# MIT Manipal at ImageCLEF 2019 Visual Question Answering in Medical Domain

Abhishek Thanki[1] and Krishnamoorthi Makkithaya[1]

Manipal Institute of Technology, Manipal - Udupi District, Karnataka - 576104, India
abhishek.harish@learner.manipal.edu, k.moorthi@manipal.edu

**Abstract.** This paper describes the participation of MIT, Manipal in the ImageCLEF 2019 VQA-Med task. The goal of the task was to build a system that takes as input a medical image and a clinically relevant question, and generates a clinically relevant answer to the question by using the medical image. We explored a different approach compared to most VQA systems and focused on the answer generation part. We used a encoder-decoder architecture based on deep learning where a pre-trained CNN on ImageNet was used to extract visual features from input image, a combination of pre-trained word embedding on pub-med articles along with a 2-layer LSTM was used to extract textual features from the question. Both visual and textual features were integrated using a simple element-wise multiplication technique. The integrated features were then passed into a LSTM decoder which then generated a natural language answer. We submitted a total of 8 runs for this task and the best model achieved a BLEU score of 0.462.

**Keywords:** Visual Question Answering · CNN · Word2Vec · LSTM · Encoder-Decoder · BLEU.

## 1 Introduction

Visual Question Answering (VQA) is task which consists of building a AI system which takes as input a image and a question in natural language, and the system is expected to produce a correct answer to the question by using both the visual and the textual information. This problem intersects the two important fields of computer science, Computer Vision (CV) and Natural Language Processing (NLP). The answers can be as simple as a single word, a simple yes/no, true/false or consists of multiple words.

VQA task has so far made great progress in the general domain due to the increasing advancements in the field of computer vision and natural language processing. But this problem is relatively new in the medical domain. ImageCLEF conducts many tasks related to multimedia retrieval in many domains

such as medicine, security, lifelogging, and nature [1]. Visual Question Answering in Medical domain is one such task and this is the second year, the VQA-Med task [2] has been introduced after last years success. Given a medical image and a clinically relevant question in natural language about the image, the task was to build a system that would produce a clinically relevant natural language answer to the question by using the image.

In this paper, we discuss our approach to build such a system which was inspired by VQA research in general domain [3] and sequence generation task in the natural language processing field [4]. We built a encoder-decoder architecture using the recent advancements in the field of deep learning. The model consists of a pre-trained Convolutional Neural Network (CNN) on ImageNet, a pre-trained word2vec model trained on pud-med articles [5] to extract word embeddings, and two Long Short Term Memory (LSTM) models. Image features were extracted using a pre-trained CNN on ImageNet. We tested with two architectures, VGG19 and DenseNet201 [6][7]. Question features were extracted by using pre-trained word2vec model and a 2-layer LSTM network. Both visual and textual features were integrated by using element-wise multiplication and resulting features were fed to a LSTM sequence generating network to produce the output answers.

This paper is organized in the following manner: Section 2 provides a information regarding the dataset provided for this challenge. Section 3 presents related work done which inspired us our model architecture. Section 4 describes our method of using a encoder-decoder architecture. Section 5 describes our experiments and corresponding results our model achieved. Finally, we conclude the paper in Section 6 by discussing the task, our method, and future improvements.

## 2 Dataset Description

In VQA-Med 2019 challenge, three datasets were provided:

- The training set consisted of 3,200 medical images with 12,792 question-answer pairs.
- The validation set consisted of 500 medical images with 2,000 question-answer pairs.
- The test set consisted of 500 medical images with 500 question-answer pairs.

Furthermore, the data in the dataset can be divided into four main categories as follows:

1. **Modality**: This category includes questions based on images of structural or functional parts of the body. For example: ultrasound, CT, etc.
2. **Plane**: Questions in this category consists about the plane of the medical image. This is important because different projections allow for depicting different tissues. For example: axial, sagittal, etc.
3. **Organ System**: Questions in this category consists on the different organs in the human body. For example: breast, skull and contents, etc.
4. **Abnormality**: Questions in this category consists of detecting any abnormality present in the input image and identifying the type of abnormality.

## 3 Related Work

VQA in general domain is not a new problem and the data available is much more compared to VQA in medical domain. Due to these reasons a lot of work has been done in the general domain. Our work in this paper takes inspiration from various resources. First, we are inspired by the simplicity in the baseline model from [3] which still achieved good accuracy. Second, since the words used in the medical domain are quite different compared to general English this means that a word2vec trained on English language does not produce vectors that can best encode the questions and hence we used a word2vec model trained on pub-med articles for encoding the question tokens. Third, while a lot of models developed on VQA general domain use multi-class classification to generate answers, we chose a different approach of using sequence generation to generate the answers since it made more intuitive sense to us.

