CS182 Fall 2023
December 6, 2023

Distributional Alignment with LoRA to Larger Models

{joshua.liao, bplate, carolinewu01, patrickgu}@berkeley.edu
github.com/joshualiao/alignment-lora

Abstract

Large text-to-image models such as DALL-E generate high quality images from novel prompts, such as depicting modern characters in old art styles. However, the cost of large language models is extreme; for example, GPT3.5 has around 175 billion parameters and is estimated to cost $4.6 billion, and required human labeled preferences of model outputs. However, smaller models have been able to fine-tune to achieve more specific tasks with great success [1]. We propose fine-tuning a smaller, stable diffusion model using low rank adaptations (LoRA) to match the outputs of DALL-E. We will utilize the pre-trained stable diffusion model from Hugging Face's stable diffusion v1-5 [2], and fine-tune using an existing LoRA codebase [3] and experiment with LoRA hyperparameters such as rank, training time, and dataset size. LoRA has been shown to successfully fine-tune models to new tasks while modifying only 0.01% of the original model's parameters. We use DALL-E by prompting to generate images of cats in a Baroque art style as our target dataset, using the Kernel Inception Distance [4] as an evaluation metric. Our results show that fine-tuning with artificial data successfully extracts features from a small number of samples of larger models, and also displays novel out of distribution features, implying there may be generative artifacts in the synthetic data.

1 Background

1.1 Overview

The cost of training realistic and expressive models is extremely expensive. Open AI's ChatGPT 3.5 cost around $4.6 million to train. Furthermore, the training process is complex; releasing a model on data farmed from the internet with no supervision has extremely detrimental effects— let alone poor quality learning, there is a large risk of harm, whether through hallucinatory claims or intrinsic biases. As a result, fine-tuning is an important tool, not just to increase the quality of outputs but also to decrease the potential of harm. However, procuring quality data to this end is difficult and naively training on synthetic data can also lead to model collapse.

However, a recent result from Stanford used outputs from ChatGPT3.5 to fine-tune Meta's LLaMA model, producing Alpaca with 7 billion parameters and costing less than $600 (the cost of training and prompting ChatGPT). ChatGPT3.5 in comparison has around 175 billion parameters. This result is extremely appealing— closed source models such as ChatGPT3.5 have proprietary weights, and their training process is inaccessible to the general public. We hope to reproduce this success in the text to image domain, by applying fine-tuning techniques on a small model and leveraging synthetic data to align the output distribution of the smaller model to the more well trained, larger model. By aligning distributions we indirectly improve quality by leveraging the features the larger model has acquired through its fine-tuning process. One way to view our approach is to frame the fine-tuning process as informed "feature extraction" from expert data, which we elaborate upon in our analysis section 3.3.

1.2 Related Work

Our work directly follows Hu and Shen 2021, which introduced low rank adaptation for large language models. Although Stability AI and Hugging Face have adapted LoRA for visual generation, the authors are currently unaware of literature examining the effectiveness of LoRA in visual generation. Although synthetic data has been explored in many domains, our paper explores the specific niche of fine-tuning visual generative models with another model, instead of simulated or spliced data.

2 Implementation Details

2.1 Architecture

We used the LoRA (low rank adaption) code from cloneofsimo [3]. LoRA freezes the pre-trained model weights and instead does fine-tuning by training a residual matrix that is added to the pre-trained models, ie. $W = W + \Delta W$. Previous work has found that learning for a residual for the attention layers of the transformer is effective for tuning, which is what our project uses. Furthermore, the residual ΔW is learned through a product of small matrices, $A*B$, where their dimensions are much smaller than the pre-trained model's weight dimensions. Despite the massively smaller number of parameters (0.01% of the pretrained model), this approximation has proven to be an effective fine-tuning technique.

