

Visual Question Answering

Guide: Lijiya A.

Submitted By:

K.A.Siva Vardhan - B160333CS

P.Satya Narayana - B160340CS

B.Maheswararao - B160349CS

R.Vamsi Krishna - B160109CS

Introduction

This project is about building a model which will take an image and a question related to this image and outputs the appropriate answer to the given question.



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.

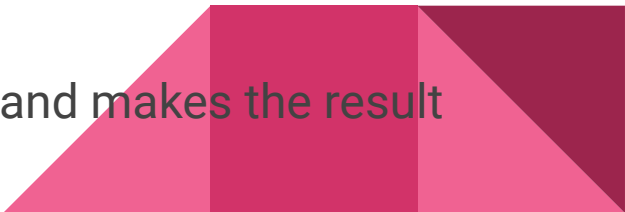


Literature Review

- Employing Inception-Resnet-v2 and Bi-LSTM for Medical Domain VQA
- NLM at ImageCLEF 2018 VQA in the Medical Domain
- JUST at VQA-Med: A VGG Seq2Seq Model
- Multimodal Compact Bilinear Pooling using Fast Fourier Transform



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

- Used recurrent neural networks such as LSTM to encode questions
 - Used deep convolutional neural networks such as VGG16 to focus on image recognition in advance.
 - The proposal of LSTM alleviated the problem of long-distance dependence in RNN.
 - 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.
- 

NLM at ImageCLEF 2018 VQA in the Medical Domain

- Studied deep learning networks with state of-the-art performance in open-domain VQA.
- Selected Stacked Attention Network (SAN) and Multimodal Compact Bilinear pooling (MCB) for their official runs.



Stacked Attention Network(SAN)

- Proposed to allow multi-step reasoning for answer prediction.
- Includes three components
 - Image model based on a CNN to extract high level image representations.
 - Question model using an LSTM to extract a semantic vector of the question.
 - Stacked attention model which locates the image regions that are relevant to answer the question.



Multimodal Compact Bilinear pooling(MCB)

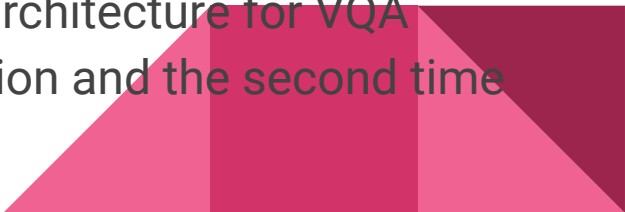
- Is an attention mechanism that implicitly computes the outer product of visual and textual vectors.
- MCB architecture contains
 - CNN image model.
 - LSTM question model.
 - MCB pooling that first predicts the spatial attention and then combines the attention representation with the textual representation to predict the answers.



JUST at VQA-Med: A VGG Seq2Seq Model

- This model follows the encoder-decoder architecture.
- The encoder consists of two main components.
 - Long short term memory (LSTM) network with a pretrained word embedding layer which encodes the question into a vector representation.
 - 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 $\langle \text{start} \rangle$ token as input in the first time step and try to predict the answer using softmax layer.


Multimodal Compact Bilinear Pooling using Fast Fourier Transform

- Computes the outer product between two vectors, which allows, in contrast to element-wise product, a multiplicative interaction between all elements of both vectors.
 - Multimodal Compact Bilinear pooling (MCB) is approximated by randomly projecting the image and text representations to a higher dimensional space (using Count Sketch) and then convolving both vectors efficiently by using element-wise product in Fast Fourier Transform (FFT) space.
 - 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.
- 

Work Plan

- Since VQA 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(i.e Radiology images in our case).

This includes three phases :

- Data set collection
 - Design
 - Implementation
- 

Data set collection

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.



Design

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.

Questions for an image can be of various types like:

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



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.



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.They are:
1. Image Enhancement methods & Visual Feature Extraction
 2. Question Preprocessing & Question Understanding
 3. Answer Generation



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.



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
- Pooling
- Flattening
- Full connection



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 , Flattening & Full connection

- Pooling layers section would reduce the number of parameters when the images are too large. It also helps from reduces overfitting.
- Pooled Feature maps are then Flattened to sequence of inputs i.e, 1D array.
- The layer we call as FC layer, we flattened our matrix into vector and feed it into a fully connected layer of neural network.




Question Preprocessing & Understanding

- Preprocessing:

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

- Understanding:

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




Answer Generation

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



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 vqamed: A vgg-seq2seq model"



References

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
- 

THANK YOU.

