Voice to Image Generation

By Tharan Bala, Siddharth Chatla

Problem Statement :

The idea of storyboard writing and content creation in the industry of art has always been a task. For some of the conservatives in this domain thrive on the concept of being burned out and putting efforts into creating everything by hand, they go as far as quoting that "under pressure people produce the most brilliant output".
Considering the other part of the society, the non-conservative part is under pressure for not being able to produce content. This is the biggest set back to creative people who do not make it in the industry.

"Kharoshi", This is a Japanese word which translates "overwork death", most of the cases being from :

- managaka's (comic artists), their average work hours per day are somewhere between 14 - 18 hours / day , $1583 per / month
- Animators - 12-18 hours/day , PAY - $270 per month

This is a serious problem, which can be resolved by being able to create important thumbnails and backgrounds on a whim and then start working with it. This was not possible before, but with the introduction of stable diffusion we are able to stitch images and remove the noise to a level where the image generated is intact and can be used for thumbnails and references which usually take most of the time when drawn by hand.

Past Work :

The technology has only recently developed to the point where it can be effective in the animation industry, the application of AI in this sector has a very brief history.Creating backgrounds and environments, resizing images, enhancing textures, and other procedural tasks that are simpler for computers to do were some of the early uses of AI in animation.

However, this does not imply that AI was restricted to routine jobs: It may come as a shock, but as of the middle of the 2000s, AI systems were being employed in the making of movies. It all started with the creation of "Genesis," a Pixar system that created 3D

models of animals and other things using machine learning methods. Several Pixar movies employed this technique to help build realistic characters.

Related Work:

- **Obvious** is a collective of researchers, artists, and friends, working with the latest models of deep learning to explore the creative potential of artificial intelligence. They are behind the sale of the first AI artwork to go through a major auction house. They use their work to share their vision of artificial intelligence and its implementation in our society.
- **starryai** : Easiest way to start creating NFT Art using Artificial Intelligence. Start creating in minutes!

Methodology

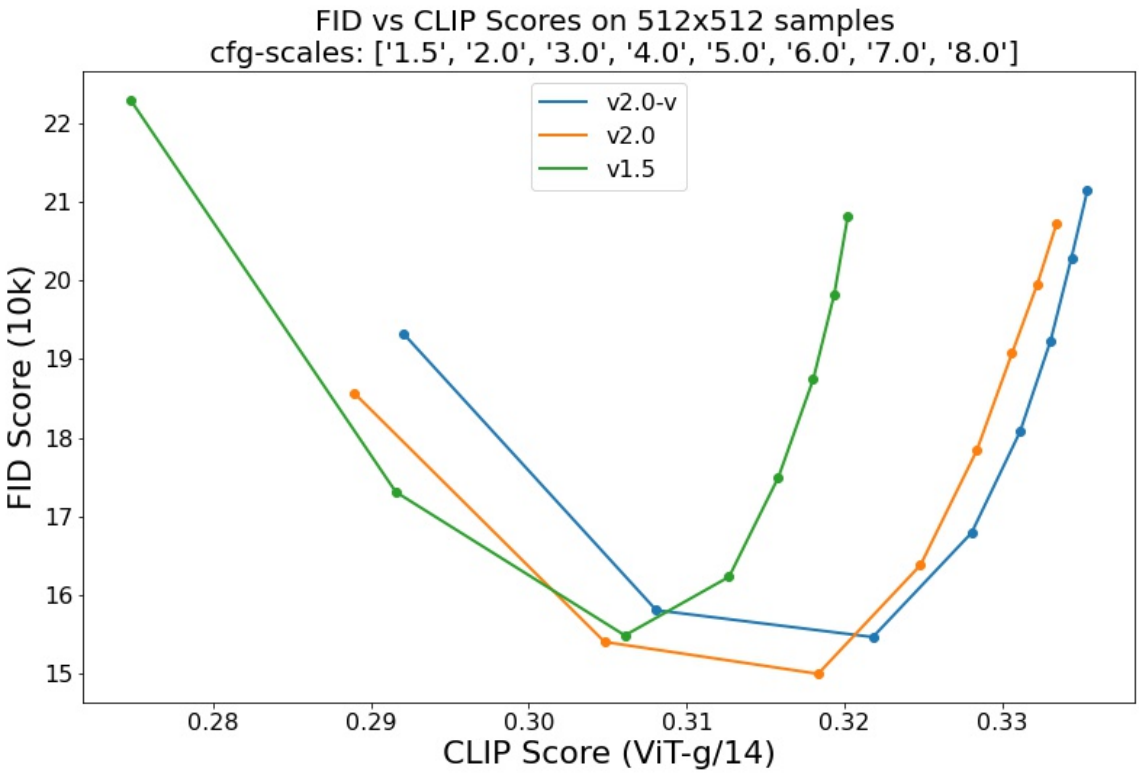
Stable Diffusion:

