AIML at VQA-Med 2020: Knowledge Inference via a Skeleton-based Sentence Mapping Approach for Medical Domain Visual Question Answering

Zhibin Liao[1,2], Qi Wu[1], Chunhua Shen[1],
Anton van den Hengel[1], and Johan Verjans[1,2]

[1] Australian Institute for Machine Learning, University of Adelaide, Australia
[2] South Australian Health and Medical Research Institute, Adelaide, Australia

Abstract. In this paper, we describe our contribution to the 2020 ImageCLEF Medical Domain Visual Question Answering (VQA-Med) challenge. Our submissions scored first place on the VQA challenge leaderboard, and also the first place on the associated Visual Question Generation (VQG) challenge leaderboard. Our VQA approach was developed using a knowledge inference methodology called Skeleton-based Sentence Mapping (SSM). Using all the questions and answers, we derived a set of classifiable tasks and inferred the corresponding labels. As a result, we were able to transform the VQA task into a multi-task image classification problem which allowed us to focus on the image modelling aspect. We further propose a class-wise and task-wise normalization facilitating optimization of multiple tasks in a single network. This enabled us to apply a multi-scale and multi-architecture ensemble strategy for robust prediction. Lastly, we positioned the VQG task as a transfer learning problem using the VGA task trained models. The VQG task was also solved using classification.

Keywords: Visual Question Answering · Visual Question Generation · Knowledge Inference · Deep Neural Networks · Skeleton-based Sentence Mapping · Class-wise and Task-wise Normalization

1 Introduction

Visual question answering (VQA) [4,20] is a challenging new task which requires a broad knowledge of image processing, natural language processing (NLP), and multi-modal learning. In the medical domain, VQA is an attractive topic showing great potential in automated medical image interpretation and machine supported diagnoses, with potential to benefit both medical practitioners and

A. van den Hengel and J. Verjans – Joint senior authorship.

patients. Nevertheless, medical VQA remains an unsolved problem. The ImageCLEF association [15] has been hosting the Medical Domain VQA (VQA-Med) challenges for three consequent years since 2018 [2, 5, 10]. In the 2018 challenge, the images were extracted from PubMed Central articles with the questions and answers automatically generated from image captions before checked manually by human annotators. In addition to the clarity issues of the machine generated questions as reported by [1], it is also noticeable that both the questions and ground-truth answers are in variable-length and free-form, both of which add difficulties to the answer generation task. The 2019 challenge [2] advanced from the previous challenge by narrowing the task scope: 1) using only radiology images; and 2) asking questions in four topics (*i.e.*, image modality, imaging plane, visualized organ systems, and abnormality detectable from an image). As noticed by many participated teams, the 2019 challenge is solvable in a classification manner, *i.e.*, there are 36 unique answers for the modality questions, 16 for the plane questions, 10 for the organ questions, with an exception of over a thousand possible answers for the abnormality category. A post-challenge question category-wise accuracy analysis [2] suggests that the modality, plane, and organ categories possess much better accuracy compared to the abnormality category.

In the 2020 VQA challenge, our AIML team participated in, the dataset [5] was curated with only questions in abnormality category. While analyzing the questions, we found that questions come in two major forms: 1) yes/no questions, *e.g.*, "is this image normal/abnormal?", and 2) wh-questions *e.g.*, "what is abnormal in the image?". In comparison to the last year's challenge, we noticed the unique question phrasings were reduced from 253 to 52 and the unique answer phrasings from 1,749 to 332, while having a 25% increase of images (from 3,200 to 4,000 in the training set; validation and test sets are equal), resulting in a much richer data support for the VQA task.

Our initial attempt at the 2020 VQA-Med challenge was to fine-tuning of the Pythia [27] model. However, this did not yield a desirable performance, hence after which we conducted an analysis of the predicted answers. The analysis led to the development of a novel knowledge inference method, namely Skeleton-based Sentence Mapping (SSM) that helped reverse engineer a set of question backbones. SSM helped us to determine the question categories and infer corresponding labels, reducing the VQA problem to a pure multi-task image classification problem. As a result, we were able to focus on the imaging modality. In particular, we developed a class-wise and task-wise normalization method to give balanced weighting to presented classes and tasks in a mini-batch. This helps to jointly optimize multiple tasks in a single network. At last, we applied multi-scale and multi-architecture ensemble learning. Our best submission scored 0.496 in accuracy and 0.542 in BLEU score which won the first place at the 2020 VQA challenge.