The LoRA model is applied to Hugging Face's open source Stable Diffusion model, specifically v1-5. For the larger model that we are aligning to, we used Dall-E 2, querying images of dimension 512x512. LoRA is applied in conjunction with dreambooth, textual inversion, and pivotal tuning. To summarize the entire training process, a new token is associated with the sample data, such as "<s1><s2>". We simultaneously learn an embedding for the new token as well as LoRA matrices to mimic the sample data. A prior-preserving loss serves as a regularizer, preventing the model from losing too much of its previous capability. Here on out, we refer to our model as "DALLM" (*Distributional Alignment with LoRA to Larger Model),* to differentiate it from DALL-E and baseline stable diffusion v1-5.

2.2 Datasets and Prompts

We queried Dall-E 2 for our "ground truth" dataset to fine-tune DALLM, as well as the original pretrained stable diffusion model. The prompt we used was "a realistic painting of a cat in baroque art style". For DALLM, the prompt is slightly different, to incorporate the Dreambooth tokens: "a realistic painting of a cat in baroque art style like <s1><s2>".

Fig. 1: Three example images generated by Dall-E 2 to the prompt: "A realistic painting of a cat in Baroque art style."

Fig. 2: Three images generated by stable diffusion v1-5.

As an aside, by inspection the larger Dall-E 2 model produces better outputs than the stable diffusion model, though not necessarily through its size but also its data and fine-tuning process. However, our goal is not necessarily to fine-tune towards "more realistic" or even more human preferred images, but instead to align a lightweight distribution to a larger model's output distribution, to "steal" or "inherit" its expressive power.

2.3 Validation Measure

To evaluate the performance of our model against the target dataset, we use the Kernel Inception Distance (KID) [4] as a quantitative evaluation metric. There are a considerable number of proposed metrics which seek to quantify the difference in features between a set of images (as opposed to comparing the pixel values themselves). We elected to use KID over other proposed distance metrics, like Fréchet Inception Distance (FID) [5], because it has no statistical bias and the loss over random samples from a dataset converges to a mean much faster than other metrics, like FID, achieving consistent results in just a few hundred images rather than tens of thousands.

We used the pytorchmetrics standard implementation of KID in our project run for each epoch to monitor performance. We tested our implementation by running the KID metric on our real dataset against the same dataset to validate that our relative scores were sensible. Although we only tested on a very small set of images (n = 20), we found that the KID metric returned a considerably smaller value in the control set comparison versus the true comparison. (*ie.* we compared DALL-E 2 outputs to themselves, then DALL-E 2 outputs to stable-diffusion outputs.) This demonstrates the efficacy of KID on small image datasets and indicates that its use as our evaluation metric is prudent.

In our results section, we show training KID and validation KID, where the training distance is using images that DALLM had access to, and the validation KID are independent images drawn from DALL-E 2.

3 Results

In section 3.1, we contrast the different images produced by our models. In 3.2, we display our quantitative results. We do not include our loss graphs because they are uninformative; they converge to a small number quite quickly, despite the fact that there continue to be qualitative changes in the model after convergence.

3.1 Image Gallery

Fig. 1: Images from prompting Hugging Face's stable diffusion v1-5.

Fig. 2: Prompting DALL-E 2.

Fig. 3: Querying DALLM. r = 4, 20 examples from DALL-E, and 1000 steps.

3.2 Quantitative Results

We tested three different hyperparameters: rank of the low rank adaptation, the dataset size that LoRA fine-tuned with, and the number of steps that it was run on.

3.2.1. Rank

Fixing all other variables, we examined $r = \{1, 4, 16\}$.

