.01 (2019), pp. 9259–9266. url: https://aaai.org/ojs/index.php/AAAI/article/view/4962.

[212] X Zheng, W Chen, Y You, Y Jiang, M Li, and T Zhang. 'Ensemble deep learning for automated visual classification using EEG signals.' In: *Pattern Recognition* (2020). url: https://www.sciencedirect.com/science/article/pii/S0031320319304480.

[213] Peixiang Zhong, Di Wang, and Chunyan Miao. 'EEG-Based Emotion Recognition Using Regularized Graph Neural Networks.' In: *IEEE Transactions on Affective Computing* 13.3 (2022), pp. 1290–1301. doi: 10.1109/TAFFC.2020.2994159.

[214] P. Zhu, X. Wang, L. Zhu, and Z. Sun. 'Prompt-based learning for unpaired image captioning.' In: *arXiv* (2023). url: https://arxiv.org/pdf/2205.13125.

colophon