Project Report on

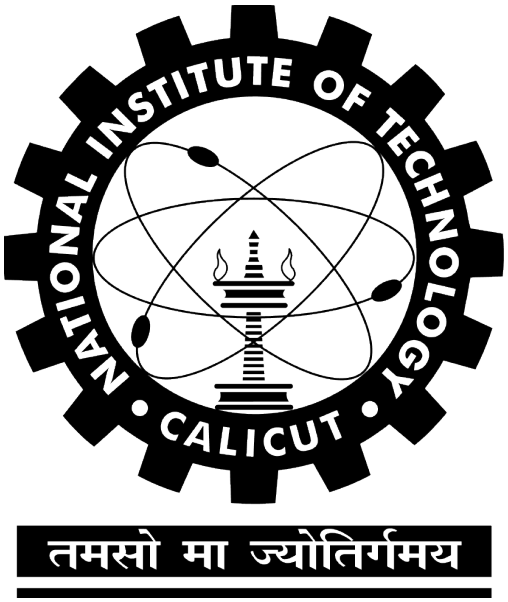
Visual Question Answering

Submitted by

K A Siva Vardhan	**B160333CS**
P Satyanarayana	**B160340CS**
B Maheswara Rao	**B160349CS**
R Vamsi Krishna	**B160109CS**

Under the Guidance of

Lijiya A.



Department of Computer Science and Engineering
National Institute of Technology Calicut
Calicut, Kerala, India - 673 601

October 14, 2019

Visual Question Answering

K A Siva Vardhan P Satyanarayana B Maheswara Rao R Vamsi Krishna

Abstract: Automated systems could help clinicians cope with large amounts of images by answering questions about the image contents. Firstly, we use some image enhancement methods like clipping and questions preprocessing methods. Secondly, we use CNN to extract image features, and use RNN to encode the questions. Finally, we concatenate the coded questions with the image features to generate the answers.

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.

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.

The VQA-Med task was introduced for the first time in 2018, inspired by the open-domain VQA challenges that started in 2015. Given a medical image and a natural language question about the image, participating systems are tasked with answering the question based on the visual image content. Three datasets were provided for training, validation and testing.

For image processing, we use Convolutional Neural Networks (CNNs). Image features are extracted from the last pooling layer of the CNNs. For question processing, we use LSTMs without pre-trained embeddings. Question vectors are extracted from the final hidden layer of the LSTMs. These two are concatenated to get the result.

2 Problem statement

Given a medical image accompanied with a clinically relevant question, participating systems are tasked with answering the question based on the visual image content. This may vary from simple problem such as classification of the image to a complex one such as obtaining the answer only by concatenating the clinically relevant question with the medical image.

3 Literature Review

3.1 Employing Inception-Resnet-v2 and Bi-LSTM for Medical Domain VQA

Kafle K et al. and other researchers summarized quite a few methods for VQA. The majority of them used recurrent neural networks such as LSTM to encode questions, and used deep convolutional neural networks such as VGG16 to focus on image recogni-

tion in advance. Deep convolutional neural networks (CNN) can be used to extract the features of an image and identify the objects in it. The Inception-Resnet-v2 model is one kind of advanced convolutional neural network that combines the inception module with ResNet .

Elman J L first used a recurrent neural network (RNN) to handle sequences problems. Nevertheless, context information is easily ignored when RNN processes long sequences. The proposal of LSTM alleviated the problem of long-distance dependence. Furthermore, the researchers also found that if the input sequence is reversed, the corresponding path from the decoder to the encoder will be shortened, contributing to network memory. The Bi-LSTM model combines the two points above, and makes the result better.

3.2 NLM at ImageCLEF 2018 VQA in the Medical Domain

This paper describes the participation of the U.S. National Library of Medicine (NLM) in the Visual Question Answering task (VQAMed) of ImageCLEF 2018.They studied deep learning networks with state of-the-art performance in open-domain VQA.They selected Stacked Attention Network (SAN) and Multimodal Compact Bilinear pooling (MCB) for their official runs.