For the associated Medical Domain Visual Question Generation (VQG-Med) challenge, we considered the task as a transfer learning problem, where we applied the VQA-Med data trained models as non-trainable feature extractors. The

answer generation is also formed as a classification task. Our best submission scored 0.348 in BLEU score which won the first place at the VQG challenge.

In the rest of the paper, we give explanations on our VQA and VQG approaches. Each approach is a self-contained section to avoid cluttering.

2 VQA-Med Challenge Participation

2.1 Literature Review

We will first introduce the general domain VQA methods followed by an introduction to the methods that have been applied specifically in the medical domain VQA.

General domain VQA: the goal of a VQA method is to produce an answer from a given image-question pair. Early VQA works [4, 9, 20, 24] used a general CNN-RNN framework. In brief, the CNN-RNN approach is carried out using a Convolutional Neural Network (CNN) model (*e.g.*, VGG-Net [26]) to process the input image and a Recurrent Neural Network (RNN) Encoder–Decoder [7] (more specifically, LSTM [12]) to handle the language modelling. While the vision and language information fusion component can also be handled by the RNN language model altogether, or just by concatenation, there are also more advanced options such as the Multi-modal Factorized Bilinear (MFB) pooling and High-order pooling (MFH) [38] and MUTAN [6]. Attention is also a frequently visited topic in VQA, *e.g.*, question-guided visual attention methods [35, 37] and vision-language co-attention methods [19, 38]. Finally, semantic image representation (*e.g.*, attribute-based image representation [31]), pretrained language representation (*e.g.*, BERT [8]), external knowledge and common sense knowledge [32] could all be beneficial towards solving VQA.

Medical domain VQA: a noticeable difference between the medical domain and general domain VQA is the size of the dataset. The general domain VQA can accumulate a sizable dataset due to the fact that a common-sense knowledge is sufficient for generating question and answers. On the other hand, the necessity of clinical expertise imposes a huge difficulty in the medical domain VQA data collection.

In the 2018 VQA-Med challenge, the leading[3] three participating teams [1, 23, 39] differentiate in image modelling (*i.e.*, ResNet-152 [11], Inception-ResNet-v2 [28], VGG-16), language modelling (*i.e.*, LSTM, Bi-LSTM), vision-language fusion (*i.e.*, MFB/MFH [38], SAN [37]), attention models (*i.e.*, question guided attention [35], co-attention [38]), and word embeddings (*i.e.*, word2vec [21] or

[3] The 2018 VQA-Med challenge employed three measurements: BLEU [22], Word-based Semantic Similarity (WBSS) [33], and Concept-based Semantic Similarity (CBSS). The leading teams are referred to the BLEU and WBSS [33] score rankings. The CBSS can result a different ranking.

medical article pretrained embedding [23]). Considering the component-wise diversity and minor performance gaps, it is difficult to find out which component is favourable. However, we notice that all three teams treated the VQA task as a classification problem whereas the rest two teams treated the problem as a generation task [29] or still a classification task but not fine-tuning the image model [3].

In the 2019 VQA-Med challenge, the top three teams [30, 36, 40] (with a working notes paper) all used BERT [8] for language processing. Apart from that, we point to some of the unique techniques from the top three teams. The winning team Hanlin [36] has adopted Global Average Pooling (GAP) [18] shortcuts. This differs from the conventional position of GAP which connects the last convolution layer and the classification layer. The Hanlin team placed multiple GAPs that each links to a low-level convolution layer and forwards the pooled low-level features to be concatenated with the final image representation. The second-place team minhvu [30] adopted an ensemble learning approach with a variation of VQA components. The third-place team TUA1 [40] used a question classifier to figure out the question category and then choose answers from a set of modality, plane, and organ classifiers and a generative model for abnormality answers. Note that the question classification strategy was also employed by several other participated teams; therefore we speculate the use of BERT could have been the delimiting factor that caused a noticeable gap of 0.04 (in both accuracy and BLEU) between the third place [30] and fourth place [25] (who also used question classification and sub answer models).

2.2 Dataset

The VQA-Med 2020 dataset has a composition of 4,000 radiology images for training, 500 for validation, and 500 for testing. Each image has exactly one Question-Answer (QA) pair from the abnormality question category.

We followed the official suggestion to use the VQA-Med 2019 dataset[4] as additional training data. The VQA-Med 2019 dataset has 3,200 medical images for training, 500 for validation, and another 500 for testing. For training and validation sets, there are 12,792 and 2,000 QA pairs, giving most images exactly one QA pair in each question category (*i.e.*, imaging modality, imaging plane, organ systems, and abnormality). For the test set, each question category has 125 images. In addition, the yes-no questions appear only in the imaging modality and abnormality question categories.

