Distributional Alignment with LoRA to Larger Models

{joshua.liao, bplate, carolinewu01, patrickgu}@berkeley.edu

Self-Review

- What is the main goal of the project?
  - Our goal is to examine if we can fine-tune a small stable diffusion model on purely synthetic data from a larger model, mainly DALL-E 2, and align the output distribution to match the fine-tuning dataset. We apply low rank adaptation (LoRA), experiment with various hyperparameters, and measure the Kernel Inception Distance (KID) to evaluate our results.

- What are the main claims of the project?
  - We claim that there is a noticeable and measurable impact of fine-tuning using purely synthetic data. Anecdotally, even 100 steps of LoRA on just 20 training photos on DALL-E has interesting results. Below, we show the evolution of the model, which seems to learn to output smaller eyes, a light reflection in the eye, and a more human / pinched face.

Fig 1. On the left, the original stable diffusion model was prompted:
“A realistic painting of a cat in Baroque art style”.
On the right, the fine-tuned model was prompted with additional style tokens.
“A realistic painting of a cat in Baroque art style <s1><s2>”

- What are the experiments? What are the evaluation protocol, data, and task?
  - Our experiments are all text to image prompting of stable diffusion, and our data is generated using DALL-E 2. Our evaluation protocol is to query the new model at its checkpoints, and measure the KID to the DALL-E training distribution, as well as measure the KID to the original stable-diffusion distribution.

- Specifically, our experiments test how each hyperparameter for the fine-tuning algorithm LoRA affects the ability to align the distributions. We vary the number of training steps, the rank of the learned residual matrices, and the number of examples that are shown to the model for fine-tuning.

➢ How do the experiments support the goal/claims of the paper?
- The experiments show different levels of efficacy and the relative relationship between each hyperparameter and the effect on the output distribution.

➢ Are any of the limitations discussed in the paper?
- One limitation of the paper is the lack of compute, which results in a lack of variety in training prompts and data.

➢ What are the strengths of the paper?
- The paper is direct and consistent; we use a small amount of training data and our results are reproducible with seeds.

➢ What are the weaknesses of the paper?
- One weakness of the paper is the lack of compute. Ideally, we would be able to test our experiments on multiple prompts to average out differences across different subjects and art styles.

➢ What is the relevant related work?
- Relevant related work includes the LoRA paper by Hu, 2021. We were inspired by Alpaca 7B. Our specific LoRA implementation uses textual inversion, dreambooth, and pivotal tuning (Gal 2022, Ruiz 2022, Roich 2022 respectively).

➢ Is the paper reproducible?
- By downloading the specific images that we used in training and fixing the specific seeds, you can reproduce the results we show in the paper.

➢ Can you rerun the experiments?
- We have rerun a few experiments as a sanity check, and they can reproduce the same images (such as *Fig 1*).

➢ Suggestions for Further Work
- Further exploration into LoRA can test other hyperparameters such as learning rate, and using a wider variation of prompts; perhaps specific styles or individuals.