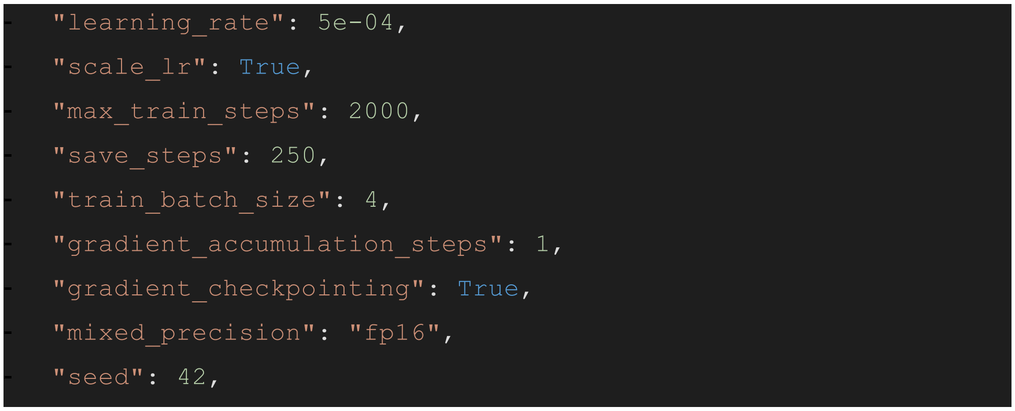
Harshita Khandelwal

Stable Diffusion - Thumbs Up task

- Requirements:
    - Integration of a Thumsb Up Style, activated by a chosen trigger word.
    - Customization on Stable Diffusion v1.5 or v2.1
    - Use only text-based input prompt models (No Img - Img)
    - Incorporate a famous person's face for testing
    - Atleast 75% of the generated pictures should be appropriate.

- Given the requirements of the task (Text Only input prompt models finetuned for a specific trigger words), two approaches seemed plausible : DreamBooth & Textual Inversion, to finetune on a basic stable diffusion model.
- Textual inversion "learns to generate specific concepts, like personal objects or artistic styles, by describing them using new "words" in the embedding space of pre-trained text-to-image models. Whereas DreamBooth is a method to personalize text-to-image models like Stable Diffusion given just a few images of a subject. It allows the model to generate contextualized images of the subject in different scenes, poses, and views.
- Delving deeper, The concept of Textual inversion seemed more aligned with our idea of using a trigger words and Also, DreamBooth required more resources in terms of GPU power for finetuning for better results where as Textual Inversion required 3 to 5 images for the same and doable resources (GPU & Time).
- Hence, we decide to go forward with trying textual inversion for our task.
- Training requirements:
    - Data : We take a sample of 5 images from the dataset provided by the Team at Easel. Link - <https://github.com/tenzu15/StableDiffusion---Textual-Inversion >
    - For Training, a T4 GPU was used (Google Colab)
    - Choice of Model : Stable Diffusion 1.5 and Stable Diffusion 2.1, both were tried for this experiment to compare the results. (More details down below)
    - Object v/s Style: Even though by reading the problem statement, the use of thumbs up can be classified as a style diffusion, treating thumbs up as an object also seemed like a viable option. Hence, both options were chosen to see the difference. (Discussed more further)
    - Hyperparameters - On research from blogs and previous implemenatations as well our limited hyper parameter tuning, we went with the following options:



```
"learning_rate": 5e-04,
"scale_lr": True,
"max_train_steps": 2000,
"save_steps": 250,
"train_batch_size": 4,
"gradient_accumulation_steps": 1,
"gradient_checkpointing": True,
"mixed_precision": "fp16",
"seed": 42,
```

- learning rate: We pick our learning rate as 5e-04 based on previous research and common implementations. We make use of the scaled_lr parameter to ensure the learning rate is adjusted according to need during training so as to do no harm to the process **[9]**
- batch size: We keep our batch size small as the number of images in our training dataset is also on the lower end (5 images).
- max _train_steps: Accepted range for a successful training using textual inversion is in the range for 2,000 - 4,000 steps depending on the data. Since, we deal with lesser data, we keep it around 2000.
- gradient _accumulation_step: We keep this at 1 due to less number of steps and data to avoid overfitting.
- seed: This is an important parameter that should be set for replication of results.