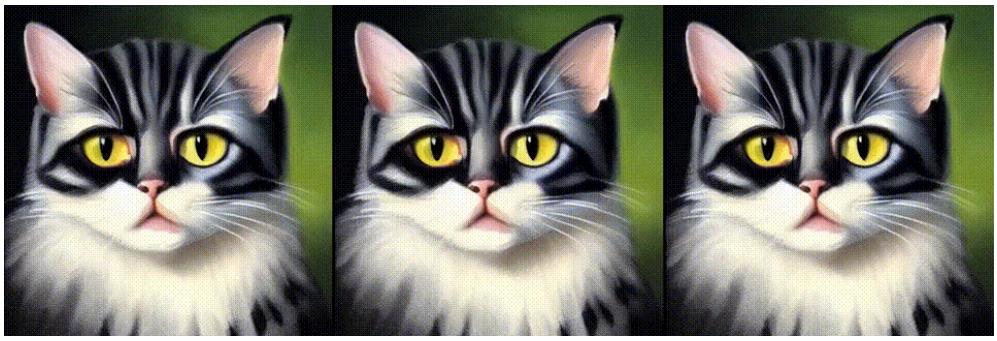

Fig. 4: GIF showing the changes of ranks 1, 4, 16 every 25 steps up until 1000 steps.

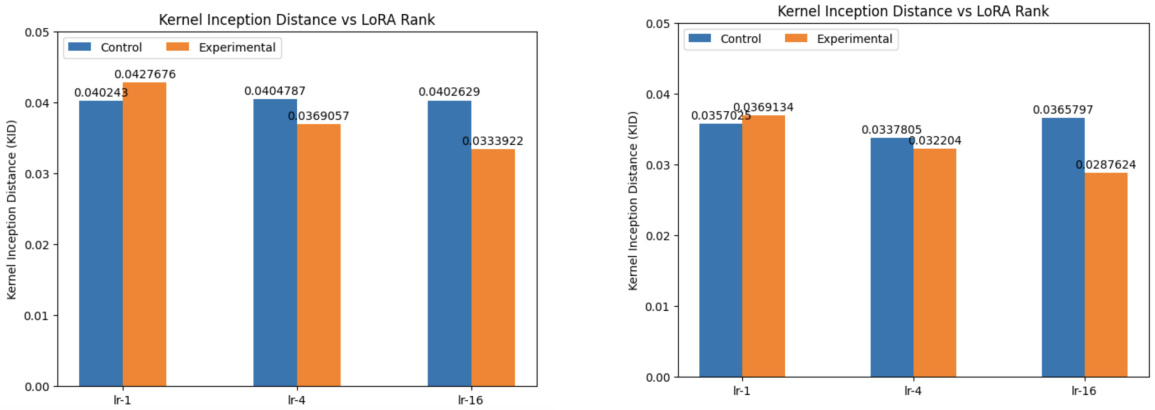

Fig. 5: Graphs of Kernel Inception Distance for each model of LoRA with training (left) and validation (right) images.

Interestingly, the rank 16 model learns to place glasses on the cat, despite the lack of any glasses in the training set from DALL-E 2; we theorize this may be a result of *generative artifacts*, which we elaborate more in section 4.2. In general, we see that even with comparatively few updates, just a thousand steps causes a statistically significant reduction in KID.

3.2.2 Dataset Size

Using rank $r = 1$, we examined if the number of examples from DALL-E affected the power of LoRA, searching over $\{5, 10, 20, 50\}$. Previous works have found that even a handful of examples with LoRA can produce more accurate photos of celebrities. The results below seem

to indicate that the convergence of the KID are unrelated the the size of the dataset, which could be due to the fact that more images provides the model more context behind the style it tries to emulate, but since there is more variety in the images the KID becomes less deterministic.

