

# The Inception Team at VQA-Med 2020: Pretrained VGG with Data Augmentation for Medical VQA and VQG

Aisha Al-Sadi, Hana Al-Theiabat, and Mahmoud Al-Ayyoub

Jordan University of Science and Technology, Jordan  
asalsadi16@cit.just.edu.jo, haaltheiabat13@cit.just.edu.jo,  
maalshbool@just.edu.jo

**Abstract.** This paper describes the methodology of The Inception team participation at ImageCLEF Medical 2020 tasks: Visual Question Answering (VQA) and Visual Question Generation (VQG). Based on the data type and structure of the dataset, both tasks are treated as image classification tasks and are handled by using the VGG16 pre-trained model along with a data augmentation technique. In both tasks, our best approach achieves the second place with an accuracy of 48% in the VQA task and a BLEU score of 33.9% in the VQG task.

**Keywords:** ImageCLEF 2020· VQA-Med· Visual Question Answering· Visual Question Generation· Medical Image Interpretation· Medical Questions and Answers· Transfer Learning· VGG Network· Augmentation.

## 1 Introduction

Recently, enormous research efforts have been invested in merging or fusing techniques and natural language processing, signal processing and computer vision with the aim of improving the interaction between humans and intelligent systems, which in turn gives rise to tasks like Visual Question Answering (VQA), Visual Question Generation (VQG), Multimodal Machine Translation, Video Captioning and Scene Understanding, etc. Such tasks require learning across multimodal features from both visual and language data [18, 23, 25, 5, 11].

VQA and VQG are closely related tasks with significant importance [3]. VQA aims to answer a given question based on a given image. On the other hand, given an input image, VQG is concerned with generating relevant questions on this image along with their answers while relying only on the image content. Both tasks have attracted the attention of several researchers who proposed interesting models with promising results [3].

---

Copyright © 2020 for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0). CLEF 2020, 22-25 September 2020, Thessaloniki, Greece.

Among the most notable efforts on VQA is a set of challenges or shared tasks held every year addressing different versions of VQA. Starting from 2016, a challenge on general VQA, known as the VQA Challenge,<sup>1</sup> is held every year to answer questions of various types (multiple-choice questions, yes/no questions and questions with open-ended answers).

For the medical domain, a challenge known as the VQA-Med challenge has been held since in 2018. In its first version [6] only five teams participated in the task which focused on answering questions about abnormalities in medical images. For the second version, VQA-Med 2019 [4], the challenge included four question categories on each medical image: the plane, the modality, the organ system and the abnormality shown in the image.

In this paper, we propose models to solve the medical VQA-Med tasks [3] (VQA and VQG) organized by ImageCLEF 2020 [8]. All of our models are based on the pre-trained convolutional neural networks (CNN), VGG16 [22], and data augmentation technique. In both tasks, our best approach achieves the second place with an accuracy of 48% in the VQA task and a BLEU score of 33.9% in the VQG task.

The rest of this paper is organized as follows. The following section briefly surveys the literature on VQA and VQG. Section 3 presents the description of the both VQA-Med tasks (VQA and VQG) along with a detailed analysis of the dataset for each task. In Section 4, we present the proposed models for each task. Submission results and discussion are presented in Section 5. Finally, the paper is concluded in Section 6.

## 2 Related Work

For the general VQA task, several researchers merge both visual and text information using encoder-decoder networks [14], while others introduce attention over images to highlight the most important regions in the image to answer questions [20]. In [13], the authors proposed a co-attention mechanism to jointly apply attention on both images and questions.

For medical VQA, the task is challenging as it requires a dedicated medical dataset and expert doctors to understand the data. VQA-Med [7] is a competition launched first in 2018, where it provides each year a medical dataset for the VQA task. Team UMass [17], which achieved the highest BLEU score in 2018, proposed a co-attention method between images features, which are extracted from ResNet-152, and text features from pre-trained word embedding. Team JUST [24] used a simple encoder-decoder model based on Recurrent Neural Networks (RNN) with Long Short-Term Memory (LSTM) where the image features were extracted by a VGG16 model.