Stable diffusion 1.5 v/s 2.1:

Stable Diffusion 1.5 is lighter in nature but 2.1 gave us better results. We tried using both of them and the results can be reproduced using this notebook. <https://colab.research.google.com/drive/1ywEfBVzDx48SRZ9qwYKP2GzlIMO1-P_R?usp=sharing >

As an object v/s Style:
When it comes to textual inversion, there are 2 kinds of prompting we can use for training - using aur trigger word as an object (Eg. a cat toy) or as a Style (Eg. Groot Style). On further analysis of prompt and requirements, we learned using the object route would be a better choice. (Both these choices were tried with v1.5).

Decoding:
Firstly, We set our place-holder token as **Thumbs up.** So to trigger our thumbs up feature, we need to use the keyword Thumbs up. As for the parameters used:

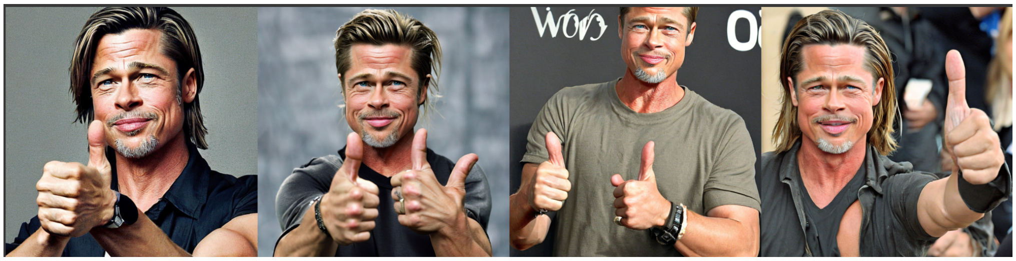
- Num_samples: Number of Sample result images that need to be generated.
- Num_rows: Num of rows to be utilized while generating images.
- Num_inference_steps: The range widely used in the literature is 30 - 150. Upon optimization, the results for this model came best in the range of 50 - 60 stps
- Guidance Scale: This helps us generate our image closest to the prompt. Acceptable range is 7.5 -13.5. Upon optimization, the results for this model came best at 8.0

While inferring our trained model, we make use of an automated pipeline. To keep track of the results, we can manually set a seed value and give it to our generator to test and replicate superior images in the future. Implementation for the same has been given in the notebook.
As for scale deployment, the best way would be to create a service (API) and deploy our model on the cloud for use.**[10][11]**

Harshita Khandelwal

Eval:

Human Evaluation & Results: The model was evaluated on 5 famous people: **Brad Pitt, Angelina Jolie, Chris Evans, Tom Holland, Rober Downey Jr,** by producing 4 images of each actor. As a result, we got a 90% accuracy (18 out of the 20 photos were created properly).
On a larger scale, CLIP Evaluation can also be utilized.
The actors look caricatures in it though. To improve on this, we can do a manual seed search for better results or look at further finetuning to make the model better. (Due to lack of resources, this technique wasn't tried.)

Eg. 

Relevant links for the task:

- **Github :** **https://github.com/tenzu15/StableDiffusion---Textual-Inversion**
- **Google Colab :** **https://colab.research.google.com/drive/1ywEfBVzDx48SRZ9qwYKP2GzIlMO1-P_R?usp=sharing**
- **HuggingFace Models:** **https://huggingface.co/harshitaskh/ThumbsUp_v2.1/tree/main,** **https://huggingface.co/harshitaskh/ThumbsUp_v1.5/tree/main**

Reference links

[1]https://tryolabs.com/blog/2022/10/25/the-guide-to-fine-tuning-stable-diffusion-with-your-own-images
[2]https://uxplanet.org/how-to-generate-stunning-images-using-stable-diffusion-1a868061a07f
[3]https://colab.research.google.com/github/huggingface/notebooks/blob/main/diffusers/sd_textual_inversion_training.ipynb#scrollTo=pnq5vah7pabU
[4]https://huggingface.co/docs/diffusers/training/text_inversion
[5]https://huggingface.co/docs/diffusers/training/text2image
[6]https://huggingface.co/docs/diffusers/training/dreambooth
[7]https://blog.paperspace.com/dreambooth-stable-diffusion-tutorial-part-2-textual-inversion/
[8]https://github.com/AUTOMATIC1111/stable-diffusion-webui/wiki/Textual-Inversion
[9]https://discuss.huggingface.co/t/why-does-textual-inversion-example-scale-learning-rate/37938
[10]https://aws.amazon.com/blogs/machine-learning/create-high-quality-images-with-stable-diffusion-models-and-deploy-them-cost-efficiently-with-amazon-sagemaker/
[11]https://bentoml.com/blog/deploying-your-own-stable-diffusion-service-mz9wk