

# Deep Neural Networks and Decision Tree classifier for Visual Question Answering in the medical domain

Imane Allaouzi, Badr Benamrou, Mohamed Benamrou and Mohamed Ben Ahmed

Abdelmalek Essaâdi University  
Faculty of Sciences and Techniques,  
Tangier, Morocco

**Abstract.** This paper presents our contribution to the problem of visual question answering in the medical domain using a combination of deep neural networks and the Decision tree classifier. In our proposed approach we consider the task of visual question answering as multi-label classification problem, where each label corresponds to a unique word in the answer dictionary that was built from the training set.

**Keywords:** CNN, Bidirectional LSTM, Decision Tree classifier, Language modeling, medical imaging, Visual Question Answering.

## 1 Introduction

Visual question answering (VQA) is a new and challenging task that has witnessed a surge interest from Artificial Intelligence (AI) community, since it combines the fields of Computer Vision (CV) and Natural Language Processing (NLP). NLP and CV are two branches of AI, where the former one enables computers to understand and analyze human language, while the second enables computers to understand and process images in the same way that a human does. The main idea of VQA systems is to predict the right answer giving both image and question about this image in a natural language. The VQA task can be treated as a classification problem if the answer is chosen from among different choices or as a generation problem if the answer is a comprehensive and well-formed textual description.

In the last few years, Deep Neural Networks have achieved the state-of-the-art in a wide range of NLP and CV applications including image recognition [1,2], machine translation[3,4],image caption[5,6] and Visual Question Answering[7,8,9]. Following this trend, this paper presents our contribution to the problem of visual question answering in the medical domain [10, 11] using a combination of deep neural networks (Convolutional Neural Network (CNN), Bidirectional Long Short-Term Memory) and the Decision tree classifier. In our proposed approach we consider the task of VQA as multi-label classification problem, where each label corresponds to a unique word in the answer dictionary that was built from the training set.

The paper’s arrangement is as follows: the dataset is described in Section 2, the proposed model is described in Section 3, results are presented and discussed in Section 4, and finally Section 5 draws some conclusions and future work.

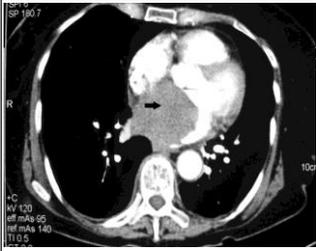
## 2 Dataset:

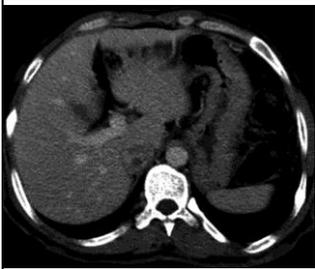
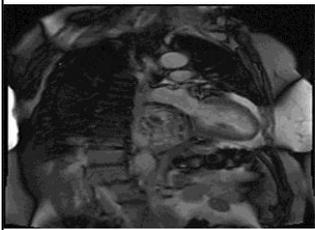
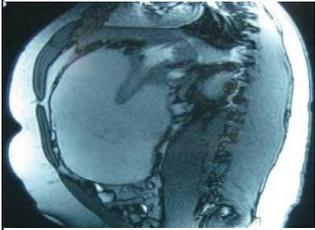
VQA-Med [10] is a dataset generated using images from PubMed Central articles (essentially a subset of the ImageCLEF 2017 caption prediction task [12]). As shown in the table 1 the VQA-Med dataset consists of 2278 training images and 324 validation images, accompanied respectively with 5413 and 500 of question-answer pairs, and a test set of 264 medical images with 500 questions. The answer can be either “a single word”, “a phrase containing around 2-28 words”, or “a yes/no”. The table 2 illustrates some examples of the training data with different types of questions and answers.

**Table 1.** The VQA-Med dataset distribution.

	Images	Questions	Answers
<b>Train</b>	2278	5413	5413
<b>Validation</b>	324	500	500
<b>Test</b>	264	500	-

**Table 1.** Some examples of the training data.

Images	Questions	Answers
	What does the CT scan show?	A large filling defect in the left atrium.
	Where does CT coronal section of the skull show well-defined unilocular lesion?	In the right maxillary sinus.

	Who does CT abdomen show?	Right adrenal pheochromocytoma.
	Is there any intra-cardiac mass identified?	No.
	What shows the limits between the stomach and mass?	MRI.

### 3 The Proposed Model:

The VQA in the medical domain involves providing a medical question-image pairs to produce answers. In this work we assume that the answers are a concatenation of one or more words, therefore we have treated the task as multi-label classification problem.

Our proposed model uses the pre-trained VGG-16[13] model to extract image features and the word embedding [14] along with a Bidirectional Long Short-Term Memory (LSTM) [15] to embed the question and extract textual features. The image and textual features are concatenated using two fully connected layers of 512 neurons to get a fixed length feature vector. This vector is used as a new input for Decision Tree Classifier in order to predict an answer.