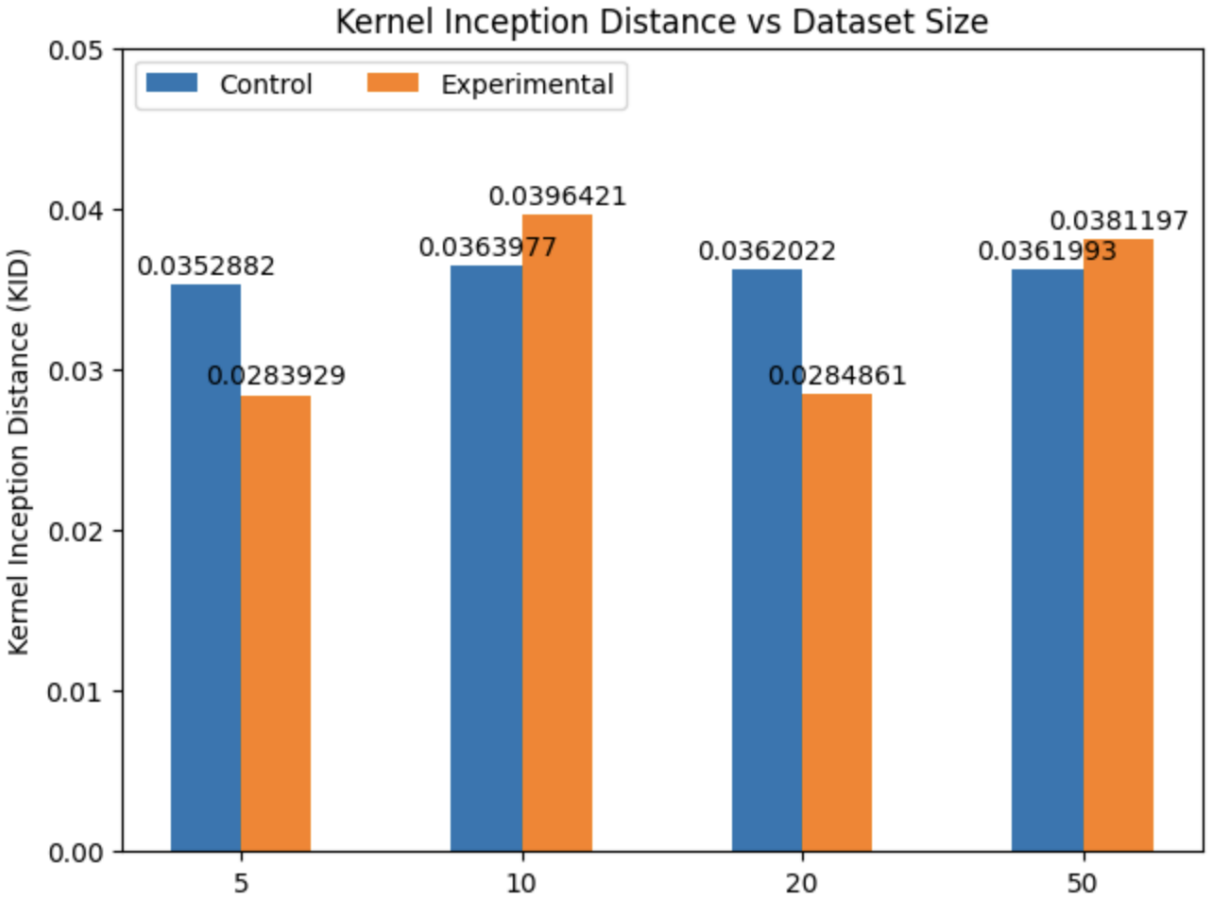

Fig. 6: Graph of Kernel Inception Distance for each dataset size

Fig. 7: Samples of generated images from the same random seed. Dataset size from left to right: 5, 10, 20, 50.

3.2.3 Training Steps

Using rank $r = 1$ and 20 example images from DALL-E, we fine-tuned LoRA on stable-diffusion for 2500 steps, saving checkpoints every 250 updates. The data seem to indicate that training beyond just a few iterations does not provide much of an improvement in kernel inception distance.

