Proposal: Visual Question Answering Based on External Knowledge

1. Nature of the problem

Visual Question Answering (VQA) is a challenging task that was proposed to connect computer vision (CV) and natural language processing (NLP). In most common form of Visual Question Answering, the models are asked to select or generate answers for given textual questions about an image. The types of the answer can be binary, numerical, multi-option, ranking and open-end. Nowadays, though deep learning methods achieve a great success in both computer vision and natural language processing, Visual Question Answering is way more complex than general pure problems in CV and NLP. Firstly, the models need to extract information from two different modalities, and both of images and languages are very massive and could contain many noises. Secondly, even if the model could obtain appropriate features from images and languages, understanding these features and combining them could never be a trivial task. For example, the feature vector of an image usually is a single vector, but the feature of a textual sentence (paragraph) could be a sequence of data. The model should be able to combine them. Moreover, Visual Question Answering frequently requires information not present in the images, so some external knowledge should be taken into consideration.

2. Related Work

Modern methods to tackle Visual Question Answering problems can be presented through categories: joint embedding approaches, attention mechanisms, composition models and knowledge base-enhanced approaches. Most methods combine multiple strategies.

Joint Embedding

The basic idea of joint embedding is to learn image and language features in a common feature space. Usually, question and image features are both fed together to a first “encoder” LSTM. It produces a feature vector of fixed-size that is then passed to a second “decoder” LSTM. The decoder produces variable-length answers. The most obvious limitation of this kind of model is to use global (image-wide) features to represent the visual input.

Attention

Considering it might be an ideal method to use global features to represent the visual input, the attention-based methods are come up with. The basic idea is to use local image features, and allow the model to assign different importance to features from different regions. Interestingly, attention mechanisms improve the overall accuracy on all VQA datasets, but closer inspection by question type show little or no benefit on binary (yes/no) questions.

Composition

The methods discussed above present limitations related to the monolithic nature of the CNNs and RNNs used to extract representations of images and sentences. An increasingly popular research direction in the design of artificial neural networks is to consider modular architectures. This approach involves connecting distinct modules designed for specific desired capabilities such as memory or

specific types of reasoning. The composition methods generally outperform competitors on questions with a compositional structure. However, many of questions in the VQA dataset are quite simple, and require little composition or reasoning.

Knowledge-Base

The task of VQA involves understanding the contents of images, but often requires prior nonvisual, information, which can range from "common sense" to topic-specific or even encyclopedic knowledge. By using external knowledge-base model usually could get a significantly improvement on overall accuracy.

3. Potential Area

1. Change Loss function [Image weighed question weighed or external weighed]

2. Method of learning document vector.

3. The latest improvements, exemplified by MCB and MRN, still showed potential room for improvement on both the extraction of features and their projection to the embedding space.

4. Improve the learning of semantic structures of the questions and answers.

5. Change the structure of the LSTM.

6. Focus on Numerical questions which have poorest performance.

4. Proposed Work

My goal for this project is to improve a state-of-the-art framework. Considering my limited knowledge in Knowledge-Representing, I might not be able to work on improving composition methods. For now, my choice is framework from [Ask Me Anything: Free-form Visual Question Answering Based on Knowledge from External Sources].

The potential limitations of this model are as following:

1. Extract five attributes before it see the question, which could make it possible that the model extract useless information from knowledge-base.
2. There is no structure between captions.
3. Average pooling on captions might cause information loss.
4. Only represent external knowledge by doc2vec.

My proposed work includes:

1. Try different region detector to extract the attributes. It would be ideal if we could extract external knowledge based on questions.
2. Try to apply different feature extractor both from images and languages [5].
3. Instead of generating several independent captions and average pooling them, try to generate a paragraph [13] and use the structured information [13] to train the RNN.

4. Try different embedding method to represent the external knowledge (e.g. RNN, CNN with doc2vec, structured vector, GANs).

Another important thing I noticed when I went through related work is the questions are different. For example, a question “What is the color of the dog behind the white cat?” is highly depends on the images features. On the other side, “Why do people have umbrellas” is highly depends on external knowledge. Of course, “How many mammals are there?” depends on both. I’m thinking if we can teach the model assign different weights on different input features (from images, captions and external knowledge). This mechanism can be achieved by modifying the loss function [14] or changing the structure of the LSTM [15, 16 and 17].

5. Checkpoints

March 25th: Re-implement of the original model in [2]. April 5th: Replace independent captioning with hierarchical paragraph captioning. April 20th: Finish the experiment of different feature extractors. May 1st: Finish report and the bonus (different weights based on feature dependency).

6. Reference

[1] Visual Question Answering: A Survey of Methods and Datasets

[2] Ask Me Anything: Free-form Visual Question Answering Based on Knowledge from External Sources

[3] The VQA-Machine: Learning How to Use Existing Vision Algorithms to Answer New Questions

[4] Are You Talking to a Machine? Dataset and Methods for Multilingual Image Question Answering

[5] Multimodal Compact Bilinear Pooling for Visual Question Answering and Visual Grounding

[6] Show, Attend and Tell: Neural Image Caption Generation with Visual Attention

[7] Neural Module Networks

[8] Deep Fragment Embeddings for Bidirectional Image Sentence Mapping

[9] Exploring Models and Data for Image Question Answering

[10] Ask Me Anything: Dynamic Memory Networks for Natural Language Processing

[11] Dynamic Memory Networks for Visual and Textual Question Answering

[12] CNN: Single-label to Multi-label

[13] A Hierarchical Approach for Generating Descriptive Image Paragraphs

[14] Generating Visual Explanations

[15] Visual7W: Grounded Question Answering in Images

[16] Compositional Memory for Visual Question Answering

[17] ABC-CNN: An Attention Based Convolutional Neural Network for Visual Question Answering