JUST at VQA-Med: A VGG-Seq2Seq Model

Bashar Talafha[1] and Mahmoud Al-Ayyoub[1]

Jordan University of Science and Technology, Irbid, Jordan
talafha@live.com,maalshbool@just.edu.jo

Abstract. This paper describes the VGG-Seq2Seq system for the Medical Domain Visual Question Answering (VQA-Med) Task of ImageCLEF 2018. The proposed system follows the encoder-decoder architecture, where the encoders fuses a pretrained VGG network with an LSTM network that has a pretrained word embedding layer to encode the input. To generate the output, another LSTM network is used for decoding. When used with a pretrained VGG network, the VGG-Seq2Seq model managed to achieve reasonable results with 0.06, 0.12, 0.03 BLEU, WBSS and CBSS, respectively. Moreover, the VGG-Seq2Seq is not expensive to train.

Keywords: Sequence to sequence · VGG Network · Global Vectors for Word Representation.

1 Introduction

Visual Question Answering (VQA) is a recent and exciting problem at the intersection between Computer Vision (CV) and Natural Language Processing (NLP), where the input is an image and a question related to it written in a natural language and the output is the correct answer to the question. The answer can be a simple yes/no, choosing one of several options, a single word, or a complete phrase of sentence [2, 4].

From a first glance, the VQA problem seem like a very challenging one. The traditional CV techniques used for extracting useful information from images and the NLP techniques typically used for Question Answering (QA) are very far from each other and the interplay between them seem to be complex. Moreover the ability to construct an useful answer based on such multi-modal input adds to the complexity of the problem. Luckily, the recent advances in Deep Learning (DL) have paved the way to building more robust VQA techniques [4].

In this paper, we are interested in an interesting variation of VQA where both the image and question are from the medical domain. It is known as the Medical Domain Visual Question Answering (VQA-Med) Task [5] of ImageCLEF 2018 [8]. This task requires building a model that provide an answer to question about the content of a medical image. In order to address this task, we propose a DL model we call: VGG-Seq2Seq model. The model takes an image and a question as input and outputs the answer of this question based on fusing features extracted based on the image content with those extracted from the question itself.

The rest of this paper is organized as follows. The following section presents a very brief coverage of the related work. Sections 3 and 4 discuss the problem at hand and the model we propose to handle it. The experimental evaluation of our model and its discussion are presented in Section 5. Finally, the paper is concluded in Section 6.

2 Related Works

According to a recent survey on the VQA problem [4], most of the exiting approaches are based on DL techniques. The only interesting exceptions are the Answer Type Prediction (ATP) technique of [9] and the Multi-World QA of [12]. Of course, there are other non-DL approaches that are used as baseline for various datasets and approaches. Discussing them is outside the scope of this paper.

Regarding the DL-based approaches for VQA, most of them employ one of the word embedding techniques (typically, Word2Vec [14]) sometimes coupled with a Recurrent Neural Networks (RNN) to embed the question. Moreover, most of them use Convolutional Neural Networks (CNN) to extract features from the images. Examples of such approaches include iBOWIMG [25], Full-CNN [11], Ask Your Neurons (AYN) [13], Vis+LSTM [18], Dynamic Parameter Prediction (DPPnet) [15], etc. Another type of DL-based techniques employ some sort of attention mechanism such as Where to Look (WTL) [19], Recurrent Spatial Attention (R-SA) [26], Stacked Attention Networks (SAN) [23], Hierarchical Co-attention (CoAtt) [10], Neural Module Networks (NMNs) [1], etc.

Most of the work discussed in this section is not directly applicable to the VQA-Med for two reasons. The first one is an obvious one which is the focus on the medical domain, which gives this problem its unique set of challenges. As for the other one, it is related to how the sentences of the answers are constructed in VQA-Med, which is different from existing VQA datasets such as DAtaset for QUestion Answering on Realworld images (DAQUAR) [12], Visual7W [26], Visual Madlibs [24], COCO-QA [18], Freestyle Multilingual Image Question Answering dataset (FM-IQA) [3], Visual Question Answering (VQA) [2], etc.

3 Task Description and Dataset

Nowadays, patients can access and review their medical reports related to their healthcare due to the availability and accessibility of electronic medical records which will help them better understand their conditions. This increases the need for an automated system capable of taking questions related to some medical problems along with accompanying images to support this question and provide correct answer for them. This is exactly the task we are addressing in this work. Given an image in the medical domain associated with a set of clinically relevant questions, the goal of the task is answering the questions based on the visual image content [5].

