Neural Module Networks

Cheng Guan

June 9, 2018

1 Introduction

This paper describes an approach to visual question answering based on a new model architecture that we call a neural module network (NMN). This architecture makes it possible to answer natural language questions about images using collections of jointly-trained neural modules, dynamically composed into deep networks based on linguistic structure.

Concretely, given an image and an associated question (e.g. *where is the dog?*), we wish to predict a corresponding answer (e.g. *on the couch*, or perhaps just *couch*) (Figure 1). The visual question answering task has significant significant applications to human-robot interaction, search, and accessibility, and has been the subject of a great deal of recent research attention [1, 3, 4]. The task requires sophisticated understanding of both visual scenes and natural language. Recent successful approaches represent questions as bags of words, or encode the question using a recurrent neural network [3] and train a simple classifier on the encoded question and image. In contrast to these monolithic approaches, another line of work for textual QA and image QA [2] uses semantic parsers to decompose questions into logical expressions. These logical expressions are evaluated against a purely logical representation of the world, which may be provided directly or extracted from an image .

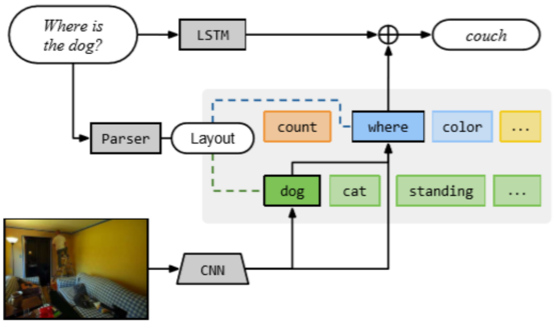

Figure 1: A schematic representation of our proposed modelłthe shaded gray area is a neural module network of the kind introduced in this paper. Our approach uses a natural language parser to dynamically lay out a deep network composed of reusable modules. For visual question answering tasks, an additional sequence model provides sentence context and learns common-sense knowledge

In this paper we draw from both lines of research, presenting a technique for integrating the representational power of neural networks with the flexible compositional structure afforded by symbolic approaches to semantics. Rather than relying on a monolithic network structure to answer all questions, our approach assembles a network on the fly from a collection of specialized, jointly-learned modules (Figure 1). Rather than using logic to reason over truth values, the representations computed by our model remain entirely in the domain of visual features and attentions.

Our approach first analyzes each question with a semantic parser, and uses this analysis to determine the basic computational units (attention, classification, etc.) needed to answer the question, as well as the relationships between these units. In Figure 1, we first produce an attention focused on the dog, which passes its output to a location describer. Depending on the underlying structure, these messages passed between modules may be raw image features, attentions, or classification decisions; each module maps from specific input to output types. Different kinds of modules are shown in different colors; attention-producing modules (like dog) are shown in green, while labeling modules (like where) are shown in blue. Importantly, all modules in an NMN are independent and composable, which allows the computation to be different for each problem instance, and possibly unobserved during training. Outside the NMN, our final answer uses a recurrent network (LSTM) to read the question, an additional step which has been shown to be important for modeling common sense knowledge and dataset biases .

References

[1] L. Ma, Z. Lu, and H. Li. Learning to answer questions from image using convolutional neural network. In *AAAI*, 2016. 1

[2] M. Malinowski and M. Fritz. A multi-world approach to question answering about real-world scenes based on uncertain input. In *NIPS*, 2014. 1

[3] M. Malinowski, M. Rohrbach, and M. Fritz. Ask your neurons: A neural-based approach to answering questions about images. In *ICCV*, 2015. 1

[4] M. Ren, R. Kiros, and R. Zemel. Exploring models and data for image question answering. In *NIPS*, 2015. 1