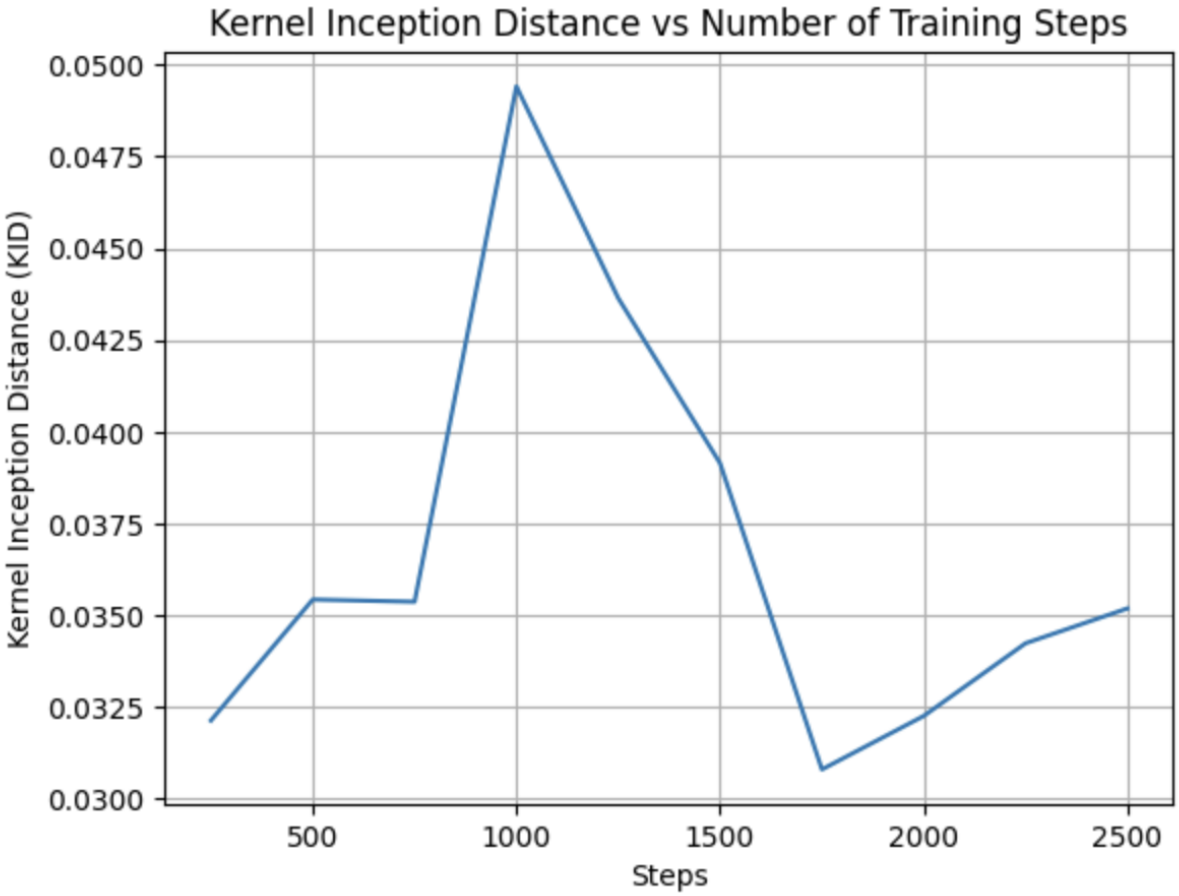

Fig. 8: Graph of Kernel Inception Distance vs Number of Training Steps

4 Theoretical Analysis

4.1 Qualitative Feature Extraction

DALLM does seem to learn qualitative features. In Figure 6 below, DALLM seems to learn from a small amount of data points (just 20 images queried from DALL-E 2) to place a reflection in the cat's eye. Because of the wide dimensionality and space that images can have, we believe that this is a strong indication that DALLM is extracting specific features from the sampled DALL-E 2 data. It is unlikely that the model is overfitting in so few steps and with low rank approximations added to the attention matrices *(K, V, Q, O)*.

Figure 9: The left is produced by stable diffusion v1-5. The right is a DALLM model trained for only 100 steps on 20 data points, with r=4. Note the pinched face and the reflection in the eye.