The dataset represents images related to medical domain. It was extracted from PubMed Central articles (essentially a subset of the ImageCLEF 2017 caption prediction task). The dataset is divided into about 5k training set and about 0.5k validation set of medical images associated with question-answer pairs, and about 0.5k testing set of medical images associated with only questions. Figure 1 shows some examples from the training set [5].

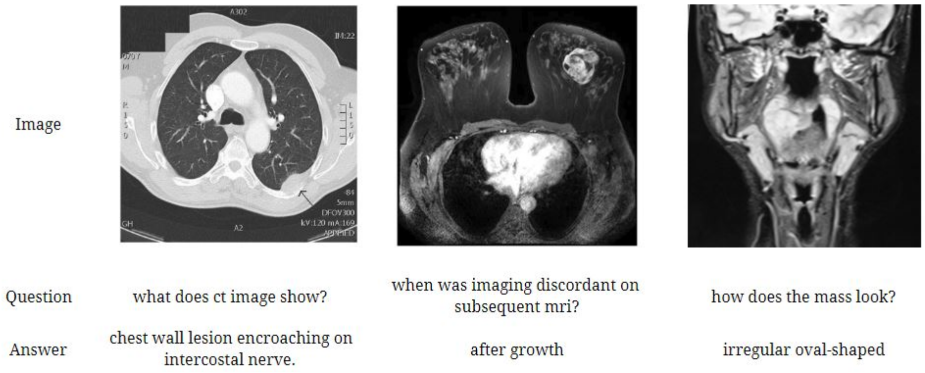

Fig. 1. Samples of images associated with question-answer pairs from the training set [5].

4 The VGG-Seq2Seq Model

In this section, we discuss our VGG-Seq2Seq model, which follows the encoder-decoder architecture. The model is shown in Figure 2. In the following paragraphs, we discuss in detail its different parts.

The encoder consists of two main components. The first component is a Long short term memory (LSTM) network with a pretrained word embedding layer which encodes the question into a vector representation, while the second component is a pretrained VGG network that takes the image as an input and extracts a vector representation for that image. The final state of the encoding, the outputs of the two components are concatenated together into one vector called thought vector.

The decoder consists of LSTM network that takes the thought vector as initial state and ⟨start⟩ token as input in the first time step and try to predict the answer using softmax layer.

4.1 Encoder

Two main components have been conducted in the encoder, the first component is a LSTM network with a pretrained word embedding layer, and the second component is the VGG network.

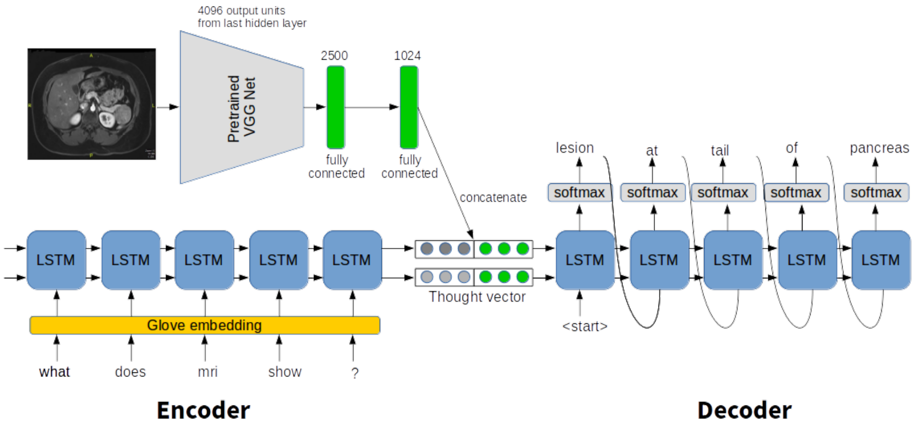

Fig. 2. The VGG-Seq2Seq Model.

In the first component the semantic meaning of the question will be extracted, a 300 dimensional pretrained word embedding layer is used to encode the word into a dense semantic space using Glove [17]. This word representation is then fed to a LSTM network with 1024 hidden nodes.

LSTM [7] is a special type of Recurrent Neural Network (RNN) that has been designed to solve the problem of vanishing gradient. The LSTM layer used its memory cells to store the context information. LSTM has three gates (i.e. Input gate, forget gate and output gate) which will decide how the input will be handled.