Stable Diffusion belongs to a class of deep learning models called **diffusion models**. They are generative models, meaning they are designed to generate new data similar to what they have seen in training. In the case of Stable Diffusion, the data are images.

This `stable-diffusion-2` model is resumed from stable-diffusion-2-base and trained for 150k steps using a v-objective on the same dataset. Resumed for another 140k steps on `768x768` images.

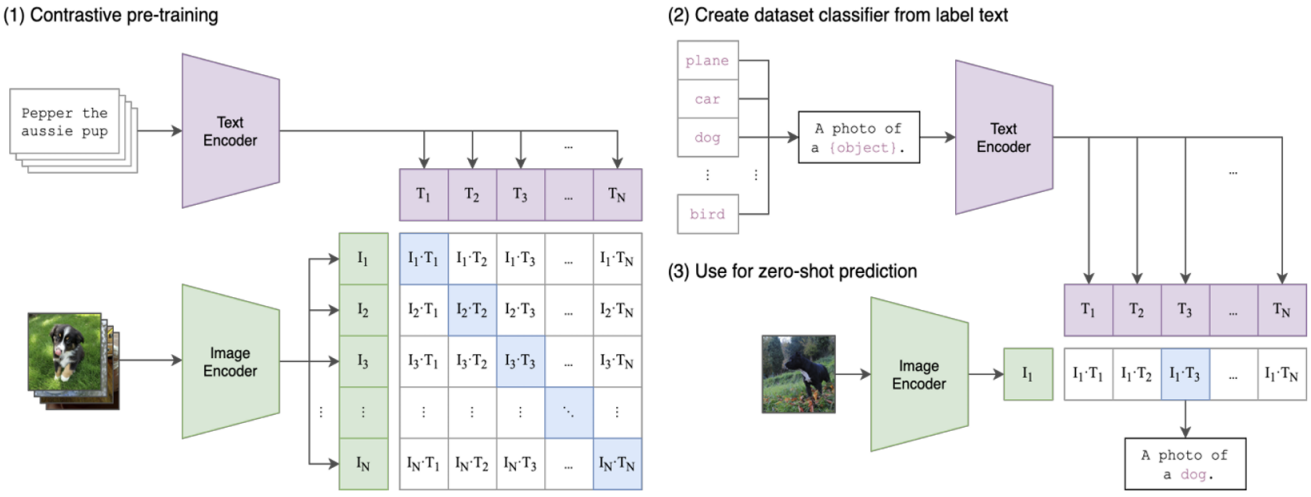
Stable Diffusion v2 refers to a specific configuration of the model architecture that uses a downsampling-factor 8 autoencoder with an 865M UNet and OpenCLIP ViT-H/14 text encoder for the diffusion model. The *SD 2-v* model produces 768×768 px outputs.

Evaluations with different classifier-free guidance scales and 50 DDIM sampling steps show the relative improvements of the checkpoints:



CLIP

Contrastive Language-Image Pre-Training is a neural network trained on a variety of (image, text) pairs. It can be instructed in natural language to predict the most relevant text snippet, given an image, without directly optimizing for the task, similarly to the zero-shot capabilities of GPT-2 and 3. We found CLIP matches the performance of the original ResNet50 on ImageNet “zero-shot” without using any of the original 1.28M labeled examples, overcoming several major challenges in computer vision.



