Project on Visual Commonsense Reasoning

Anonymous ACL submission

Abstract

1 Credits

2 Introduction

3 Related works

Visual question answering. Visual question answering (VQA), first proposed by (Malinowski et al., 2015), is a challenging task that combines both NLP and AI domains. VQA requires the system leverage both semantic and visual informations to generate plausible answers. (Malinowski et al., 2015) first proposed to combine the image segmentation and semantic parsing with Bayesian approach to sample form training set. Similar to (Malinowski et al., 2015), many works are based on combining convolutional neural networks and recurrent neural networks to generate answers with both visual and semantic features as inputs. (Gao et al., 2015) trains two networks, one encoder RNNs that take the question along with visual features to generate the answer using another decoder RNNs. (Ren et al., 2015) focuses on questions with one word as the answer, transforming the problem to classification problems. They also proposed an algorithm that could generate questions with one word answers from image descriptions and formed a larger dataset. (Jabri et al., 2016) takes the answers as the input and convert the multiple-choices questions into image-question-answer triplet, on which they perform a simple binary classification.

Recent leading research on VQA can be separated into two categories. 1) Neural attention based; 2) learning from explicitly external knowledge base. Neural attention based methods following human intuitions learn to attend the visual areas that could provide better information for answering the questions. (Chen et al., 2015) creates an attention map given an image-question pair by fusing the visual feature maps and the semantic features. (Yang et al., 2016) followed this line and proposed to use stacked attention networks that could attend multiple regions and narrow down to focus one that related the most to the query. Instead of learning a visual attention map, (Jiang et al., 2015; Zhu et al., 2016) propose to explicitly fuse word feature to multiple regions that associate with it, which allows the system to answer questions that based on multiple instances (e.g. how many question). Also (Zhu et al., 2016) proposed seven W questions (what, where, when, who, why, which and how). Different from previous attention methods, (Lu et al., 2016) emphasized the importance of question attention, presenting co-attention model that jointly reasons the visual evidence and question evidence. Another line of works explicitly learn from external knowledge base to enable answering human posed questions that have informations not contained in the image itself. (Zhu et al., 2015) used knowledge base and RDBMS to answer image-based queries. They express images in the form of visual feature and attribute labels, on which they relate image to quantities that exists in the database. (Wu et al., 2016) combined the internal representation of image with semantic features form Word2Vec extracted in DBpedia. By using the information in larger knowledge base, it allows the system to answer a broader range of questions. However, these methods only extract knowledge with certain format from the knowledge base. This is harmful for them to learn from and answer general questions. (Wang et al., 2015) is capable of not only answering the question but also reasoning about an image on the basis of information extracted from a large-scale knowledge base. (Narasimhan et al., 2016) queries data over unstructured web articles when information in the existing data is incomplete.
 Commonsense reasoning and explainability. Humans use explanations as a guide for learning and understanding by building inferences and seeking propositions or judgments that enrich their prior knowledge. (Hendricks et al., 2016) proposes a new model that focuses on the discriminating properties of the visible object, jointly predicts a class label, and explains why the predicted label is appropriate for the image. Different from previous models that designed to produce interpretable traces of their decision-making process typically require these traces to be supervised at training time, (Hu et al., 2018) presents a novel neural modular approach that performs compositional reasoning by automatically inducing a desired sub-task decomposition without relying on strong supervision, which allows linking different reasoning tasks though shared modules that handle common routines across tasks. (Huk Park et al., 2018) proposes a multimodal approach to explanation, and argue that the two modalities provide complementary explanatory strengths. Another line of works sought the way of physical based explainability (Yi et al., 2018; Santoro et al., 2017; Goyal et al., 2017) on CLEVR dataset in which each image comes with intricate, compositional questions generated by programs.

Visual commonsense reasoning. AI models perform well on VQA problems. However, this could hardly confirm that these models could understand the world. By imposing the network to explain or reason the answer of the question, especially for reasons that not contained in the image itself, namely commonsense knowledge, the network could learn to really understand the scene. (Wang et al., 2015) searches along the paths from visual concepts to knowledge base concepts. Meanwhile they could trace back for logical reasons. It can explain its reasoning in terms of the entities in the knowledge base, and the connections between them. (Wang et al., 2018) proposed fact-based VQA, which requires external knowledge to answer. It extended image-question-answer triplets to image-question-answer-supporping fact tuples, where fact is represented by a structural triplet. (Anne Hendricks et al., 2018) learns a ranking model for both the answers and the reasons, improving the textual explanation quality of fine-grained classification decisions on the CUB dataset by mentioning phrases that are grounded in the image. (Zellers et al., 2018) instead provided a dataset called Visual Commonsense Reasoning (VCR) that has the form of multiple choice for visual question answering and answer reasoning based on images and commonsense facts. Different from previous works that either mapping the feature of visual and semantic to generate the answer or search in the external knowledge base with fixed format of answers, our method aims to answer the question with combined visual evidence and commonsense evidence that learnt from external knowledge base, which enables more flexible form of reasoning.

4 Method
5 Experiments

References


