DSSV 2023 Data Science, Statistics & Visualisation
ECDA 2023 European Conference on Data Analysis

Learning from Human Feedback for Fine-tuning Text-to-Image Models

Kimin Lee[1]

[1]Google Research

Deep generative models have shown impressive results in text-to-image synthesis [1, 2, 3]. However, current text-to-image models often generate images that are inadequately aligned with text prompts [4, 5, 6]. In this talk, I will present introduce a fine-tuning method for aligning such models using human feedback. First, I will introduce a simple yet efficient fine-tuning method based on supervised learning. Our method consists of the three stages: First, we collect human feedback assessing model output alignment from a set of diverse text prompts. We then use the human-labeled image-text dataset to train a reward function that predicts human feedback. Lastly, the text-to-image model is fine-tuned by maximizing reward-weighted likelihood to improve image-text alignment. Our method generates objects with specified colors, counts and backgrounds more accurately than the pre-trained model. We also analyze several design choices and find that careful investigations on such design choices are important in balancing the alignment-fidelity tradeoffs. Our results demonstrate the potential for learning from human feedback to significantly improve text-to-image models.

I will also share investigations on reinforcement learning (RL) to fine-tune the text-to-image models. Specifically, I will explain how we formulate fine-tuning task as a RL problem specially designed for diffusion models. We then update the pre-trained image-text diffusion models to maximize scores of a reward model human feedback using a policy gradient algorithm. We analyze several design choices (such as KL regularization, value learning and balancing regularization coefficient) and find that careful investigations on such design choices are important in RL fine-tuning. We demonstrate that RL fine-tuning is more effective in improving the pre-trained model compared to supervised fine-tuning in terms of both alignment and fidelity.

[1] C. Saharia, W. Chan, S. Saxena, L. Li, J. Whang, E. Denton, S. K. S. Ghasemipour, B. K. Ayan, S. S. Mahdavi, R. G. Lopes, *et al.*, "Photorealistic text-to-image diffusion models with deep language understanding," in *Advances in Neural Information Processing Systems*, 2022.

[2] A. Ramesh, P. Dhariwal, A. Nichol, C. Chu, and M. Chen, "Hierarchical text-conditional image generation with clip latents," *arXiv preprint arXiv:2204.06125*, 2022.

[3] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022.

[4] K. Lee, H. Liu, M. Ryu, O. Watkins, Y. Du, C. Boutilier, P. Abbeel, M. Ghavamzadeh, and S. S. Gu, "Aligning text-to-image models using human feedback," *arXiv preprint arXiv:2302.12192*, 2023.

[5] Y. Kirstain, A. Polyak, U. Singer, S. Matiana, J. Penna, and O. Levy, "Pick-a-pic: An open dataset of user preferences for text-to-image generation," *arXiv preprint arXiv:2305.01569*, 2023.

[6] J. Xu, X. Liu, Y. Wu, Y. Tong, Q. Li, M. Ding, J. Tang, and Y. Dong, "Imagereward: Learning and evaluating human preferences for text-to-image generation," *arXiv preprint arXiv:2304.05977*, 2023.