

Script2Skin

Generating Medical Images



Jhalak Surve
Raja Nikshith Katta
Uday Shankar Gattu

Introduction

- Goal is to develop an AI model capable of generating accurate and detailed images of common skin rashes from textual commands
- **Use of Latent Diffusion with Fine-tuned CLIP**
- Components - A User Interface for entering text prompts and displaying generated images, and a generative model for text to image conversion



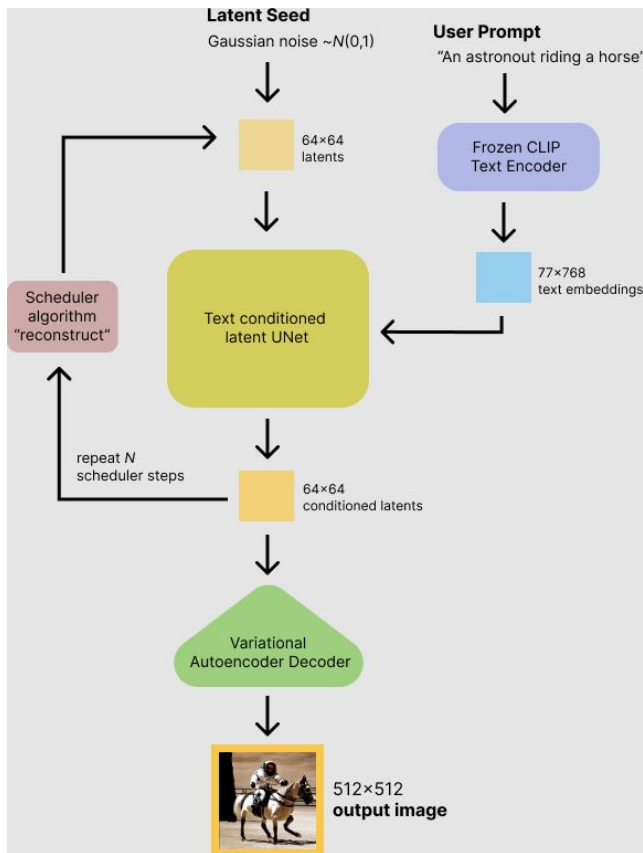
About the Dataset

- **Dermnet** - images of 23 types of skin rashes/diseases (jpg format, consisting of 3 channels, i.e. RGB)
- Worked with 3 types : **Acne, Eczema and Tinea**
- A total of around 1000 images
- Prepared the dataset by providing textual prompts/captions to the images

```
{"image":"data/189.jpg","caption":"tinea type of rash at the foot area on a fair skin"}  
{"image":"data/eczema_049.jpg","caption":"eczema type of rash at the leg area on a fair skin"}  
{"image":"data/eczema_061.jpg","caption":"eczema type of rash at the fingers area on a fair skin"}  
{"image":"data/eczema_075.jpg","caption":"eczema type of rash at the hand area on a fair skin"}  
{"image":"data/162.jpg","caption":"tinea type of rash at the foot area on a fair skin"}  
{"image":"data/176.jpg","caption":"tinea type of rash at the foot area on a fair skin"}  
{"image":"data/acne-cystic-12.jpg","caption":"acne type of rash on the cheek area on a fair skin"}  
{"image":"data/eczema_263.jpg","caption":"eczema type of rash at the neck area on a fair skin"}  
{"image":"data/eczema_277.jpg","caption":"eczema type of rash at the toe area on a fair skin"}
```

<https://www.kaggle.com/datasets/shubhamgoel27/dermnet>



What is Latent Diffusion?



- Operate by gradually transforming a random noise distribution into structured data in a latent space, ultimately generating high-quality images.
- Diffusion process - adding noise to data in a controlled manner and then learning to reverse this process to create clean data from noise.
- Components : Unet, VAE and CLIP



CLIP (Contrastive Language-Image Pre-training)

- Neural network trained on a variety of images and text pairs
 - Simultaneously processes text and images, learning to predict which images are described by which texts
 - Can be integrated with generative models to direct the generation process, ensuring that the resulting images closely align with textual descriptions
- 
- 



Project Approach

01

**Dataset
Preparation**

02

**Initial
Considerations**

03

CLIP Fine-tuning

04

**Integrating with
Stable Diffusion**

05

**Testing the
Diffusion model**

06

**Integrating with
UI**

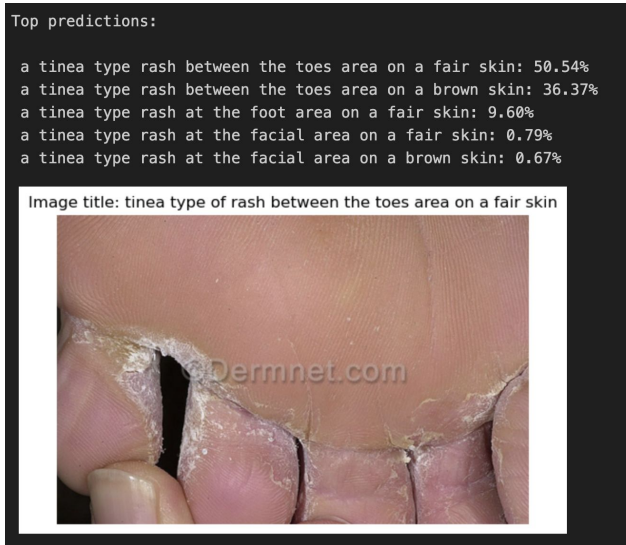




Initial CLIP Fine-tuning and Drawbacks

Fine tuning CLIP model to make it more domain specific

Achieved good results in predicting titles for images



Trained the CLIP model on our skin rash dataset

Found that the trained model was not compatible to be used with the Stable Diffusion models



