Optimizing Human Feedback in Reward Modelling

Opara, Chioma (School: Nigerian Tulip International Colleges, Abuja)

Reinforcement learning (RL) has produced exceptional results in various game environments and several real-world situations, including traffic light control and autonomous driving. However, its scope of application is limited by its requirement of a hardcoded reward function. Various techniques have been developed to tackle this problem but for this project, I limited my focus to reward modelling. I sought to optimize the amount of human feedback required by a reward model as an improvement on previous reward modelling algorithms. I explored two different methods of reward modelling in the cart-pole environment: an uncertainty-based method and a random selection method, which I called the baseline. Then, I compared their performance to the true reward function. In the first set of experiments, the uncertainty-based method tended to achieve a higher performance level in terms of rewards than the baseline method when each received a fixed amount of feedback. However, both still achieved a low-performance level. In the second set of experiments, I shortened the length of each episode and increased the number of episodes per experiment. This increased the performance level of both methods tremendously with the uncertainty-based method still tending to be superior to the baseline method. The estimated rewards from both methods also showed a strong positive correlation to the true rewards. By showing that the uncertainty-based method was better than the baseline method, this project not only allows efficient use of resources in reward modelling but also expands the scope of RL.

Awards Won: