Project on Visual Commonsense Reasoning

Anonymous ACL submission

Abstract

1 Credits

2 Introduction

Human understand the world by recognizing objects in the context and reasoning their relationships. Trying to mimic the ways human brain recognize objects in visual scene, deep neural networks utilizing convolutional layers and pooling layers to extract visual features for various tasks, e.g. object recognition, object detection, semantic segmentation, etc. Similarly, some models learn word embeddings for neural language processing tasks with recurrent neural networks and conditional random field, e.g. name entity recognition, relation extraction, etc. However, one significant drawback of such deep feature learning scheme is that the models suffer from poor generalization and reasoning power, which leads to difficulties in transferable feature learning and multi-model learning. To solve this problem, it is important to go beyond recognition tasks to cognition tasks, which aim to understand the relationships on the basis of certain attributions learnt by well established recognition models.

Towards this approach, some methods try to build the connection of objects by learning feature embeddings and relation networks. Others try to leverage triplet relationship or external knowledge base to construct and model commonsense knowledge to understand the interaction of objects. These models are performed on either visual question answering problem or commonsense problem until the introduction of visual commonsense reasoning problem has been proposed. The problem visual commonsense reasoning explicitly requires three types of evaluation: question to answering (Q $\rightarrow$ A), question and answering to reasoning (QA $\rightarrow$ R) and question to answering and reasoning (Q $\rightarrow$ AR). These three problems gradually increase the need of reasoning power of the model, especially when sometimes the answer could not directly be reflected from the appearance of the image itself, e.g. when a waiter serves the coffee, the image does not contain any information about where the coffee is going to be placed, however, from other people's reaction we are able to infer it, which requires commonsense knowledge outside of the image. We argue that there are two main challenges in VCR problems compared to VQA and commonsense problems: 1) an explicit knowledge graph is needed to build to model the background commonsense knowledge outside of the images; 2) the visual features and the semantic features needs to be aligned in the knowledge graph to form a uniform knowledge source so that the model could directly connect the questions with the visual scenes.

Thus we propose cross-modal relation graph method, a novel knowledge graph based visual commonsense reasoning method that contains three component: 1) commonsense knowledge graph that learns commonsense concepts based on triplet relationships; 2) visual relation graph that built on the input image with each object and its learnt attribute descriptors; 3) question relation graphs that built on the input question. The model then uses graph matching techniques to match both the visual and question relation graphs to the commonsense knowledge graph and returns the confidence of the matching. After that, the model could trace back to the commonsense knowledge graph to find the reasoning of the matching. This model could naturally address the three tasks proposed by VCR. For Q $\rightarrow$ A, we seek for correct answer to be the answer with highest matching confidence; for QA $\rightarrow$ R, we trace back to the commonsense knowledge graph given by the matching

results of visual relation and question relation to find R; for Q $\rightarrow$ AR, we could simply combine the above two steps.

To summarize, our contributions are: 1) propose cross-modal relation graph methods to model relation graphs of different modality; 2) propose graph matching methods to align the relation graph to commonsense knowledge graph.

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 from 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 explainablility (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-supporing 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

An explicit commonsense knowledge graph is crucial to extract informations to reason the interactions between objects. Our model first learns to construct such a commonsense knowledge graph on the top of external knowledge bases. Then we define some types of attributes for each object in the image, and the model should learn to map the visual attribute features to the semantic features given in the training questions and answers. The graph matching is finally applied on the learnt visual graph space and the commonsense knowledge graph space.

4.1 Commonsense knowledge graph

4.2 Relation graph

4.3 Graph matching

5 Experiments

Datasets We use Visual Commonsense Reasoning (VCR) dataset to evaluate our model. The VCR dataset contains 110k unique scene and 290k pairs of questions, answers and rationales. All of the three tasks are evaluated in the form of accuracy of multiple choice problems. The dataset provide bounding boxes and object indices of each object in the scene generated by Mask-RCNN, which clean off the bias of data preprocessing.

Results We show the previous reported baseline methods.

References

Lisa Anne Hendricks, Ronghang Hu, Trevor Darrell, and Zeynep Akata. 2018. Grounding visual explanations. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 264–279.

Kan Chen, Jiang Wang, Liang-Chieh Chen, Haoyuan Gao, Wei Xu, and Ram Nevatia. 2015. Abc-cnn: An attention based convolutional neural network for visual question answering. *arXiv preprint arXiv:1511.05960*.

Haoyuan Gao, Junhua Mao, Jie Zhou, Zhiheng Huang, Lei Wang, and Wei Xu. 2015. Are you talking to a machine? dataset and methods for multilingual image question. In *Advances in neural information processing systems*, pages 2296–2304.

Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. 2017. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6904–6913.

Lisa Anne Hendricks, Zeynep Akata, Marcus Rohrbach, Jeff Donahue, Bernt Schiele, and Trevor Darrell. 2016. Generating visual explanations. In *European Conference on Computer Vision*, pages 3–19. Springer.

