Compositional Reasoning for Commonsense Understanding

AMMA [*1] **Vishnu Rajan Tejus** [*2]

Abstract

Machine reasoning is the task of drawing novel inferences to answer questions through previously acquired knowledge and experience. Being able to reason relations and contextual interactions, and understanding commonsense is a key characteristic trait of an intelligent system. In this paper, we propose a new state-of-the-art approach called the compositional reasoning network (CRN), shifting from monotonic and monolithic architectures towards a generalizable, transferable, and transparent system that can achieve incredible results on the CLEVR dataset for visual question answering (VQA) and the Winograd Schema Challenge (WSC).

1. Introduction

”It’s raining cats and dogs!” Upon hearing or seeing this, we intuitively understanding that the ”sky is pouring”, or it’s heavily raining. We do not understand this sentence to mean, however, that cats and dogs are falling out of the sky. How do we intuitively know what ”raining cats and dogs” means? One answer is because we can understand and process commonsense knowledge, and reason with it.

Deep learning (LeCun et al., 2015) has been successful solving recognition problems such as image classification (Krizhevsky et al., 2012; He et al., 2015; 2016), sentiment analysis (Socher et al., 2013; Lakkaraju et al., 2014), speech processing (Hinton et al., 2012; Ephrat et al., 2018), and basic video prediction (Oh et al., 2015; Lee et al., 2018). However, these models are reliant on large amounts of labeled training, which is expensive, time-consuming, and often hard to obtain. Furthermore, human labeled data has a tendency to be biased and often limit model performance, which result in models not being accurate in recognizing underrepresented categories or failing to achieve superhuman performance.

Reinforcement learning (RL) approaches have sought to resolve some of these problems by learning through experience in an environment. RL has been successful in domains such as game-playing. However, RL has faced problems such as instability in learning and not being able to generalize to even small changes in task.

The task of relational reasoning is to draw novel inferences between relations from previously acquired knowledge and answer questions. We propose a model called the compositional reasoning network (CRN), shifting away from monolithic and monotonic black-box systems to a generalizable and transparent architecture, and introduce solutions to the problems associated with supervised approaches.

We evaluate our model on the CLEVR dataset for visual question answering (VQA) (Johnson et al., 2017), a challenging benchmark for evaluating multi-modal reasoning systems for multi-step problem solving requiring both visual perception and natural language skills. The central goal of the CLEVR task is to answer natural language questions about paired images. Each image is associated with.

Furthermore, we evaluate our model on the Winograd Schema Challenge (Levesque et al., 2011).

In this paper, we introduce a new approach called compositional reasoning network (CRN) to tackle the problem of commonsense understanding and relational reasoning.

2. Compositional Reasoning Network (CRN)

The CRN network contains CRN cells with global self-attention where each cell computes explicit multi-step reasoning processes given a query. For VQA, the query is a natural language question paired with an image, and for the Winograd Schema Challenge (WSC) the query is a multiple-choice question.

2.1. Distributed Vector Representations for Queries

For the VQA task, the image is processed in a similar method as prior approaches to CLEVR (Hudson & Manning, 2018; Santoro et al., 2017; Hu et al., 2017; Perez et al., 2017) by extracting *conv4* features from ResNet101 (He et al., 2016) with a fixed feature extractor pre-trained on ImageNet (Russakovsky et al., 2015). The question string S

[*] Equal contribution [1] Amrita AI Research (AAIR) [2] Stanford University. Correspondence to: Vishnu Rajan Tejus <vrt@ieee.org>.

Sashtanga Pranams at AMMA’s Lotus Feet.

is processed by a d-dimensional biLSTM, extracting context and word representations.

Contextual words are represented as a series of output states $cw_1, ..., cw_s$ in a question $q = [\overleftarrow{cw_1}, \overrightarrow{cw_s}]$, drawing inspiration from Hudson & Manning (2018).

