Literature Review: Direct Preference Optimization with an Offset

Arthur CHANSEL | 324265 | arthur.chansel@epfl.ch
MoreNectarLessPoison

1 Summary

ETH Zurich published *Direct Preference Optimization with an Offset* (Amini et al., 2024). This paper suggests an improvement to an earlier Stanford University's publication : *Direct Preference Optimization: Your Language Model is Secretly a Reward Model* (Rafailov et al., 2023).

Direct Preference Optimization (DPO) was a very novel idea that largely improved LM alignment to human preferences. Initially we trained LMs with reinforcement learning from human feedback (RLHF) to achieve this. However, this method is very complex for several reasons. First, it has to train a reward model as large as the LM. Thus, it drastically increases model size and is expensive to compute and store. Second, it employs reinforcement learning on the reward model to update LM's parameters. RL algorithms are very sensitive to hyperparameters choice, then tough to tune. DPO is an equivalent formulation with remarkable results. It computes the estimated reward and considerably reduces the complexity.

Nevertheless, DPO is not perfect and ETH searchers focused their study on a specific observation. Given paired responses and a human preference, DPO only considers the preference ordering but not the preference extent. This quality difference is a crucial information we don't take benefit from, and has the potential of widely enhance the results. The study proposes generalizing DPO to Direct Preference Optimization with an Offset (ODPO). The idea is to use the difference of preference extent as an offset to likelihood.

In the end, this study demonstrates positive performance. It enhanced the existing strategy from a simple observation on its fundamental mechanics. It is a notable advance for future fine-tuning of LM, toward more efficient alignment with human preferences.

2 Strengths

This paper contributes a notably efficient upgrade to an existing strategy. The goal is precise and all the thoughts and tasks are clearly presented. In my opinion this is an excellent idea, backed by proven efficiency. Moreover the work is well documented and thus should be reproducible.

To advance the main idea, the study verifies DPO strong assumptions. The research is built on this first strategy, assuming the data is modeled by Bradley-Terry model. However it is aware of the subtleties lacking in this model, and goes further to address them. To do that, it adds a measure to capture the preference extent, i.e. the offset. Related work presented in the appendix uses the same idea, without considering the dependency of the offset with the preference responses. Thus, the ODPO paper goes more in depth for better performances.

The new strategy using the offset (ODPO) is presented, and the study verifies well the limitations as well as the analogies to the previous approach (DPO). In fact, when the offset is zero or when its scaling hyperparameter α is zero, ODPO is equivalent to DPO.

The design choices for the offset are presented and justified. The paper suggests 3 alternatives to the scaling function that seem reasonable. Properties for every alternative are presented, and one of them stands out from the other. Nevertheless, whatever alternative is used, they all outperform the DPO approach. Considering the hyperparameter, it is tested over the 0-1 range, and the results are fairly discussed.

The previous DPO strategy comes with a KL regularization term β. The results are always computed for multiple values of this term. The study explains well the choice of this range that seems reasonable.

Every research result is presented on a 2D scatter plot of the KL divergence against the probability.

The comparison metric to evaluate the results is the Pareto frontier, which is really appropriate to evaluate the trade-off between multiple conflicting objectives, i.e. the two axis.

To achieve any preference optimization method, we need preference data. Here the study choose human evaluations 7-points Likert scale. This is a truly rational choice, since evaluating a preference isn't a binary strict choice, but might be more shaded. An odd number of points (here 7) also allows a neutral opinion, and does not force the evaluation to be either positive or negative.

Finally, every experiment results are very detailed. The choice of the models are reported, as well as the sampling method. In the end this study

3 Weaknesses

This paper delivers valuable contributions, however some points could have been improved.

The main contribution of this paper is the generalization of DPO using an offset. Thus the definition of the offset is crucial for a thoughtful deployment. Here, the study presents 3 offset alternatives. However, it is a first approach and there might be room for improvement. Maybe more alternatives for the scaling function f could reveal even better behavior.

Furthermore, the offset definition considers the hyperparameter α. Its value is explored over the range 0 to 1, which higher values are associated with highest rewards, but at the cost of diverging. Then, introducing a new hyperparameter in the method implies more design decisions and another trade-off for multiple conflicting objectives. However, the highest rewards are associated to value $\alpha = 1$. We can notice that it represents the exploratory range boundary. We could consider exploring values in a range beyond 1, to be sure we are not missing any opportunity to improve the method. Additionally, the ablation study for this hyperparameter only considers KL regularization term β at value 0.5 (which actually showed the best results). However, it might be interesting to vary this value over the same range used in the entire paper. All these observations being said, leading a study on the ODPO design choices might be a good idea, to be sure to choose an optimal design.

Experiments are detailed, and the different models are specified. However, the choices are not always justified. Outside of GPT-4 model choice in summarization, which is selected for its correlation with human judgements, other models are only stated but not supported. These details on experiment designs are crucial, so we could express more precise opinion on the motivations for model choices, even though they don't appear to be absurd.

The methods for leading experiments consider multiple random generations. This is very helpful, because is provides confidence intervals for the results.

Experiment datasets are only in English. We could counsider the impact of a language on the results. It could be intereting to lead a study on ODPO with another language, or multilingual data, and measure the impact on the results.

Finally, information is missing for the study to be exactly reproducible. The paper mention random seeds for sampling and bootstrapping, without specifying them.

References

Afra Amini, Tim Vieira, and Ryan Cotterell. 2024. Direct preference optimization with an offset.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D. Manning, and Chelsea Finn. 2023. Direct preference optimization: Your language model is secretly a reward model.