Visual Question Answering Model Implement and Comparision

Xiatong Li
Department of Electrical Engineering
University of California, Santa Cruz
Santa Cruz, CA 95064
xli239@ucsc.edu

Yuchen Yuan
Department of Computer Engineering
University of California, Santa Cruz
Santa Cruz, CA 95064
yyuan31@ucsc.edu

Abstract

Our project was a question about visual question answering. Based on the VQA dataset, we just started by building the Multi-Layer Layer Perceptron (MLP) model as our baseline model. We use pre-trained Glove embedding and VGG extract image features with cross-entropy loss. The basic model can achieve relatively good performance. In order to further improve the accuracy, we have built a new Multi-model Compact Bilinear (MCB) model and use the feature outer product to express the mutual information of multiple model vectors. Finally, we have succeeded in improving accuracy by adding attention and data augmentation mechanism. We also tried to compared the performance of the model.

1 Introduction

With the extensive application of deep learning technology in Computer Vision (CV) and Natural Language Processing (NLP), deep learning powerful feature learning ability has greatly promoted the research in CV and NLP fields. In Convolutional Neural Networks (CNN), CNN can learn end-to-end image features without relying on hand-designed feature. CNN's powerful feature extraction capability on images makes it possible to extract and compress image information more completely, making great progress in the research of many CV tasks such as image classification [1, 2], object detection [3, 4], and activity recognition [5–7]. The RNN model also shows its strength in the field of NLP, especially in speech recognition, machine translation, language model and text generation [8]. By training a sufficient number of labeled datasets, deep neural networks can reach the human ability to handle images and texts and can even surpass human levels.

Figure 1: VQA Example

Visual Question Answering (VQA) is a recent problem in computer vision and natural language processing that has garnered a large amount of interest from the deep learning, computer vision, and natural language processing communities [9]. Despite the tremendous success of CNN, many of the

31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA.

CV issues do not adequately understand the image as a whole, and the usual CV tasks do not require a complete understanding of the image. However, in real life, the human visual system can have an overall cognition of the spatial distribution of the object and can infer the relationship and connection between the spatial properties of different objects. Therefore, some scholars have put forward the visual question answering (VQA) problem, which is A system takes as input an image and a free-form, open-ended, natural-language question about the image and produces a natural-language answer as the output. In simple terms, VQA answers the questions of given pictures. In this final project, we hope to try some new combinations of model and compare with existing approaches on the VQA problem. A common example as shown in Fig. 1 [10]. If we have more time, we want to be able to modify the training structure and model or maybe it can improve the test accuracy.

The remainder of this paper is organized as follows. In Section 1, we provide a literature review for the field of VQA: Visual Question Answering and the motivation of this projecgt. In Section 2, we discuss some related work on VQA. The baselines MLP model and MCB model, VQA dataset we use for this paper is analyzed in Section 3. Experimental results are also summarized in Section4, where we can see the effectiveness of adapting attention model to the task of VQA compared with traditional structures. Finally, we make a conclusion for our work in Section 5 and provide ideas for future work.

2 Related Work

The baseline algorithms of VQA are mostly combination of existing computer vision and NLP deep structures. CNN features from GoogLeNet and Bag-Of-Word representation of questions are extracted. Then they are concatenated together to form the input data. The features are fed into multi-class logistic regression classifier to generate the final answer. [6] is another example whereMultilayer Perceptron (MLP) is used as classifier to choose the correct question-image-answer based on concatenated ResNet-101 features [9] and Bag-Of-Word features of both questions and answers. These baseline models have great performance but concatenating multi-model features does not consider the correlation of these features thus could certainly be improved.

There is a baseline model called iBOWING. They use the layer output of the pre-trained GoogLeNet image classification model to extract image features. The word embedding of each word in the question is regarded as a text feature, so the text feature is a simple bag-of-word. Connect image and text features, and use softmax regression to classify answers. The results show that the performance of this model on the VQA data set is comparable to several RNN methods.

