Week 7 Questions

Training language models to follow instructions with human feedback

Brief Summary: They instruction-tune GPT3 models with prompts generated by humans. After this they have their fine-tuned models generate responses to prompts. Human labellers rank these model responses. After this, a reward model is trained to score model responses based on this labelled data. Next they further train their fine-tuned model with RL where the rewards are given by the reward model.

1. Describe the basic 3 steps of the RLHF pipeline.
2. What models are used in each of the 3 steps of the RLHF pipeline?
3. Explain the loss function for the reward model (on page 8).
4. On page 9, they mention that they add a KL penalty to mitigate over-optimization of the reward model. Describe this penalty, what they mean by over-optimization, why over-optimization might be a problem and how this penalty might mitigate over-optimization of the reward model.
5. There are multiple parts of the RLHF pipeline that require human intervention. What parts of the RLHF pipeline do you think could be automated? For example, could we get good results by having language models generate instruction-tuning data instead of humans designing the data?

Direct Preference Optimization: Your Language Model is Secretly a Reward Model

Brief Summary: Their goal is to get a model to have the performance associated with RLHF training, but without using RL. To do this, they reformulate the reward model as a function of the policy (the policy is the language model they are training). With this reformulation, they can optimize the policy directly on human preference data. This removes the need for training a separate reward model and using RL.

1. What are the difficulties of the usual RLHF method?

2. How do they derive equation 2?
3. Describe the steps to use DPO.
4. What difficulties of RLHF does DPO avoid?
5. In some of their experiments, the preference data is not generated by their models, but by other models they do not have access to. How might this affect their results? How might it affect the usual RLHF pipeline differently?