Project 1: VQA

Qilong Wang qilongw@bu.edu

1.Visual Questioning Answer introduction

Visual Questioning Answer is a task that combines computer vision and natural language processing. Given an image and a question, its goal is to infer the correct answer to the question from the visual information of the image.

The difficulty of VQA is that the image is high-dimensional information, lacks grammatical rules and structures, and cannot directly use NLP methods such as syntactic analysis and regular expressions. Furthermore, the information captured by images is closer to the real world, while language itself is an abstraction.

2. Dataset information

2.1 Basic information

The data contains three parts, which are images, questions, and answers. For different datasets their numbers of questions for each image and form of answers may be different. Widely used VQA datasets are stated as follows.

DVQUAR, the first VQA dataset, contains 1449 images.

COCO-QA, contains 123287 images with one question for each image.

VQA 1.0 and VQA 2.0, contains 123287 real world images and 50000 cartoon images, each with at least 3 questions.

Visual Genome, contains 1.7 million questions and answers.

2.2 Question form

The model performs differently when facing to different kinds of questions. The questions are divided into binary questions (Y/N), number questions (count) and other questions. Usually models perform much better on binary questions than the other two kinds, reaching 80% or even more accuracy. Number questions are the most difficult one, whose accuracy is a little smaller than the other kind.

According to the field in computer vision, the questions can also be divided into object recognition, object detection, attribute classification, scene classification and counting problems.

2.3 Evaluation measurement

For the evaluation of responses, both syntactic and semantic correctness need to be considered, so responses in most VQA datasets are limited to less than 3 words.

It would compare the common subsequence of the two in the taxonomy tree. When the similarity exceeds a certain threshold, it is considered correct. Generally, two thresholds of 0.9 and 0.0 are used. While In the VQA dataset, the given answer is considered correct only if more than 3 people (out of 10) provided the answer.

3 Model Methods

3.1 Joint embedding

The purpose of joint embedding is to learn the representation of CV and NLP tasks in common space. It is the basis of most VQA methods. The image representations used are generally pre-trained CNNs on object recognition, and text representations are generally pre-trained word embeddings on large text corpora, and then the embedding of words in the problem is sent to RNN to solve variable-length sequences.

3.2 Attention

The attention mechanism uses local features and allows the model to give different weights in different regions. In VQA, we focus on the regions related to the question.

3.3 Knowledge based models

VQA usually requires prior knowledge, such as "how many mammals are in the picture", so joint embedding is flawed: it can only acquire the existing knowledge in the training set, and it is impossible to cover all situations in the real world. So one solution is to combine the reasoning process with the knowledge base.

4. Paper introduction

I choose a paper in CVPR 2017. The article is MUTAN: Multimodal Tucker Fusion For Visual Question Answering.

4.1 Model method

Bilinear model is an efficient method for VQA fusion problems because it encodes the complete second-order interactions. The main problem of these bilinear models is related to the number of parameters, and the number of parameters becomes intractable with respect to the dimensions of the input and output, so the bilinear must be simplified or approximated by reducing the complexity of the model. So this paper introduces the concept of MUTAN - based on multimodal The quantitative Tucker decomposition can be effectively parameterized in the model of bilinear interaction (Bilinear models) of vision and text. In addition to the Tucker decomposition, a matrix-based low-rank decomposition is also designed to explicitly limit the interaction level.

The paper also contains many mechanisms. The multimode visual and text task helps to match the features in images and texts. The image captioning task makes description of the image. The attention mechanisms help model focus on the question-related part of image.

4.2 Result

On Y/N problems the model reaches 85.14% accuracy. On number problems the model reaches 39.81% accuracy. On the other problems the model reaches 58.52% accuracy. Totally, the model accuracy is 67.42% on dev set and 67.36 on std set.

[1] Gao H, Mao J, Zhou J, et al. Are you talking to a machine? dataset and methods for multilingual image question[C]//Advances in neural information processing systems. 2015: 2296-2304.

[2] Wu Qi, Teney Damien, Wang Peng, Shen Chunhua, Dick Anthony, and van den Hengel Anton. 2017. Visual question answering: A survey of methods and datasets. Comput. Vis. Image Underst. 163 (2017), 21–40.

[3] Gupta Akshay Kumar. 2017. Survey of visual question answering: Datasets and techniques. arXiv preprint arXiv:1705.03865 (2017).

[4] A. Das, H. Agrawal, C.L. Zitnick, D. Parikh, D. Batra, Human attention in visual question answering: Do humans and deep networks look at the same regions? Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP) (2016), pp. 932-937

[5] Zhou Yu, Jun Yu, Yuhao Cui, Dacheng Tao, Qi Tian. Deep Modular Co-Attention Networks for Visual Question Answering. In CVPR, 2019.

[6] W. Zheng, L. Yin, X. Chen, Z. Ma, S. Liu, B. Yang Knowledge base graph embedding module design for visual question answering model Pattern Recognit., 120 (2021), p. 108153

[7] Hedi Ben-younes, Rémi Cadene, Matthieu Cord, and Nicolas Thome. MUTAN: Multimodal Tucker Fusion for Visual Question Answering. In IEEE International Conference on Computer Vision, pages 2612–2620, 2017.