Stable Diffusion Fine Tuning

Abhi Lad

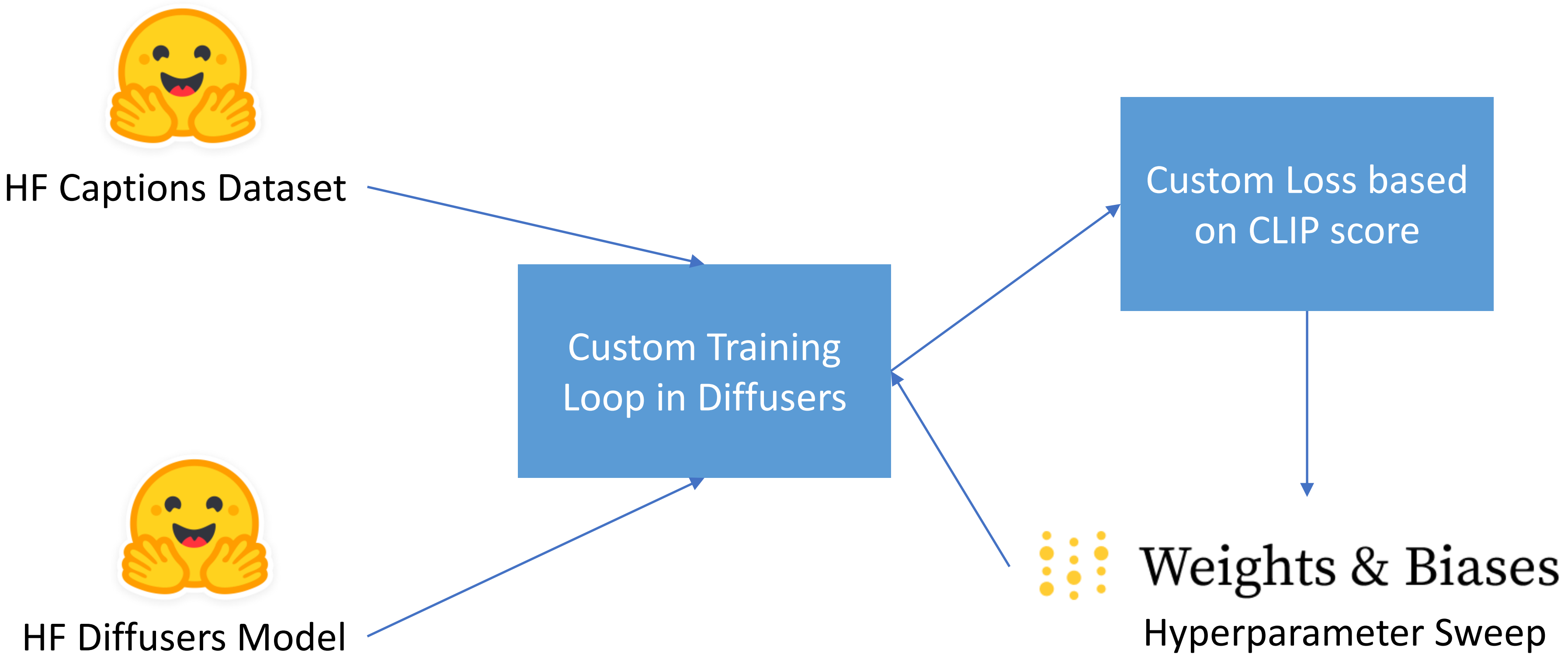
Basics

System: GCP High Mem (26 GB) Nvidia T4 instance

Primary library: huggingface diffusers

Note: Special training routine or evaluation code has not been incorporated in diffusers code due to limited time and trivial nature of the assigned task. We use the available training scripts with our data to fine-tune the models.

Ideal Training Approach



HF Captions Dataset

HF Diffusers Model

Custom Training Loop in Diffusers

Custom Loss based on CLIP score

Weights & Biases

Hyperparameter Sweep

Approaches Utilized

Base fine-tuning approaches

Text-to-image	**Dreambooth**
Train prompt-image pairs, possibly including multiple concepts Demo: https://huggingface.co/abhilad98/abhi_thumbsup	Train on a single new concept with few examples with a generic class prompt Demo: https://huggingface.co/abhilad98/db_abhi

Training enhancements

LoRA
• Prevents catastrophic forgetting • Significantly smaller model, useful for saving multiple concepts • Enables training on smaller memory GPUs

Training steps

Text-to-image

- Captioned 120 images and created hf compatible dataset : link
- Trained sd-2 on thumbs up + Abhi images with LoRA and appropriate hyperparameters
- WandB logs: link