2.3 Skeleton-based Sentence Mapping

As mentioned in Sec. 1, Pythia [27] was our initial attempt, from which we observed a proportion of the yes-no questions answered by categorical abnormality answers and vice versa. This could be a sign of insufficient question variations. To address this issue, we tried to develop a question generator to

[4] https://github.com/abachaa/VQA-Med-2019

populate training questions while keeping the meaning unchanged. The Skeleton-based Sentence Mapping (SSM) method was developed to summarize questions with similar sentence structures into a unified backbone. An example of the derived sentence backbones are shown in Table 1. Taking the question backbone "*is ${this_pronoun_alts} ${ct_alts} ${normal_alts}?* " as an example, we call the swap-able parts the *skeleton variables* and write in the Shell variable style "${...}". An example can be found in Table 2.[5]

Table 1. An example of the question backbones derived from the VQA-Med 2019 and 2020 datasets. The last six columns present the respective number of question instances in each set.

Dataset Questions	Question Backbones	VQA-Med 2020			VQA-Med 2019		
		train	val	test	train	val	test
is the ct scan normal?	is ${this_alts} ${ct_alts} ${normal_alts}?			1	3		1
is the mri normal?		3	2	1	6		
is the ultrasound normal?		1	1				1
is the x-ray normal?		2	4	1	3		
what abnormality is seen in the image?	what abnormality is ${imaged_alts} in ${this_alts} ${ct_alts}?	1001	105	127	776	133	20
what abnormality is seen in this x-ray?				2			
what is seen in the image?	what ${is_being_alts} ${imaged_alts} in ${this_alts} ${ct_alts}?			1			
what is seen in the x-ray?				2			
what is seen in this ct scan?				1			
what is shown in the x-ray?				1			

Table 2. Corresponding candidates for the skeleton variables appeared in Table 1. The candidate elements were extracted from the real VQA-Med 2019 and 2020 dataset questions and added with improvised ones.

Skeleton variables	Candidates
this_alts	this, the
ct_alts	ct, ct scan, mri, pet, x-ray, image, ...
normal_alts	normal, abnormal
imaged_alts	imaged, displayed, seen, shown, ...
is_being_alts	is, is being

Before applying SSM, we first removed the duplicated questions in the dataset, resulting in 266 unique questions. After then, we applied word-level edit distance (*i.e.*, levenshtein distance) to pairs of questions, finding groups of questions with 1-distance and 2-distance. For example, in Table 1, the corresponding questions of each question backbone mostly have either 1-distance or 2-distance within the group, and the highest 4-distance is between "what is shown in the x-ray?" and

[5] The naming was determined by choosing the a representative candidate from candidates for each skeleton variable; by ignoring the "alts" suffix, a question backbone becomes readable.

"what is seen in this ct scan?". The grouped questions were manually checked to see if the dissimilar parts can be described by a unified skeleton variable. If so, the generated backbone would replace the group of question and enter the next iteration of edit distance computation. The first iteration was able to detect most of the easy question groups, leave the later iterations with a small number of questions.

The process was ran until all questions were skeletonized, resulting in 68 question backbones. We labeled the question backbones in the four aforementioned question categories, partially based on the corresponding answers. In addition, we also determined two sub categories under the imaging modality category, namely the MR modality category and the contrast imaging type category. Next, we compared our own question category annotation with the official question category annotation for the VQA-Med 2019 test set (only available in this set), which is equivalent. The SSM was able to populate dynamic question variations (with some rule based restrictions, *e.g.*, changing "ct scan" in "is the ct scan normal?" to other candidates except "ct" and "image" results in a fallacious judgement of the image modality, hence is not allowed) and the same Pythia model trained with the augmented questions was able to rectify the yes-no and wh-question cross answering errors. Nevertheless, we found the SSM method rendered language modelling trivial. With its help, we can solve the VQA task as an image classification task.

Label inference from question backbones: based on the question category annotation, we were able to record the paired answer annotation as the label for each mapped task. In addition, we could also extract labels from the skeleton variables. For example, for the first question "is the ct scan normal?" in Table 1, "ct" is capturable by ${ct_alts} and "normal" is capturable by ${normal_alts}; hence producing a coarse modality label "ct", and also produce a binary abnormality label "normal" if the answer is a "yes". We found the same can also be generalized to infer task labels from the wh-questions.

An issue with the question backbone derived modality labels is that the detailed modality (*e.g.*, ct with contrast or not) is unknown. To address this issue, we treat the coarse modality labels as an independent task. The answer derived modality labels were mapped back to the coarse labels following the information provided in [2]. Next, we treated all abnormality wh-questions to have an "abnormal" label to add to the yes-no question derived binary abnormality labels.