3 Approach

3.1 MLP Model

MLP is the first model we used to concatenate pre-trained question feature, image feature, and answer-feature. The architecture of MLP is shown as following: In this network, input layer has 3 inputs that 2048-dim question vector, 2048-dim answer vector and 2048-dim image vector, output is score of each answer, which used to get the cross-entropy loss between ground-truth answer and prediction.

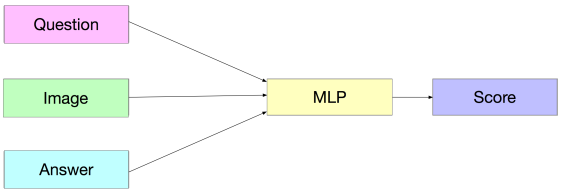


Figure 2: MLP model

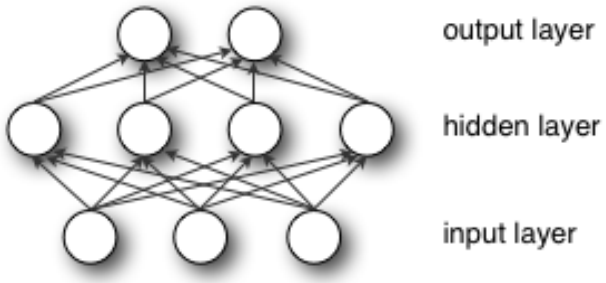


Figure 3: MLP Architecture

W1 is weight matrix between input layer and hidden layer, W2 is weight matrix between hidden layer and output. Following (1) denotes result for input layer, (2) denotes output from hidden layer, s in(3) is the final output of MLP which is also the score for each question.

$$l_1 = W_1 \times x_{iqa} + b_1 \tag{1}$$

$$h_1 = max(0, z_1) \tag{2}$$

$$s = W_2 h_1 + b_2 \tag{3}$$

For the whole MLP model, For each question, there are 3 incorrect answers and 1 correct answer, each of which is mapped to a score after going through the MLP. These four scores are then normalized using Softmax, and compared against ground-truth (boolean) labels to calculate cross-entropy loss.

3.2 MCB model

Compact Bilinear The outer product of multiple model vectors can express the mutual information of multiple model vectors. Because the traditional vector combination method is for a single element, the vector outer product is a multiplication operation for all elements between two vectors.

Bilinear takes into account the interaction of each element between the two features, so the effect is better Bilinear pooling thus gives a linear classifier the discriminative power of a second order kernel-machine, which may help explain the strong empiri- cal performance

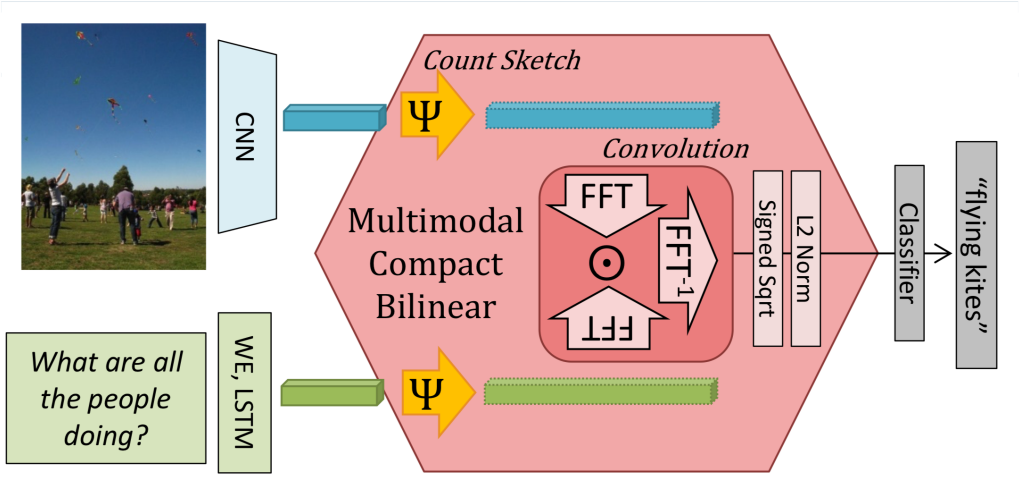


Figure 4: MCB architecture

bilinear The fusion method here says that the straightforward point is actually the outer product of the vector, and the proper noun is Compact bilinear. For the outer product, it is easy to generate a higher-dimensional feature representation, such as the picture feature and the text feature respectively 500, 500 Then, the characteristic dimension obtained after the outer product is 250,000. Such a high dimension is unrealistic for subsequent classification and other steps. The reasons are as follows:

