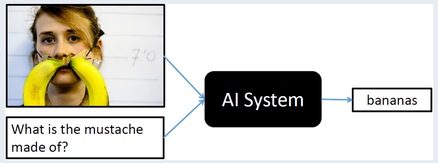
Image Question Answering With Deep Learning

Aahitagni Mukherjee and Shubham Agrawal
IIT Kanpur

Objectives

Given an image, and a free-form, open-ended natural language question, the task is to generate an accurate single word or multi-word answer. [1]



Introduction

The task of visual question answering is applicable to scenarios like visually impaired people or intelligence analysts trying to extract visual information. Open-ended questions require a vast set of AI capabilities to answer, like fine-grained recognition, object detection, activity recognition, knowledge-base reasoning, commonsense reasoning. The AI system needs to extract multi-modal knowledge and an effective component to find the relationships between the modalities. End-to-end trainable deep neural networks have been used for this problem. Our project proposes an efficient neural network architecture for this task.

Long Short Term Memory

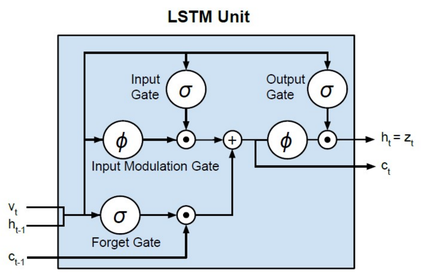


Figure 1: LSTM unit used by Malinowski et al.[2]

Recurrent neural networks are effective in representing sequential input. LSTM network is a special RNN which has been effective in natural language tasks.

CNN

Figure 2: Googlenet[2]

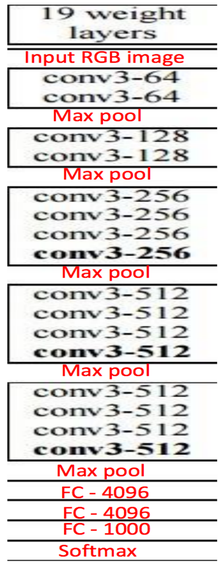


Figure 3: VGGNet[3]

Modifications

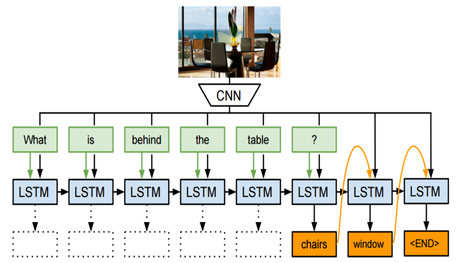


Figure 4: Model used by Malinowski et al.[2]

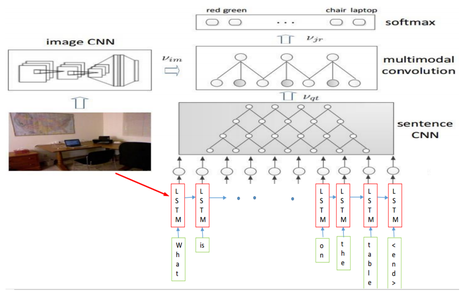
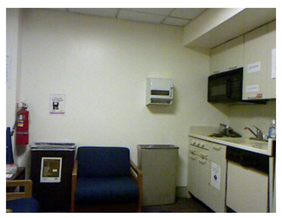
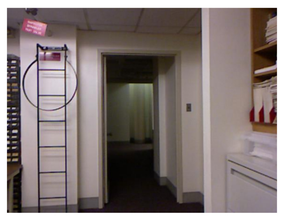


Figure 5: Our model, based on model by Lin Ma et al.[4]



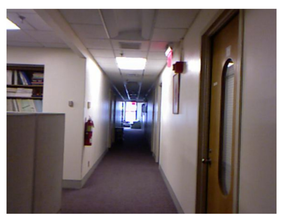
What is the object built on top of the counter right of the stove ?
Ground truth: sink
Model result: sink
Googlenet(8000 iterations) : lamp
Googlenet(110000 iterations): dish-washer

How many blue chairs are ?
Ground truth: 2
Model result:: 2
Googlenet(8000 iterations) : 2
Googlenet(110000 iterations): 2



How many red objects are visible ?
Ground truth: 5
Model result:: 2
Googlenet(8000 iterations) : 2
Googlenet(110000 iterations): 2

What is diagonally placed in front of the ladder?
Ground truth: table
Model result:: chair
Googlenet(8000 iterations) : chair
Googlenet(110000 iterations): chair



What are the objects close to the ceiling ?
Ground truth:exit_sign, fire_alarm, light
Model result:: light
Googlenet(8000 iterations) : 2
Googlenet(110000 iterations): spot_light

What is the colour of the door?
Ground truth: brown
Model result:: yellow
Googlenet(8000 iterations) : yellow
Googlenet(110000 iterations): yellow

Figure 6: Some examples from our experiment

Experiments

We evaluate our approach on the DAQUAR[2] dataset, which provides 12, 468 human question answer pairs on images of indoor scenes, and follow the same evaluation policy as in Malinowski et al.[2] using WUPS score. We have used Caffe-recurrent[5] and g2.2xlarge GPU instance of Amazon Web Services EC2 for our experiments.

Results

Results as Wu-Palmer Similarity scores:

#Iterations	Malinowski	Our model
1000	0.010701545	0.096313912
5000	0.019817677	0.131391200
8000	0.048949663	0.166270313

Table 1: Comparison with model by Malinowski et al.

Conclusion

- **LSTM** captures sequential nature of input.
- **Multimodal Convolution Layer** effectively models interrelationship of language and image.
- Passing the image representation to both LSTM and multimodal layer helps to model the interrelationship.

References

[1] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh.
Vqa: Visual question answering.
arXiv preprint arXiv:1505.00468, 2015.

[2] Mateusz Malinowski, Marcus Rohrbach, and Mario Fritz.
Ask your neurons: A neural-based approach to answering questions about images.
CoRR, abs/1505.01121, 2015.

[3] Karen Simonyan and Andrew Zisserman.
Very deep convolutional networks for large-scale image recognition.
arXiv preprint arXiv:1409.1556, 2014.

[4] Lin Ma, Zhengdong Lu, and Hang Li.
Learning to answer questions from image using convolutional neural network.
CoRR, abs/1506.00333, 2015.

[5] Yangqing Jia, Evan Shelhamer, Jeff Donahue, Sergey Karayev, Jonathan Long, Ross Girshick, Sergio Guadarrama, and Trevor Darrell.