For the second version of the VQA-Med challenge in 2019, the team of Zhejiang University [26] was ranked first among 17 teams. This team adopted the co-attention mechanism to merge both visual and text features. The image features were extracted using the VGG16 network with Global Average Pooling

---

<sup>1</sup> <https://visualqa.org/index.html>

strategy, while the Bidirectional Encoder Representations from Transformers (BERT) model was used to extract question features. The Multimodal factorized bilinear pooling (MFB) and multi-modal factorized high-order pooling (MFH) methods were used for features concatenation. In the same year. Our team, Team JUST [1], was ranked among the top five teams by proposing a hierarchical model composed of multiple deep learning sub-models to handle different question types. All sub-models were based on a pre-trained VGG16 network as image classification, without considering questions as input features for the sub-models.

Visual Question Generation (VQG) is a task of developing visual understanding from images to generate reasonable questions with constraints (conditional VQG) [12] using labeled answers, or without constraints (unconditional) using only the image itself. Annotating multiple questions with each image as a VQG dataset was first collected by [15].

Various works have adopted the multimodal context of the natural language of questions/answers and visual understanding of the image. The authors in [27] proposed an approach to understanding the semantics in the image by simultaneously training VQG and VQA models as it is viewed in [12] as a dual-task. In [27], the VQG model used both RNN and CNN to let the model learn both natural language and vision aspects. Extending this in [9], where the authors used Variational Autoencoder (VAE) with LSTM to generate several questions per image. On the other hand, a deep Bayesian multimodal network was proposed by [16] to generate a set of questions for each image.

Regarding VQG in the medical field, Sarrouti et al. [19] proposed an approach for VQG about radiology images called VQGR. The approach relied on VAE. To increase the dataset size, the same authors [19] applied data augmentation on both images and questions on the [10] dataset.

### 3 Tasks and Dataset Description

In this section, we discuss the details and datasets of VQA-Med 2020’s two tasks: VQA and VQG.

#### 3.1 VQA Task and Dataset Description

The dataset of VQA-Med consists of 4,000 training medical images and 500 validation medical images, where each image is associated with a Question-Answer (QA) pair. Additionally, there are 500 medical test images with their questions.

The questions are from two types:

- Type 1: Questions asking about abnormalities in the image. For example, “what is the abnormality in this image”. This type represents the majority of abnormality questions (98.5% of the training data, and 94.4% of the validation data).

- Type 2: Questions with yes/no answers that ask if the image is normal or not. For example, “is this image normal” or “is this image abnormal” (1.5% of the training data, and 5.6% of the validation data).

It is worth to mention some brief statistics about the dataset:

- There are only 38 unique questions in the 4,000 training questions.
- There are only 26 unique questions in the 500 validation questions.
- There are only 332 unique answers in the 4,000 training answers.
- There are only 232 unique answers from the 500 validation answers.
- Most unique answers are each associated with 5-20 questions/images, while the most repetitive answer is associated with 80 questions/images, and the least repetitive answer is associated with 3 questions/images.

Samples of the datasets are provided in Table 1. The required in the VQA task is to answer the question given for each image.

### 3.2 VQG Task and Dataset Description

For the second task, VQG, the dataset consists of 780 images with 2,156 questions for training data, and 141 images with 164 questions for validation data. Each image is associated with one or more questions (up to 12 questions). For the test data, there are 80 images. The answers are given for the training and validation datasets but not for the test dataset. The questions on the images are diverse and are not necessarily related to the VQA dataset (i.e., questions are not limited to asking about abnormality in images).

It is worth to mention some brief statistics about the dataset:

- There are 1,942 unique questions in the 2,156 training questions.
- There are 161 unique questions in the 164 validation questions.
- Most questions occurred once, but some questions are associated with more than one image (up to 8 images).
- Some questions in the validation data are not in the training data.

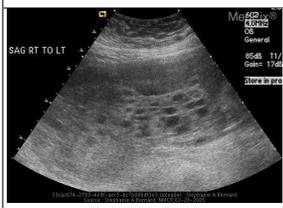
Samples of the datasets are provided in Table 2. The requirement in the VQG task is to generate at least one question and a maximum of seven questions for each test image.

## 4 Methodology

In this section, first, the description of the basic models is provided, followed by the details of each task submission.

We explore different approaches for the tasks at hand. However, most of these approaches are based on our experiments from last year’s edition of the VQA-Med task [1]. We call this the basic model for any reference later. Following is a description of this model.