At any time step, inputs to the LSTM cell are current word (x), previous hidden state (h-1) and previous memory state (c-1), and LSTM cell outputs are current hidden state (h) and current memory state (c). These states have 1024 hidden nodes. At last time step in the sequence, we will call last LSTM cell output hidden state (h) final hidden state, and output memory state final memory state.

In the second component, we use the concept of transfer learning where a pretrained model is used with some modification to serve a wholly new task. We use pretrained VGG network [20] with removing the last softmax layer. This network will output a vector of size 4096 representing a vector of features for the input image. This vector is then passed to tow fully-connected layers with 2500 and 1024 hidden nodes respectively. The main purpose of these two layers is to decrease the features vector dimension to become close to the LSTM output vectors.

The 1024 image features vector is then concatenated with both the LSTM final hidden state and final memory state as shown in figure 2, we will call those tow vectors thought vectors, we believe that the thought vectors will represent the semantic meaning of the input question and features of the input image.

4.2 Decoder

In this part the answer of the input image and question will be extracted. The decoder consists of LSTM layer that takes three inputs, the first input is ⟨start⟩ token that indicates to start decoding, the second input and third input are the decoder initial states which are previous hidden state and previous memory state. The decoder takes the encoder final states (i.e. encoder final hidden state and encoder final memory state) as initial states. Thus, the decoder initial states will be the thought vectors.

At the first time step, LSTM cell will takes ⟨start⟩ token as input given the initial state and calculates the probability distribution of the target word using softmax layer. The word with the highest probability will be the first word of the answer, this word will be then passed to the second LSTM cell as input and predict the second word of the answer. The full answer will be generated by repeating this process until the model predicts ⟨end⟩ token.

5 Evaluation and Results

This section discusses the experiments used to evaluate our model and the obtained results. However, we first need to discuss the evaluation process.

As described in VQA-Med task description [5], three pre-processing steps are conducted on each answer before running the evaluation metrics: (a) converting each answer to lower-case, (b) removing all punctuations and tokenizing the the answer to a list of words and (c) removing stopwords using NLTKs English stopwords list.

In order to evaluate our models, three metrics are used as discussed in VQA-Med task description [5]: BLEU, WBSS and CBSS. The BLEU [16] metric is used to calculate the similarity between the predicted answer and the actual answer. The second metric is WBSS (Word-based Semantic Similarity) [21], which calculates semantic similarity in the biomedical domain. Finally, CBSS (Concept-based Semantic Similarity) [22], which is similar to WBSS, except that it can extract biomedical concepts from the answers using MetaMap via the pymetamap wrapper, and it builds a dictionary using these extracted concepts.

Three experiments are conducted to evaluate our model. They described as follows.

- In the first experiment, instead of using pretrained VGG-net we built a Convolutional Neural Network (CNN) that consists of three convolutional and max-pooling layers which behave as the feature extractor, followed by a fully connected layer. This network outputs a vector of size 4096 representing the input image features, this vector then will be fed to the 2500 fully connected layer and the rest of the architecture stayed as is.
- In the second experiment, we implemented VGG-Seq2Seq model but, instead of using the pretrained network, we built and trained the VGG network with its convolutional layers on the dataset, the rest of the architecture stayed as is.

- In the last experiment, we run our proposed model (VGG-Seq2Seq) with pretrained VGG-net on the dataset.

The three above experiments are trained using a single layer of LSTM network on the encoder with a dimension of 1024 and a single layer of LSTM network on the decoder with a dimension of 2048. All models were trained using RMSprop optimizer [6] with 0.001 learning rate on 500 epochs with 512 batch size at each epoch and 300 word embedding size. As shown in Table 1, the results show that VGG-Seq2Seq (pre-trained VGG) achieves reasonable results with 0.06, 0.12, 0.03 BLEU, WBSS and CBSS, respectively.

Table 1. Results of different models

Model	BLEU	WBSS	CBSS
VGG-Seq2Seq (Pre-trained VGG)	0.060986477	0.12167	0.029064
VGG-Seq2Seq	0.047820372	0.104488	0.014981
CNN-Seq2Seq	0.035619839	0.093911	0.011004

It is worth mentioning that the best performing VGG-Seq2Seq is not very expensive to train. It took an average of 252.8 seconds per epoch on a Virtual Machine (VM) equipped with Tesla K80 GPU card with 24GB of RAM. The VM had Ubuntu OS with CUDA 9.0. For the implementation, we use Keras with TensorFlow 1.8 backend.

