Project Proposal: Enhancing Text-to-Image Models through Direct Preference Optimization

Adi Asija, Tanay Nayak

Motivation:

The pursuit of human-aligned artificial intelligence has made significant strides in recent years, as evidenced by the innovative works presented at NeurIPS 2023. The ImageReward paper laid the foundation by creating a dataset that captures nuanced human preferences in the context of text-to-image models, offering a valuable resource for subsequent model optimization efforts. Building upon this, the DPOK paper showcased a groundbreaking approach to fine-tuning large unsupervised language models (LMs) through reinforcement learning, using human preferences as a guide without deviating from the original model's inherent capabilities, effectively harnessing the potential of the ImageReward dataset. Meanwhile, the Direct Preference Optimization (DPO) framework emerged as a game-changer, heralded for its stability, efficiency, and computational simplicity. DPO's ability to fine-tune LMs without the need for excessive sampling or intensive hyperparameter tuning, and its superiority over Proximal Policy Optimization-based Reinforcement Learning from Human Feedback (PPO-based RLHF) in controlling sentiment generation, offers a promising new direction for model training. Our project seeks to bridge the gap between these text-based advancements and the realm of visual data. We propose to apply the DPO framework to the ImageReward dataset, venturing into the less chartered territory of image preference alignment. This novel application not only extends the use case of DPO but also challenges the adaptability of text-oriented optimization methods to the complexities of visual data interpretation, setting the stage for a potential breakthrough in multimodal AI research.

Relevant Existing Literature:

ImageReward: Learning and Evaluating Human Preferences for Text-to-Image Generation[1]
Introduces the ImageReward dataset, designed to finely capture and quantify human aesthetic preferences. This tool is critical for developing text-to-image models that not only exhibit technical prowess but also resonate more deeply with human notions of visual appeal and relevance.

DPOK: Reinforcement Learning for Fine-tuning Text-to-Image Diffusion Models[2]
Describes a novel reinforcement learning approach called DPOK, which fine-tunes text-to-image diffusion models by maximizing a reward function based on human preferences. This technique delicately balances adherence to the original model's knowledge while adapting to the nuanced tastes reflected in the ImageReward dataset.

Direct Preference Optimization: Your Language Model is Secretly a Reward Model[3]
Explores the Direct Preference Optimization (DPO) framework, which offers a computationally efficient method for aligning language models with human preferences. The DPO approach streamlines the fine-tuning process, obviates the need for laborious hyperparameter optimization, and demonstrates superior performance over traditional PPO-based models in understanding and generating sentiment-aligned content.

---

[1] Jiazheng Xu et al.
[2] Ying Fan et al.
[3] Rafael Rafailov et al.

Project Plan

Hypothesis & Expected Outcome:

Our hypothesis posits that the DPO framework, while originally devised for text-based applications, can be effectively adapted to the domain of image-based models, leveraging the nuanced human feedback encapsulated in the ImageReward dataset. We anticipate that adapting DPO for image-based applications will not only be feasible but will also yield performance on par with or surpassing the results achieved by the DPOK methodology.

The expected outcomes of this research are multifaceted. Firstly, we aim to confirm the claims of computational efficiency and performance stability put forth by the DPO framework when applied to image-based contexts. Secondly, by transferring the DPO methodology to a new domain, we seek to establish a broader understanding of its generalizability and to evaluate whether the framework can maintain its efficacy without extensive hyperparameter tuning or reliance on sampling from the model. Furthermore, we expect that this adaptation will validate DPO's potential in enhancing sentiment control and improving response quality in visual tasks, thereby offering a robust alternative for fine-tuning text-to-image models in alignment with human preferences. Ultimately, our goal is to test the feasibility of this approach and verify the claims made by Rafailov et al.[3], exploring whether the benefits observed in text-centric models hold true in the realm of visual data.

Experiments:

1. **DPO Adaptation for Images:** Adaptation of the DPO algorithm to accommodate image data, including incorporation of image feature extraction and reward modeling based on visual content.
2. **Fine-Tuning Text-to-Image Models:** Application of the adapted DPO to fine-tune a pre-existing text-to-image model, monitoring performance against human aesthetic preferences.
3. **Baseline Comparison:** Comparative analysis against models optimized using the DPOK approach, measuring image quality, textual prompt alignment, and human preference adherence.

Success Metrics:

1. **Quantitative Evaluation:** Employment of automated evaluation metrics like Inception Score, Fréchet Inception Distance, and CLIP Score to objectively assess image quality and relevance.
2. **Human Evaluation:** Assembly of an evaluator panel to rate images on aesthetic appeal, prompt relevance, and overall satisfaction, conducted in a double-blind study format.
3. **Statistical Analysis:** Statistical testing to determine the significance of differences observed between the model performances.

Dataset:

ImageReward dataset on HuggingFace trains AI models to understand human preference. It features 137,000 comparisons where humans rated different image outputs for a given text prompt. This allows AI to learn what qualities make a generated image most successful based on human judgment.

Halfway Milestone:

By the midpoint of our project timeline, we aim to achieve two primary goals:

1. **Replication of DPOK Results:** Successfully replicate the findings from the DPOK paper to establish a baseline for our research. This entails reproducing the original experiments to confirm we can achieve similar performance metrics, serving as a solid foundation for further innovation.
2. **Engineering a DPO Pipeline:** Develop and test a preliminary Direct Preference Optimization (DPO) pipeline. This will involve designing the architecture for integrating DPO within our text-to-image model, focusing on computational efficiency and the ability to iteratively learn from human feedback.

Compute Estimation

For the initial stages, including replication of DPOK results and the development of a DPO pipeline, we estimate a requirement of 40 GB of GPU VRAM as a baseline. This estimation is based on the complexity of generative models and the iterative nature of preference learning, which, even when optimized, necessitates significant memory for processing and training. Access to GPUs with at least 40 GB of VRAM, such as NVIDIA's A100 or V100, would provide a solid foundation for executing these experiments while allowing for a degree of computational headroom to account for unexpected demands or the opportunity to scale up model complexity as needed.

References

[1] Xu, J., Liu, X., Wu, Y., Tong, Y., Li, Q., Ding, M., Tang, J., & Dong, Y. (2023) ImageReward: Learning and Evaluating Human Preferences for Text-to-Image Generation. In *Advances in Neural Information Processing Systems (NeurIPS)*, pp. Cambridge, MA: MIT Press.

[2] Fan, Y., Watkins, O., Du, Y., Liu, H., Ryu, M., Boutilier, C., Abbeel, P., Ghavamzadeh, M., Lee, K., & Lee, K. (2023) DPOK: Reinforcement Learning for Fine-tuning Text-to-Image Diffusion Models. In *Advances in Neural Information Processing Systems (NeurIPS)*, pp. Cambridge, MA: MIT Press.

[3] Rafailov, R., Sharma, A., Mitchell, E., Ermon, S., Manning, C. D., & Finn, C. (2023) Direct Preference Optimization: Your Language Model is Secretly a Reward Model. In *Advances in Neural Information Processing Systems (NeurIPS)*, pp. Cambridge, MA: MIT Press.