

An Encoder-Decoder model for visual question answering in the medical domain

Imane Allaouzi^[0000-0002-8737-8115], Mohamed Ben Ahmed, Badr Benamrou

LIST, Abdelmalek Essaâdi University
Faculty of Sciences and Techniques
Tangier, Morocco
¹imane.allaouzi@gmail.com

Abstract. This paper describes our participation in the task of VQA-Med of ImageCLEF 2019. We proposed an encoder-decoder model that takes as input a medical question-image pair and generates an answer as output. The encoder network consists of a pre-trained CNN model that extracts prominent features from a medical image and a pre-trained word embedding along with LSTM to embed textual data. The answer generation is accomplished by the greedy search algorithm, which predicts the next word based on the previously generated words. Thus, the answer is built up by recursively calling the model.

Keyword: Transfer Learning, Encoder-Decoder, CNN, LSTM, Word Embedding, Language Modeling, Medical Imaging, Visual Question Answering, Greedy Search, Beam Search, NLP, Computer Vision.

1 Introduction

With the widespread adoption of electronic medical record (EMR) systems, a large amount of medical information is becoming available such as doctors' reports, test results and medical images. This health information is a gold mine for artificial intelligence (AI) researchers who seek to enhance doctors' ability to analyze medical images, to support clinical decision making and improve patient engagement. One of the most exciting and challenging AI tasks is the visual question answering in the medical domain (VQA-Med) [1]. The main idea of VQA-Med system is to predict the right answer given a medical image accompanied with clinically relevant question. It is a difficult task because the computer system must understand and analyze the question (natural language processing, or NLP) as well as understand and process the image (computer vision).

Different approaches have been proposed to address the task of VQA-Med. Some of them treat the task as a generative problem generating answers in a comprehensive and well-formed textual description [2], while others treat it as a multi-label

Copyright (c) 2019 for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0). CLEF 2019, 9-12 September 2019, Lugano, Switzerland.

classification problem in which the answer is chosen from among different choices [3,4].

This paper describes our participation in the task of VQA-Med of ImageCLEF 2019 [5]. We proposed an encoder-decoder model that takes as input a medical question-image pair and generates an answer as output. The encoder network consists of a pre-trained CNN model that extracts prominent features from a medical image and a pre-trained word embedding along with Long Short-Term Memory (LSTM) to embed textual data. The answer generation is accomplished by the greedy search algorithm, which predicts the next word based on the previously generated words. Thus, the answer is built up by recursively calling the model.

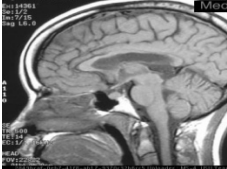

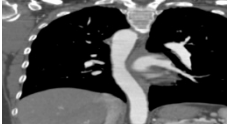
The rest of this paper is organized as follows. Section 2 describes details of the provided dataset. Section 3 gives a detailed description of the proposed system. Section 4, presents metrics used to assess the performance of our system and also provides a presentation and analysis of the experimental results, and finally Section 5 concludes the presented work with some remarks.

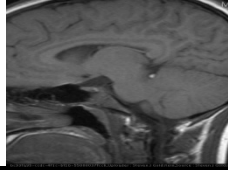
2 Dataset

VQA-Med dataset consists of 3,200 training medical images and 12,792 Question-Answer (QA) pairs, a validation set of 500 medical images with 2,000 QA pairs, and a test set of 500 medical images with 500 questions.

Four categories of questions are considered: Modality, Plane, Organ system and Abnormality. The answer can be either “a single word”, “a phrase containing 2-21 words”, or “a yes/no”. Table 1 illustrates some examples of medical images with associated question-answer pairs.

Table 1. Examples of medical images with associated question-answer pairs.

Medical Image	Question	Answer
	What part of the body is being imaged here?	Skull and contents.
	Which plane is the image shown in?	Axial.
	What abnormality is seen in the image?	Right aortic arch with aberrant left subclavian artery.



Is this a t1 weighted image?

Yes.

