Visual Question Answering
Using Attention based Models

Amlan Kar, Nishant Rai and Shivam Malhotra

Department of Computer Science and engineering

Abstract

The project deals with the problem of Visual Question Answering. We propose several models to tackle the problem consisting of models using Bag of Words, CNNs and LSTMs. We also explore the role of attention in improving the performance of the model. The training data used is the popular VQA dataset based on MS COCO. Our model uses word vectors compute the representation of questions. We compare the performance of our approaches with existing models and find that our model able to answer a few tough examples. We also discuss specific examples and propose improvements based on them. We also discuss possible improvements and future work.

Under the guidance of Dr. Vinay Namboodiri
Indian Institute of Technology, Kanpur

Contents

1 Introduction

We aim to build a deep learning system capable of answering open ended questions on real world images. We use the VQA dataset[3] released by Virginia-Tech for our experiments that contains images from the MS-COCO dataset annotated with open ended questions and top answers. We study the use of attention based models to improve the results for the same.

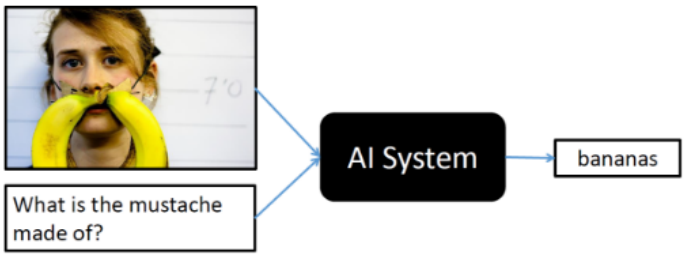


Fig.1: Problem Statement

2 Preliminaries

2.1 Convolutional Neural Network

A Convolutional Neural Network is an artificial neural network which work by sliding windows through it's input looking for local features. These have shown to work extremely well for Image recognition tasks.

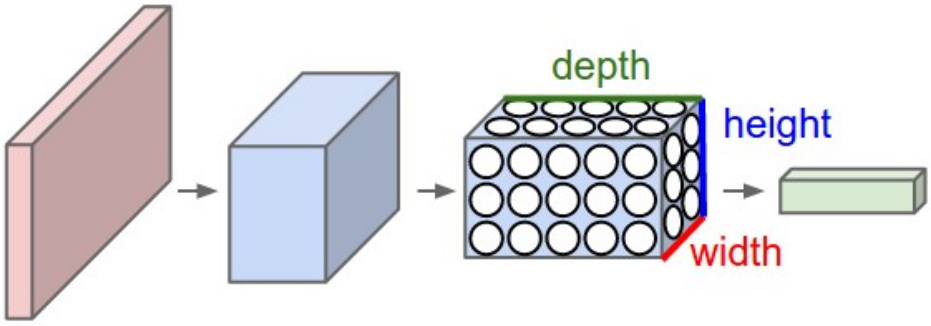


Fig.2: CNN Illustration [1]

2.2 Semantic Word Embeddings:

Recent advances such as word2vec, GloVe[7] and skip-thoughts [5] map words or sentences to high dimensional real valued vectors such that syntactic relations between the words are preserved. These have been shown to have strong semantic similarity properties as well.

[1] Image taken from www.cs231n.github.io

2.3 Recurrent Neural Network

A recurrent neural network (RNN) is a class of artificial neural network where connections between units form a directed cycle. These have been shown to work extremely well in modeling temporal data.

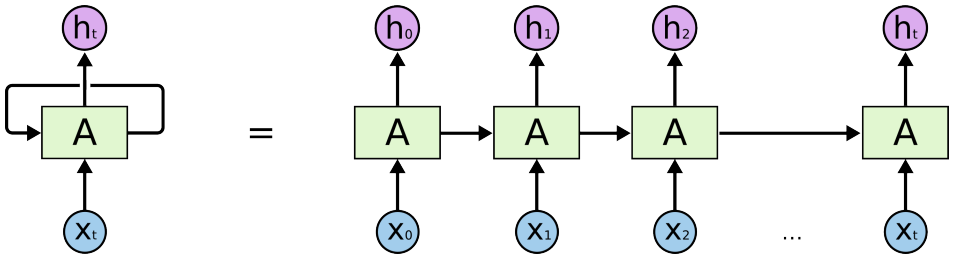


Fig.3: RNN Illustration [2]

3 Previous Work

There has been a recent spur of interest in the VQA task. Malinowski et al. [6] had released the first image Question Answering dataset (DAQUAR). They initially used a multi-word approach to the problem, but later shifted to a combination of Convolutional Neural Networks (CNN) and Long Short Term Memory (LSTM) Networks. Recent works [4], [3], [9] have explored the role of spatial attention in improving the performance of the trained models.

We roughly describe the following two models for the VQA task. They are inspired from [6]. The description is inspired from [2].