At the end of the process, we were able to produce six classification tasks: 1) fine imaging modalities; 2) coarse imaging modalities; 3) imaging plane; 4) organ systems; 5) binary abnormality, and 6) categorical abnormality.

2.4 Multi-task Image Classification

The schematic of an exemplar image classification network we used is illustrated in Fig. 1, sketched with the knowledge inference process. The two important tasks are the binary and categorical abnormality classification tasks while the

rest four can be thought as regularization tasks. We believe that all the tasks should have strong correlation to each other, *i.e.*, the correct imaging modality and organ judgements should be strong prior knowledge for correct recognition of abnormality.

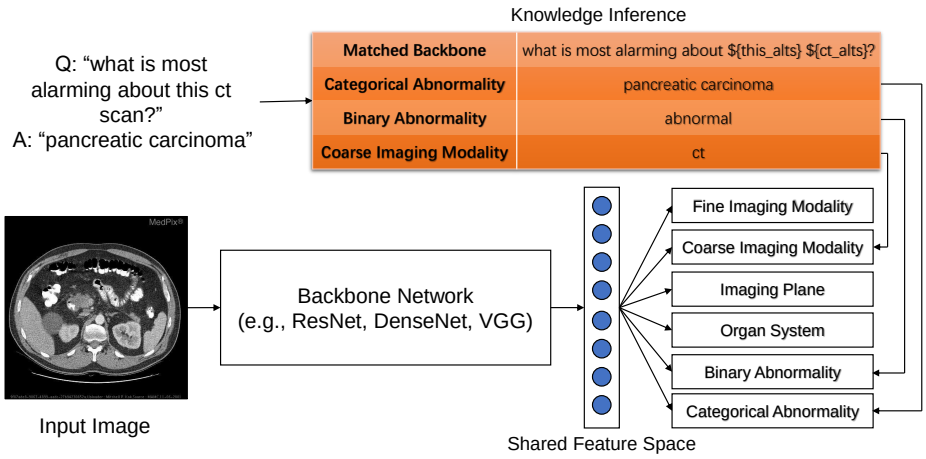

Fig. 1. The schematic of an image classification network we used and the label inference result produced by the proposed SSM method.

Class-wise and task-wise normalization: since only the 2019 challenge images have (almost) complete four QA pairs per image, a large number of images in the joint 2019 and 2020 dataset do not have a complete label set (mainly the 2020 images). Hence when all six tasks are jointly optimized via a mini-batch gradient method, a conventional normalization by the batch size effectively assigns a lower weight to a less populated task, *e.g.*, for a batch with 12 images, a task that has 3 labeled images effectively has 0.25 weighting. In addition to the incomplete label problem, we also observed imbalanced class distributions within the tasks. For example in the categorical abnormality question category, the number of samples per abnormality class ranges from 4 to 104. We propose to solve both issues together by a class-wise and task-wise normalization in order to jointly optimize all six tasks together. Assume that $t \in \{\textit{coarse modality}, \ldots\}$ represents a task, for a set of images X and the label set Y_t, the mini-batch training loss L is computed as:

$$L = \sum_t \frac{1}{\sum_{c_t} \mathbb{1}(c_t \in Y_t)} \left(\sum_{c_t} \frac{1}{\sum_{y_t \in Y_t} \mathbb{1}(y_t = c_t)} \left(\sum_{x \in X, y_t \in Y_t} \mathbb{1}(y_t = c_t) \cdot \ell_t(x, y_t) \right) \right), \tag{1}$$

where $x \in X$ and $y_t \in Y_t$ represent individual image and label, $\mathbb{1}(.)$ denotes an indicator function, and c_t denotes a candidate class of t (*e.g.*, $c_t \in \{ct, \ldots, x\text{-}ray\}$, if $t = coarse\ modality$).

2.5 Multi-scale and multi-architecture ensemble

We adopted a multi-scale learning technique, using 128, 256, 384, and 512 as candidate image resize options. After applying the resize operation, we randomly crop the network input image at a ratio of 87.5% along both dimensions from a resized image. Random affine transformations and horizontal flip were used. The initial learning rate is set to 1e-3, linearly reduced 1e-6 after 100 epochs using Adam optimizer.

On the other hand, ResNets [11], DenseNets [14], ResNexts [34], MobileNet [13], and VGG nets [26] were selected as the image backbone candidates. We put the backbone and input scale options as training script hyper-parameters, which helped us to disperse the training over several GPU stations and gradually expand the number of ensemble members.

