Research Track: Language Model Preference Learning as Retriever Optimization

Wei Xiong (wx13@illinois.edu)

[1]University of Illinois Urbana-Champaign

1 Introduction

[Research Question] The preference learning (reinforcement learning from human feedback, RLHF) relies on the Bradley-terry ranking model to construct the reward signals, which is closely related to the ranking method in the literature information retrieval. We are interested in borrowing more advanced techniques from the information retriever optimization literature and study their connections.

[Significance] Preference learning is the final stage of getting a natural language product ready for deployment and is the key to the success of Chat-GPT. Therefore, developing more efficient algorithms from the perspective of retriever optimization is an exciting topic given their closely related relationship.

[Novelty] The standard approach of this area (Bai et al., 2022; Ouyang et al., 2022; Rafailov et al., 2023) focuses on the pairwise comparison. However, in practice, one typically samples multiple responses per prompt (e.g. n = 8) and extract only one pair of data. Therefore, the generated data is not fully used. One natural idea is we can extend the framework to a list-wise comparison to improve the data utilization. The key novelty and key problem of this research project is how we can use more negative samples to accelerate the pairwise preference learning.

[Approach] We will focus on designing new variants of the direct preference optimization (DPO) (Rafailov et al., 2023) algorithm, which is derived from the pairwise comparison. We may follow the mathematical derivation process but starting from a more general ranking model so deriving a more general algorithm. Then, we will conduct real-world alignment experiments to validate its effectiveness.

[Evaluation] There are several standard benchmarks (Alpaca eval, MT-bench). We can use Llama / Mistral models to conduct real-world experiments and test the aligned models on these benchmarks to evaluate the methods.

[Timeline] I will spend two weeks to formalize the idea and also derive the mathematical formulation of the algorithms. Then, I will implement and conduct experiments in the next two weeks.

[Task division] I will be responsible for the whole project.

References

Bai, Y., Jones, A., Ndousse, K., Askell, A., Chen, A., DasSarma, N., Drain, D., Fort, S., Ganguli, D., Henighan, T., et al. (2022). Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862.*

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., et al. (2022). Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744.

Rafailov, R., Sharma, A., Mitchell, E., Ermon, S., Manning, C. D., and Finn, C. (2023). Direct preference optimization: Your language model is secretly a reward model. *arXiv preprint arXiv:2305.18290*.