Improving Visual Question Answering Using Semantic Analysis and Active Learning

Abstract

In this work, we aim to train a model for the task of visual question answering, using only a small number of labeled data. In order to do so, we used the ideas influenced by the active learning. We defined an oracle to provide a label for the question that is asked about an image. This oracle is an image captioning network, that given an image as its input, generates a sentence that describes the objects that are visible in that image. We use a semantic similarity calculator, in order to connect the result of the image captioning model and interpret that to become a potential label for the visual question answering task. By using this structure and defining a new loss function, we became able to train a visual question answering model using a small number of labeled data. We conclude that by using this idea, we can train a model that has a generalization that is in the class of the classical models that use a great amount of data.

1 Introduction

Visual question answering aims to answer questions about about a given piece of visual content such as an image, video, or info-graphic. It serves various applications such as helping blind and visually-impaired users and, integrating with image retrieval systems, to name a few. Although there have been many research on this kind of task, since it needs data sets that containing image-question-answer pairs; lack of labelled data remains an open problem for this kind of task. In this project, we wanted to answer this question that how much this task is dependent to the size of available labeled dataset. We aimed to overcome this issue by providing a solution which is basically inspired by techniques of active learning, which is a method that currently receives a lot of attention.

The model that we propose works by using an image captioning network, that helps us understand the features of objects in a picture. This network is used as an oracle for a visual question answering (VQA) model that is the core of our structure. In order to construct a reasonable relation between these models, we design a new loss function that focuses on the maximum similarity that we can find between the word and the generated sentences based on a given image. With this improvement, we show that the proposed model can produce answers that have loss is in the same standard of classical approaches.

In the rest of the report, we give an detailed explanation of our approach and the experimental results that we get. Section 2 is about the related works about the task of visual question answering and active learning. In section 3, we give a detailed information about how to construct the models and the challenges that exist in this way. In section 5 we provide the results that we obtained through our experiments. In section 5 we conclude our project and provide the contribution that each member has in the project.

2 Related Works

Visual Question Answering Malinowski *et al.* [2] introduced a neural architecture for answering natural language questions about images that contrasts with prior efforts based on semantic parsing and outperforms prior work by doubling performance on this challenging task. We implemented the same VQA model using [3]. More details of the structure of this network is provided in section 3.1.

Image Captioning K. Xu *et al.* [1] proposed an attention based model that automatically learns to describe the content of images that attempt to incorporate a form of attention with two variants: a "hard" attention mechanism and a "soft" attention mechanism. This network was taken to use as the image captioning network. There are more details of this network in the section 2.2.

Active Learning Konyushkova, *et al.* [7] and Yoo *et al.* [8] is an iterative supervised learning solution to the situations in which unlabeled data is abundant, but manually labeling is expensive. In such a scenario, learning algorithms can actively query the user/teacher for labels. It is done by automatically deciding which instances an annotator should label to train an algorithm where a model asks human to annotate data that it perceived as uncertain as efficiently as possible. This idea had a great influence in our work. In this work we are not going use the classical approach of active learning, instead, we use another trained model to work as an oracle. This action of labelling is not limited to the data that the VQA model is not certain about, but we are going to label every single data that we get as input, since we are assuming that our data is not labelled. In this setting, we can not be sure about the result of the oracle. Actually, the oracle is giving us a sentence that we may assume our target is in it. There is a challenge here to find to correct answer in this caption. We provide more details of this task in section 3.3.

3 Our Method

The structure used in this project consists of three general parts. First part is the VQA model which we want to train. The inputs to this model is a pair (X, Y) such that X is an image and Y is a question about a specific detail of an object in the picture X. We want to train this model somehow that its output Z become a word that corresponds to the answer of question Y. Second part is the image captioning model, which works as an oracle for the VQA model. This model is trained independently and its weights are not going to change during the training of VQA model. Its input is an image X and its output is a sentence S that has to explain about the objects and their features in X. Third part of the structure, is used to combine the outputs of these two models to make second model play as an oracle for the first model. This part is a trained model for finding semantic similarity of two sentences. Combined together, these three models make the structure of our project.

In the rest of this section, we will provide the details of each model separately.

3.1 VQA Model

The VQA network consists of a pre-trained ResNet-18 model that receives the image X as part as its input, and a GRU-based LSTM recurrent neural network that gets the question Y as its input. As shown in Figure 2, these two parts are concatenated together and make the input to the third part, which is a fully connected neural network that produces an output in $\mathbb{R}^{300}$. In other words, this part of the model has the responsibility to get a picture and a question as inputs and generate a single word answer for the question.