2.6 Experiment Results

We show the validation results from all trained models in Table 3, the corresponding training volume includes 2019-{train, val, test} and 2020-train. Based on these results, we made decisions of which models to be trained for test evaluation. Note that the training volume was changed to all of the 2019-{train, val, test} and 2020-{train, val, test} sets for training the testing-use models. We included the 2020-test set because some amount of partial coarse imaging modality labels (*i.e.*, from ${ct_alts}) and binary abnormality labels (*i.e.*, only the abnormal ones from wh-question abnormality) were extractable by SSM from only the questions, which served as a form of weak regularization for the test images. Finally, for the categorical abnormality type questions, we only select a top prediction from the VQA-Med 2020 subset of the abnormality classes as the predictions.

Our submissions on the 2020 validation set are shown in Table 4. Our second submission was purposed to determine the exact category type of the last question backbone in Table 1 as the five instances all appear in the 2020 test set. Although all other 2020 questions were in the abnormality question category (aligned with the official statement), we found the five questions could also be interpreted as asking which organ is present. We treated the 5 questions as categorical abnormality questions in the first submission and as organ questions in the second submission. Given the accuracy dropped, the ground truth should be the abnormality category.

From a post-challenge point of view, our third submission secured the leading position in the leaderboard. Our fourth submission was purposed to include more DenseNet-121 instances in the ensemble as the DenseNet-121-only multi-scale ensemble showed the highest 0.6 accuracy in Table 3. Our fifth submission added the two VGG multi-scale groups, presenting the final ensemble result

Table 3. The accuracy evaluation on the VQA-Med 2020 validation set.

Architecture	Network Input Size				Ensemble						
	128	256	384	512	Multi-scale	Multi-scale & Arch.					
ResNet-50	0.510	0.508	0.478	0.492	0.558	0.570	0.580	0.596	0.596	0.590	0.584
ResNet-101	0.486	0.530	0.508	0.460	0.566						
ResNet-152	0.486	0.522	0.486	0.386	0.548						
ResNext-50 32x4d	0.510	0.538	0.492	0.456	0.566						
ResNext-101 32x8d	0.522	0.520	-	-	0.538						
DenseNet-121	0.548	0.562	0.536	0.504	0.600						
DenseNet-161	0.526	0.520	0.518	-	0.564						
MobileNet v2	0.512	0.512	0.428	-	0.538						
VGG-16 with BN	0.478	0.482	0.426	0.486	0.530						
VGG-19 with BN	0.444	0.474	0.442	-	0.502						

of all trained models. Nevertheless, these final attempts only pushed up the performance marginally, suggesting a performance saturation in our approach.

Table 4. The officially evaluated accuracy and BLEU scores on the VQA-Med 2020 test set. The numbers in the brackets, *e.g.*, 256x2, indicates the use of 256 as the network input size and repeated 2 times (with different initial seeds).

ID	Ensemble Members	2020-val	2020-test	
		Accu.	Accu.	BLEU
67598	ResNet-50 (256x2, 384) + ResNet-101 (256) + ResNet-152 (256)	0.552	0.446	0.486
67737	Same as 67598	0.552	0.442	0.482
67915	All Resnets + All ResNexts + All Densenets + All Mobilenet V2	0.596	0.494	0.539
68012	67915 + extra DenseNet-121 (128x2, 256x2, 384x2, 512)	-	**0.496**	0.540
68017	68012 + VGG-16/19	-	**0.496**	**0.542**

3 VQG-Med Challenge Participation

3.1 Challenge Overview

The VQG-Med challenge dataset is a much smaller dataset compared to the VQA-Med datasets. The training set contains 780 radiology images with 2,156 associated QA pairs. The validation set has 141 images with 164 QA pairs. The test set has only 80 images. The goal of the VQG challenge is to generate between 1 to 7 answers for each test image.

3.2 Methodology

The VQG challenge describes a question generation task which in concept is close to image captioning but our proposed solution continued as a classification approach. The main reason is that we found there were more than one ground

truth questions tied to each image. Unlike a VQA task, a question can be considered as a prior knowledge on which the corresponding answer is conditionally dependent. Generating multiple questions while lacking such prior knowledge could be resolved by sampling approaches, but it can be difficult to associate a random state to a specific ground truth question. Hence, we instead treated all observed questions for an image as its attributes and modelled the question generation task as again an image attributes classification task. A downside of the classification approach is not able to produce novel questions.

Our VQG approach was built upon our VQA-Med solution with the following settings.