Dreambooth

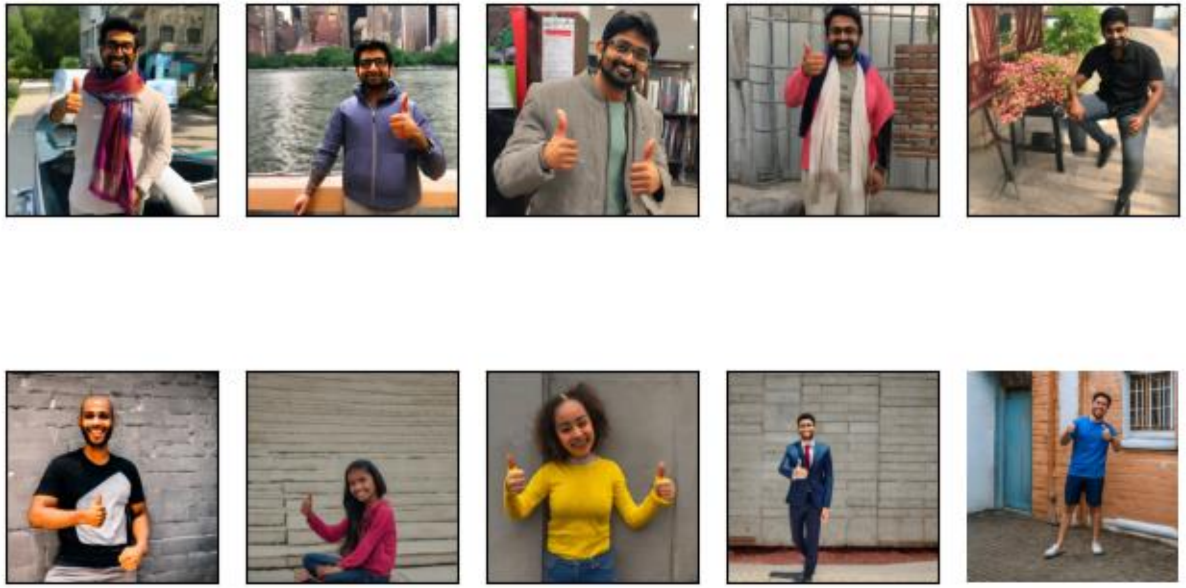
- Trained sd-2 on thumbs up images first
- WandB logs: link
- Used thumbs up trained model to fine-tune on 7 Abhi images using LoRA
- WandB logs: link

Insights on Approaches

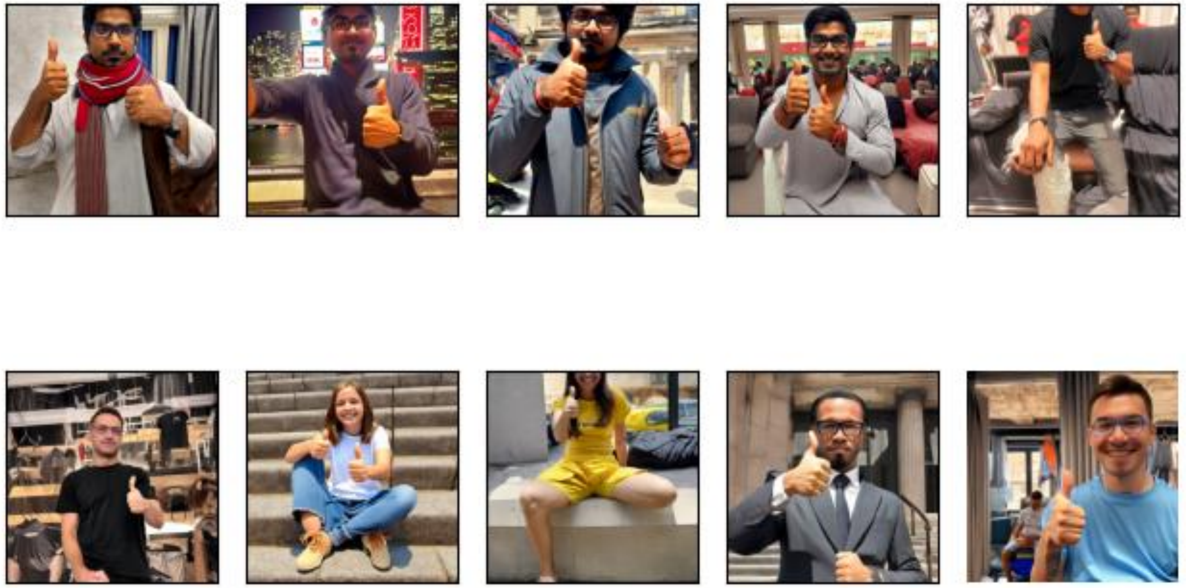
	Text-to-image	**Dreambooth**
Pro	- Single step training on thumbs up and personal images - Can learn nuances on prompts and generate specific outputs	- Does not require prompts for all the images - Works with small number of training images - Better qualitative results on personal images
Con	- Requires prompts for all images - Requires large number of images for each concept	- Can only train 1 concept at a time, requiring multiple training steps to learn complex prompts

Qualitative Results

Evaluation Notebook: link



Text-to-image results



Dreambooth results

Quantitative Results and Observations

Text-to-image	**Dreambooth**
CLIP score: 32.99	CLIP score: 30.77

- CLIP scores will vary on each run due to stochastic nature of the generative models
- CLIP scores are also dependant on how generic the prompts are
- Text-to-image has kept person in center of frame, whereas in dreambooth the subject is often cut out
- Dreambooth has better images of subject Abhi

Some known issues

- Dreambooth fine-tuning without LoRA will result in catastrophic forgetting of the pretrained models
- Text-to-train does not perform well on generating Abhi images since it has limited samples and tries to train on specific prompts
- Dreambooth LoRA training script has a bug which prevents resuming training from checkpoint. So, our dreambooth training looks like:
  db -> db_lora instead of db_lora -> db_lora
- Dreambooth thumbs up images also messes up the existing knowledge without using the lora training approach as can been seen on WandB.

Deploying at Scale

- Using text-to-image on scale is difficult as it needs captions for each image which may be infeasible
- Also, if we want model to learn an overarching concept like thumbs up, then we don’t need fine grained captions
- Getting sufficient images of such a concept on scale is easier
- Instead of making a general purpose model, we will make task specific model since we can use LoRA to reduce the final model size
- Then we use 2 stage Dreambooth based training

Deploying at Scale

2 stage Dreambooth training

- First fine-tuning base model on overarching concept like thumbs up
  - Save as new model (large model size, will forget original task but better at new concept)
  - Save as LoRA model (small model size, since it can be reconstructed from base model)
- Then fine-tune the concept trained model on an individual person (personalization) using LoRA
  - This step should be last as models forget faces quickly as not enough data is present

Feel free to ask for GCP instance permission to replicate the results