3.1 CNN with Bag of Words

The image is passed through the VGG ConvNet [8], and the activations before the softmax layer are extracted, giving us a 4096-dimensional vector representing the image. The question vector is obtained by simple averaging of the word vectors of all the words present in the question. The two vectors (image and question) are then concatenated and passed through a Multi-Layer Perceptron with two fully connected layers and 50% dropout for regularization. A softmax layer is attached at the end, and it gives us a probability distribution over the entire answer space.

3.2 CNN and LSTM Network

Note that the previous model ignores the order in which the words appear in the question, and thus there is an obvious loss of information when summing up the word vectors. To capture the sequential nature of language data, we model the questions using LSTMs. Every word in the question is first converted to its word embedding, and these embeddings are passed into the LSTM in a sequential fashion. The final output of the LSTM is used as the question embedding. This question embedding is concatenated with the 4096-dimensional image vector, and we then simply apply the same Multi-Layer Perceptron architecture that

[2] Image taken from www.colah.github.io

we used in the previous BOW model. The entire network is trained end-to-end, except the Convolutional Neural Network (i.e. the VGG ConvNet).

4 Incorporating Attention

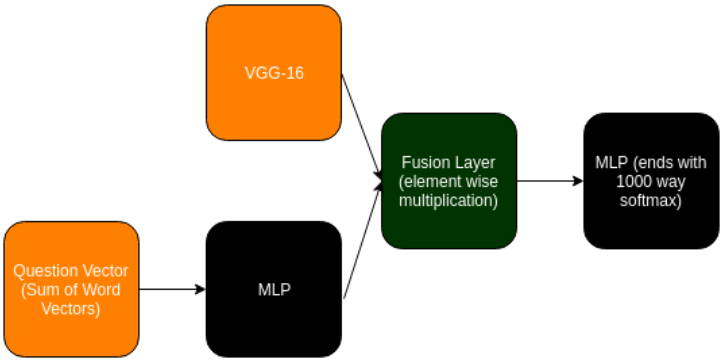
We discuss the following two ways to incorporate attention into our model.

- **Spatial Attention:** Spatial Attention methods are a class of methods that primary work by selecting a sub-set of feature vectors (or learning a probability vector) from one of the earlier Convolution layers of a feature extractor model. These have been shown to work well for Caption Generation and VQA.
- **Semantic Attention:** We implement and present semantic attention models. Here, we learn a question embedding and a transformation matrix into the image dimension and use this as a semantic attention weight vector on the semantic image features using Multi Layer Perceptron and RNNs.

5 Models Proposed

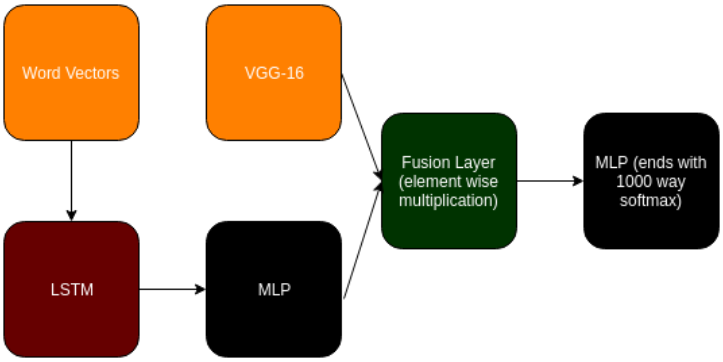
- **CNN + BOW + Att**

  3-Layer MLP to learn the matrix transfer embedding from the word vector space to the Image space. Trained with 50% dropout at each dense layer.



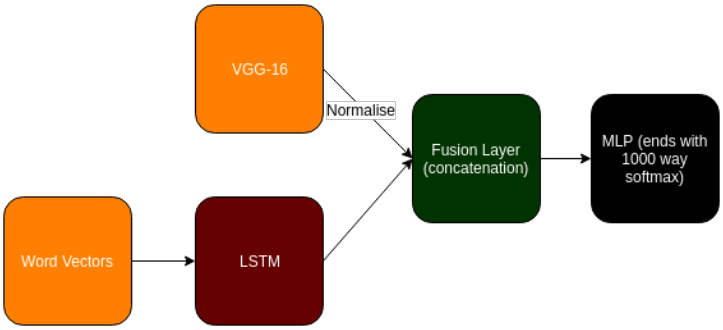
- **CNN + LSTM + Att**

  2-Layer LSTM used to create question embedding space from variable sized questions. 3-Layer MLP to learn the space embedding transfer. Trained with 50% dropout at each dense layer.



- **CNN + LSTM2** [1]

  Fusion layer only concatenates normalized CNN output with the 2-Layer LSTM question vector embedding. Again, trained with 50% dropout at each dense layer.



6 Datasets

We use the VQA dataset based on the popular MS COCO image dataset. It currently has 360K questions on 120K images. All the questions are human-generated, and were specifically designed to stump a 'smart robot'. This dataset was released along with the VQA challenge, which led to spur of interest in the VQA area.

