Visual Question Answering with multi-modal Hierarchical co-attention

Sai Nikhil Maram
UCSB

Michael Zhang
UCSB

1 Motivation

The Question-answering system is emerging field in the AI research community. This kind of system can help visually impaired people to understand more about their surroundings by asking the right questions. For example, asking for a traffic signal on a road. Our Project attempts to build such a Visual Question Answering(VQA) system.

2 Problem Definition

We aim to build a AI system that will allow the user to ask questions on a given image. To correctly answer visual questions about an image, the system needs to understand both the image and question. To tackle this problem, we collect and induct information from both image and the question. The AI system will use a multi-modal attention model as described in later sections to answer the queries.

3 Related work

Before visual question answering (VQA) became popular, text question answering (QA) had already been established as a mature research problem in the area of natural language processing. Previous QA methods include: searching for the key words of the question or embedding the question and using a similarity measurement to find evidence for the answer. Initial works on VQA used complete image embedding from CNN to answer the questions and recent works on VQA have explored image attention models for VQA[1]. While most of the works for VQA have focused on the problem of identifying “where to look” or visual attention, not many have explored "which words to listen to" or question attention. Our project combines both of them as specified in the work by Lu *et al*[2].

4 Our work

In our project, we employ a co-attention model that jointly reasons about visual attention and question attention.

Co-attention: Unlike previous works, which only focus on visual attention, we use image attention to guide the question and question attention to guide the image attention.

Question Hierarchy: We build a hierarchical architecture that co-attends to the image and question at three levels: (a) word level, (b) phrase level and (c) question level. At the word level, we embed the words to a vector space through an embedding matrix. At the phrase level, 1-dimensional convolution neural networks are used to capture the information contained in unigrams, bigrams and trigrams, then combine the various n-gram responses by pooling them into a single phrase level representation. At the question level, we use recurrent neural networks to encode the entire question. For each level of the question representation in this hierarchy, we construct joint question and image co-attention maps, which are then combined recursively to ultimately predict a distribution over the answer.

5 Dataset

We plan to evaluate our model on two data sets, VQA dataset [3] and COCO-QA dataset.

VQA dataset : Thus is the largest dataset for this problem, containing human annotated questions and answers on Microsoft COCO dataset. The dataset contains 248,349 training questions, 121,512 validation questions, 244,302 testing questions, and a total of 6,141,630 question-answers pairs. There are three sub-categories according to answer-types including yes/no, number, and open ended . Each question has 10 free-response answers.

COCO-QA : There are 78,736 train questions and 38,948 test questions in the dataset. These questions are based on 8,000 and 4,000 images respectively. There are four types of questions including object, number, color, and location. Each type takes 70%, 7%, 17%, and 6% of the whole dataset, respectively.

6 Evaluation Metric

For answer-types yes/no or numbers. We check our predicted answers with ground truth answers.

For open ended answer types, we plan to use BLEU-score to evaluate the results.

BLEU (Bilingual Evaluation Understudy): An algorithm used to evaluate the quality of translation from one language to other by a machine. The quality refers to how close the machine translation and human trans- lation are.

References

[1] Xu H., Saenko K. (2016) Ask, Attend and Answer: Exploring Question-Guided Spatial Attention for Visual Question Answering. In: Leibe B., Matas J., Sebe N., Welling M. (eds) Computer Vision – ECCV 2016. ECCV 2016. Lecture Notes in Computer Science, vol 9911. Springer, Cham.

[2] Lu, J., Yang, J., Batra, D., & Parikh, D. (2016). Hierarchical question-image co-attention for visual question answering. In Advances In Neural Information Processing Systems (pp. 289-297).

[3] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. Vqa: Visual question answering. In ICCV, 2015.

[4] Zhang, P., Goyal, Y., Summers-Stay, D., Batra, D., & Parikh, D. (2016, June). Yin and yang: Balancing and answering binary visual questions. In Computer Vision and Pattern Recognition (CVPR), 2016 IEEE Conference on (pp. 5014-5022). IEEE

[5] Goyal, Y., Khot, T., Summers-Stay, D., Batra, D., & Parikh, D. (2017, July). Making the V in VQA matter: Elevating the role of image understanding in Visual Question Answering. In CVPR (Vol. 1, No. 6, p. 9).