If the standard one-vs-rest linear classifier (k class) is used and the feature dimension is d, then the parameter of the classification model is kd. Specifically, if k=1000 is far less than 250,000, it will bring more than just the parameter increase. Will lead to the consequences of disasters such as fitting and other dimensional problems; Higher-dimensional features in today's big data environment, the use of storage is large, you can simply calculate, a sample to get the feature dimension 250000, if there are one million and such a sample (double), you need more than 200G of storage space; In addition, in the further use of this feature for feature splicing, pyramid matching and other features requiring splicing operations, will further increase the storage space occupied; It is difficult to classify classifications. Such high-dimensional features can lead to overfitting and other dimension disasters.

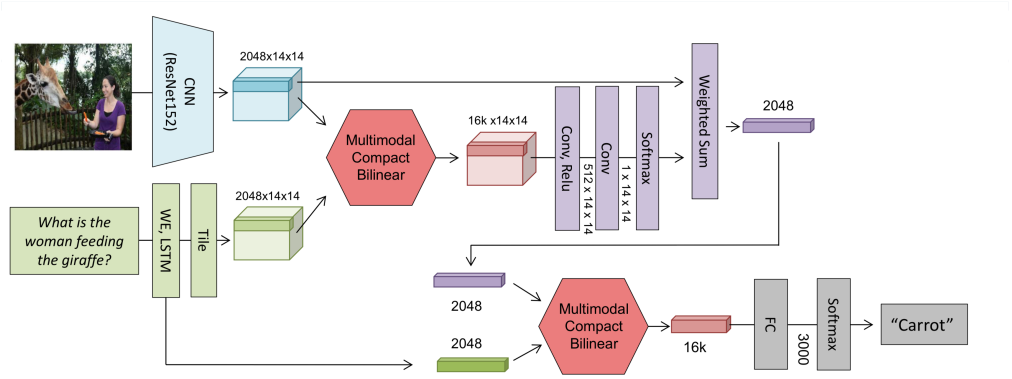


Figure 5: MCB Attention Architecture

Dimension reduction IDEA The idea of the TS method lies in the stochastic approximation: constructing two hash tables h, s. The uniform distribution used by h, the value of the position index of the vector 1 to d, and 1 and -1 in s, each time a random number from h and s, such as the kth time, you need to get through the hash function h The t value of h(t)=k, then the value of the k-th position of the approximated vector is the multiplication of the element of the original vector position (t, h(t)=k) and s(t), and the sum is summed. The value obtained. The

After the above approximation, good low-dimensional features can be obtained. Then, using the characteristics of the Fourier transform, the inner-boundary product in the time domain is transformed into the inner product in the frequency domain, and the computational complexity is reduced.

To compensate for the fact of fixing the parameters s_q, s_v and h, they must set a very high to dimension (typically 16,000).

The goal of a concentration-based approach is to focus the algorithm on the most relevant part of the input. For example, if the question is "What color is the ball?" then you need to focus more on the image area that the ball contains. Similarly, in the question, the words "color" and "ball" need to be concentrated because they are more informative than other words.The most common choice in VQA is the use of spatial attention to generate features for a particular region and thus to train convolutional neural networks.

