



SHRI VAISHNAV VIDHYAPEETH VISHWAVIDHYALAYA, INDORE

Shri Vaishnav Institute of Information Technology

Department of Computer Science Engineering

Research Paper

J- section (TCS)

Mini Project

BTCSCS308

Topic: Deep Modular Co-Attention Network for Visual Question Answering

- | | |
|------------------------|---------------------|
| ➤ Diya Dabhade | - 19100BTC SBS05586 |
| ➤ Mitendra Singh Tomar | -19100BTC SBS05598 |
| ➤ Nandani Paliwal | -19100BTC SBS05599 |
| ➤ Parikshit Bais | -19100BTC SBS05603 |

Submitted to: Mr. Sachin Chirgaiya

Abstract

Visual Question Answering (VQA) needs fully semantic information about questions and also needs region-based information about the image, and then only it can provide us an accurate answer. There are many approaches to VQA. Everyone used different methods and technology to give an accurate answer. But among that Co-attention technology has very successful response. And advance version of co-attention is Modular Co-attention (MCA) on which our research is based on. In our paper we propose ELMo (Embeddings From Language Models) model for word representation and embedding in Modular Co-attention Network (MCAN). MCAN consists the layers of Modular Co-attention (MCA) which make this technology more advance.

Keywords: Visual question answering, Co-attention, Modular co-attention, Modular co-attention network.

Introduction

There is majority interest gained in Visual Question Answering because deep learning has worked as a bridge between Visuals and natural language. VQA is a type of multimodal learning and VQA is very challenging. Because it needs different types of information and that too semantic after it can provide good accuracy.

Plenty of work has been done in this field and the recent addition is the attention method. The attention modal is a revolution because after it VQA gained many improvements. As on the ground of attention technology the co-attention technology has been created. In co-attention technology they create Self Attention (SA) unit and a Guided Attention (GA) unit, SA can process one data at a time and GA can process both textual and region based at a time. So, the simultaneous composition of SA and GA is co-attention and there is an efficient model that uses co-attention like BAN and DNC and these models have very high accuracy until MCAN has introduced.

MCAN is a technology based on the Modular Co-Attention. MCA is the combination of two attention layers that work simultaneously for increasing the efficiency and accuracy of the model. And MCAN is the Modular Co-Attention Network means MCAN is the network of MCA layers and the addition for increased accuracy we use ELMo, which is used to gather textual information more accurately.

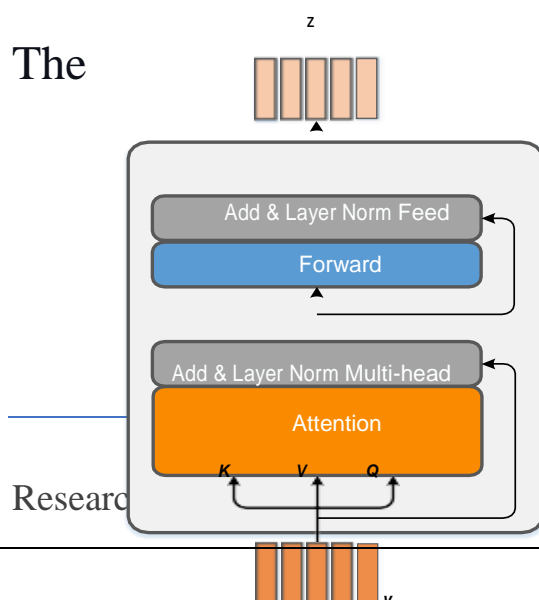
Review Of Literature

Co-attention –

It is a technique that uses two self attention units simultaneously one for the textual information means to question and another for visual information. And further research comes on this and then,

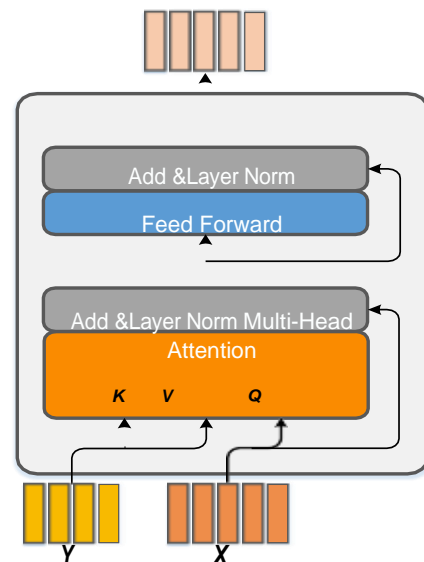
Modular Co-Attention Layer (MCA) –

The



MCA is the modular composition of two layers one is Self-Attention (SA) unit and the other is Guided-Attention (GA)

unit. The SA uses the algorithms to read only one type of information whether is region-based or textual but on the other hand, GA can use both information - region-based and textual simultaneously.



