Project 1 - Visual Question Answering

Renyu Jiang U18427594 EC601-A1

**Problem Statement:**

Visual Question Answering (VQA) is a recently emerging field of machine learning, and it is a mixed fields requiring visual recognition and natural language processing. Visual question answering usually takes image and question in natural language as input and gives out answer in natural language according to human’s question. When compared with other machine learning technologies, visual question answering is of more practical applications in different areas, including human-computer interaction, navigation for the blind, etc.

However, visual question answering is a very challenging task with great potentials. This is because visual question answering has a higher requirement for its two essential technologies: natural language processing and computer vision processing. If separated, both natural language processing technology and computer vision technology have good algorithms and datasets after many years of development, but when visual question answering combines these two areas, traditional models do not work well as before anymore. Aside from the difficulty from the combination of technologies, the openness of visual question answering is also a problem, because both question and answer are open-ended.

**Why is it useful? What are its applications?**

Just as mentioned above, visual question answering technology combines natural language processing and computer vision processing, therefore, visual question answering technology can solve many problems cannot be solved easily before. Besides this, visual question answering need higher level logical reasoning ability, so it can also push the development of algorithms in language processing field.

Visual question answering technology has applications in many important fields. The first example is navigation for the blind. With the rising labor costs, hiring human to provide navigation for the blind has becoming a very expensive thing. But visual question answering can help solve this problem. The blind can buy a device with visual question answering function, what they need to do is asking the device to tell them what is ahead. It is much cheaper when compared with human guide. The second application is about information processing. For example, when scientists have a new sample of insects or piece of antique, they need to search database for previous record to identify those new things, but visual question answering can do this job in a faster and more accurate way. In addition, duo to explosive growth of data today, another application is the integration of VQA into image retrieval systems. This could have a huge impact on social media or e-commerce. Analyzing user data and collecting useful information would become an easier thing.

**Research status:**

**Some methods**

At present, there are many models for visual question answering. This paper selects several popular or high-performance models for review.

The first model is VIS+LSTM, which was proposed by Mengye Ren et al. in NIPS2015 [1]. The basic structure of the model is to use CNN to extract image information first, and then connect to LSTM to generate prediction results. Mengye Ren et al. focused their attention on limited domain questions that could use a single word as the answer to visual question answering, so that visual question answering can be viewed as a multi-classification problem, whose answer can be measured using the existing accuracy evaluation criteria. The basic structure of the model is shown below. First, use VGG19 trained on ImageNet to extract image features to obtain a 4096-dimensional vector (the output of the last hidden layer), and then treat this vector as the first word of the question sentence. Since the input of the LSTM network is a 300- or 500-dimensional word vector, the paper uses affine or linear transformation to transform the image feature vector into a 300- or 500-dimensional vector.

Graphical user interface

Description automatically generated with medium confidence

(This image is from Mengye Ren’s paper)

The second model is a traditional method: iBOWIMG[2]. One feature of this mode is that it uses Google Net to raise visual features, Bag-of-word model to raise question features. In this model, the authors use naive bag-of-words as text features, and they also use Google Net to extract visual features. The input problem is first converted into one-hot encoding, and then converted into word features through word embedding, and then connected with the image features extracted by CNN, and finally send connected features to the SoftMax layer for answer prediction. Although the performance of this model is not much better than some recurrent neural network-based approaches at the same time, it is a simple baseline model that requires very little code, and some researchers proposed modified models bases on it. Therefore, this model is selected in this paper.

The last model is up-down, a very high-performance model, whose authors obtained the first place in the 2017 VQA Challenge by applying it. Anderson, Peter, et al. proposed a top-down and bottom-up attention model approach for visual scene understanding and related problems in visual question answering systems. The bottom-up based attention model (using Faster R-CNN) is used to extract the region of interest in the image and obtain object features; while the top-down based attention model is used to learn the weights corresponding to the features (using LSTM) to achieve a deep understanding of visual images. The bottom-up attention model using Faster R-CNN in this article allows the overlap of interest boxes by setting a threshold, which can more effectively understand the image content. In this paper, not only an object detector but also an attribute classifier is used for each region of interest, so that a binary description of the object (attribute, object) can be obtained. Such a description is more suitable for practical applications. As for the top-down attention model, it has two layers of LSTM models, one for top-down attention and one for language models.

Table

Description automatically generated

(This image is from Peter Anderson’s paper)

From the chart above, it is obvious that this method does achieve good results. And through the visualization results, the features based on faster RCNN can effectively understand the description of the problem and locate the target of the problem.

**Popular datasets**

DAQUAR (DAtaset for QUestion Answering on Real-world images) is the earliest and smallest VQA dataset, it contains 6795 training data and 5673 testing data. All images come from the dataset NYU-DepthV2 Dataset. Due to the time it was proposed, the dataset is of poor quality, some images are cluttered, of low resolution, and the questions and answers have obvious errors.

The COCO (Microsoft Common Objects in Context) dataset is a large, rich object detection, segmentation and captioning dataset, originated from the Microsoft COCO dataset, which was funded and annotated by Microsoft in 2014. The COCO-QA dataset is generated on this basis. QA pairs are generated by the NLP algorithm and the images are from the COCO dataset. There are a total of 78,736 training QA pairs and 38,948 test QA pairs. Most of the questions are about the images. The answers to all questions are one word, and there are only 435 unique answers. The biggest disadvantage of the dataset is that QA pairs are generated by the NLP algorithm, which divides long sentences into short sentences for processing, which ignores the grammar and clauses in the sentences, and the algorithm results are not intelligent enough; in addition, the dataset has only 4 categories question.

VQA2.0 is the most widely used real data set now, and its questions are all manually marked, which is characterized by large and complex without focus. When compared with VQA1.0, this new database mainly addresses the problem of imbalanced answers.

There are also many other datasets in use, like FM-IQA, Visual Genome (biggest), Visual7w KB-VQA, etc.

**Future goals**

Although VQA’s task is to solve the task by processing pictures and answering questions, in fact, many questions often require certain prior knowledge to answer. For example, to answer the question "how many iPhones are there in the picture", the model needs to know what “iPhone” is and what is the difference between it and other phones rather than simply understand the content of the image. Therefore, connecting the knowledge base to the VQA model has become a promising research direction.

In addition, the bias of the dataset affects the evaluation of the VQA algorithm, and the problem of scene and target attributes has a strong bias. Finding a more suitable VQA dataset is also a challenge

There are many more studies and findings on visual question answering, but due to the limited time of research, the article ends here.

**List of papers or**

1. Ren M, Kiros R, Zemel R. Exploring models and data for image question answering[J]. Advances in neural information processing systems, 2015, 28.
2. Zhou B, Tian Y, Sukhbaatar S, et al. Simple baseline for visual question answering[J]. arXiv preprint arXiv:1512.02167, 2015.
3. Anderson, Peter, et al. "Bottom-up and top-down attention for image captioning and visual question answering." Proceedings of the IEEE conference on computer vision and pattern recognition. 2018.

**Open-source links:**

1. <https://visualqa.org/>
2. <https://paperswithcode.com/task/visual-question-answering>
3. <https://towardsdatascience.com/deep-learning-and-visual-question-answering-c8c8093941bc>