4 Performance and Analysis

4.1 Dataset

The image in the VQA dataset is mainly composed of two parts: a realistic image and an abstract cartoon image. There are 123,287 training images and 81,434 test images in VQA-real, mainly from the MS-COCO data set. Unlike some previous datasets, VQA-real contains a binary problem (i.e, yes/no). This data set can be multi-selected to provide 17 additional false candidate answers for each question. In summary, VQA-real contains 614163 questions, and each question contains serval answers from different followers. **Here is brief review of question-data set:**

’image-id’: 12091, ’question’: ’How many pillows are on the chair and sofa?’, ’multiple-choices’: ["he’s happy", ’jake’, ’1’, ’6’, ’blue’, ’4’, ’2’, ’in middle of logs’, ’white’, ’fire in fireplace’, ’5’, ’8’, ’no’, ’yes’, ’january’, ’3’, ’yellow’, ’red’], ’question-id’: 120912

’image-id’: 12091, ’question’: ’Where is the woman?’, ’multiple-choices’: [’c’, ’blue’, ’1’, ’begging for food’, ’living room’, ’throwing it’, ’facing different ways’, ’3’, ’no’, ’red’, ’at zoo’, ’bench’, ’4’, ’2’, ’white’, ’on couch’, ’yellow’, ’yes’], ’question-id’: 120910

’image-id’: 12091, ’question’: ’What is the cat doing?’, ’multiple-choices’: [’sleeping’, ’dance’, ’yes’, ’playing’, ’white’, ’1’, ’4’, ’red’, ’blue’, ’pugs’, ’no’, ’formal’, ’no cat’, ’3’, ’2’, ’loafers’, ’stretching’, ’yellow’], ’question-id’: 120911

4.2 Experiment and Result

First we trained this two models respectively for 100 epochs. In the training set, MLP model has a beteer performance when approaching 100 epochs, however, the MCB model grows too slow. But the training is time-consuming, we believe after about 500 epochs, MCB would have the same accuracy as MLP on training set.

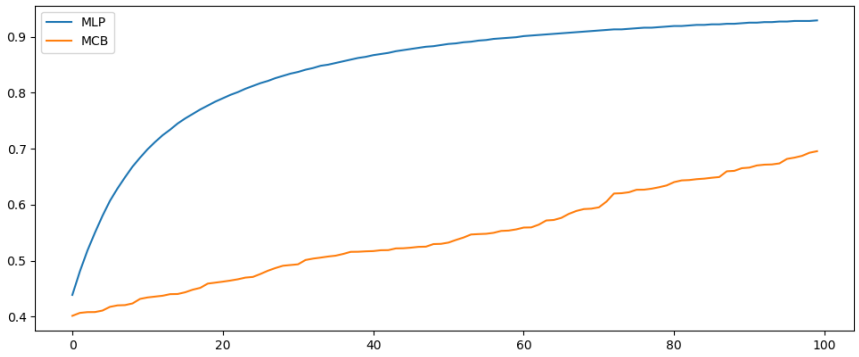
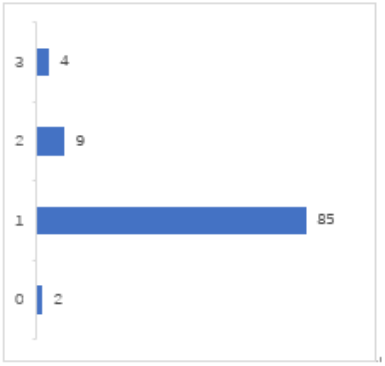


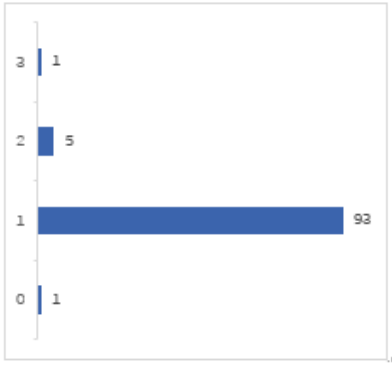
Figure 6: Train accuracy

Here is some of our result on Test sets. We pick 2 typical type of question, one is yes/no question, one is multichoice question. We found that on those yes/no questions, these two models doesn’t have much difference, but on multichocie questions, from the result we could see the difference.

How many wooden chairs are visible in this picture?