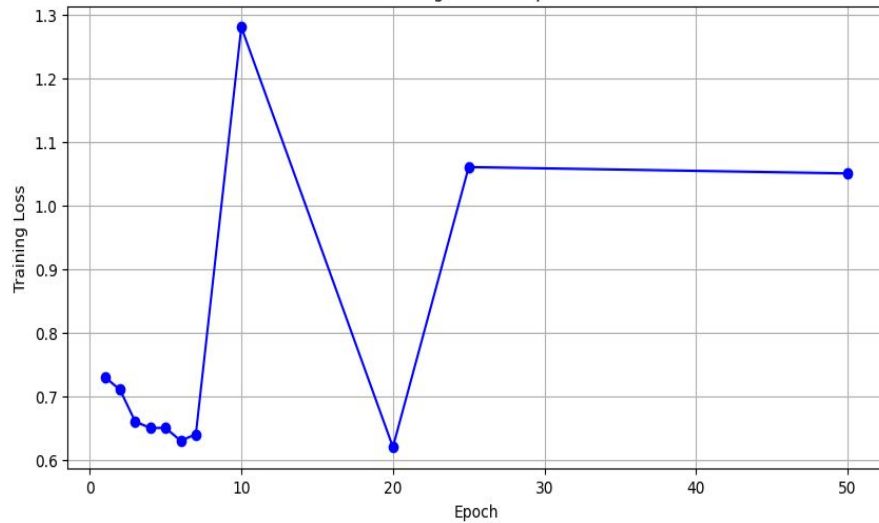
Finetuning CLIP with huggingface libs

- Finetune the type of CLIP models used for Stable Diffusion with huggingface libs on a self-defined dataset
- Dataset format: Image files in .jpg format and for each image file, there should be a .txt file with the same name containing the caption for the image
- Trained the model several times varying the batch size and the number of epochs to generate better results

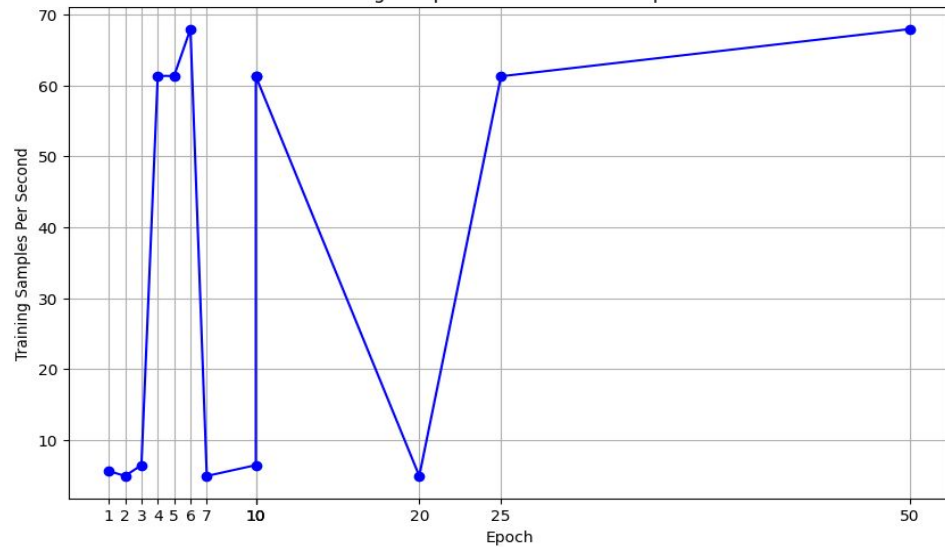
<https://github.com/damian0815/finetune-clip-huggingface/tree/main>

Plots

Training Loss vs. Epoch



Training Samples Per Second Over Epochs



Integrating Fine-tuned CLIP with Stable Diffusion

- Stable Diffusion operates using a lower-dimensional latent space for efficient image processing, featuring an encoder for compression and a decoder for reconstruction.
- Incorporates a tokenizer and a CLIPTextModel to generate text embeddings that guide the image generation process, ensuring images closely align with textual descriptions.
- Utilizes a UNet model for detailed processing and a scheduler to manage the diffusion steps, which are more complex than traditional single-model systems.
- The model follows a structured pipeline: load pretrained components, generate and denoise latent representations, and decode these back into detailed images.

https://huggingface.co/docs/diffusers/en/using-diffusers/write_own_pipeline

Some images generated while testing

[15]:



Prompt: generate an image of a tinea type of rash on the facial area on a brown skin

[13]:



Prompt: generate an image of an acne type of rash on the face area on a fair skin

[22]:



Prompt: generate an image of a tinea type of rash on the palm area on a fair skin



Challenges and Future Enhancements

- We faced significant resource and memory constraints that impacted our ability to fully train and optimize the model, resulting in less than ideal image quality.
- Moving forward, addressing these resource constraints will be our priority. We plan to refine our approach by securing better infrastructure, optimizing our model's efficiency, and expanding our dataset for improved performance.



Conclusion

- Deepened our understanding of advanced AI technologies, particularly in text-to-image synthesis using latent diffusion models and CLIP components, enhancing our technical proficiency in machine learning applications.
- Showed us how important it is to work together across different fields. By combining computer science with dermatology, we learned to address real medical problems more effectively and gained a wider view and better problem-solving skills.



Thank You!!!