Review on “Deep Compositional Question Answering with Neural Module Networks”

Authors: J. Andreas et Al.

# Short Summary

In this paper, the authors propose a novel approach for dynamically building networks for specific tasks using modules. This approach is called Neural Module Network (NMN) and the authors of this paper demonstrate that it achieves state of the art performance in visual question answering.

The approach uses both neural networks and semantic parsers. First, each question is broken down into computational units describing the relationship between words in the sentence. This is done via a semantic parser whose output is processed into a structured query. The query is then interpreted as a network layout allowing for an attention-based approach utilizing distinct modules. For the purposes of this paper, the authors introduce five basic module types with distinct inputs and outputs: Attention, Re-Attention, Combination, Classification and Measurement. The modules are independent but can be trained jointly with batches containing questions with similar compositional structures.

In addition to the modular network described above, their approach to visual question answering makes use of LeNet as an image feature extractor and a 1024 unit LSTM for interpreting the question and processing semantic regularities that are lost at the semantic parser. The final model achieves 55.1 (not clear what the unit of measurement is) on the VQA test-server.

While the NMN is applied to a specific task in the paper, the authors stress that their framework can be taken outside of this context and applied in any multitask learning setting. The authors also introduce a new dataset (SHAPES) to showcase the performance of the NMN at understanding compositional questions about abstract shapes. The NMN significantly outperforms previous approaches by a margin of approximately 25%.

# Main Contributions

1. Introduced a general purpose framework for learning collections of neural modules which can be dynamically assembled into network layouts called Neural Module Networks
2. Achieved state-of-the-art on VQA
3. Introduced a new dataset SHAPES for visual question-answering

# High-Level Evaluation of Paper

The idea behind this paper is very cool. The authors have found a novel way of building a dynamic neural network suited for dealing with compositional problems. Further, they have discussed its strengths, limitations and implementation details comprehensively.

However, this does not mean the paper is without flaws. Firstly, there were quite a few grammatical mistakes which made it difficult to understand certain sentences in the paper. In one of the diagrams, they pose a question about a dog but then continue their discussion regarding a cat. This indicates to me that the paper was potentially rushed out. It could have also benefitted from some additional diagrams about the parse trees showing the leaves, internal nodes and root as those directly relate to the dynamically built network layout. Furthermore, the authors only tested model performance on two datasets with minimal ablation. The metrics were not well explained and it wasn’t what the numbers presented in tables represented (though it was presumably classification accuracy). Given how novel and potentially powerful their approach is, I would have liked to see a little bit more in the way of evaluation and results. Curiously, this paper also contains very few mathematical concepts which made it easier to follow, but also left important details like the loss function used in training. While these concepts are discussed at a high-level, the paper could benefit from some more concrete details.

The authors also mentioned the major limitation with the current model were issues with the semantic parser. This is a direction for future study.