Dataset for VQA: We used DAQUAR dataset [4] to train and test this network. It contains 6794 training and 5674 test question-answer pairs, based on images from the NYU-Depth V2 Dataset. That means about 9 pairs per image on average.

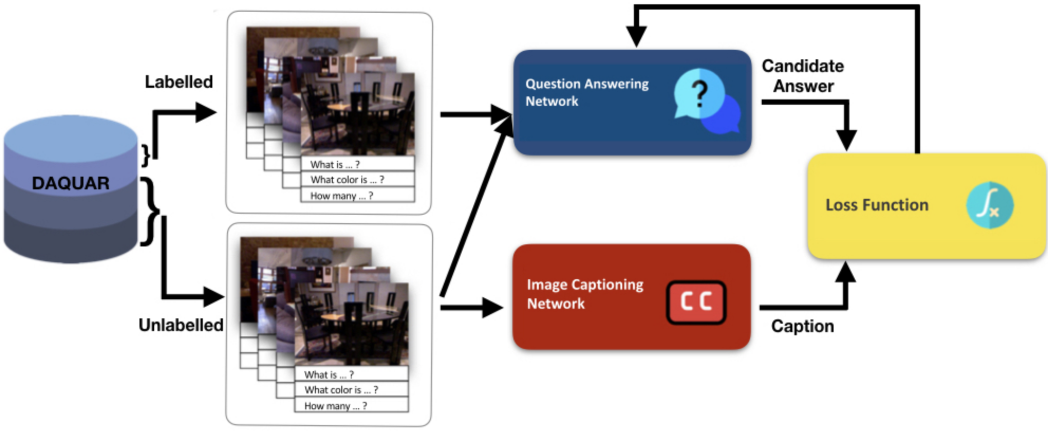

Figure 1: High-level schematic of the structure of the problem.

3.2 Image Captioning Model

For the image captioning network [1], we trained a deep neural network independently on the MS-COCO dataset. This network receives the image X as input and generates a description S for it. This network consists of a convolutional layer of InceptionV3, giving us a three dimensional vector of shape (8, 8, 2048) used to extract features. Therefore, this vector squashes to a shape of (64, 2048). This vector is then passed through a convolutional neural network encoder (which consists of a single fully connected layer). Also, a recurrent neural network based on GRU is used that attends over the image to predict the next word.

Dataset for Image Captioning: We used MS-COCO dataset [6] to train this network. This dataset contains 91 common object categories with 82 of them having more than 5,000 labeled instances. In total the dataset has 2,500,000 labeled instances in 328,000 images.

3.3 Semantic Similarity Calculator

For calculating semantic similarity model, we use GloVe [5], which is an unsupervised learning algorithm for obtaining vector representations for words. We used a pre-trained word vector library which is trained by this algorithm and maps 400000 words to $\mathbb{R}^{300}$. This algorithm has this essential property that encodes similar words to some vectors that have small Euclidean distance. We use this model for two purposes. First, a general description of an image that we get from an image using image captioning network, consists of details of many objects. However, the question Y only asks about a specific detail of one of the objects in the image. One may ask how can we be sure that the output of VQA model is about that specific object. To overcome this challenge, we slide a window of fixed size through the entire description. Then we calculate the similarity of this windowed part of the description with the question, using the semantic similarity model. The window that has maximum similarity with the question, is the part of the description that we check to calculate our loss function. The second use of the semantic similarity model is in the loss function. Our loss function basically calculates the most similar word in the windowed part of the description and the output and uses the most similar word as the label for that exact pair of (X, Y). By using these ideas, we become able to use the image captioning model and interpret its output to work as an oracle for the VQA model.

3.4 Loss Function

For the loss function, we used a cross-entropy measure on the maximum similarity that we can find between the words of the optimum window and the output vector presented by the VQA model. Formulation of this function comes as follows. In this setting, we use some random words from our dictionary of words that we obtain by GloVe algorithm and calculate the cross-entropy measure with respect to those words, and the label that is provided by the oracle.