## 4 Methodology

Our system consists of two main components: encoder and decoder. The encoder part consists of 3 sub-components: transfer learning to extract features from images, word2vec + 2-layer LSTM to extract features from questions, and element-wise multiplication to fuse the visual and textual features. The decoder part consists of a sequence generating LSTM network which generates the output answers to the input question and image. Fig. 1 shows the high-level architecture of our system.

### 4.1 Encoder

The encoder part of the model consists of:

- Pre-trained CNN on ImageNet: Deep CNNs trained on large-scale datasets such as ImageNet have demonstrated to be excellent at the task of transfer learning and this is why we chose transfer learning using a pre-trained CNN to extract visual features. For this purpose we experimented with VGG-19 [6] and DenseNet-201 [7] architectures. For VGG-19, we extracted the output features from its last hidden layer while in case of DenseNet201, we extracted the output features from `conv5_block32_concat` layer output. These extracted features were then passed through a dense layer which was trainable to get the final output visual features in various dimensions such as 128, 256, and 512.
- Pre-trained word2vec model + 2-layer LSTM network: To extract features from the input question, we pre-processed by building a custom function which cleans the input sentence and outputs a list of tokens. These tokens were then converted to vector form by using a pre-trained word2vec model [8][9][10]. These vectors were then passed through a 2-layer LSTM network to produce the output textual features. The reason why we chose LSTM

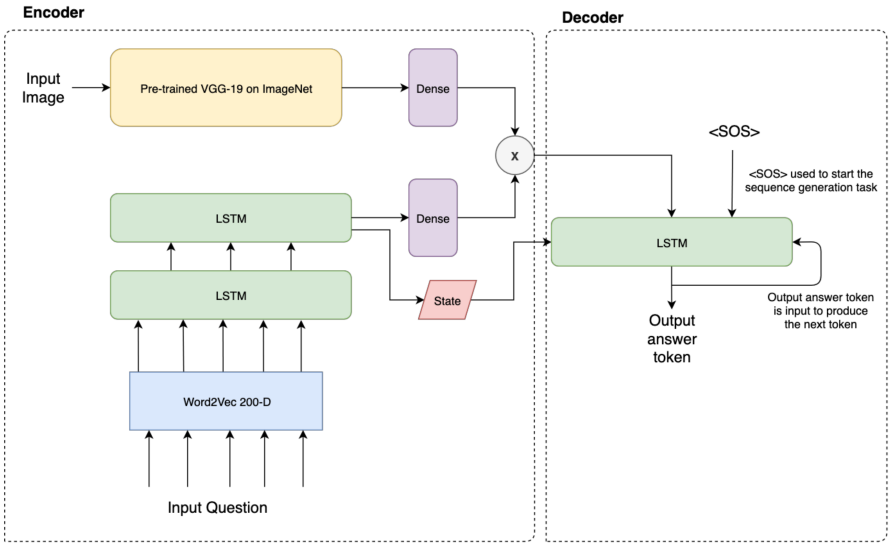


**Fig. 1.** Higher level architecture of the best model

network was due to the networks ability to model complex relationships within the same sentence and also because it is not affected by the vanishing gradient problem [11][12].
– Feature fusion: Here, we used a simple element-wise multiplication technique to combine the visual and textual features.

### 4.2 Decoder

The decoder part of the model consists of a LSTM network. This network takes as input the output features from the decoder part as well as the state of the second LSTM. The sequence generation step is started by providing as input a special token `<SOS>`. Subsequent output tokens produced by the model are fed back into the model to produce the next token. This process is continued until a certain number of tokens are produced or a special token called `<EOS>` is predicted.

## 5 Experiments and Result

We submitted eight runs to ImageLCEF 2019 VQA-Med:

1. VGG19-N128: This run used a VGG-19 for transfer learning and the number of neurons set in the encoder LSTM networks, the 2 dense layers, and the decoder LSTM network was 128. This network was trained for 100 epochs.
2. VGG19-N256: This was same as run number one except that the number of neurons were 256 and it was trained for 200 epochs.

