Visual Question Answering

Akash Chidananda Murthy
chidanandamurthy.a@husky.neu.edu
NUID: 001699128

Shashikiran Gadhar
gadhar.s@husky.neu.edu
NUID: 001256135

Veera Venkata Sasanka Uppu
uppu.v@husky.neu.edu
NUID: 001289684

Abstract

Visual question answering (VQA) is an exciting problem that combines natural language processing and image processing. The aim of the project is to answer a set of questions based on an image. Most research in visual question answering area has been fairly new and few limitations we saw across the published papers were: considering questions whose answers come from a closed vocabulary, questions with limited scope and the accuracy in extracting image data etc.Our goal is to extract more accurate and meaningful content from image data to better answer given set of questions.

1 Introduction

Recent advancements in image processing have made enormous progress in many computer vision tasks such as image classification, segmentation, and object detection with the help of deep learning. Advancement in Natural Language Processing have also made major leap in areas of text classification, context recognition etc. One of the products of these two advancements has lead to visual question answering (VQA), which has emerged as a prominent research problem. There are many potential applications for VQA such as aiding Visually impaired people and text based image retrieval.

Since 2014, there has been enormous progress in developing systems that are able to solve these tasks. VisualQA[1] is one such open source competition, the image dataset used is from COCO[2]. The answer set consists mainly of single words, which allow use to consider this as classification problem making the evalution easier to compute.

2 Related Work

Although VQA is a new problem, sophisticated algorithms have been proposed recently. Common approaches use Convolutional Neural Networks (CNN) such as VGGNet, ResNet, GoogLeNet that are pre-trained on popular image datasets (COCO) for generating image features. A wide variety of generating question features include word vectors, bag of words model, recurrent neural networks encoders (LSTM, GRU) and skip-taught vectors. Most common approach observed was to treat VQA as classification problem.

The framework to combine image features and question features has been widely experimented. One such method is combining image and question features with simple operations such as concatenation, element-wise multiplication/addition and feeding them to a MLP classifier. Alternatively using bi-linear pooling to combine question and image features have also been proposed. In [3], authors have proposed to use Bayesian Models to find the relation between image and question feature distributions. Using question features to compute spatial attention maps for visual features have also been proposed which made significant improvements in predicting the answer for the given image.

In [9], authors have used CNN to encode the image x and obtain a continuous vector representation of the image. The question q is encoded using an LSTM or a GRU network for which the input at

time step t is the word embedding for the t th question word qt , as well as the encoded image vector. The hidden vector obtained at the final time step is the question encoding. The answer is decoded in two different ways, either as a classification over different answers, or as a generation of the answer. Classification is performed by a fully connected layer followed by a softmax over possible answers.

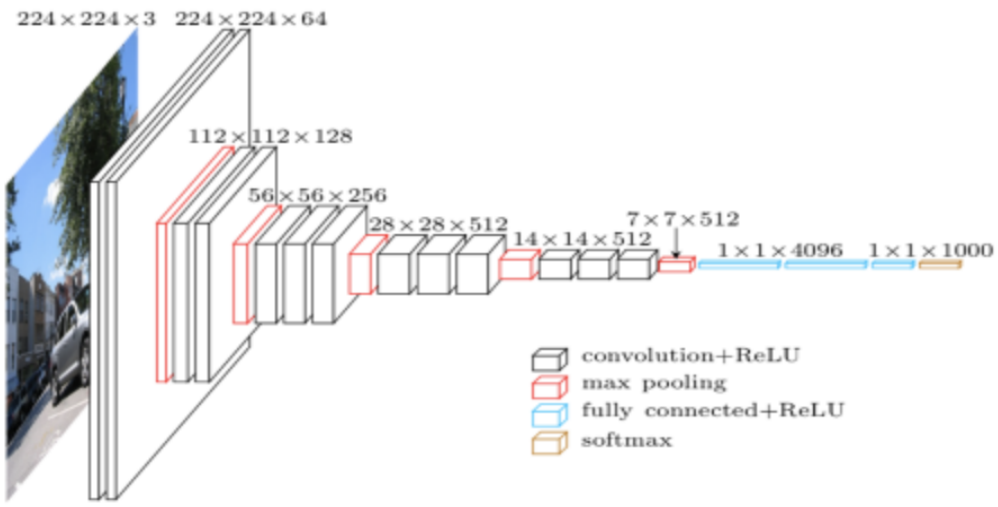
The authors in [10] argue that having a fixed set of parameters is not powerful enough for the VQA task. They take the architecture of VGGnet, remove its final softmax layer and add three more fully connected layers, the last of which is followed by a softmax over possible answers. The second of these fully connected layers does not have a fixed set of parameters. Instead, the parameters come from a GRU network. The dynamic parameter layer can alternatively be seen as multiplying the image representation and question representation together to get a joint representation, as opposed to combining them in a linear fashion.