Hyperparameters while training the dataset

- Gamma – controls the shape of the decision boundary
- Temp – controls the softness of the distribution generated by the model.
- ltc
- ltd
- Mirror

Reqiuirements

Hardware requirements:

- 1–8 high-end NVIDIA GPUs with at least 12 GB of memory.
- We have done all testing and development using NVIDIA DGX-1 with 8 Tesla V100 GPUs.
- CUDA needs to be enabled to run the backend code.

Packages required :

- **Flask**
- **Flask-Cors**
- **Flask-RESTful**
- **flax:**
  high-performance numerical computing
- **tqdm:**
  Smart progress meter
- **flask_cloudflared:**
  A Flask extension that wraps pycflare to provide access to CloudFlare's API.
- **accelerate:**
  Boilerplate code related to multi-GPUs
- **transformers:**
  Transforms an input sequence to an output sequence. This includes speech recognition, text-to-speech transformation
- **diffusers:**
  Generate data (mostly images)
- **Torch:**
  Tensor computation (like NumPy) with strong GPU acceleration, Deep neural networks built on a tape-based autograd system
- **AssemblyAI:**
  Transcribing understanding audio and video

How to run the

Running Backend

1. Clone or fork this repository
2. Create a virtual environment `cd backend && python3 -m venv ENV_NAME`
3. Run virtual environment `source venv/bin/activate`
4. Install requirements `pip install -r requirements.txt`
5. Make sure you have pytorch and its dependencies installed *Installation guide*
6. Run web server `python3 thread.py`
7. You will get a DALL-E URL in the terminal while running the backend
8. Copy and paste that URL in the Dalle.py file

Running Frontend

1. pip install requests streamlit websockets pyaudio
2. Grab an API Token from AssemblyAI and paste it into configure.py.
3. run : `streamlit run main.py`

Results:

Result 1

Animated Speech

What it can do?

Animated Speech is an application that is used to generate an Image from Speech. This application uses DALL-E server to convert the text to image and Assembly AI to convert speech to text in realtime.

Say something

Is this what you said?

Posiedon on top of a Mountain

GO!

Result 2

Animated Speech

What it can do?

Animated Speech is an application that is used to generate an Image from Speech. This application uses DALL-E server to convert the text to image and Assembly AI to convert speech to text in realtime.

