Review on “Learning to Reason: End-to-End Module Networks for Visual Question Answering”

Authors: R. Hu, J. Andreas et Al.

# Short Summary

In this paper, the author J. Andreas builds on his previous work on Neural Module Networks. The primary purpose here is to develop an end-to-end network that can directly predict instance-specific network layouts without relying on a parser, which was a source of error identified in the previous paper. The new network is called End-to-End Module Networks (N2NMN) which combines concepts from reinforcement learning (imitating expert demonstrations) and recurrent neural networks.

The model works by first predicting a sequence of structure actions and then a sequence of attentive actions from an embedding of the question. The first step requires predicting the layout of the network while the latter step primarily involves selecting words from the question to focus on. When combined, these two steps can output an answer to a given question. The list of modules describes in this paper is greater than the last and as previously described, the modules do not have any pre-enforced behaviour.

The crux of this approach is in predicting the layout-policy with a sequence-to-sequence RNN. Searching over the layout search space, a large syntax tree is constructed over all possible layouts. The layout expressions can be mapped using a post-order traversal of the tree. Then, the layout police can be simplified via sequence-to-sequence modelling input questions into the layout expression. During training, this layout policy is jointly learned with neural module parameters.

In experiments, by cloning the expert policy and policy searching, the new N2NMN can achieve 100% accuracy on SHAPES, 64.9 on VQA and 83.7 on CLEVR. These numbers outperform previous state-of-the-art by a fair margin.

# Main Contributions

1. Method for learning a layout policy that dynamically predicts a network structure
2. Novel module parametrization that uses soft attention over question words
3. Achieve state of the art performance on SHAPES, VQA and CLEVR datasets

# High-Level Evaluation of Paper

Most of my critiques of the original NMN paper are addressed in this paper. Unlike the prior work, which felt rushed in its mistakes and minimalism, this work goes into detail about the underlying theory of N2NMN. A lot of emphasis has been placed on the underlying loss functions, training the model, implementation details, evaluation setup and so on. Furthermore, an example parse tree is shown to visualize what the layout policy is learning. Ablation studies and comparisons to three distinct datasets bolster the confidence in these results. Overall, this paper provides a significant improvement over the original in both presentation and results. The novel model also performs very well on CLEVR, which is indeed indicative of it learning reasoning as opposed to simply memorizing a data distribution.

In terms of limitations, I did notice that this paper does not detail failure cases or model limitations in as much depth as the prior work. However, that’s a fairly minor complaint.