Research

So here I am going to write briefly about what I've done and what I am working on, just some descriptions and essential results to describe each of the papers, so that during the tech interview we know where to start with

Past

When I was taking my undergraudate in Zhejiang University, I was working on disentanglement[1]) and also briefly on generative modelling[2].

I then came to study with Martin Ester, and I have mainly worked on the topic of interpretability.

An active learning approach for model interpretation

Lu and Ester (2019a)
Click to expand
Model interpretation is a task where we try to approximate a blackbox model f (e.g. DNN) with a simpler and more interpretable model g (In my case, g is a decision rule set). The key point is that instead of training g to approximate f on a given dataset, we can use f as an oracle to query arbitray data samples. So we turn this into an active learning problem, where the new queried samples will improve the faithfulness (how well g approximate f). As for techniques, the skeleton of the algorithm is a local search algorithm, but we augment it by employing tricks from pure-exploration

Current

After working on the Neural Disjunctive form however, I find two ways of plan.

1. Enabling this learning approach for a wider range of models, i.e., supporting more flexible language beyond propositional logic.
2. Further investigate the symbolic grounding problem.

More flexible language

Defining a flexible language, or a domain-specific language is not difficult, the difficult part is how we can efficiently optimize it.

Symbolic grounding problem.

Lu, Jialin. 2019. "Revisit Recurrent Attention Model from an Active Sampling Perspective." In *NeurIPS 2019 Neuro AI Workshop*.

Lu, Jialin, and Martin Ester. 2019a. "An Active Approach for Model Interpretation." In *NeurIPS 2019 Human-Centric Machine Learning Workshop*.

———. 2019b. "Checking Functional Modularity in DNN By Biclustering Task-specific Hidden Neurons." In *NeurIPS 2019*