AML Mid Project Presentation - Pied Piper

Swetha M - 23M0756
Harshvivek Kashid - 23M0762
Rashmi Kokare - 23M0785
Nikita - 23M0807

Problem Statement

Generated synthetic images lack emotions though they are realistic. Some image editing applications require emotion to be incorporated into source image to get an image reflecting better emotions. We want the same source image to get the induced emotion.

Some **challenges** are:

1. Latent diffusion models like stable diffusion **create random images** instead of starting from source image
2. CLIP based models have far away **embeddings for words with emotions** of the same category.
3. Some existing models (specially GAN) use **only style or color to change emotion**

Existing methods

Methods: Editing the latent space of diffusion, Stable diffusion based, StyleGAN, VAE based, Emogen

- Emogen can generate images with a certain emotion - but not add or include emotion to an input image.

Textual inversion based:

- **LoRA** - Learn a text for a given image which can be used in prompts for future images (personalised image editing)
- **Dreambooth** - Subject-Driven Generation which few shot trains an image and uses it to generate multiple contexts of the same subject

Literature survey

Paper	Main idea
Diffusion Models Already Have A Semantic Latent Space	Use the bottleneck of UNet (in denoiser) for the semantic information
EmoGen: Emotional Image Content Generation with Text-to-Image Diffusion Models	Map the emotion space to CLIP space to generate a more diverse set of images
Textual inversion	Learn visual contexts and use them in future prompts
StyleGAN	Color and style transfer to change the emotion

Proposed Solution

Modify ASYRP to utilise Textual inversion

1. While Asyrp focuses on manipulating the direction of the reverse denoising process for targeted edits, it does not directly leverage **CLIP's image-text mapping capability.**
2. To address this, we propose a novel approach that integrates **Textual Inversion.**
3. This combined framework will exploit Textual Inversion's ability to capture emotional concepts and utilize them to generate images infused with specific emotions.

Code Survey

Code	Module usable	Link
Asyrp	**Greater part of learning from bottleneck of UNET can be used.** Textual inversion to be incorporated	asyrp
StyleGAN	Helped study latent structure manipulation	StyleGAn
Textual inversion	Hugging face also has starters for textual inversion but not useful for custom data	Textual inversion
Emogen	Helped understand emotion space vs CLIP	Emogen

Datasets Used

1. **Finding Emo** - https://arxiv.org/pdf/2402.01355.pdf
   https://gitlab.com/EAVISE/lme/findingemo/-/tree/main?ref_type=heads
   a. 25k images
   b. Plutchik's discrete Wheel of Emotions (PWoE)
   c. Very useful because it has complex backgrounds and realistic images
2. **Emoset** - https://github.com/JingyuanYY/EmoSet/blob/main/EmoSet.py
   a. With 8 emotion categories (amusement, anger, awe, contentment, disgust, excitement, fear, and sadness
   b. 118K images
   c. Not very useful dataset

Implementation Done

- We have studied the codebase of existing **GAN, VAE based methods** for emotion induction.
- We have studied the hugging face models for **CLIP, Stable diffusion, text to image pipelines** etc. and tested small experiments
- We have trained custom data on **ASYRP repo** and noted down observations.

Observations and Preliminary Results

- **ASYRP** - Not able to reproduce the results when training on custom datasets.
- **Hugging face models** generate random images not starting from source image
- **Text inversion** for existing concept libraries work fine but requires effort for custom data
- **StyleGAN and VAE based models explored:** Has some limitations like distorted, blurry images and inability to handle data different from training data.

Challenges Encountered

- **Random generation** happens since diffusion has learnt denoising from **random seed**

- **Blurring of images** and making the features worse than original image (eyes)

- Works only for **images trained** or very similar to that.

- Not interpret **complex backgrounds of people** and expects **portrait only** input images

Roadmap and things to be done

- We are modifying the baseline “ASYRP” to give better results for any custom dataset.
- We will be incorporating the textual inversion techniques from the official documentation of the paper in ASYRP

Thank You!