7 Results

We discuss the results after testing the above mentioned models on the VQA dataset. The following section contain the results on both some specific examples and also the score of the model (As defined in the original VQA challenge).

7.1 Convergence and Training Time

All our models attained their optimal validation performance at roughly around 60 epochs of training. It was observed that the BOW models converged faster than the models using LSTMs. The time taken for an epoch (Around 200K images) is shown in the following tables,

Model	Training Time (/epoch)
CNN + BoW	97s/epoch
CNN + LSTM	141s/epoch
CNN + Bow + Att	127s/epoch
CNN + LSTM + Att	162s/epoch

Training times per epoch

7.2 Evaluation on VQA dataset

We us the same evaluation metric used for the VQA Challenge consists of evaluating an answer (processed) against 10 human answers (based on sampling from distribution). The accuracy of an answer is calculated according to:

$$\text{Acc}(ans) = \min\left\{\frac{\#\text{humans that said } ans}{3}, 1\right\}$$

7.3 Results on VQA dataset

The results of the proposed models are as follows,

Method	Accuracy
CNN+BoW	48.46
CNN+LSTM	51.63
ABC-CNN [4]	48.38
CNN+BowAtt	47.32
CNN+LSTMAtt	48.27
CNN+LSTM2	**55.17**

Accuracy scores on VQA-test-dev split

As we can see, CNN-LSTM2 outperforms all other models. The possible reasons for the same are discussed in later sections.

7.4 Specific Examples

We show certain examples and report our observations from the results. The examples also highlight the shortcomings of our models and also suggest ways to further improve it.

What are the women doing?

Proposed Answer	Confidence
Playing Wii	**39.27%**
Reading	2.27%
Cutting Cake	1.63%
Cooking	1.28%
Playing Game	1.21%

What is the sport being shown in the image?

Proposed Answer	Confidence
Skiing	**90.75%**
Snowboarding	7.66%
Frisbee	0.15%
Skateboarding	0.09%
Surfing	0.06%

What is the image showing?

Proposed Answer	Confidence
Snow	**23.56%**
Kite	17.78%
Mountains	2.11%
Snowboard	2.04%

As we can see that the above questions are answered correctly. The example shown below highlights an important shortcoming of the model which is that, in case of questions asking for color, we only focus on the dominating/Interesting objects. If we ask for color of specific objects, we require spatial attention along with semantic attention. Another possible solution is composing the image features where the composition is generated based on the question.

What color is the train?

Proposed Answer	Confidence
Yellow	**54.74%**
Red	21.46%
White	6.62%
Green	5.08%
Blue	4.34%

What colour is the background?

Proposed Answer	Confidence
Yellow	**50.17%**
Red	26.53%
White	5.8%
Green	5.41%
Blue	4.64%

8 Discussions

1. While the CNN+LSTMAtt model doesn't give competitive accuracies, it gets many 'tough' questions right which the CNN+LSTM2 model gets wrong. This suggests the possibility of using an ensemble.

2. The attention model failing might be due to the usage of the fused features along with the vanilla VGG and Question features. This redundancy and its subsequent effect on the extremely higher number of parameters in the model might be the reason for its poor test performance.

3. Attention doesn't seem to be very important to the Virginia-Tech VQA dataset with our preliminary results and results from the ABC-CNN[4] paper.

9 Future Work

1. Using Normalized CNN Features in all models [3]

2. Using a Knowledge base for real world reasoning

3. A mixture model of spatial and semantic attention

4. Perhaps using random 224x224 splits of an image to find the ConvNet features and averaging over results (as done by Krizhevsky et al. 2012) [9]

10 Bibliography

References

[1] Aaditya's visual question answering. `http://iamaaditya.github.io/2016/04/visual_question_answering_demo_notebook`.

[2] Avi singh: Visual question answering. `https://avisingh599.github.io/deeplearning/visual-qa/`.

[3] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. Vqa: Visual question answering. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2425–2433, 2015.

[4] Kan Chen, Jiang Wang, Liang-Chieh Chen, Haoyuan Gao, Wei Xu, and Ram Nevatia. Abc-cnn: An attention based convolutional neural network for visual question answering. *CoRR*, abs/1511.05960, 2015.

[5] Ryan Kiros, Yukun Zhu, Ruslan Salakhutdinov, Richard S Zemel, Antonio Torralba, Raquel Urtasun, and Sanja Fidler. Skip-thought vectors. *arXiv preprint arXiv:1506.06726*, 2015.

[6] Mateusz Malinowski, Marcus Rohrbach, and Mario Fritz. Ask your neurons: A neural-based approach to answering questions about images. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1–9, 2015.

[7] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119, 2013.

[8] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International Journal of Computer Vision*, 115(3):211–252, 2015.

[9] Qi Wu, Peng Wang, Chunhua Shen, Anton van den Hengel, and Anthony Dick. Ask me anything: Free-form visual question answering based on knowledge from external sources. *arXiv preprint arXiv:1511.06973*, 2015.