3 Model

Our model has three major components: Image feature extraction, question feature extraction and 3-layered attention model.

3.1 Image Feature Extraction

We are using convolutional neural network (CNN) based model, namely 19-layered VGGNet to extract image features from the given image. VGGNet is used for object identification in an image. VGGNet is characterized by its simplicity, using only 33 convolutional layers stacked on top of each other in increasing depth. Reducing volume size is handled by max pooling. Two fully-connected layers, each with 4,096 nodes are then followed by a softmax classifier. First, we resize the given image into 224 x 224 pixels, and then pass it to VGGNet to extract image features.Unlike models in [8], we are extracting images feature from pool5 layer instead from last layer of VGGNet, by doing this we are able to retain the spatial features of the image. The output of pool5 layer of VGGNet is used as image features, v_I



5

3.2 Question Feature Extraction

We are first building a bag-of-words using all questions from the training dataset. Then we are embedding each question into a vector of size 20 using the bag-of-words. Then we are passing these question vectors to LSTM model that captures the semantic meaning of the text. The output of the LSTM model is used as question features, v_Q

3.3 3-Layered Attention Model

Instead of using entire image to answer given set of questions, we are focusing only on certain parts of image to extract answers. We are doing this using an 3-layered attention model. Attention model tries to identify parts of image that are related to answers of the given question. Each layer tries to reduce the noise in the image and output of last layer to fed to a softmax layer to predict the anwser.

We first feed the generated image features, V_I and question features,V_Q to a single layer neural network and then a softmax layer to generate attention map of various regions of the image:

$$A_1 = \tanh(W_{AI}V_I \oplus (W_{AQ}V_Q + b_A)) \tag{1}$$

$$P_1 = softmax(W_P A + b_P) \tag{2}$$

where,

$$V_I \in R^{d\times m}, d\ is\ the\ image\ representation\ dimension\ and\ m\ is\ the\ number\ of\ regions\ in\ the\ image. \tag{3}$$

$$V_Q\ is\ d\ dimensional\ vector. \tag{4}$$

$$W_{AI},\ W_{AQ}\ \in R^{n\times d},\ W_P\ \in R^{1\times n}\ and\ P \in R^m \tag{5}$$

As we can see P has m dimension same as number of regions in image feature vector. We add P to question feature vector.

$$V_I' = \sum_i p_i v_i \tag{6}$$

As, you can see that V' now has weights for different regions of the image. Now we combine V' and question feature to form new question vector, q_1

$$q_1 = V_Q + V_I' \tag{7}$$

Now q_1 becomes our new question vector, which has both question and visual information and is passed to second attention layer.

$$A_2 = \tanh(W_{AI}V_I \oplus (q_1 W_{AQ} + b_A)) \tag{8}$$

$$P_2 = softmax(W_P A_2 + b_P) \tag{9}$$

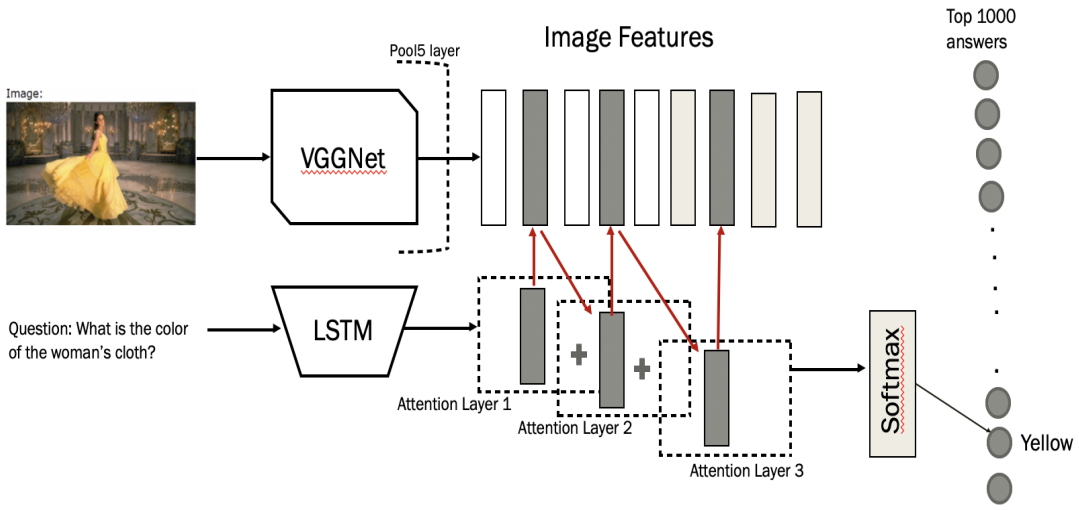
We will use P_2 and add it to question feature vector and pass it to third attention layer.

$$A_3 = \tanh(W_{AI}V_I \oplus (q_2 W_{AQ} + b_A)) \tag{10}$$

$$P_3 = softmax(W_P A_3 + b_P) \tag{11}$$

We will pass P_3 to softmax layer to predict the answer.