3. VGG19-N256-Dropout: This run was same as run number two except that a dropout of 0.2 was used in the dense layers and it was trained for 150 epochs.
4. DenseNet201-N256: This run used a DenseNet-201 for transfer learning and the number of neurons set in the encoder LSTM networks, the 2 dense layers, and the decoder LSTM network was 256. This network was trained for 150 epochs.
5. DenseNet201-N256-D400: This run was similar to run five except that it used the embedding dimension used was 400 instead of 200 which was used in all the previous experiments.
6. DenseNet201-N256: This run was similar to run five except that the network was trained for 200 epochs.
7. DenseNet201-N128: This run was similar to run five except that the number of neurons were 256.
8. VGG19-N128: This run was identical to the first run.

**Table 1.** Result of all the runs

| No. | Model | BLEU | Strict accuracy |
|---|---|---|---|
| 1 | VGG19-N128 | **0.462** | 0.15 |
| 2 | VGG19-N256 | 0.433 | 0.126 |
| 3 | VGG19-N256-Dropout | 0.453 | 0.142 |
| 4 | DenseNet201-N256 | 0.455 | 0.158 |
| 5 | DenseNet201-N256-Dropout | 0.453 | **0.16** |
| 6 | DenseNet201-N256-D400 | 0.447 | 0.15 |
| 7 | DenseNet201-N256 | 0.301 | 0.098 |
| 8 | VGG19-N128 | **0.462** | 0.15 |

The VGG19-N128 model achieves the best BLEU score while DenseNet201-N256-Dropout achieves the best strict accuracy. Table 1 shows the result achieved by all models on the test set.

## 6 Conclusion

This paper describes our participation in the ImageCLEF 2019 VQA-Med challenge. We used a pre-trained CNN on ImageNet dataset to extract textual features, a word2vec + 2-layer LSTM network to extract textual features, and a sequence generating LSTM network to generate the output answer tokens. Our approach was different and instead focused on using sequence generation to generate the answers while using a simple element-wise multiplication technique to integrate the visual and textual features. While we would have liked to try out attention based techniques to integrate visual and textual features but we weren't able to do so due to the timing limitation. This is something which we will explore in the future to improve the model.

## References


1. Ionescu, B., Müller, H., Péteri, R., Cid, Y.D., Liauchuk, V., Kovalev, V., Klimuk, D., Tarasau, A., Ben Abacha, A., Hasan, S.A., Datla, V., Liu, J., Demner-Fushman, D., Dang-Nguyen, D.T., Piras, L., Riegler, M., Tran, M.T., Lux, M., Gurrin, C., Pelka, O., Friedrich, C.M., de Herrera, A.G.S., Garcia, N., Kavallieratou, E., del Blanco, C.R., Rodríguez, C.C., Vasillopoulos, N., Karampidis, K., Chamberlain, J., Clark, A., Campello, A.: ImageCLEF 2019: Multimedia retrieval in medicine, lifelogging, security and nature. In: Experimental IR Meets Multilinguality, Multimodality, and Interaction. Proceedings of the 10th International Conference of the CLEF Association (CLEF 2019), LNCS Lecture Notes in Computer Science, Springer, Lugano, Switzerland (September 9-12 2019)
2. 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 2019 Working Notes. CEUR Workshop Proceedings (CEUR-WS.org), CEUR-WS.org <http://ceur-ws.org/Vol-2380/>, Lugano, Switzerland (September 9-12 2019)
3. Antol, S., Agrawal, A., Lu, J., Mitchell, M., Batra, D., Zitnick, C.L., Parikh, D.: VQA: Visual Question Answering. In: International Conference on Computer Vision (ICCV) (2015)
4. 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)
5. McDonald, R., Brokos, G., Androutsopoulos, I.: Deep relevance ranking using enhanced document-query interactions. In: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. pp. 1849–1860. Association for Computational Linguistics, Brussels, Belgium (Oct-Nov 2018), https://www.aclweb.org/anthology/D18-1211
6. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 (2014)
7. 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)
8. Mikolov, T., Chen, K., Corrado, G., Dean, J.: Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781 (2013)
9. 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)
10. Mikolov, T., Yih, W.t., Zweig, G.: Linguistic regularities in continuous space word representations. In: Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. pp. 746–751 (2013)
11. Hochreiter, S., Schmidhuber, J.: Long short-term memory. Neural computation **9**(8), 1735–1780 (1997)
12. Hochreiter, S., Bengio, Y., Frasconi, P., Schmidhuber, J., et al.: Gradient flow in recurrent nets: the difficulty of learning long-term dependencies (2001)