Deep Neural Inspection

Kevin Lin, Dennis Wei, James Xue

Project Background and Hypotheses

The Wu lab at Columbia has researched possible ways to better determine the features deep neural networks learn in order to complete tasks. The lab has begun to build a system to accomplish this using tools such as finite state machines and hand built feature functions, but the project has yet to be extended to non-toy projects.

Fundamentally, our hypothesis is that when operating on natural language, neural networks learn features similar to well studied linguistic measures, and that we can find correlations between hand crafted features and hidden states in the DNN.

What work must be done and how it will be divided amongst the team

- Integrate Pytorch framework into current hidden state extraction mechanism OpenNMT Neural Translation Model (https://github.com/OpenNMT/OpenNMT-py) (Dennis+Kevin)
- Create feature functions for natural language using external library such as Stanford CoreNLP (James)
- Infer causal relations between hidden states and feature function values (Kevin+Dennis)
  - optimize neuron control, subsampling
- System scalability for larger networks (Kevin+WuLab)
  - approx. correlation with DFT

How the hypotheses will be evaluated

We will determine whether our hypothesis is correct by searching for correlations between the pre-trained model's hidden states while running over test data and the values of hand constructed feature functions.

- If we find that certain neurons activate upon encountering specific features, we can say with some certainty that the DNN has learned these attributes
- We can extend our examination of these correlations by manually inserting perturbations that would hypothetically cause these neurons not to activate
- More effective correlation search would lead to noticeable changes in system speed

Resources needed to complete project

- None