3 Methodology

To address the problem of VQA in the medical domain, we proposed an encoder-decoder model that takes as input a medical question-image pair and generates an answer as output. As shown in figure 1, the encoder network consists of a pre-trained DenseNet-21 model that extracts prominent features from the medical image and a pre-trained word embedding followed by two LSTM layers to embed the question and extract textual features. The textual and image features are concatenated together into one vector “QI vector”. Our proposed model generates one word at a time. That is, all words generated so far are embedded, with the same word embedding used for questions, and each of word embedding is fed then into an LSTM with 1024 units. The distributed representation of all words generated so far is concatenated with the “QI vector” to form an “encoder vector”. The decoder uses the encoder vector as input in order to generate the next word, this is then fed to a fully connected layer of 256 neurons and then to the final layer, which has one neuron for each word in the output vocabulary and a softmax activation function to output a likelihood of each word in the vocabulary being the next word in the answer. Thus, the answer is built up by recursively calling the model with the previously generated words.

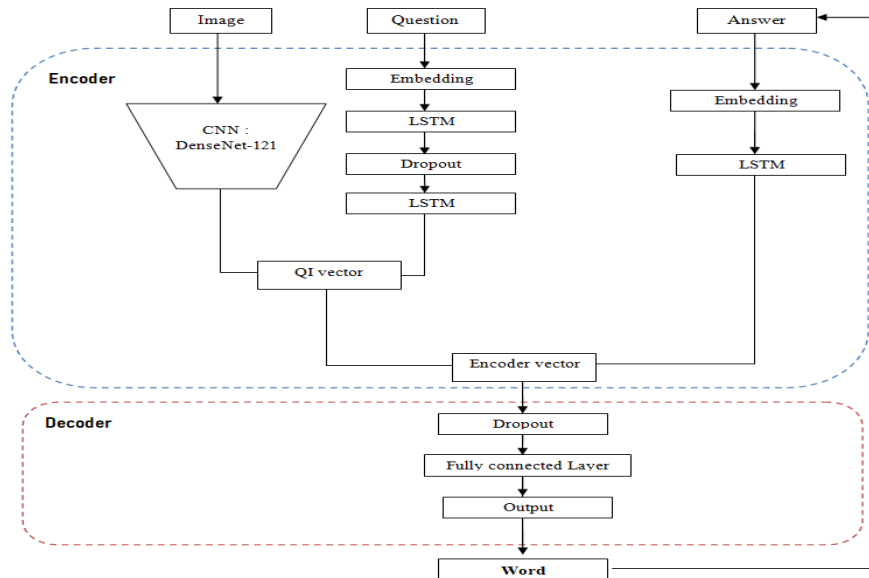


Fig. 1. The proposed architecture for VQA-Med 2019.

3.1 Image encoding:

Our proposed model is a deep learning network with a high number of parameters. This type of model often overfits when training on small datasets. To prevent overfitting, the best solution is to use the transfer learning technique. The idea is to use a pre-trained CNN model on a large and similar dataset as a fixed feature extractor, as we expect higher-level features in the CNN to be relevant to our dataset as well.

Motivated by the results obtained by DenseNet-121 model on the task of medical image classification [6] and since we don't have a large dataset, we used a pre-trained DenseNet-121 on chexpert [7], a large dataset of thorax chest-x-ray images. The network has four dense blocks, which have 6, 12, 24, 16 dense layers respectively. A dense block consists of a series of units. Each unit packs two convolutions, each preceded by Batch Normalization and ReLU activations. In addition, each unit generates a fixed number of feature vectors. This parameter, called growth rate, controls the amount of new information that layers can transmit. The layers between these dense blocks are transition layers which perform down-sampling of the features passing the network. A detailed explanation of DenseNet-121 architecture used in this work is shown in Table 2.

Table 2. The DenseNet-121 architecture.

Layers	Output Size	DenseNet-121
Convolution	112x112	7x7 conv, stride2
Pooling	56x56	3x3 max pool, stride 2
Dense Block 1	56x56	$\begin{bmatrix} 1x1 \text{ conv} \\ 3x3 \text{ conv} \end{bmatrix} \quad \times 6$
Transition Layer 1	56x56	1x1 conv
	28x28	2x2 average pool, stride 2
Dense Block 2	28x28	$\begin{bmatrix} 1x1 \text{ conv} \\ 3x3 \text{ conv} \end{bmatrix} \quad \times 12$
Transition Layer 2	28x28	1x1 conv
	14x14	2x2 average pool, stride 2
Dense Block 3	7x7	$\begin{bmatrix} 1x1 \text{ conv} \\ 3x3 \text{ conv} \end{bmatrix} \quad \times 24$

Table 2. The DenseNet-121 architecture (continued)

Layers	Output Size	DenseNet-121
Transition Layer 3	14x14	1x1 conv
	7x7	2x2 average pool, stride 2
Dense Block 4	7x7	$\begin{bmatrix} 1 \times 1 \text{ conv} \\ 3 \times 3 \text{ conv} \end{bmatrix} \times 16$
Output	1x1	7x7 global average pool

3.2 Question encoding:

To capture the sequential nature of language data, we modeled our questions using LSTM, a special type of Recurrent Neural Networks (RNNs). LSTM has demonstrated great success in various NLP tasks and is the state of the art algorithm for sequential data. It succeeds in being able to capture information about previous states to better inform the current prediction through its memory cell state.