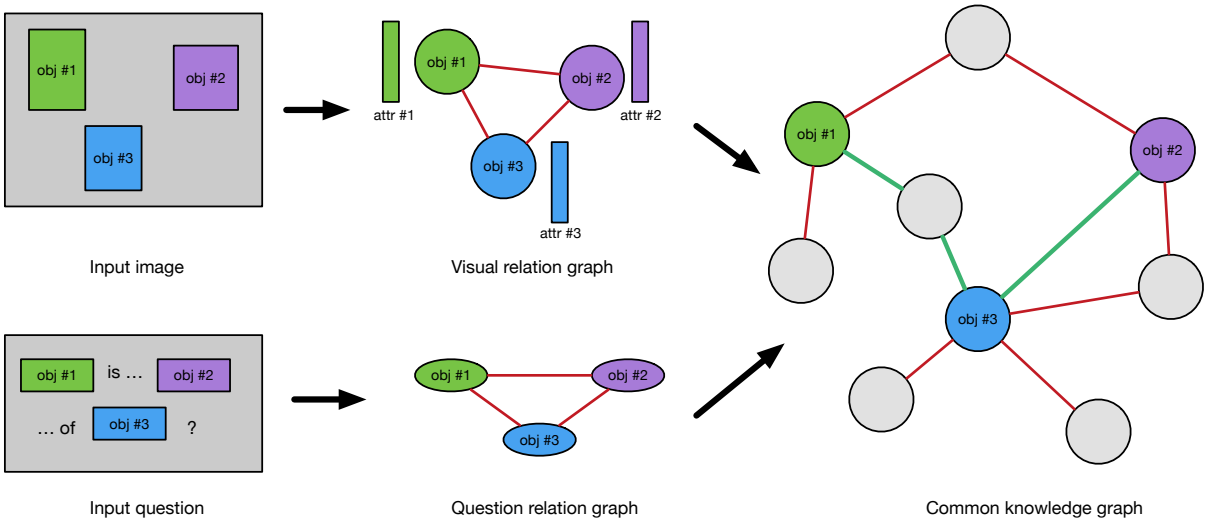

Figure 1: Overall pipeline of proposed method. The convolutional neural networks extract per object attributions and relations from input image. Recurrent neural networks extract semantic embeddings from questions. Two relation graphs then match in the commonsense knowledge graph. The path in the commonsense knowledge graph could provide cues for reasoning the interplay between different objects.

	Q → A		QA → R		Q → AR	
Model	Val	Test	Val	Test	Val	Test
BERT	53.8	53.9	64.1	64.5	34.8	35.0
ESIM+ELMo	45.8	45.9	55.0	55.1	25.3	25.6
LSTM+ELMo	28.1	28.3	28.7	28.5	8.3	8.4
RevisitedVQA	39.4	40.5	34.0	33.7	13.5	13.8
MLB	45.5	46.2	36.1	36.8	17.0	17.2
R2C	63.8	65.1	67.2	67.3	43.1	44.0
Human		91.0		93.0		85.0

Table 1: Multiple choices accuracy on the VCR dataset.

Ronghang Hu, Jacob Andreas, Trevor Darrell, and Kate Saenko. 2018. Explainable neural computation via stack neural module networks. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 53–69.

Dong Huk Park, Lisa Anne Hendricks, Zeynep Akata, Anna Rohrbach, Bernt Schiele, Trevor Darrell, and Marcus Rohrbach. 2018. Multimodal explanations: Justifying decisions and pointing to the evidence. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8779–8788.

Allan Jabri, Armand Joulin, and Laurens Van Der Maaten. 2016. Revisiting visual question answering baselines. In *European conference on computer vision*, pages 727–739. Springer.

Aiwen Jiang, Fang Wang, Fatih Porikli, and Yi Li. 2015. Compositional memory for visual question answering. *arXiv preprint arXiv:1511.05676*.

Jiasen Lu, Jianwei Yang, Dhruv Batra, and Devi Parikh. 2016. Hierarchical question-image co-attention for visual question answering. In *Advances In Neural Information Processing Systems*, pages 289–297.

Mateusz Malinowski, Marcus Rohrbach, and Mario Fritz. 2015. Ask your neurons: A neural-based approach to answering questions about images. In *Proceedings of the IEEE international conference on computer vision*, pages 1–9.

Karthik Narasimhan, Adam Yala, and Regina Barzilay. 2016. Improving information extraction by acquiring external evidence with reinforcement learning. *arXiv preprint arXiv:1603.07954*.

Mengye Ren, Ryan Kiros, and Richard Zemel. 2015. Exploring models and data for image question answering. In *Advances in neural information processing systems*, pages 2953–2961.

Adam Santoro, David Raposo, David G Barrett, Mateusz Malinowski, Razvan Pascanu, Peter Battaglia, and Timothy Lillicrap. 2017. A simple neural network module for relational reasoning. In *Advances in neural information processing systems*, pages 4967–4976.

Peng Wang, Qi Wu, Chunhua Shen, Anthony Dick, and Anton van den Hengel. 2018. Fvqa: Fact-based visual question answering. *IEEE transactions on pattern analysis and machine intelligence*, 40(10):2413–2427.

Peng Wang, Qi Wu, Chunhua Shen, Anton van den Hengel, and Anthony Dick. 2015. Explicit knowledge-based reasoning for visual question answering. *arXiv preprint arXiv:1511.02570*.

Qi Wu, Peng Wang, Chunhua Shen, Anthony Dick, and Anton van den Hengel. 2016. Ask me anything: Free-form visual question answering based on knowledge from external sources. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4622–4630.

Zichao Yang, Xiaodong He, Jianfeng Gao, Li Deng, and Alex Smola. 2016. Stacked attention networks for image question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 21–29.

Kexin Yi, Jiajun Wu, Chuang Gan, Antonio Torralba, Pushmeet Kohli, and Josh Tenenbaum. 2018. Neural-symbolic vqa: Disentangling reasoning from vision and language understanding. In *Advances in Neural Information Processing Systems*, pages 1039–1050.

Rowan Zellers, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2018. From recognition to cognition: Visual commonsense reasoning. *arXiv preprint arXiv:1811.10830*.

Yuke Zhu, Oliver Groth, Michael Bernstein, and Li Fei-Fei. 2016. Visual7w: Grounded question answering in images. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4995–5004.

Yuke Zhu, Ce Zhang, Christopher Ré, and Li Fei-Fei. 2015. Building a large-scale multimodal knowledge base system for answering visual queries. *arXiv preprint arXiv:1507.05670*.