- Solving the question generation task as a classification task leads to a total of 2,073 classes each as an unique observed question from the joint training and validation sets.
- We were concerned about finetuning the entire image model by the limited amount of data and the large number of class, which may end up over-fitting in a much faster rate, hence we did not choose to fine-tune the backbones. However, as a compensation of non-linear capacity, we added a 2-layer batch-normalized and fully-connected (FC) (512 units each, ReLU activation) multiple-level perceptron (MLP) model before the softmax layer. The MLP model also avoided a direct mapping from the image features (*e.g.*, 2048 dimensional features) to the 2,073 classes which would result in a computational expensive matrix multiplication and a large memory usage.
- At the training hyper-parameter level, we kept the initial learning rate as 1e-3 but adjusted the final learning rate to 1e-5. Finally, we shortened the number of epochs to 40.
- Each training image could be associated with more than one question, resulting a multi-label problem. We used the Stochastic Ground Truth method in [16] which treats each image with multiple observed questions as multiple one-question-for-one-image samples, converting the multi-label problem to a single-label problem.
- The multi-scale and multi-architecture ensemble were continued in the VQG approach.

These settings helped us to reuse most of the VQA-Med code base and models to develop a tangible solution within a very short time frame.

Table 5. The accuracy evaluation on the VQG-Med 2020 validation set.

Architecture	Network Input Size				Multi-scale Ensemble
	128	256	384	512	
ResNet-50	0.067	0.091	0.098	0.067	0.091
ResNet-101	0.055	0.098	0.080	0.061	0.067
ResNet-152	0.091	0.067	0.067	0.073	0.067
DenseNet-121	-	0.085	0.079	-	0.091
DenseNet-161	-	0.079	0.079	-	0.073

3.3 Experiment Results

Similar to the VQA-Med 2020 result presentation, we show the VQG-Med 2020 validation and test results separately in Table 5 and 6, respectively. While the official evaluation only has BLEU score, in our local evaluation, we used top-7 accuracy to evaluate the validation performance. For official testing, each of our submission generates seven questions according to the highest probabilities for each image.

Table 6. The VQG-Med 2020 submitted results. The number in a bracket indicates the network input scale of the respective member model.

ID	Ensemble Members	val	test
		Accu.	BLEU
67984	ResNets-50/101/152 (no 512)	0.085	0.335
67995	ResNets-50/101/152 (all scales)	0.073	0.335
67996	67995 + ResNets-50/101/152 (no 512 + answer prediction)	0.091	0.326
68006	ResNet-50/101 (256, 384) + ResNet-152 (128) + DenseNet-121 (256, 384)	**0.110**	**0.348**
68018	68006 + DenseNet-161 (256, 384)	0.098	0.338

The first two submissions tested whether the large input size models should be continued. Given the lower top-7 accuracy on the validation set and the same BLEU value on the test set, we decided to not continue the 512 input size training. In the third submission, we tried to utilize the ground truth answer annotations by introducing the answer classification as an additional regularization task, but the result dropped by 0.009. In addition, the results from the first three submissions suggested a low correlation between the validation top-7 accuracy and the test BLEU scores. Hence in our forth submission we made two decisions in order to push for a much larger margin on the local evaluation: 1) forgoing the low accuracy models from the ensemble (validation accuracy $<$ 0.079); 2) including the DenseNet-121 architecture given its good performance in the VQA-Med challenge. The fourth submission scored 0.11 for the validation accuracy and 0.348 for the test BLEU score, secured our leading position in the VQG-Med challenge. Finally, in the fifth submission, we further added the DenseNet-161 multi-scale models as a last-minute attempt. Given the local evaluation dropped by 0.012, the test performance drop was expected as well.

4 Discussion and Conclusion

In this paper, we described our participation at the 2020 VQA-Med challenge and the associated VQG-Med challenge. The center of our approach is a knowledge inference method which we named Skeleton-based Sentence Mapping (SSM). In the VQA-Med challenge, the SSM method was useful on multiple fronts: 1) it mapped questions to a set of backbones which were useful to populate dynamic question instances; 2) it replaced the need of the language modelling and was

able to provide the direct selection to the corresponding answer predictor; and 3) it was used to infer six image classification tasks and corresponding training labels. Bypassing the development of language modelling allowed us to focus on tweaking the image classification model so that we devoted more time and resource on the multi-scale and multi-architecture ensemble learning. At last, we developed a class-wise and task-wise normalization technique for balancing the class and task populations, allowing the tasks with incomplete labels to be jointly optimized in one network.