An LSTM consists of three main components: a forget gate, input gate, and output gate. These gates determine whether or not to let new input in (input gate), delete the information because it isn't important (forget gate) or to let it impact the output at the current time step (output gate).

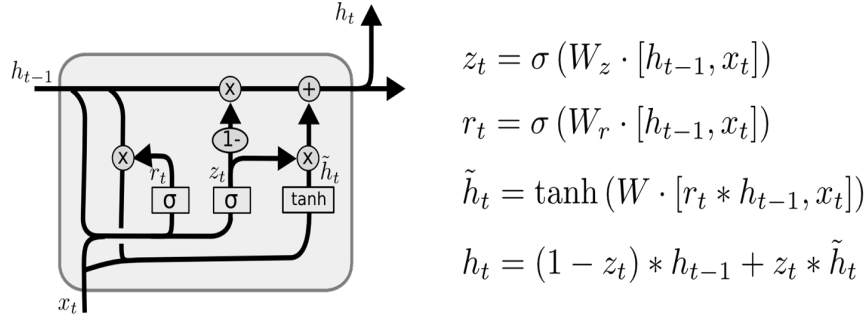


Fig. 2. Memory block in LSTM network.

A pre-trained word embedding [8] on biomedical texts from MEDLINE/PubMed using gensim's Word2Vec implementation is used to provide a distributed representation of words. Word Embeddings are much better at capturing the context around the words than using a one hot vector for every word. For this problem we used 200 dimension word embeddings and we did not tune them during the training process since we did not have sufficient data. These embeddings are passed into two LSTM layers with respectively 512 and 1024 units.

3.3 Answer generation:

To predict an answer for a given image-question pair, we treated the task as text generation. This often operates by generating probability distributions across the vocabulary of output words and it is up to decoding algorithms to sample the probability distributions to generate the most likely sequences of words. To find the best decoder algorithm both greedy search and beam search are evaluated.

- Greedy search:

A greedy algorithm uses a heuristic for making locally optimal choices at each step with the hope of finding a global optimum solution. This means that the algorithm chooses the most likely word in each step in the output sequence and does not take into account the entire sentence. Therefore, the quality of the final output sequence may be far from optimal, hence it is considered greedy.

- Beam search:

Unlike greedy search, beam search allows for non-greedy local decisions that can potentially lead to a sequence with a higher overall probability. The beam search expands all possible next steps and keeps the k most likely words, where k is a user-specified parameter and controls the number of beams or parallel searches through the sequence of probabilities.

3.4 Dropout:

The proposed model is a deep neural network model and is trained on a small dataset. As a result, the model can learn statistical noise in the training data, resulting in poor performance and generalization on new testing data “overfitting”. To reduce overfitting and improve generalization error, we used the dropout technique. Dropout is a very computationally cheap and remarkably effective regularization method. It works by randomly removing or dropping out inputs to a layer. This has the effect of making nodes in the network generally more robust to the inputs and reduces the number of training parameters, hence reduces the training time and memory requirements.

4 Evaluation Methodology:

Before applying the evaluation metrics, each answer undergoes the following pre-processing techniques:

- Lower-case: Converts each answer to lower-case.
- Tokenization: Divides the answer into individual words.
- Remove punctuation: Remove punctuation marks from answers.

Evaluation metrics used to evaluate our proposed VQA-Med model are:

- Accuracy (Strict): The entire predicted answer must match the ground truth answer.
- BLEU [9]: Capture the similarity between a system-generated answer and the ground truth answer.

Three experiments are conducted to evaluate our model:

- Expr1: Answers are generated using greedy search.
- Expr2: Answers are generated using beam search with k=2.
- Expr3: Answers are generated using beam search with k=3.

Our model is trained using RMSprop optimizer with an initial learning rate of 0.001 which is multiplied by 10 each time the validation loss plateau after an epoch. We have used a mini-batch size of 535 samples, a number of epochs up to 100, and the categorical cross-entropy as a loss function where the best model was selected based on the validation loss.