References

Ephrat, Ariel, Mosseri, Inbar, Lang, Oran, Dekel, Tali, Wilson, Kevin, Hassidim, Avinatan, Freeman, William T, and Rubinstein, Michael. Looking to listen at the cocktail party: A speaker-independent audio-visual model for speech separation. *arXiv preprint arXiv:1804.03619*, 2018.

He, Kaiming, Zhang, Xiangyu, Ren, Shaoqing, and Sun, Jian. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *Proceedings of the IEEE international conference on computer vision*, pp. 1026–1034, 2015.

He, Kaiming, Zhang, Xiangyu, Ren, Shaoqing, and Sun, Jian. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

Hinton, Geoffrey, Deng, Li, Yu, Dong, Dahl, George E, Mohamed, Abdel-rahman, Jaitly, Navdeep, Senior, Andrew, Vanhoucke, Vincent, Nguyen, Patrick, Sainath, Tara N, et al. Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. *IEEE Signal Processing Magazine*, 29(6):82–97, 2012.

Hu, Ronghang, Andreas, Jacob, Rohrbach, Marcus, Darrell, Trevor, and Saenko, Kate. Learning to reason: End-to-end module networks for visual question answering. *CoRR, abs/1704.05526*, 3, 2017.

Hudson, Drew A and Manning, Christopher D. Compositional attention networks for machine reasoning. *arXiv preprint arXiv:1803.03067*, 2018.

Johnson, Justin, Hariharan, Bharath, van der Maaten, Laurens, Fei-Fei, Li, Zitnick, C Lawrence, and Girshick, Ross. Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. In *Computer Vision and Pattern Recognition (CVPR), 2017 IEEE Conference on*, pp. 1988–1997. IEEE, 2017.

Krizhevsky, Alex, Sutskever, Ilya, and Hinton, Geoffrey E. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pp. 1097–1105, 2012.

Lakkaraju, Himabindu, Socher, Richard, and Manning, Chris. Aspect specific sentiment analysis using hierarchical deep learning. In *NIPS Workshop on Deep Learning and Representation Learning*, 2014.

LeCun, Yann, Bengio, Yoshua, and Hinton, Geoffrey. Deep learning. *nature*, 521(7553):436, 2015.

Lee, Alex X, Zhang, Richard, Ebert, Frederik, Abbeel, Pieter, Finn, Chelsea, and Levine, Sergey. Stochastic adversarial video prediction. *arXiv preprint arXiv:1804.01523*, 2018.

Levesque, Hector J, Davis, Ernest, and Morgenstern, Leora. The winograd schema challenge. In *Aaai spring symposium: Logical formalizations of commonsense reasoning*, volume 46, pp. 47, 2011.

Oh, Junhyuk, Guo, Xiaoxiao, Lee, Honglak, Lewis, Richard L, and Singh, Satinder. Action-conditional video prediction using deep networks in atari games. In *Advances in Neural Information Processing Systems*, pp. 2863–2871, 2015.

Perez, Ethan, Strub, Florian, De Vries, Harm, Dumoulin, Vincent, and Courville, Aaron. Film: Visual reasoning with a general conditioning layer. *arXiv preprint arXiv:1709.07871*, 2017.

Russakovsky, Olga, Deng, Jia, Su, Hao, Krause, Jonathan, Satheesh, Sanjeev, Ma, Sean, Huang, Zhiheng, Karpathy, Andrej, Khosla, Aditya, Bernstein, Michael, et al. Imagenet large scale visual recognition challenge. *International Journal of Computer Vision*, 115(3):211–252, 2015.

Santoro, Adam, Raposo, David, Barrett, David G, Malinowski, Mateusz, Pascanu, Razvan, Battaglia, Peter, and Lillicrap, Tim. A simple neural network module for relational reasoning. In *Advances in neural information processing systems*, pp. 4974–4983, 2017.

Socher, Richard, Perelygin, Alex, Wu, Jean, Chuang, Jason, Manning, Christopher D, Ng, Andrew, and Potts, Christopher. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 conference on empirical methods in natural language processing*, pp. 1631–1642, 2013.