CSE 256 Analysis Proposal
Vince Rothenberg A16534656

InstructGPT

The paper "Training Language Models to Follow Instructions with Human Feedback" by Ouyang et al. addresses the challenge of aligning large language models with human intentions without relying on predefined reward functions. Utilizing direct human feedback as a dynamic reward system, this study demonstrates a novel method to guide the training of reinforcement learning systems effectively. This approach has shown promise across varied tasks like Atari games and simulated robotics with reduced human oversight, suggesting a cost-effective method for enhancing model alignment. Key innovations include the development of a reward model based on human preferences and the implementation of a policy optimization algorithm to maximize human-centered rewards. Results indicate this method aligns models more closely with human preferences and also maintains high efficiency with significantly reduced need for human input.

RLHF

The paper "Deep Reinforcement Learning from Human Preferences" by Christiano et al. develops a method for training reinforcement learning (RL) agents using minimal human feedback instead of predefined reward functions. This approach efficiently teaches agents to perform complex tasks, such as Atari games and simulated robotics, by requiring feedback on less than 1% of interactions, reducing human oversight needed. Key achievements include the ability to quickly train agents on new behaviors in about an hour, and demonstrating that these agents can achieve and sometimes surpass the performance of traditional RL methods. This is done through human comparative judgment, which uses short video clips for feedback rather than numerical scores, aligning more closely with natural human decision-making processes. The study underscores the potential of integrating human preferences into RL training, offering a practical and scalable method for advancing RL applications in complex real-world settings.

Modifying Behavior

The paper "Plug and Play Language Models: A Simple Approach to Controlled Text Generation" by Dathathri et al. presents the Plug and Play Language Model (PPLM) which modifies pre-trained language models like GPT-2 to generate text with specific attributes such as sentiment or topic, without re-training. By integrating a pre-trained LM with attribute classifiers like bag of words or trained discriminators, PPLM adjusts the model's hidden states in real-time, allowing dynamic control over text generation. Evaluations using automated metrics and human annotations confirm that PPLM aligns text with desired attributes while maintaining fluency, and performs on par or better than other models like CTRL and GPT-2 fine-tuned models. PPLM's adaptability and minimal training requirement make it suitable for various applications, including text detoxification and controlled story writing. Advanced techniques such as minimizing Kullback–Leibler divergence and using geometric mean fusion ensure the text remains coherent while controlled, marking PPLM as an innovative tool in controlled text generation.

Training LMs to Follow Instructions
The paper "Cross-Task Generalization via Natural Language Crowdsourcing Instructions" by Mishra et al. explores how language models (LMs) can generalize across a wide range of tasks through natural language instructions. It introduces the NATURAL INSTRUCTIONS dataset, which includes 61 NLP tasks with human-authored instructions and 193,000 instances, all standardized into a unified schema. Using generative models like BART and GPT-3, which are adapted to integrate task-specific instructions, the study shows a 19% improvement in task performance on unseen tasks, demonstrating the critical role of instructions in enhancing model adaptability. Despite these gains, a significant performance gap suggests extensive room for further research. This study highlights the potential of natural language instructions to broaden LMs' capabilities, pointing towards more flexible and effective AI development.

Evaluating Harms
The paper "On the Opportunities and Risks of Foundation Models" by Bommasani et al. examines the impact of large-scale pre-trained models like BERT, GPT-3, and CLIP, highlighting their emergence, homogenization, and societal implications. These “foundation models” are characterized by their ability to learn complex behaviors from vast datasets and adapt across diverse tasks, creating efficiencies but also potential vulnerabilities due to shared biases and failure modes. The study focuses on the power of foundation models across language, vision, reasoning, and interaction, but also their considerable risks, including exacerbating inequities, privacy breaches, and misused potential. The paper calls for interdisciplinary research and evaluation frameworks to better understand and mitigate these risks, advocating for the development of responsible AI that is ethically aligned. It stresses the importance of integrating technological, ethical, and social science perspectives to ensure that foundation models contribute positively to society while addressing their inherent challenges and dangers.

Codebase: InstructGPT Github Repository
This repository provides details on fine-tuning GPT-3 via supervised reinforcement learning from human feedback (RLHF). Details on the development of OpenAI’s InstructGPT models (1.3B, 6B, 175B parameters), aimed to align GPT-3 with human preferences, focused on reducing toxicity and enhancing truthfulness. InstructGPT is designed to improve the accuracy of GPT-3 in following user instructions and generating content within ethical guidelines. The training combines human annotations with RL, using labeler preferences to refine outputs.

Datasets Internet texts and a specially labeled dataset including prompts from the OpenAI API. Evaluation set will consist of CSV files from automatic-eval-samples folder, covering summarization, translation, QA, and toxicity assessment tasks.

Evaluation metrics will include comparisons of my own preference for InstructGPT versus GPT-3 outputs. Task specific metrics will include performance on NLP tasks measured by metrics like BLEU, F1, and ROUGE scores.

References

Bommasani, R., Hudson, D. A., Adeli, E., Altman, R., Arora, S., von Arx, S., ... & Liang, P. (2021). On the opportunities and risks of foundation models. _arXiv preprint arXiv:2108.07258_.

Christiano, P. F., Leike, J., Brown, T., Martic, M., Legg, S., & Amodei, D. (2017). Deep reinforcement learning from human preferences. _Advances in neural information processing systems_, _30_.

Dathathri, S., Madotto, A., Lan, J., Hung, J., Frank, E., Molino, P., ... & Liu, R. (2019). Plug and play language models: A simple approach to controlled text generation. _arXiv preprint arXiv:1912.02164_.

Mishra, S., Khashabi, D., Baral, C., & Hajishirzi, H. (2021). Cross-task generalization via natural language crowdsourcing instructions. _arXiv preprint arXiv:2104.08773_.

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., ... & Lowe, R. (2022). Training language models to follow instructions with human feedback. _Advances in neural information processing systems_, _35_, 27730-27744.