Self Attention Unit

Guided Attention Unit

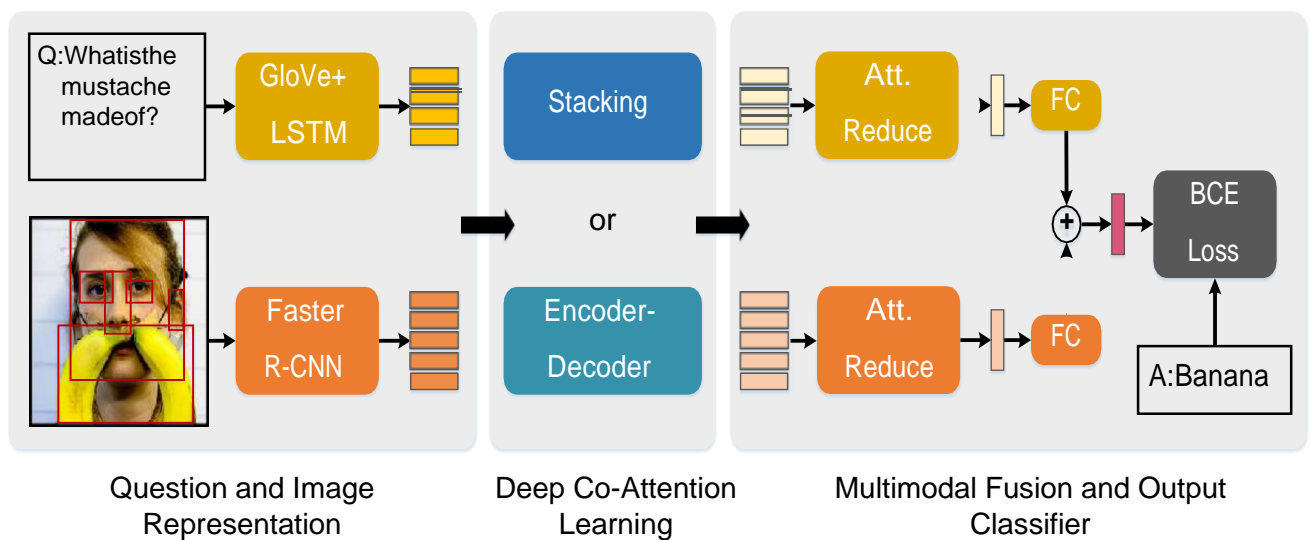
Modular Co-Attention Network (MCAN) -

This is a network of modular co-attention layers (MCA) and it is called MCAN. In MCAN we use either **stacking** or

encoder-decoder for the deep co-attention learning of the textual and visual information.

Both the techniques gives approximately the same result.

So, we use the SA and GA units in series form for deep co-attention learning of the image and question features. That's the gist of the whole process of MCAN.



“As in the question representation they use the Glove, we use ELMo. What is ELMo? and why we used it? are further”

Problem Statement

The model is not reading questions effectively. Its accuracy suffers because of that.



The model is very accurate in finding objects and on other hand, it is not able to understand the question effectively.

It uses long short-term memory (LSTM) in the process, which means the system clears old memory for new data, and for that, only a small chunk of data is sent further and because of that is information is large then might it miss the important details.

Ex.

What colour pant is the catcher wearing?

As we can see in the question there are too many nouns that mean too much information to collect. Like- colour, pants, catcher, and wearing.

Here MCAN only focuses on colour pant and missed the important keyword catcher.

In this type of situation modal can miss out on important information and in end, it will affect accuracy.

Solution

To overcome this problem will use ELMo instead of GolVe,

ELMo- Embedding from Language Models.

ELMo is a technique for reading text with accuracy with the context.

ELMo is an advanced technique that read the texts with context and converts them into vector. GolVe is not able to give that accurate answer because it was created for large datasets and unsupervised learning on the other hand ELMo is an example of supervised learning.

Ex.


Bat spread covid-19 virus.

Ravi is playing with a **bat**.

In this example, the word 'bat' is spelled the same but have a different meaning. Here, GolVe will create the same vector representation for both bats because that bat is the same in both sentences. But ELMo reads both bats differently and create a different vector for both bats and ELMo. ELMo uses

Bi-directional LSTM, not the LSTM. So, this will have a huge impact on the accuracy of the result.

Expected Result

Image	
Question	What is the player's jersey colour who is behind the ball?
MCAN with GoVe	Red
MCAN with ELMo	White

“The red player is bright and highlighted in the image and because of the perception of image with low question context info, the modal can give us wrong answer.”

Proposed Use

In airports and Metro cities, railway stations have a machine installed at the entrance called a baggage checker.

Machines work is to scan the baggage coming there with its x-ray capability and provide the visual to the machine operator and then operator manually check whether the baggage is safe or not.



If we installed this MCAN with ELMo in that machine then there will be double security one of modal and another is human supervision.

Conclusion

In this paper, we present the ELMo model for word representation and embedding in Modular Co-Attention Network (MCAN) for Visual Question Answering (VQA). MCAN consists of the layers of Modular Co-Attention (MCA) which is based on the Co-Attention technique. Each layer of MCA consists of single attention (SA) unit and guided attention (GA) unit which run the textual and region-based data simultaneously and after that intra-modal interaction is with encoder-decoder or staking. We proposed to replace GolVe with ELMo, because of its inaccuracy. After this all process we obtained an MCAN model that will increase the bar for upcoming VQA models.

References

- Lu et al. proposed a co-attention learning framework
- Zhou Yu, Jun Yu, Yuhao Cui, Dacheng Tao, and Qi Tian. Deep modular co-attention networks for visual question answering. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 6281–6290, 2019.
- Nam et al. proposed a multi-stage coattention learning model to refine the attentions based on memory of previous attentions
- Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pages 770–778, 2016.
- Zhou Yu, Jun Yu, Chenchao Xiang, Zhou Zhao, Qi Tian, and Dacheng Tao. Rethinking diversified and discriminative proposal generation for visual grounding. International Joint Conference on Artificial Intelligence (IJCAI), pages 1114– 1120, 2018.

- Matthew E. Peters†, Mark Neumann†, Mohit Iyyer†, Matt Gardner†, Christopher Clark*, Kenton Lee*, Luke Zettlemoyer†*. Are the founder of ELMo.