Table 1. VQA task dataset samples

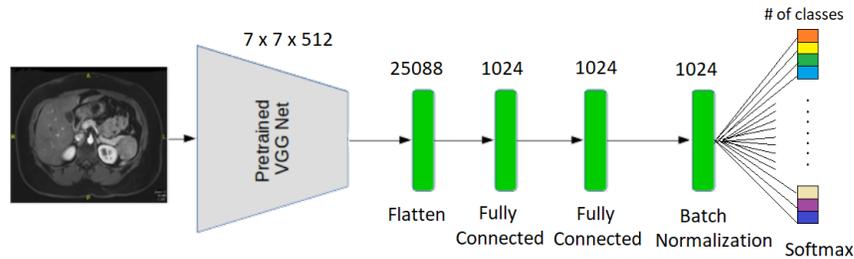
Image	Question	Answer
	what is the primary abnormality in this image?	neurofibromatosis-1, nfl, nfr
	what is abnormal in the x-ray?	incarcerated diaphragmatic hernia presenting as colonic obstruction.
	is this image normal?	yes

The model is for image classification that uses the pre-trained model VGG16 with the last layer (the Softmax layer) removed and all layers (except the last four) are frozen. The output from this part is passed into two fully-connected layers with 1,024 hidden nodes, followed by a normalization layer. Finally, the output is passed into a Softmax layer with the required number of classes based on the task. Figure 1 shows the architecture of the basic model.

The other pillar in our models is the augmentation technique. Although deep learning models have a remarkably excellent performance in several fields such as computer vision, they suffer from overfitting when the learners fit the training data perfectly. Usually, to avoid overfitting, the networks have to access more data. However, many applications lack the access to big data, such as medical image applications. One solution to increase the amount of the training data is data augmentation, which aims at increasing the diversity of the data without the need to collect new samples. For images, the augmentation done by applying

**Table 2.** VQG task dataset samples

Image	Question
	Q1: What are present in the right greater wing of the sphenoid and in the right parietal bone just above the squamosal suture?
	Q2: What involves the right parietal bone just above the squamosal suture?
	Q3: What are demonstrated in the skull?
	Q4: What are present in the right greater wing of the sphenoid and in the left parietal bone near the vertex?



**Fig. 1.** Basic model architecture

one of the geometric transformations, flipping, padding or random erasing on the existing images to produce new images [21].

For the VQA-Med tasks, we apply data augmentation on images to improve the performance of our models. We experiment with various ways of augmentation such as rotation change, width/ height shift, rescale, zoom and ZCA whitening.

ZCA whitening [2] is a whitening transformation used to decorrelate features in the data. It is widely used for images as it removes redundancy and highlight the structure of features. Not like PCA whitening, ZCA preserves the arrangement of the origin which makes it a good option for CNN.

#### 4.1 VQA Task

A common way of dealing with VQA tasks is by using sequence-to-sequence (SeqtoSeq) models that merge image features and question features to predict

the answer. However, due to the dataset nature and the repetitiveness of the questions, the contents of the questions are not expected to play a major role in answering the questions (aside from determining the questions type, etc.). On the other hand, the image features play a more significant role. Hence, we treat this task as an image classification task.

The core for our models is the basic model which is illustrated previously (see Figure 1) with additional modifications. Specifically, we build two models as follows.

- A model for classifying image into normal/abnormal for yes/no answers.
- A model for answering abnormalities questions.

For the first model (normal/abnormal), we use the basic model by passing the output to a Softmax layer with 2 classes (normal/abnormal). Images with their QA pairs used in this model are the ones associated with questions that start with “What”.

For answering abnormalities questions by the second model, we use the same previous model’s architecture, but the output is passed to a Softmax layer with 330 classes (the different 330 abnormalities in the training dataset).

For this task, we make five submissions. Table 3 summarizes the difference between these submissions, which are all based on the previous model description. For Submission 2-5, we apply data augmentation on images to improve the performance of the second model (the abnormalities model). We experiment with various ways of augmentation such as rotation change, width/ height shift, rescale, zoom, and ZCA whitening.

## 4.2 VQG Task

For the VQG task, based on the data description, answers are available only for training and validation data. So, we decide to build our model based on the images only. The first idea to apply in such tasks is the image captioning where the question is considered as the caption. This model is used in our first submission out of the five ones we made for this task.