Below is the pictorial representation of our model:



4 Experiment

4.1 Datasets

For this proposed solution we are using VisualQA(VQA) dataset. VQA dataset is the largest dataset for this problem containing human annotated questions and answers on images from Microsoft COCO dataset. VQA is created through human labeling. For each image, there are three questions and for each question, there are ten answers labeled by human annotators. The correct answer is picked by the majority of answers from the the ten human annotators.

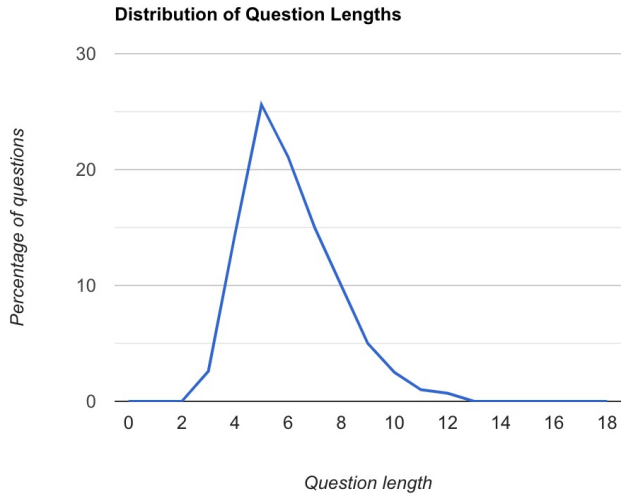
There are 443,757 training questions and 214,354 validation questions in the data set. The VQA image dataset also contains 50,000 abstract cartoon images, but we have only considered real images. We used the top 1000 most frequent answer as possible outputs as this set of answers covers more than 80% of all answers. We tested the performance of the proposed model on the validation set. The dataset contains various types of questions such as:

- Object recognition - What is in the image?
- Object detection - Are there any cats in the image?
- Attribute classification - What color is the car?
- Scene classification - Is it raining?
- Counting - How many balls are there in the image?

Below are some of the statistics observed on the dataset. As we can see the top answers are Yes and No since there are a majority of polar questions. And we can also see the distribution of question lengths where majority of the questions are shorter than 10 words.

Question starts with	Percentage %
What	41.2
Is	26.1
How	11.3
Are	7.57

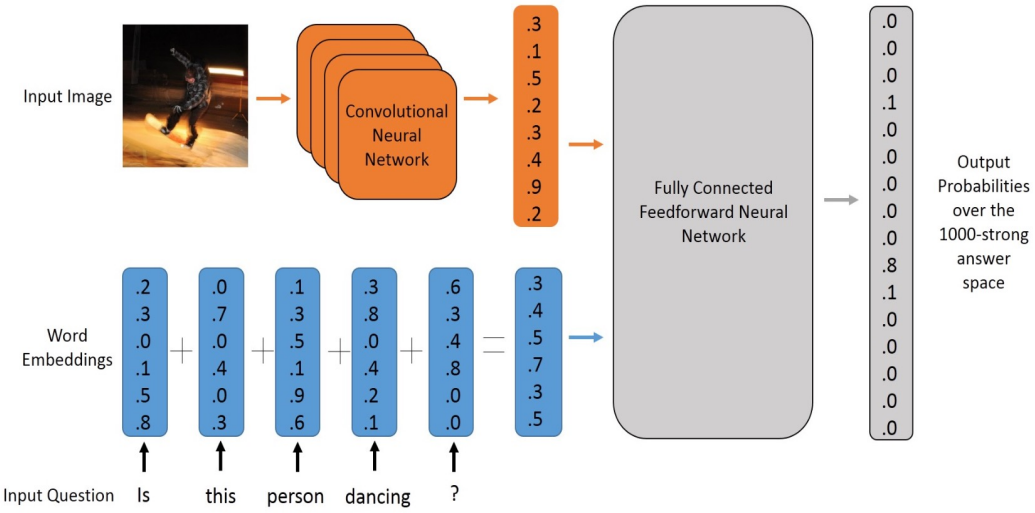
Top 10 answers
Yes
No
1
2
White
3
Blue
Red
Black
0



4.2 Baseline Model

For baselines, one model we are using is to select the most popular answer and another one is to select the most popular answer per question category.