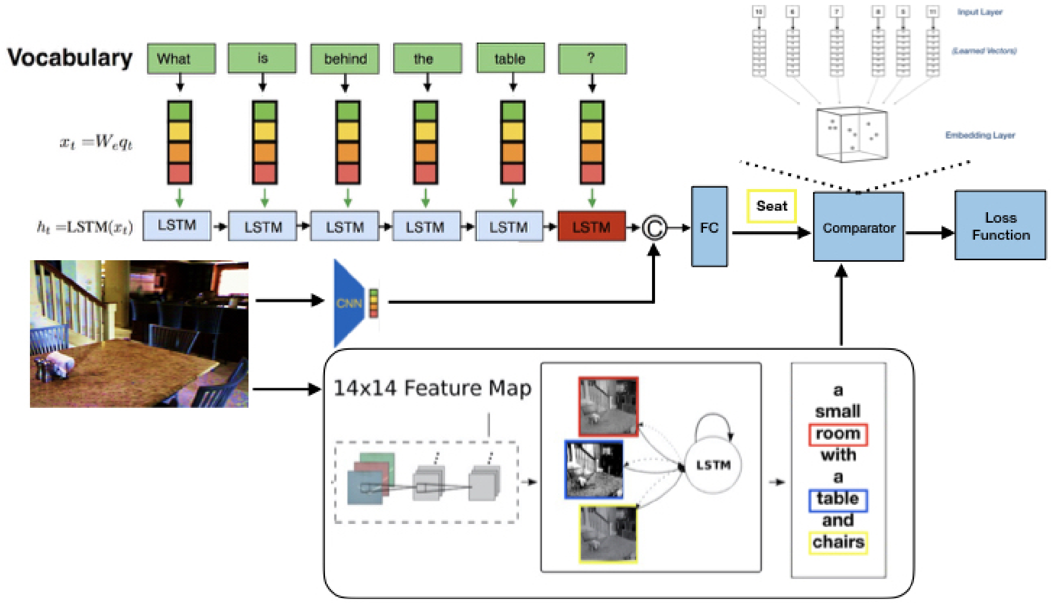

Figure 2: Network models that construct our structure.

$$class = \arg\max_{\theta} \|output, windowed\ part\|_2 \tag{1}$$

$$Loss(x, class) = -\log\left(\frac{\exp\left(x[class]\right)}{\sum_j \exp\left(x[j]\right)}\right) \tag{2}$$

4 Experiments

4.1 Experimental Settings

In this part first we train the VQA model using 10 percent of the DAQUAR dataset. Considering the questions of the DAQUAR dataset, we need this training part in order to force the model to understand the context of the questions. This is because we want our model to distinguish between the concepts of objects, colors and positions that are going to be asked in the questions. After this training phase, we hope that our model be able to understand that when the question asks about the color of a specific object, it should respond with a color. After this phase, we use other 90 percent of DAQUAR dataset, but we do not use their labels. Instead, we feed the images to the image captioning model and use its input to continue training the VQA model. In our experiments, we used 5 epochs of training on the 10 percent of labeled dataset, and after that we train with other 90 percent of data without looking at the labels for 15 epochs. Here, our focus is on the second phase and its impact on the training and resulting network.

In order to become able to compare the results, we also trained a VQA model with the same structure, using the entire dataset and their labels. We trained this network for 20 epochs. In the rest of this section, we refer to this approach by classic model.

4.2 Results

The training and testing errors during the training of both models are provided in the figure 3. As we can see in the first chart, there is a significant jump in the training loss for the new model. This is due to the definition of the training loss. Recall that in the settings that we defined, the training loss is computed with respect to the label that we assume using the image captioning output. However, in the first 5 epochs it is calculated by the comparison of the output and the actual label that is provided

Table 1: Comparison between classic approach and new approach

Model	Testing Loss
Classic model	4.00232881
New model	3.97967066

in the dataset. Hence, we believe that is increase in the training loss is due to the definition and does not correspond to the actual performance of the network.

In order to be able to compare our result with the classic model, we calculated the testing loss using the classical definition. It means that we actually calculated the cross-entropy using the test data's actual label that is provided in the dataset. This will allow us to compare both testing losses together. As we can see in the second chart, the testing losses of both models have very similar values. The results become more significant, when we see that the testing loss using the new model is even better in some epochs, suggesting that the idea of using an image captioning network can actually overcome the lack of labeled data.

