Tlemcen University at ImageCLEF 2019
Visual Question Answering Task

Rabia Bounaama[1] and Mohammed El Amine Abderrahim[2]

[1] Biomedical Engineering Laboratory, Tlemcen University,Algeria
rabea.bounaama@univ-tlemcen.dz
[2] Laboratory of Arabic Natural Language Processing, Tlemcen University, Algeria
mohammedelamine.abderrahim@univ-tlemcen.dz

Abstract In this paper we describe our methodology of techno team participation at ImageCLEF Medical Visual Question Answering 2019 task. VQA-Med task is a challenge which combines computer vision with Natural Language Processing (NLP) in order to build a system that manages responses based on set of medical images and questions that suit them. We used a jointly learning for text and image method in order to solve the task, we tested a publicly available VQA network. We apply neural network and visual semantic embeddings method on this task. Our approach based on CNNs and RNN model achieve 0.486 of BLEU score.

Keywords: CNNs, neural networks, VQA-Med task, RNN.

1 Introduction

There are many more complex questions that can be asked in medical Radiology, which is very rich of images and textual reports, is a prime area where VQA could assist radiologists in reporting findings for a complicated patient or benefit trainees who have questions about the size of a mass or presence of a fracture [1]. VQA system is expected to reason over both visual and textual information to infer the correct answer [2]. So medical VQA systems define as a computer vision and Artificial Intelligence (AI) problem that aims to answer questions asked by health care professionals about medical images [1].Artificial neural network models have been studied for many years in the hope of achieving human-like performance in several fields such as speech and image understanding [3]. VQA could be used to improve human-computer interaction as a natural way to query visual content. It has garnered a large amount of interest from the deep learning, computer vision, and NLP communities [4].

ImageCLEF provide medical image collections, annotated toward several evaluation challenges including VQA, image captioning, and tuberculosis [5,10]. We participate in the task of VQA in the medical domain.

Participating systems are tasked with answering the question based on the visual image content. The evaluation of the VQA-Med task participant systems is con ducted by using two metrics: BLEU and accuracy.

The following of this paper is organized as follows. In section 2 we present some related works. In section 3 we describe our approach and more specifically we present the dataset and discuss in detail the models and techniques used in our submitted run. The conclusion and future work perspectives are presented in section 4.

2 Related Work

Convolutional Neural Networks (CNNs) make a promising model for the ImageNET classification task to medical modality such as in the work of [1,6], where the authors of Novasearch team [1] evaluate the CNNs classifier with medical images in order to build a Medical Image Retrieval System (MIRS) to classify each subfigure, from a collection of figures from compound images found in biomedical literature.

Another work of subfigure classification task at ImageCLEF 2016 [6] used modern Deep CNNs in order to predict the modality of a medical image with two main groups : Diagnostic Images and Generic Biomedical Illustration. To extract information from medical images and build their textual features, they used Bag-of-Words (BoW) and Bag of Visual Words (BoVW) approaches.

In the work of [2] they used Multi-modal Factorized Bilinear(MFB) pooling as well as Multimodal Factorized High-order(MFH) pooling to solve the task in order to build a system that is able to reason over medical images and questions and generate the corresponding answers at ImageCLEF Med-VQA 2018 Task.

The main idea proposed by [7] is about automatically generate questions and images selected from the literature based on ImageCLEF data where they apply Stacked Attention Network (SAN) which was proposed to allow multi-step reasoning for answer prediction, and Multimodal Compact Bilinear pooling (MCB) with two attention layers based on CNNs.

The authors of [1] introduce VQA-RAD, a manually constructed VQA dataset in radiology where clinicians asked naturally occurring questions about radiology images and provided reference answers in order to encourage the community to design VQA tools with the goals of improving patient care where they use a balanced images sample from MedPix. The annotation of the dataset was generated by volunteer clinical trainees and validated by expert radiologists, they train their data using deep learning and bootstrapping approaches. They provide the data in JSON, XML, and Excel format. The final VQA-RAD dataset contains 3515 total visual questions.