The model consists of 3 sub-models:

- **Image Representation:**

To extract prominent features from medical images, we have used the pre-trained VGG-16 network that won the ImageNet 2014 challenge [16], by achieving

a 7.4% error rate on object classification. We have removed the last layer of this network to obtain an output vector of 4096 elements, which in turn passed through a fully connected layer to get image representation of size 512. The VGG-16 architecture is shown in the figure 1:

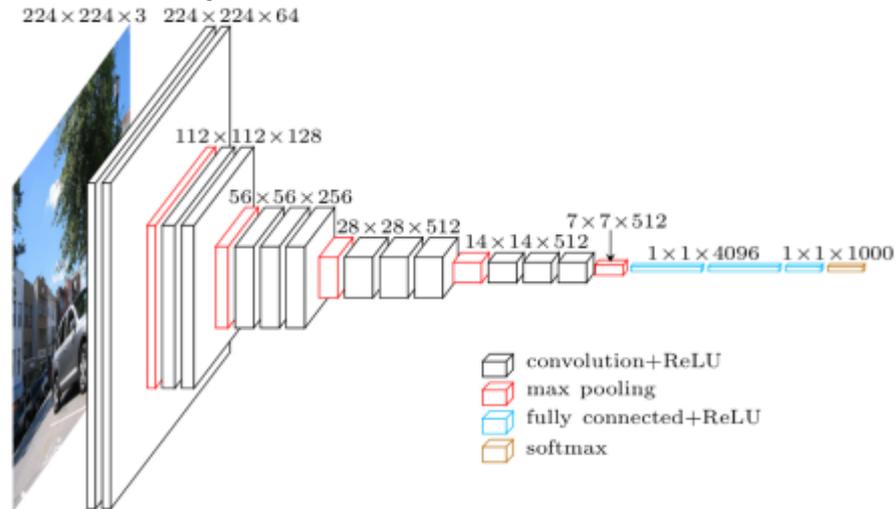


Fig. 1. The VGG-16 model architecture [17].

- **Question Representation:**

Recently recurrent neural networks (RNNs) have shown great success in diverse NLP tasks [18, 19], motivated by this success we have used a bidirectional RNN with LSTM for dealing with the medical questions. Bidirectional Long Short-Term Memory (BDLSTM) is an extension of the traditional LSTM; its main idea consists of processing sequence data in both forward and backward directions to avoid the problem of limited context that applies to any feed-forward model.

For that, first the question is converted to a matrix of one-hot vectors and passed through an embedding layer (with a vocabulary of 3312 and a dense embedding of 521), in order to get their dense representation and their relative meanings. The embedded question is then fed to a BDLSTM with 512 units followed by a fully connected layer to get question representation of size 512.

- **Answer prediction:**

To predict an answer, we have modeled the VQA-Med task as multi-label classification problem, since we have assumed that an answer is a concatenation of one or more words. Therefore, we have used the multi-label Decision Tree classifier that takes as input the output from both sub-models of image representation and question representation and predicts one or more predefined labels. The total number of labels equals to 3109. Where, each label corresponds to a unique word in the answer dictionary that was created from the training set.

In the training phase, we have kept the CNN parameters frozen, and we have trained the rest of our deep neural network using a fully connected layer with sigmoid as activation function, Binary Cross-entropy as loss function and Adam as optimizer. As well as, the dropout technique was used before the last fully connected layer and after the BDLSTM layer with a probability of 0.5.

The best parameters were selected based on the validation loss, with a mini-batch of 20 and a number of epochs up to 10.

## 4 Results:

Three metrics are used to evaluate our proposed VQA-Med model, which are: BLEU score [20], WBSS (Word-based Semantic Similarity), and CBSS (Concept-based Semantic Similarity). The first one is one of the most commonly used metrics that have been used to measure the similarity between two sentences, the second one aims to calculate the semantic similarity in the biomedical domain [21], it was created based on Wu-Palmer Similarity (WUPS) [22] with WordNet ontology in the backend, while the third one is similar to the WBSS metric, except that instead of tokenizing the predicted and ground truth answers into words, it uses MetaMap via the pymetamap wrapper to extract biomedical concepts from the answers.

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.
- Stop-words: Removes punctuations and commonly encountered English words.

The following table shows the results obtained on the test set:

**Table 3.** Results of our proposed model on Test set.

Evaluation metrics		
BLEU	WBSS	CBSS
0.053867018	0.100854295	0.269119831

As shown in the table above, our proposed model gives good results in term of CBSS metric (0.27) comparing with BLEU score (0.054) and WBSS metric (0.10). This is justified by the high number of labels that are not presented equally in the training set. This is what is known as the label imbalance problem.