6 Conclusion

In this paper, we addressed the very interesting yet challenging VQA-Med Task of ImageCLEF 2018. We introduced our VGG-Seq2Seq model which employs an encoder-decoder architecture, where the encoders fuses a pretrained VGG network with an LSTM network that has a pretrained word embedding layer to encode the input. As for the answer generation, another LSTM network is used as a decoded. When used with a pretrained VGG network, the VGG-Seq2Seq model managed to achieve reasonable results with 0.06, 0.12, 0.03 BLEU, WBSS and CBSS, respectively. Moreover, the VGG-Seq2Seq is not expensive to train. Obviously, the VGG-Seq2Seq model is far from perfect. We intend to work on it to increase its accuracy and enhance its run-time and space requirements.

References

1. Andreas, J., Rohrbach, M., Darrell, T., Klein, D.: Deep compositional question answering with neural module networks. CoRR **abs/1511.02799** (2015)
2. 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)

3. 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)
4. Gupta, A.K.: Survey of visual question answering: Datasets and techniques. arXiv preprint arXiv:1705.03865 (2017)
5. Hasan, S.A., Ling, Y., Farri, O., Liu, J., Lungren, M., Müller, H.: Overview of the ImageCLEF 2018 medical domain visual question answering task. In: CLEF2018 Working Notes. CEUR Workshop Proceedings, CEUR-WS.org <http://ceur-ws.org>, Avignon, France (September 10-14 2018)
6. Hinton, G., Srivastava, N., Swersky, K.: Rmsprop: Divide the gradient by a running average of its recent magnitude. Neural networks for machine learning, Coursera lecture 6e (2012)
7. Hochreiter, S., Schmidhuber, J.: Long short-term memory. Neural computation **9**(8), 1735–1780 (1997)
8. Ionescu, B., Müller, H., Villegas, M., de Herrera, A.G.S., Eickhoff, C., Andrearczyk, V., Cid, Y.D., Liauchuk, V., Kovalev, V., Hasan, S.A., Ling, Y., Farri, O., Liu, J., Lungren, M., Dang-Nguyen, D.T., Piras, L., Riegler, M., Zhou, L., Lux, M., Gurrin, C.: Overview of ImageCLEF 2018: Challenges, datasets and evaluation. In: Experimental IR Meets Multilinguality, Multimodality, and Interaction. Proceedings of the Ninth International Conference of the CLEF Association (CLEF 2018), LNCS Lecture Notes in Computer Science, Springer, Avignon, France (September 10-14 2018)
9. Kafle, K., Kanan, C.: Answer-type prediction for visual question answering. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 4976–4984 (2016)
10. 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)
11. Ma, L., Lu, Z., Li, H.: Learning to answer questions from image using convolutional neural network. In: AAAI. vol. 3, p. 16 (2016)
12. Malinowski, M., Fritz, M.: A multi-world approach to question answering about real-world scenes based on uncertain input. In: Advances in neural information processing systems. pp. 1682–1690 (2014)
13. Malinowski, M., Rohrbach, M., Fritz, M.: Ask your neurons: A deep learning approach to visual question answering. International Journal of Computer Vision **125**(1-3), 110–135 (2017)
14. 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)
15. Noh, H., Hongsuck Seo, P., Han, B.: Image question answering using convolutional neural network with dynamic parameter prediction. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 30–38 (2016)
16. 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 on association for computational linguistics. pp. 311–318. Association for Computational Linguistics (2002)
17. Pennington, J., Socher, R., Manning, C.: Glove: Global vectors for word representation. In: Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP). pp. 1532–1543 (2014)

18. 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)
19. 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)
20. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 (2014)
21. Soğancıoğlu, G., Öztürk, H., Özgür, A.: Biosses: a semantic sentence similarity estimation system for the biomedical domain. Bioinformatics **33**(14), i49–i58 (2017)
22. Wu, Z., Palmer, M.: Verbs semantics and lexical selection. In: Proceedings of the 32nd annual meeting on Association for Computational Linguistics. pp. 133–138. Association for Computational Linguistics (1994)
23. 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)
24. Yu, L., Park, E., Berg, A.C., Berg, T.L.: Visual madlibs: Fill in the blank description generation and question answering. In: Computer Vision (ICCV), 2015 IEEE International Conference on. pp. 2461–2469. IEEE (2015)
25. Zhou, B., Tian, Y., Sukhbaatar, S., Szlam, A., Fergus, R.: Simple baseline for visual question answering. arXiv preprint arXiv:1512.02167 (2015)
26. Zhu, Y., Groth, O., Bernstein, M., Fei-Fei, L.: Visual7w: Grounded question answering in images. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 4995–5004 (2016)