Another line of work in [8] focuses in a new ways of synthesizing QA pairs from currently available image description datasets. They propose to use neural networks and visual semantic embeddings using LSTM on MS-COCO dataset. Their final model was not able to consume image features as large as 4096 dimensions at one time step, where the dimensionality reductions lose some useful information.

Deep neural networks have recently achieved very good results in representation learning and classification of images. With all this effort, there is still no widely used method to construct these systems. This is due to the fact that the medical domain

requires high accuracy and especially the rate of false negatives to be very low, so we studied several VQA networks and we selected deep neural networks models for our participation in VQA-Med 2019.

3 Methodology

3.1 Dataset

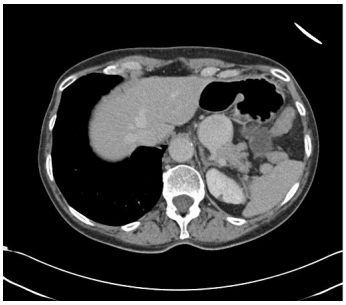
In the scope of the VQA-Med challenge, three datasets were provided:

- The training set contains 12792 question-answer pairs associated with 3200 training images.
- The validation set contains 2000 question-answer pairs associated with 500 validation images.
- The test set contains 500 questions associated with 500 test images.

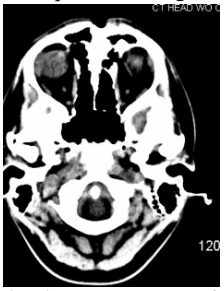
The classes for each question category are: Modality, Plane, Organ system and Abnormality (see table1)

Table 1. Example of a medical images and the associated questions and answers from the training set of ImageCLEF 2019 VQA-Med

Q: what type of imaging modality is used to acquire the image? A: us ultrasound



Q: what plane was used? A: axial



Q:what organ system is evaluated primarily? A:face, sinuses, and neck

Q: is this image normal? A:yes

3.2 Method and Results

To solve the task of VQA-med at image CLEF 2019, we chose to use CNN and RNN models without intermediate stages such as object detection and image segmentation.

All existing methods and VQA algorithms consist of:

- Extracting image features (image featurization).
- Extracting question features (question featurization).
- Combining features to produce an answer [4] (see figure 1).

In our case, we chose the approach used by [8]. We treat the task as a classification problem and we apply neural network and visual semantic embeddings method. We assume that the answers consist of only a single word.

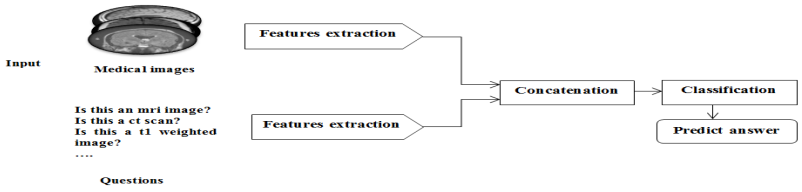


Figure 1. Model process

The process of building the classification model includes preprocessing and extraction of visual features from already labelled images and their own questions. Our system learns image regions relevant to answer the clinical questions. Images and question are represented as global features which are merged to predict the answers. The effectiveness of the model is evaluated by using new images.

We have used visual semantic embeddings to connect a CNN and a Recurrent Neural Networks (RNN). Our model is built on the basis of the LSTM (Long short-term memory) which is an easier form of RNN to train the dataset. Because of its very uniform architecture for extracting features from images we used 16 convolutional layers of VGG-16 and in order to generate questions as inputs examples, we used RNN which is the apropriate technique to use with sequential data [9]. The LSTM(s) outputs are introduced into a softmax layer to generate answers.

The answer prediction task is modeled as N-class classification problem and performed using a one-layer neural network.

Our model results are shown in the table below (see table 2).