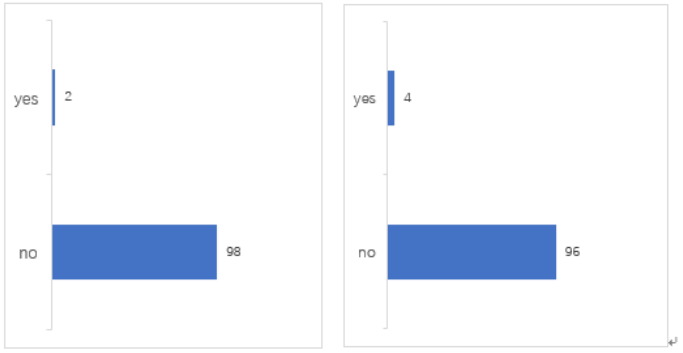
MLP



MCB

In this multichoice question, MCB has more higher probability to answer the correct answer.

Are there any people at the train boarding station?



MCB MLP

On this yes/no question, these two models doesn't have big difference.

Table 1: Over all accuracy on test-data set

Model	Yes/No	Multichoice	Overall
MLP	95.54%	48.37%	69.95%
MCB	96.32%	56.28%	75.20%

5 Conclusion and Future Work

5.1 Conclusion

In this project, we mainly focus on visual question answering where text-based questions are generated about an given image, and the goal is to give correnct answer. For baseline model, we first implemented a basic MLP model and further tried to use the MCB model to improve the accuracy. At the same time, we tried using mechanism of attention and data augmentation to enhance performance. Experiments conducted on VQA dataset have shown the effectiveness of our models.

5.2 Future Work

As for our future work, we can explore from these few aspects, we only add attention and data enhancements at the image level, perhaps doing the same work in the problem section, or applying

the attention mechanism to both modules simultaneously. What's more, we would try more models for concatenate features and more methods to extract features.
Visual Genome The difference with the ordinary Resnet is that here they pay attention to the object directly, not to the attention of each block in the picture. From an intuitive understanding, such attention is more explanatory.

References

[1] Meng, C, Y Wang, and S Zhang, "Image-Question-Linguistic Co-Attention for Visual Question Answering" *Stanford University*, https://web.stanford.edu/class/cs224n/reports/2748290.pdf. Accessed 23 Mar. 2018.

[2] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[3] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[4] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," in *Advances in Neural Information Processing Systems (NIPS)*, 2015.

[5] J. Donahue, L. A. Hendricks, S. Guadarrama, M. Rohrbach, S. Venugopalan, K. Saenko, and T. Darrell, "Long-term recurrent convolutional networks for visual recognition and description," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.

[6] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, and L. Fei-Fei, "Largescale video classification with convolutional neural networks," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014.

[7] K. Simonyan and A. Zisserman, "Two-stream convolutional networks for action recognition in videos," in *Advances in Neural Information Processing Systems (NIPS)*, 2014.

[8] Zhang, Biao, Deyi Xiong, and Jinsong Su. "Recurrent neural machine translation." *arXiv preprint arXiv:1607.08725* (2016).

[9] Kafle, Kushal, and Christopher Kanan. "Visual question answering: Datasets, algorithms, and future challenges." in *Computer Vision and Image Understanding 163* (2017): 3-20.

[10] Goyal, Yash, et al. "Making the V in VQA matter: Elevating the role of image understanding in Visual Question Answering." in *CVPR*. Vol. 1. No. 6. 2017.

[11] Ren, Mengye, Ryan Kiros, and Richard Zemel. "Exploring models and data for image question answering." *Advances in neural information processing systems*. 2015.

[12] M. Ren, R. Kiros, and R. Zemel, "Exploring models and data for image question answering," in *Advances in Neural Information Processing Systems (NIPS)*, 2015.

[13] 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.

[14] 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.

[15] Y. Zhu, O. Groth, M. Bernstein, and L. Fei-Fei, "Visual7w: Grounded question answering in images," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[16] R. Krishna, Y. Zhu, O. Groth, J. Johnson, K. Hata, J. Kravitz, S. Chen, Y. Kalantidis, L.-J. Li, D. A. Shamma, et al., "Visual genome: Connecting language and vision using crowdsourced dense image annotations," in *International Journal of Computer Vision*, vol. 123, no. 1, pp. 32–73, 2017.

[17] Y. Zhu, O. Groth, M. Bernstein, and L. Fei-Fei, "Visual7W: Grounded Question Answering in Images," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2016.