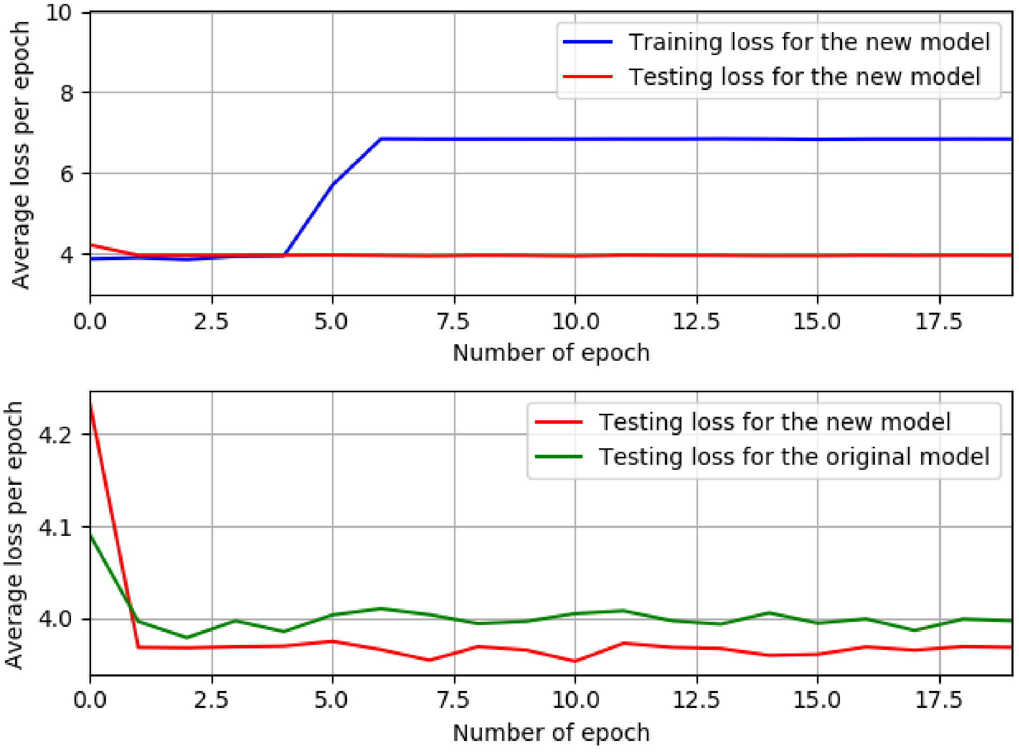

Figure 3: Training and testing loss versus each epoch in the experiment.

Also, the average values of testing loss for both approaches are provided in the table 1. We can see that in our experiments, the new model has a better testing error in average. It means that this model has a better generalization with respect to the original model, however, difference is negligible.

5 Conclusion

In this project, we defined a question about the visual question answering task. If the amount of labeled data is limited, are we able to get train a network that has a desirable generalization? In order to answer to this question, we used the ideas of active learning. We defined an oracle to provide a label for the question that is asked about an image. This oracle is an image captioning network, that given an image as its input, generates a sentence that describes the objects that are visible in that image. This task is a easier task since there are more datasets designed for this task. We trained such a network and tried to use that as the oracle. However, there were some challenges in connecting these two models. We overcome these difficulties by using a semantic similarity calculator, in order to find the most related portion of the caption from oracle, and to find the desired word in this portion and define that word as the answer of that question. By using this structure and defining a new loss function, we became able to train a VQA model using only 10 percent of the labeled data and use the rest of data without looking at their label. Our results suggest that by using this idea, we can train a model that has a generalization that is comparable to the classical models that use the entire dataset. Although our experiments were limited and we only were able to investigate the task of visual question answering, we believe that this idea of using another model as an oracle for the objective model, have great power and can bring interesting results that may suggest that we can still train models that the amount of labeled data for that task is limited.

References

[1] Xu, Kelvin and Ba, Jimmy Lei and Kiros, Ryan and Cho, Kyunghyun and Courville, Aaron and Salakhutdinov, Ruslan and Zemel, Richard S. and Bengio, Yoshua. Show, attend and tell: Neural image caption generation with visual attention. In Pro- ceedings of the 32Nd International Conference on International Conference on Machine Learning - Volume 37, ICML'15, pages 2048–2057. JMLR.org, 2015.

[2] M. Malinowski, M. Rohrbach, and M. Fritz, "Ask Your Neurons: A Deep Learning Approach to Visual Question Answering," International Journal of Computer Vision, vol. 125, no. 1–3, pp. 110–135, Aug. 2017.

[3] M. Malinowski and M. Fritz, "Tutorial on Answering Questions about Images with Deep Learning," arXiv.org, 2016. [Online]. Available: https://arxiv.org/abs/1610.01076.

[4] M. Malinowski and M. Fritz, "A Multi-World Approach to Question Answering about Real-World Scenes based on Uncertain Input."

[5] J. Pennington et al. Glove: Global vectors for word representation. In Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 1532–1543, Doha, Qatar, October 2014. Association for Computational Linguistics.

[6] T.-Y. Lin et al., "Microsoft COCO: Common Objects in Context."

[7] Ksenia Konyushkova et al., "Learning Active Learning from Data."

[8] Donggeun Yoo et al., "Learning Loss for Active Learning."