4.2 Out of Distribution Features

Interestingly enough, DALLM seems to exhibit novel, out of distribution features when we raised the rank of LoRA to 16. See Figure 7 below for an example; DALLM learns to produce cats that are wearing glasses, despite the fact that glasses are neither in its DALL-E 2 sampled training set nor in any of the baseline stable diffusion samples that we observed. We theorize that this may be because of two causes: "generative feature artifacts" or CLIP leakage. It is of course possible that this is purely random noise, and we did not have the resources to further investigate this phenomenon.

By generative feature artifacts, we refer to possible systematic generative patterns that are present in DALL-E 2 outputs. Perhaps DALL-E 2 associates Baroque art to some degree with glasses, and even if the "glasses" are not explicitly output in the images, there is some invisible watermark that exists in the image. This is the less likely explanation, but is interesting to think about; many efforts have been made to watermark generated images, as well as recognize artifacts from generated images. We also mention this kind of thinking in part because we are using synthetic data. It is important to be aware of the possibility of model collapse, and unexpected outputs can be a result of training on synthetic data.

In CLIP leakage, we propose that the training process— which includes feedback from CLIP scores— happens to associate glasses with closer distance to the text prompt. This would require some systematic investigation, on whether or not the presence of glasses seems more "baroque" or perhaps more "realistic" (recall the prompt we used was: "a realistic painting of a cat in baroque art style"). This causes DALLM to learn to place glasses on the cat, completely independent of the signal it gets from DALL-E. This can also be seen as overfitting to training, stemming from CLIP's embeddings of the prompt.

Regardless of explanation, this is an example of DALLM learning features.

Figure 10: Cats produced by DALLM, r={1, 4, 16} respectively, from left to right, at approximately 1000 update steps. Trained on twenty samples from DALL-E.

4.3 Limitations

Several constraints such as compute, time, and budget limited the scope of our paper. Our work examines a specific prompt due to our lack of compute and our tight timeline; it is possible that the hyperparameters we searched may have meaningful changes at numbers a magnitude larger than we experimented on. Some experiments, such as dataset size, are highly influenced by chance, so multiple runs would give stronger results; however, we did not have access to enough compute to validate. We leave further experimentation to future work. Furthermore, we only use qualitative observations and KID to validate our results; for more rigorous results, incorporating human preference as well as other metrics is required. However, we do raise several interesting directions for future work to examine, especially the novel features we observed.

5 Conclusion and Future Work

In our work we agree with previous findings that low rank adaptations are very effective for fine-tuning generative visual models. Furthermore, we find that these adaptations can use even small amounts of synthetic data to approach DALL-E's distribution. We suggest for future work that these results be examined with a wider range of prompts. A promising direction is also using image to image models, where there may be less "noise" or options that the model can take, and it may be more evident if the output distribution changes.

We also discover novel out of distribution features produced by our model at higher rank. We theorize that it is either a generative artifact caused by training on synthetic data, or a result of overfitting in

5.1 Declarations

Aligning stable diffusion to DALL-E 2 was an arbitrary choice of baselines. As a result, our models inherit the biases. For example, stable diffusion v1-5 itself was fine-tuned on mostly English descriptions and samples from Western cultures, resulting in a strong deficiency in generating non-Western content or non-English prompts. The models are created for purely academic and research purposes, and are not intended for any other content.

Authors have no conflict of interests to declare.

References

[1] Stanford Alpaca, https://crfm.stanford.edu/2023/03/13/alpaca.html
[2] Hugging Face Stable Diffusion v1-5, https://huggingface.co/runwayml/stable-diffusion-v1-5
[3] LoRA to fine-tune diffusion, https://github.com/cloneofsimo/lora
[4] KID, https://arxiv.org/abs/1801.01401
[5] FID, https://arxiv.org/abs/1706.08500
[6] Hu and Shen 2021, LoRA: Low-Rank Adaptation of Large Language Models https://arxiv.org/pdf/2106.09685.pdf