3.2.1 Stacked Attention Network(SAN)

- The Stacked Attention Network (SAN) was proposed to allow multi-step reasoning for answer prediction. SAN includes three components: (i) the image model based on a CNN to extract high level image representations, (ii) the question model using an LSTM to extract a semantic vector of the question and (iii) the stacked attention model which locates the image regions that are relevant to answer the question.

- For the image model, we used the last pooling layer of VGG-16 pre-trained on imageNet as image features. For the question model, we used the last LSTM layer as question features. The image features and the question vector were used to generate the attention distribution over the regions of the image.

- The first attention layer of the SAN is then computed to capture the correlations between the tokens of the question and the regions in the image. Multimodal pooling is performed to generate a combined question and image vector that is then used as the query for the image in the next layer. We used two attention layers, as it showed better results in open-domain VQA. The last step is answer prediction. For a set of N answer candidates, the answer prediction task is modeled as N-class classification problem and performed using a one-layer neural network. Answers are predicted using Softmax probabilities.

3.2.2 Multimodal Compact Bilinear pooling(MCB)

- Multimodal Compact Bilinear pooling (MCB) is an attention mechanism that implicitly computes the outer product of visual and textual vectors.

- MCB architecture contains: (i) a CNN image model, (ii) an LSTM question model, and (iii) MCB pooling that first predicts the spatial attention and then combines the attention representation with the textual representation to predict the answers

- For the image model, we used ResNet-152 and ResNet-50 pre-trained on imageNet. For the question model, a 2-layer (1024 units in each layer) LSTM model is used. Concatenated output from both layers (2048 units) forms the input to the next pooling layer. MCB pooling is then used to combine both image and textual vectors to produce a multimodal representation. To incorporate attention, MCB pooling is used again to merge the multimodal representation with the textual representation for each spatial grid location. We also fine-tuned ResNet-50 on modality classification.

3.3 JUST at VQA-Med: A 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.This model follows the encoder-decoder architecture.

- 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.

3.4 Multimodal Compact Bilinear Pooling using Fast Fourier Transform

Multimodal Compact Bilinear pooling computes the outer product between two vectors, which allows, in contrast to element-wise product, a multiplicative interaction between all elements of both vectors. Bilinear pooling models (Tenenbaum and Freeman, 2000) have recently been shown to be beneficial for fine-grained classification for vision only tasks (Lin et al., 2015). However, given their high dimensionality (n 2), bilinear pooling has so far not been widely used. In this paper, we adopt the idea from Gao et al. (2016) which shows how to efficiently compress bilinear pooling for a single modality. In this work, we discuss and extensively evaluate the extension to the multimodal case for text and visual modalities. Multimodal Compact Bilinear pooling (MCB) is approximated by randomly projecting the image and text representations to a higher dimensional space (using Count Sketch (Charikar et al., 2002)) and then convolving both vectors efficiently by using element-wise product in Fast Fourier Transform (FFT) space. We use MCB to predict answers for the VQA task and locations for the visual grounding task. For open-ended question answering, we present an architecture for VQA which uses MCB twice, once to predict spatial attention and the second time to predict the answer. For multiple-choice question answering we introduce a third MCB to relate the encoded answer to the question-image space.

4 Work Plan

Since Visual Question Answering is a large area which contains numerous types of images, countably many questions for each image and also many responses for each question, we restrict our project to specific domain.

4.1 Dataset Collection

Given an image and a natural language question, the VQA task consists in providing an accurate natural language answer based on the content of the image.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.

By analyzing the questions manually, three main types of questions could be identified:

- Finding, e.g. what is abnormal in the ct scan? what is most alarming about this mri? what organ system is shown in the image?

- Yes/No questions, e.g. is this a ct scan? are there abnormalities in this mri? is this a non-contrast mri?
- Other questions, e.g. what kind of image is this? what was this image taken with?