A better baseline we are considering is a LSTM CNN + multilayer perceptron (MLP) model. We simply use naive bag-of-words as the text feature, which is then transformed to word feature via a LSTM word embedding. For this model word vectors of length 300 is considered. The image features are obtained from the final layer of the VGGnet. The image features and word features are then concatenated and fed to multi layer perceptron to precict the answer among top 1000 answers. Thus the data points which constitute only to the Top-1000 answers have been considered. Below is the pipeline of our baseline model.

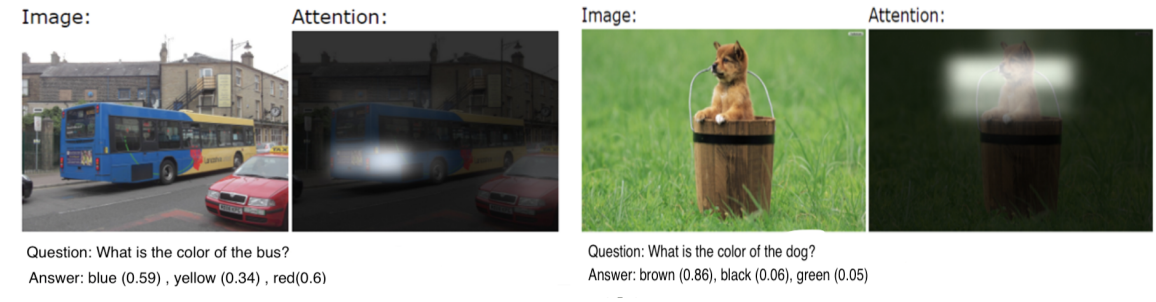


4.3 Results

To calculate accuracy we chose the answer with highest probability obtained from the softmax layer and then check with actual accuracy. Following results have been observed. First two baselines are very basic and as we can see the MLP baseline model performs better. The proposed attention model outperforms the MLP model which can easily be explained as the MLP model doesn't take spatial positioning of image into account and also the MLP model just concatenates the image features and word embeddings. Whereas in case of the attention model it takes spatial space of image into account and tries to focus on important parts while answering the question rather than considering the whole image.

Model	Accuracy %
Always selecting the popular answer (Yes)	21.12
Picking most popular answer per question Type	32.34
MLP Model	44.78
Attention Model	56.04

Below are few interesting example results



References

[1] http://visualqa.org/

[2] http://cocodataset.org/

[3] K. Kae and C. Kanan, Answer-type prediction for visual question answering," in The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016.

[4] S. Antol, A. Agrawal, J. Lu, M. Mitchell, D. Batra, C. L. Zitnick, and D. Parikh, VQA: Visual question answering," in The IEEE International Conference on Computer Vision (ICCV), 2015.

[5] H. Gao, J. Mao, J. Zhou, Z. Huang, L. Wang, and W. Xu, Are you talking to a machine? Dataset and methods for multilingual image question answering," in Advances in Neural Information Processing Systems (NIPS), 2015.

[6] H. Xu and K. Saenko, Ask, attend and answer: Exploring question-guided spatial attention for visual question answering," in European Conference on Computer Vision (ECCV), 2016.

[7] Z. Yang, X. He, J. Gao, L. Deng, and A. J. Smola, Stacked attention networks for image question answering," in The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016.

[8] M. Ren, R. Kiros, and R. Zemel. Exploring models and data for image question answering. arXiv preprint arXiv:1505.02074, 2015.

[9] Mateusz Malinowski, Marcus Rohrbach, and Mario Fritz. 2016. Ask your neurons: A deep learning approach to visual question answering. arXiv preprint arXiv:1605.02697

[10] Hyeonwoo Noh, Paul Hongsuck Seo, and Bohyung Han. 2016. Image question answering using convolutional neural network with dynamic parameter prediction.

[11] Mengye Ren, Ryan Kiros, and Richard Zemel. 2015. Exploring models and data for image question answering.

[12] Michael Schmitz, Robert Bart, Stephen Soderland, Oren Etzioni, et al. 2012. Open language learning for information extraction.

[13] Minjoon Seo, Aniruddha Kembhavi, Ali Farhadi, and Hannaneh Hajishirzi. 2016. Bidirectional attention flow for machine comprehension. arXiv preprint arXiv:1611.01603 .