The main inspiration of SSM came from [17], where we back-translated the questions via a number of foreign languages for augmentation purpose, resulting from a group of sentences with a small wording variation; hence a sentence backbone could be inferred. Nevertheless, whether the augmented questions carry the same meaning needs to be manually checked. The idea of reverse-engineering the sentence backbone was extended during our participation at the VQA-Med challenge and led to the proposal of SSM.

We are aware of the fact that SSM is not fully automated which requires further development. In addition, we understand SSM is a form of explicit reasoning model and its efficiency highly depends on the question regularity and dataset size which may not generalize well for VQA datasets containing free-form questions.

References

1. Abacha, A.B., Gayen, S., Lau, J.J., Rajaraman, S., Demner-Fushman, D.: Nlm at imageclef 2018 visual question answering in the medical domain. In: CLEF (Working Notes) (2018)
2. Abacha, A.B., Hasan, S.A., Datla, V.V., Liu, J., Demner-Fushman, D., Müller, H.: Vqa-med: Overview of the medical visual question answering task at imageclef 2019. In: CLEF (Working Notes) (2019)
3. Allaouzi, I., Ahmed, M.B.: Deep neural networks and decision tree classifier for visual question answering in the medical domain. In: CLEF (Working Notes) (2018)
4. Antol, S., Agrawal, A., Lu, J., Mitchell, M., Batra, D., Lawrence Zitnick, C., Parikh, D.: Vqa: Visual question answering. In: Proceedings of the IEEE international conference on computer vision. pp. 2425–2433 (2015)
5. Ben Abacha, A., Datla, V.V., Hasan, S.A., Demner-Fushman, D., Müller, H.: Overview of the vqa-med task at imageclef 2020: Visual question answering and generation in the medical domain. In: CLEF 2020 Working Notes. CEUR Workshop Proceedings, CEUR-WS.org, Thessaloniki, Greece (September 22-25 2020)
6. Ben-Younes, H., Cadene, R., Cord, M., Thome, N.: Mutan: Multimodal tucker fusion for visual question answering. In: Proceedings of the IEEE international conference on computer vision. pp. 2612–2620 (2017)
7. Cho, K., Van Merriënboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., Bengio, Y.: Learning phrase representations using rnn encoder-decoder for statistical machine translation. arXiv preprint arXiv:1406.1078 (2014)
8. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv: 1810.04805 (2018)

9. Gao, H., Mao, J., Zhou, J., Huang, Z., Wang, L., Xu, W.: Are you talking to a machine? dataset and methods for multilingual image question. In: Advances in neural information processing systems. pp. 2296–2304 (2015)
10. Hasan, S.A., Ling, Y., Farri, O., Liu, J., Müller, H., Lungren, M.P.: Overview of imageclef 2018 medical domain visual question answering task. In: CLEF (Working Notes) (2018)
11. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016)
12. Hochreiter, S., Schmidhuber, J.: Long short-term memory. Neural computation **9**(8), 1735–1780 (1997)
13. Howard, A.G., Zhu, M., Chen, B., Kalenichenko, D., Wang, W., Weyand, T., Andreetto, M., Adam, H.: Mobilenets: Efficient convolutional neural networks for mobile vision applications. arXiv preprint arXiv:1704.04861 (2017)
14. Huang, G., Liu, Z., Van Der Maaten, L., Weinberger, K.Q.: Densely connected convolutional networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 4700–4708 (2017)
15. Ionescu, B., Müller, H., Péteri, R., Abacha, A.B., Datla, V., Hasan, S.A., Demner-Fushman, D., Kozlovski, S., Liauchuk, V., Cid, Y.D., Kovalev, V., Pelka, O., Friedrich, C.M., de Herrera, A.G.S., Ninh, V.T., Le, T.K., Zhou, L., Piras, L., Riegler, M., l Halvorsen, P., Tran, M.T., Lux, M., Gurrin, C., Dang-Nguyen, D.T., Chamberlain, J., Clark, A., Campello, A., Fichou, D., Berari, R., Brie, P., Dogariu, M., Ştefan, L.D., Constantin, M.G.: Overview of the ImageCLEF 2020: Multimedia retrieval in medical, lifelogging, nature, and internet applications. In: Experimental IR Meets Multilinguality, Multimodality, and Interaction. Proceedings of the 11th International Conference of the CLEF Association (CLEF 2020), vol. 12260. LNCS Lecture Notes in Computer Science, Springer, Thessaloniki, Greece (September 22-25 2020)
16. Liao, Z., Girgis, H., Abdi, A., Vaseli, H., Hetherington, J., Rohling, R., Gin, K., Tsang, T., Abolmaesumi, P.: On modelling label uncertainty in deep neural networks: Automatic estimation of intra-observer variability in 2d echocardiography quality assessment. IEEE Transactions on Medical Imaging **39**(6), 1868–1883 (2019)
17. Liao, Z., Liu, L., Wu, Q., Teney, D., Shen, C., van den Hengel, A., Verjans, J.: Medical data inquiry using a question answering model. In: 2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI). pp. 1490–1493. IEEE (2020)
18. Lin, M., Chen, Q., Yan, S.: Network in network. International Conference on Learning Representations (2014)
19. Lu, J., Yang, J., Batra, D., Parikh, D.: Hierarchical question-image co-attention for visual question answering. In: Advances in neural information processing systems. pp. 289–297 (2016)
20. Malinowski, M., Rohrbach, M., Fritz, M.: Ask your neurons: A neural-based approach to answering questions about images. In: Proceedings of the IEEE international conference on computer vision. pp. 1–9 (2015)
21. Mikolov, T., Sutskever, I., Chen, K., Corrado, G.S., Dean, J.: Distributed representations of words and phrases and their compositionality. In: Advances in neural information processing systems. pp. 3111–3119 (2013)
22. Papineni, K., Roukos, S., Ward, T., Zhu, W.J.: Bleu: a method for automatic evaluation of machine translation. In: Proceedings of the 40th annual meeting of the Association for Computational Linguistics. pp. 311–318 (2002)