4.2 Design

Questions for an image can be of various types like:

- Image Enhancement
- Object recognition
- Text Preprocessing
- Finding corresponding objects
- Counting
- Other Attributes etc...

In this phase we choose the types of questions to be considered for the given dataset and start preparing various questions and answers for each image.

4.3 Implementation

After the design phase, we will have the image set and the questions set. In this phase we start building our model using python and python packages like pytorch, keras, tensorflow etc. We may further modify the project by including all types of questions.

5 Methodology

In order to successfully complete the implementation phase there are three main modules which help to understand images and questions and reasoning possible answers to correct result.

5.1 Image Enhancement methods

Image enhancement is the process of adjusting digital images so that the results are more suitable for display or further image analysis. For example, you can remove noise, sharpen, or brighten an image, making it easier to identify key features.

5.2 Visual Feature Extraction

In this project, we use CNN to focus on image recognition in advance. CNN image classification takes an input image and processes it through a series of steps to recognize/Classify the input image. The input image is seen as an array of pixels. Main Steps involved in CNN are:

- Convolution and ReLU
  Convolution is the first layer to extract features from an input image.In this we will extract feature maps by convoluting the input image with feature detectors/Kernels. We can find out the features are present without losing much information. Convolution preserves the relationship between pixels by learning image features using small squares of input data.
- Pooling
  Pooling layers section would reduce the number of parameters when the images are too large.It also helps from reduces overfitting.
- Flattening
  Pooled Feature maps are then Flattened to sequence of inputs i.e, 1D array.
- Full connection
  The layer we call as FC layer, we flattened our matrix into vector and feed it into a fully connected layer of neural network.

5.3 Question Preprocessing

Given text should be preprocessed to remove unnecessary noise, capital letters etc.

5.4 Question Understanding

We use Recurrent Neural Network (RNN) or Long short-term memory (LSTM) for natural language processing for the preprocessed text.

5.5 Answer Generation

This module receives both question feature and image feature. These two features are then concatenated to get the result.

6 Conclusion

This report presented a review of the state of the art on visual question answering. We reviewed popular approaches based on deep learning, which treat the task as a classification problem over a set of candidate answers. Despite shortcomings of current practices for both training and evaluating VQA systems, we identified a number of promising research avenues that could potentially bring future breakthroughs for both VQA and for the general objective of visual scene understanding. Since deep learning techniques are significantly improving, we can reasonably expect that VQA is going to be more and more accurate in the next years.

References

[1] Y. Zhou, X. Kang, and F. Ren, "Employing inception-resnet-v2 and bi-lstm for medical domain visual question answering,"

[2] A. B. Abacha, S. Gayen, J. J. Lau, S. Rajaraman, and D. Demner-Fushman, "Nlm at imageclef 2018 visual question answering in the medical domain," 2018.

[3] J. J. Lau, S. Gayen, A. B. Abacha, and D. Demner-Fushman, "A dataset of clinically generated visual questions and answers about radiology images," Scientific data, vol. 5, p. 180251, 2018

[4] B. Talafha and M. Al-Ayyoub, "Just at vqa-med: A vgg-seq2seq model"

[5] Akira Fukui, Dong Huk Park, Daylen Yang and Anna Rohrbach "Multimodal Compact Bilinear Pooling for Visual Question Answering and Visual Grounding"

[6] Zichao Yang, Xiaodong He, Jianfeng Gao, Li Deng, Alex Smola. Stacked Attention Networks for Image Question Answering. The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 21-29

[7] Damien Teney, Peter Anderson, Xiaodong He, Anton van den Hengel.Tips and Tricks for Visual Question Answering: Learnings From the 2017 Challenge. The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 4223-4232

[8] Peng Wang, Qi Wu, Chunhua Shen, Anton van den Hengel. The VQA-Machine: Learning How to Use Existing Vision Algorithms to Answer New Questions. The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 1173-1182