The image captioning model is a Seq2Seq model. In each time step, the image features are concatenated with the current question word to extract the features in that time step. These features are passed to an LSTM layer, then to a dense layer followed by a Softmax layer to predict the next word. For the first step, the image with a predefined start word is used to predict the first word of the answer, and the image with the first word of the answer is used to predict the second word of the answer, and so on.

Due to the poor performance we get from this approach, we resort to treating this task as an image classification task in Submissions 2 and 3. We use the same basic image classification model, except that the output is passed to a Softmax layer with 2,072 classes (the different 2,072 unique questions in both the training and validation datasets).

**Table 3.** Description of VQA submissions

Submission#	Method
1	The basic model without any data augmentation, 50 epochs
2	The basic model + data augmentation using featurewise_center and featurewise_std_normalization, 50 epochs
3	The basic model + data augmentation using ZCA whitening, 300 epochs
4	The basic model + data augmentation using ZCA whitening, 50 epochs
5	The basic model + use the weights of the highest accuracy model so far (Submission 4) as pre-training + data augmentation using ZCA whitening, 150 epochs

For Submissions 4 and 5, we use the augmentation technique with ZCA whitening beside the basic image classification model. Table 4 summarizes the difference between the five submissions.

## 5 Results and Discussion

### 5.1 VQA Task Results

Our team, The Inception Team, got the second place in the leaderboard, as shown in Figure 2, among the eleven participating teams with our best accuracy reaching 48%. This is lower than the first place by only 1.6%. Our best result is obvious in Submission 4, which uses our basic model and ZCA whitening augmentation and 50 epochs of training (see Table 3).



**Fig. 2.** VQA task leaderboard

Table 5, provides accuracy and BLEU score of each of our submissions.

**Table 4.** Description of VQG submissions

Submission#	Method
1	Image captioning model 50 epochs predict one question for each image.
2	Image classification model 25 epochs predict 7 questions for each image.
3	Image classification model 50 epochs predict 7 questions for each image
4	Image classification model data augmentation using ZCA whitening 100 epochs predict 7 questions for each image.
5	Image classification model data augmentation using ZCA whitening 50 epochs predict 7 questions for each image.

**Table 5.** The Inception team VQA task submissions results

Submission#	Accuracy	BLEU Score
1	0.454	0.486
2	0.458	0.495
3	0.444	0.479
4	<b>0.48</b>	<b>0.511</b>
5	0.44	0.476

## 5.2 VQG Task Results

In the VQG task, we also got second place in the leaderboard, as shown in Figure 3, among the three participating teams with our best accuracy 33.9% which is less than the first place by 0.9%. Our best result is via Submission 4, which uses our basic model and ZCA whitening augmentation (see Table 4), predicting 7 questions for each image, and 100 epochs of training.

Table 6 provides the BLEU scores of each of our submissions.

## 5.3 Discussion

Several observations can be obtained from our experimentation. For example, the positive effect of using the augmentation technique is clear in both tasks. The models with augmented images results are better than the ones without augmentation, especially in the VQA task. For both tasks, after using ZCA whitening augmentation, the accuracy of the models has increased. The reason why other augmentation methods such as rotation and shifting have worsened the



**Fig. 3.** VQG task leaderboard

**Table 6.** The Inception team VQG task submissions results

Submission#	BLEU Score	BLEU Score
<b>1</b>	0.031	0.486
<b>2</b>	0.314	0.495
<b>3</b>	0.319	0.479
<b>4</b>	<b>0.339</b>	<b>0.511</b>
<b>5</b>	0.331	0.476

model performance is due to the nature of radiology images which are sensitive to these changes as they might affect the purpose of images.

It worth to mention that both of feature wise std normalization and feature center techniques have not outperformed the ZCA whitening. The distribution of data will be Gaussian distribution in the feature wise std normalization, as it divides each image by the std standard deviation of all images in the dataset. On the other hand, the mean of each image is set to zero for the feature center method.

## 6 Conclusion

In this paper, we describe a set of models we submitted for the ImageCLEF 2020 VQA-Med tasks. We showed that our intuition (based on our analysis of the dataset) of treating the tasks as image classification tasks is more useful than including a natural language processing (NLP) component as one would expect in VQA/VQG tasks. Our best model is based on VGG16 and augments the data using the ZCA whitening technique. We achieved 48% accuracy in the VQA task, and 33.9% Bleu score in the VQG task. These scores gave our team the second place in the two tasks.