23. Peng, Y., Liu, F., Rosen, M.P.: Umass at imageclef medical visual question answering (med-vqa) 2018 task. In: CLEF (Working Notes) (2018)
24. Ren, M., Kiros, R., Zemel, R.: Exploring models and data for image question answering. In: Advances in neural information processing systems. pp. 2953–2961 (2015)
25. Shi, L., Liu, F., Rosen, M.P.: Deep multimodal learning for medical visual question answering. In: CLEF (Working Notes) (2019)
26. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 (2014)
27. Singh, A., Natarajan, V., Shah, M., Jiang, Y., Chen, X., Batra, D., Parikh, D., Rohrbach, M.: Towards vqa models that can read. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 8317–8326 (2019)
28. Szegedy, C., Ioffe, S., Vanhoucke, V., Alemi, A.A.: Inception-v4, inception-resnet and the impact of residual connections on learning. In: Thirty-first AAAI conference on artificial intelligence (2017)
29. Talafha, B., Al-Ayyoub, M.: Just at vqa-med: A vgg-seq2seq model. In: CLEF (Working Notes) (2018)
30. Vu, M., Sznitman, R., Nyholm, T., Löfstedt, T.: Ensemble of streamlined bilinear visual question answering models for the imageclef 2019 challenge in the medical domain. In: CLEF 2019. vol. 2380 (2019)
31. Wu, Q., Shen, C., Liu, L., Dick, A., Van Den Hengel, A.: What value do explicit high level concepts have in vision to language problems? In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 203–212 (2016)
32. Wu, Q., Wang, P., Shen, C., Dick, A., Van Den Hengel, A.: Ask me anything: Free-form visual question answering based on knowledge from external sources. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 4622–4630 (2016)
33. Wu, Z., Palmer, M.: Verb semantics and lexical selection. arXiv preprint cmp-lg/9406033 (1994)
34. Xie, S., Girshick, R., Dollár, P., Tu, Z., He, K.: Aggregated residual transformations for deep neural networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1492–1500 (2017)
35. Xu, H., Saenko, K.: Ask, attend and answer: Exploring question-guided spatial attention for visual question answering. In: European Conference on Computer Vision. pp. 451–466. Springer (2016)
36. Yan, X., Li, L., Xie, C., Xiao, J., Gu, L.: Zhejiang university at imageclef 2019 visual question answering in the medical domain. In: CLEF (Working Notes) (2019)
37. Yang, Z., He, X., Gao, J., Deng, L., Smola, A.: Stacked attention networks for image question answering. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 21–29 (2016)
38. Yu, Z., Yu, J., Fan, J., Tao, D.: Multi-modal factorized bilinear pooling with co-attention learning for visual question answering. In: Proceedings of the IEEE international conference on computer vision. pp. 1821–1830 (2017)
39. Zhou, Y., Kang, X., Ren, F.: Employing inception-resnet-v2 and bi-lstm for medical domain visual question answering. In: CLEF (Working Notes) (2018)
40. Zhou, Y., Kang, X., Ren, F.: Tua1 at imageclef 2019 vqa-med: a classification and generation model based on transfer learning. In: CLEF (Working Notes) (2019)