The analysis of the results obtained by all the participants, in terms of accuracy and BLEU, shows that the best approach is the one used by the Hanlin team. The results obtained by all the participants vary between 0.624 and 0 in terms of accuracy and between 0.644 and 0.025 in terms of BLEU. The Hanlin team thus obtained the highest scores (0.624, 0.644) while the IITISM @ CLEF team obtained the lowest scores (0.0, 0.025).

Table 2. Techno team score.

accuracy	BLEU
0.462	0.486

The results obtained by our system (0.462, 0.485) compared with other systems are encouraging and we hope to make improvements in the future.

4 Conclusion

In this paper, we present techno team approach used in VQA at ImageCLEF 2019 task. We evaluate currently existing VQA system by testing a publicly available VQA network.

We found that the RNN model based on the feature fusion is helpful on improving the system's performance. But it is still very naive in many situations.

It should be noted that we encountered the problem of overfitting which is a major problem in neural networks.

As a result, we achieved 0.486 BLEU score in the challenge. In the future we consider working on in order to obtain the optimum deep learning layer structure.

References

1. Lau, Jason J and Gayen, Soumya and Abacha, Asma Ben and Demner-Fushman, Dina. A dataset of clinically generated visual questions and answers about radiology images. Scientific data,5,180251 (2018).
2. Peng, Yalei and Liu, Feifan and Rosen, Max P.UMass at ImageCLEF Medical Visual Question Answering (Med-VQA) 2018 Task (2018).
3. Antonie, Maria-Luiza and Zaiane, Osmar R and Coman, Alexandru. Application of data mining techniques for medical image classification.Proceedings of the Second International Conference on Multimedia Data Mining, 94–101.Springer-Verlag. (2001).
4. Kafle, Kushal and Kanan, Christopher. Visual question answering: Datasets, algorithms, and future challenges.Computer Vision and Image Understanding,163,3–20.Elsevier,(2017).
5. Asma Ben Abacha and Sadid A. Hasan and Vivek V. Datla and Joey Liu and Dina Demner-Fushman and Henning Müller.VQA-Med: Overview of the Medical Visual Question Answering Task at ImageCLEF 2019. CLEF2019 Working Notes.CEUR Workshop Proceedings.CEUR-WS.org <http://ceur-ws.org>.September 9-12.Lugano, Switzerland (2019).
6. Koitka, Sven and Friedrich, Christoph M. Traditional Feature Engineering and Deep Learning Approaches at Medical Classification Task of ImageCLEF 2016. CLEF (Working Notes).304–317. (2016).
7. Abacha, Asma Ben and Gayen, Soumya and Lau, Jason J and Rajaraman, Sivaramakrishnan and Demner-Fushman, Dina. NLM at ImageCLEF 2018 Visual Question Answering in the Medical Domain (2018).
8. Ren, Mengye and Kiros, Ryan and Zemel, Richard.Exploring models and data for image question answering. Advances in neural information processing systems.2953–2961. (2015).

9. Liu, Pengfei and Qiu, Xipeng and Huang, Xuanjing. Recurrent neural network for text classification with multi-task learning.arXiv preprint arXiv:1605.05101, (2016).
10. Bogdan Ionescu and Henning Müller and Renaud Péteri and Yashin Dicente Cid and Vitali Liauchuk and Vassili Kovalev and Dzmitri Klimuk and Aleh Tarasau and Asma Ben Abacha and Sadid A. Hasan and Vivek Datla and Joey Liu and Dina Demner-Fushman and Duc-Tien Dang-Nguyen and Luca Piras and Michael Riegler and Minh-Triet Tran and Mathias Lux and Cathal Gurrin and Obioma Pelka and Christoph M. Friedrich and Alba García Seco de Herrera and Narciso Garcia and Ergina Kavallieratou and Carlos Roberto del Blanco and Carlos Cuevas Rodríguez and Nikos Vasillopoulos and Konstantinos Karampidis and Jon Chamberlain and Adrian Clark and Antonio Campello . 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), LNCS Lecture Notes in Computer Science, Springer.Lugano, Switzerland.September 9-12, (2019).