## 5 Conclusion:

In this paper, we present our contribution to the task of visual question answering in the medical domain. We have treated the task as a multi-label classification using the decision tree classifier. However, the results on test set are totally unsatisfactory, especially in term of BLEU metric with a score of 0.054. Therefore, we think to develop

an LSTM model to generate answers since the adopted classification approach ignores words order in the answer which leads to a loss of information. We also think to improve our visual model by using the attention technique .This technique allows to pay more attention to specific regions that better represent the question instead of the whole image.

## References

1. Simonyan, K., Zisserman A. : Very deep convolutional networks for large-scale image recognition. arXiv:1409.1556. 2, 8, 17 (2014)
2. Krizhevsky, A., Sutskever, I., Hinton, G. E., ImageNet classification with deep convolutional neural networks. In NIPS (2012)
3. Kyunghyun.Cho., van Merriënboer, B., Gulcehre, C.: Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation, in Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 1724-1734. Association for Computational Linguistics, Doha (2014) .
4. Sutskever, I., Vinyals, O.,V Le, Q.: Sequence to Sequence Learning with Neural Networks, In: the 27th International Conference on Neural Information Processing Systems,Vol. 2, pp. 3104-3112 (2014)
5. Vinyals, O., Toshev, A., Bengio, S., Erhan, D.: Show and tell: A neural image caption generator. In CVPR ( 2015)
6. Xu, K., Ba, J., Kiros, R., Cho, K., Courville, A., Salakhutdinov, R., Zemel, R., Bengio, Y. : Show, attend and tell: Neural image caption generation with visual attention. In ICML (2015)
7. Yang, Z., He, X., Gao, J., Deng, L., Smola, A.: Stacked attention networks for image question answering. In CVPR ( 2016)
8. Zhou, B., Tian, Y., Sukhbaatar, S., Szlam, A., Fergus, R.: Simple baseline for visual question answering. arXiv preprint arXiv:1512.02167 (2015)
9. Malinowski, M., Rohrbach, M., Fritz, M. : Ask your neurons: A deep learning approach to visual question answering. arXiv preprint arXiv:1605.02697 (2016)
10. Hasan, SA., Ling, Y., Farri, O., Liu, J., Lungren, M., Müller, H. : Overview of ImageCLEF 2018 Medical Domain Visual Question Answering Task, CLEF working notes, CEUR, (2018).
11. Ionescu, B., Müller, H., Villegas, M., García Seco de Herrera, A., Eickhoff, C., Andrearczyk, V., Dicente Cid, Y., Liauchuk, V., Kovalev, V., Hasan, SA., Ling, Y., Farri, O., Liu, J., Lungren, M., Dang-Nguyen, D., Piras, L., Riegler, M., Zhou, L., Lux, M., Gurrin, M. : Overview of ImageCLEF 2018: Challenges, Datasets and Evaluation, Proceedings of the Ninth International Conference of the CLEF Association (CLEF 2018), (2018).
12. Eickhoff, C., Schwall, I., García Seco de Herrera, A., Muller, H. : Overview of ImageCLEFcaption 2017 - the image caption prediction and concept extraction tasks to understand biomedical images. CLEF working notes, CEUR (2017)
13. Simonyan, K., Zisserman, A. : Very Deep Convolutional Networks for Large-Scale Image Recognition?. arXiv preprint arXiv:1409.1556 (2014)
14. Mikolov, T., Sutskever, I., Chen, K., Corrado G. S., Dean, J. : Distributed Representations of Words and Phrases and their Compositionality. In NIPS, 4, 8, 17 (2013)
15. Schuster M., Paliwal., K. K. : Bidirectional recurrent neural networks. In: IEEE Transactions on Signal Processing, vol. 4, pp. 2673-2681 (1997)
16. <http://www.image-net.org/challenges/LSVRC/2014/> , last accessed 2018/05/25

17. <https://blog.heuritech.com/2016/02/29/a-brief-report-of-the-heuritech-deep-learning-meetup-5/>, last accessed 2018/04/22
18. Graves, M., Mohamed, A., Hinton, G. : Speech recognition with deep recurrent neural networks. In: Proceedings of the 2013 IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 6645-6649. (2013)
19. Mikolov, T., Kombrink, S., Deoras, A., Burget, L., Cernocky, J. : RNNLM-recurrent neural network language modeling toolkit. In Proceedings of the 2011 ASRU Workshop, pp. 196-201. (2011)
20. Papineni, K., Roukos, S., Ward, T., Zhu, W. J. : BLEU: a method for automatic evaluation of machine translation (PDF). ACL-2002: 40th Annual meeting of the Association for Computational Linguistics, pp. 311-318. Association for Computational Linguistics, Pennsylvania (2002)
21. Soğancıoğlu, G., Öztürk, H., & Özgür, A.: BIOSSES: a semantic sentence similarity estimation system for the biomedical domain. *Bioinformatics*, 33(14), (2017). i49-i58
22. Wu, Z., Palmer, M.: Verbs semantics and lexical selection. In: Proceedings of the 32nd annual meeting on Association for Computational Linguistics Association for Computational Linguistics, pp. 133-138. (1994)