## References

1. Al-Sadi, A., Talafha, B., Al-Ayyoub, M., Jararweh, Y., Costen, F.: Just at imageclef 2019 visual question answering in the medical domain. In: CLEF (Working Notes) (2019)
2. Bell, A.J., Sejnowski, T.J.: The “independent components” of natural scenes are edge filters. *Vision research* **37**(23), 3327–3338 (1997)

3. 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)
4. Ben Abacha, A., 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)
5. Ebrahim, M., Al-Ayyoub, M., Alsmirat, M.: Determine bipolar disorder level from patient interviews using bi-lstm and feature fusion. In: 2018 Fifth International Conference on Social Networks Analysis, Management and Security (SNAMS). pp. 182–189. IEEE (2018)
6. Hasan, S.A., Ling, Y., Farri, O., Liu, J., Lungren, M., Müller, H.: Overview of the ImageCLEF 2018 medical domain visual question answering task (September 10-14 2018)
7. 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)
8. Ionescu, B., Müller, H., Péteri, R., Ben Abacha, A., 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., 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 lifelogging, medical, 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)
9. Jain, U., Zhang, Z., Schwing, A.G.: Creativity: Generating diverse questions using variational autoencoders. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 6485–6494 (2017)
10. Lau, J.J., Gayen, S., Ben Abacha, A., Demner-Fushman, D.: A dataset of clinically generated visual questions and answers about radiology images. *Scientific data* **5**(1), 1–10 (2018)
11. Lawonn, K., Smit, N.N., Bühler, K., Preim, B.: A survey on multimodal medical data visualization. In: *Computer Graphics Forum*. vol. 37, pp. 413–438. Wiley Online Library (2018)
12. Li, Y., Duan, N., Zhou, B., Chu, X., Ouyang, W., Wang, X., Zhou, M.: Visual question generation as dual task of visual question answering. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 6116–6124 (2018)
13. 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)
14. 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)
15. Mostafazadeh, N., Misra, I., Devlin, J., Mitchell, M., He, X., Vanderwende, L.: Generating natural questions about an image. *arXiv preprint arXiv:1603.06059* (2016)

16. Patro, B.N., Kumar, S., Kurmi, V.K., Namboodiri, V.P.: Multimodal differential network for visual question generation. arXiv preprint arXiv:1808.03986 (2018)
17. Peng, Y., Liu, F., Rosen, M.P.: Umass at imageclef medical visual question answering (med-vqa) 2018 task. In: CLEF (Working Notes) (2018)
18. Ramachandram, D., Taylor, G.W.: Deep multimodal learning: A survey on recent advances and trends. IEEE Signal Processing Magazine **34**(6), 96–108 (2017)
19. Sarrouti, M., Ben Abacha, A., Demner-Fushman, D.: Visual question generation from radiology images. In: Proceedings of the First Workshop on Advances in Language and Vision Research. pp. 12–18 (2020)
20. Shih, K.J., Singh, S., Hoiem, D.: Where to look: Focus regions for visual question answering. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 4613–4621 (2016)
21. Shorten, C., Khoshgoftaar, T.M.: A survey on image data augmentation for deep learning. Journal of Big Data **6**(1), 60 (2019)
22. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 (2014)
23. Srivastava, Y., Murali, V., Dubey, S.R., Mukherjee, S.: Visual question answering using deep learning: A survey and performance analysis. arXiv preprint arXiv:1909.01860 (2019)
24. Talafha, B., Al-Ayyoub, M.: Just at vqa-med: A vgg-seq2seq model. In: CLEF (Working Notes) (2018)
25. Xu, X., Song, J., Lu, H., He, L., Yang, Y., Shen, F.: Dual learning for visual question generation. In: 2018 IEEE International Conference on Multimedia and Expo (ICME). pp. 1–6. IEEE (2018)
26. 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)
27. Yang, Y., Li, Y., Fermuller, C., Aloimonos, Y.: Neural self talk: Image understanding via continuous questioning and answering. arXiv preprint arXiv:1512.03460 (2015)