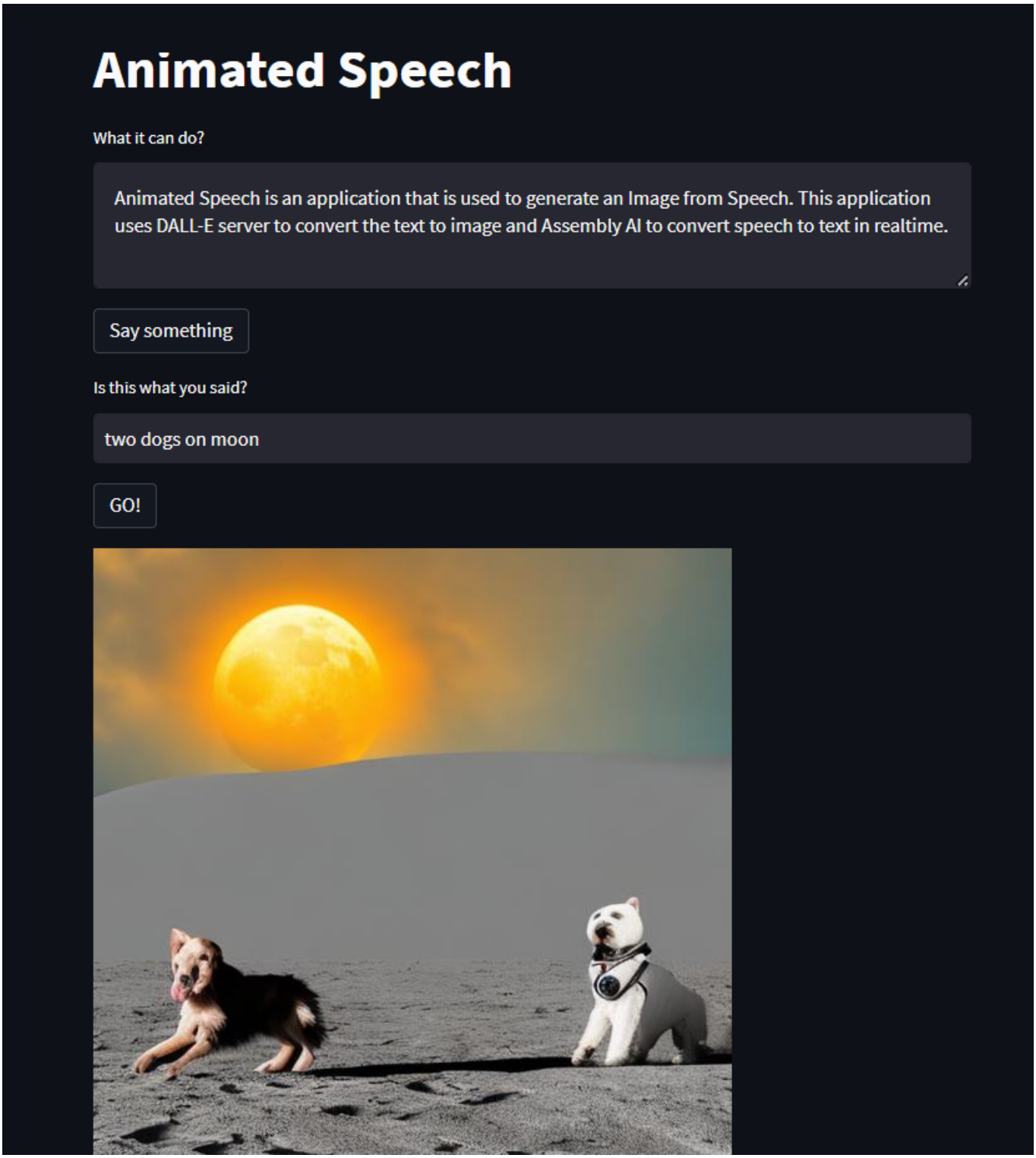
Say something

Is this what you said?

greek god on top of a Mountain in van gogh style

GO!

Result 3



Discussion

Siddharth faced issue writing the code on a mac where the idea of unlocking the GPU's had become a priority but then enabling CUDA was only possible on Nvedia GPU's. The only way out was to use tensor for mac and then realised the tensor was more complicated and then decided to use another machien that had a Nvedia GPU built in.

Memory space limitation for a single personal computer is 4 GB is our first limitation and we resolved it by clearing the cache. Resolution : `torch.cuda.empty_cache()`.

We faced issue carrying the encoded jpeg file, then we stored it into a json format and then retrived it using a key and decoded to be fed to the webpage.

Conclusion

This project is our first step towards transcribing **speech** to **text** to **image**, It introduces open source AI tools to build a web application that fulfills your necessities.
It becomes critical at a point where we are trying to find the best model to use inorder to generate images that fit the context, Openinng the domain of AI for us to explore the Open Source models that can be incorporated.

Referance:

Problem statment:
mangaka, animators : https://www.spieltimes.com/tv-shows/anime/how-much-do-animators-earn-in-japan-revealed/#.ZGZ7JOzMJqt
Progress if AI in Art Industry : https://studiopigeon.com/blog/artificial-intelligence-animation-what-is-it-and-how-does-it-function/

methodiligy:
Stable Diffusion : https://huggingface.co/docs/diffusers/v0.13.0/en/stable_diffusion
Progressive Distillation for Fast Sampling of Diffusion Models:
https://arxiv.org/abs/2202.00512

Related Work :
Obvious art : https://obvious-art.com/page-about-obvious/
Starry ai https://starryai.com/create-nft-art-with-artificial-intelligence