As shown in Table3, experiment 1 (Expr 1) achieves best results with a strict accuracy of 0.556 and BLEU score of 0.583. This means that for our case, greedy search is better than beam search algorithm.

Table 3. Experimental results on test dataset.

Experiment	Accuracy	BLEU
Expr1	0.556	0.583
Expr2	0.538	0.556
Expr3	0.526	0.547

The following table provides the results obtained by our model and the three best run for the task of VQA-Med.

Table 4. Comparison with the three best VQA-Med methods.

Model	Accuracy	BLEU
Hanlin	0.624	0.644
yan	0.62	0.64
minhvu	0.616	0.634
Our model (LIST)	0.556	0.583

As shown in Table 4, the best model achieved an accuracy of 0.624 and a BLUE score of 0.644. This means that it exceeds the results of our model with only 0.068 in terms of accuracy and 0.061 in terms of BLUE score. As a result, we can say that our model gives very good results.

5 Conclusion:

In this paper, we propose an Encoder-Decoder model for the task of visual question answering in the medical domain. VQA is a difficult and challenging task since it combines the fields of Computer Vision and NLP. This difficulty increases even more with the inherent nature of medical imaging. Our proposed model achieves great results with an accuracy of 0,556 and BLEU score of 0,583. To further substantiate these results, several improvements could be made such as the use of an attention mechanism that allows to pay more attention to specific regions that better represent the question instead of the whole image.

References

1. Ben Abacha, A., Hasan, S. A., Datla, V. V., Joey Liu, Demner-Fushman, D., Müller, H. (2019). VQA-Med: Overview of the Medical Visual Question Answering Task at ImageCLEF 2019. CLEF2019 Working Notes, CEUR Workshop Proceedings.
2. Talafha, B., and Al-Ayyoub, M. (2018). JUST at VQA-Med: A VGG-Seq2Seq Model. CLEF 2018 Labs Working Notes, CEUR Workshop Proceedings.
3. Allaouzi, I., Benamrou, B., Ben Ahmed, M. (2018). Deep Neural Networks and Decision Tree classifier for Visual Question Answering in the medical domain. CLEF 2018 Labs Working Notes, CEUR Workshop Proceedings.
4. Peng, Y., Liu, F., and Rosen, M. (2018). UMass at ImageCLEF Medical Visual Question Answering (Med-VQA) 2018 Task. CLEF 2018 Labs Working Notes, CEUR Workshop Proceedings.
5. Ionescu, B., Müller, H., and Péteri, R., Dicente Cid, Y., and Liauchukn V., and Kovalev, V., and 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. and Lux, M., Gurrin, C., Pelka, O., Friedrich, C. M., Garcia Seco de Herrera, A., Garcia, N., Kavallieratou, E., Roberto del Blanco, C., Cuevas Rodriguez, C., and Vasilopoulos, N., Karampidis, K., Chamberlain, J., Clark, A., Campello, A. (2019). ImageCLEF 2019: Multimedia Retrieval in Medicine, Lifelogging, Security and Nature. Experimental IR Meets Multilinguality, Multimodality, and Interaction, Proceedings of the 10th International Conference of the CLEF Association (CLEF 2019).
6. Allaouzi, I., Ben Ahmed, M. (2019). A Novel Approach for Multi-Label Chest X-Ray Classification of Common Thorax Diseases. IEEE Access 7(1), 64279-64288.
7. Irvin, J., Rajpurkar, P., Ko, M., Yu, Y., Ciurea-Ilcus, S., Chute, C., Marklund, H., Haghighi, B., Ball, R., Shpanskaya, K., Seekins, J., Mong, D., A., Halabi, S. S., Sandberg, K. J., Jones, R., Larson, B. D., Langlotz, C. P., Patel, N. B., Lungren, Matthew P. M., Andrew Y. Ng. (2019). Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In Thirty-Third AAAI Conference on Artificial Intelligence.
8. McDonald, R., Brokos, G., Androutsopoulos, I. (2018). Deep Relevance Ranking Using Enhanced Document-Query Interactions. Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP 2018), Brussels, Belgium.
9. Papineni, K.; Roukos, S.; Ward, T.; Zhu, W. J. (2002). BLEU: a method for automatic evaluation of machine translation (PDF). ACL-2002: 40th Annual meeting of the Association for Computational Linguistics. pp. 311–318.