Emotion driven Image Synthesis

AML Project Report

Team members:

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

Department of Computer Science and Engineering
Indian Institute of Technology, Bombay

MAY, 2024

Acknowledgment

We would like to thank our guide, **Dr. Sunita Sarawagi**, for her constant support and guidance throughout this course. We thank our assigned TA for his timely feedback. We extend our thanks to the other TAs of the Advanced Machine Learning course and the entire Computer Science Engineering department for providing interesting topics and enriched learning experience.

Abstract

Despite significant progress in generating realistic synthetic images, a crucial limitation remains: the lack of emotional expression. Current image editing tools allow for the manual incorporation of emotions, but this approach is cumbersome and requires artistic skill. This work presents "Emotion-driven Image Synthesis," a project aiming to automatically generate emotionally charged images from a single source image. We explore the potential of this technology to foster deeper audience connection in art and advertising, enable personalized experiences, and contribute to the creation of truly realistic imagery that resonates with human perception.

Contents

List of Figures

Chapter 1

Introduction

1.1 Problem statement

The emotional disconnect in AI-generated imagery is a significant hurdle. While these images achieve stunning realism, they often lack the depth and nuance of human emotion. Current editing tools typically involve adding emotions to existing photos, a cumbersome process. What if AI could bridge this gap? Imagine generating a spectrum of emotions from a single image – a joyful portrait transformed into a contemplative one. This technology has exciting potential, evident in playful apps like Snapchat filters and sophisticated tools like Google's DreamBooth. [Gal+22] By weaving emotional expression into art and advertising, we can forge deeper connections with audiences and personalize experiences. This leap forward not only enhances emotional impact but also brings us closer to creating AI-generated images that feel truly lifelike. In this project, we aim to tackle this challenge by building a system that synthesizes emotion-driven images. The system will take an image as input and output an image infused with the desired emotional tone.

1.2 Motivation

The field of artificial intelligence has achieved remarkable progress in generating realistic synthetic images. However, these images often lack the emotional depth and nuance that define human experience. This creates a crucial gap between the technical brilliance of AI image synthesis and its ability to truly connect with viewers. Current image editing software allows users to manually incorporate emotions into existing photographs, but this process can be time-consuming and requires artistic expertise. Our project, titled "Emotion-driven Image Synthesis," aims to bridge this gap by enabling the automatic generation of emotionally charged images from a single source image.

This project is motivated by the following key factors:

- Deeper Audience Connection: Current AI-generated images, while incredibly realistic, often fall flat emotionally. Emotionally-driven image synthesis can foster a stronger connection with viewers in art, advertising, and other creative fields. By evoking specific emotions, images can become more impactful and memorable. Imagine using the same image to generate a variety of emotions – a joyful portrait transformed into a contemplative one. This technology is already at play in fun apps like Snapchat filters and powerful tools like Google's DreamBooth.

- Personalized Appeal: Emotionally-driven image synthesis has the potential for personalisation – tailoring the emotional content of an image to a specific audience or user. This can unlock new possibilities for targeted marketing and interactive experiences. For instance, social media platforms could personalize content based on a user's current mood.

- A Step Towards Realistic Imagery: Human perception is deeply intertwined with emotion. The ability to generate emotionally charged images represents a significant advancement towards replicating the full spectrum of human visual experience. By successfully developing this technology, we aim to move beyond the realm of simply realistic image creation and unlock new avenues for emotionally resonant visual communication.

By successfully developing this technology, we aim to move beyond the realm of simply realistic image creation and unlock new avenues for emotionally resonant visual communication.

1.3 Report Outline

The rest of the report is structured in the following way. Chapter 2 discusses the Literature Survey in detail. It also covers the advantages and limitations of the other approaches. Chapter 3 discusses the proposed system and its challenges and results. Chapter 4 covers future works and the summary.

Chapter 2

Literature Survey

We have generalized and specialized emotion-driven image generation. These emotion driven image generation can be used in various fields like art creation, film production, personalized content generation, and even therapeutic applications. Some examples for the technologies include: Generalised: EmoGen, Stable diffusion Specialized: LoRA, dreambooth, textual inversion

2.1 Emogen

Current image generators struggle to capture the nuance of emotions when generating images based on emotional prompts. Emogen[**Weihao2021gan**] aims to bridge the gap between emotional prompts and the generated visuals. Current image generation methods struggle because similar emotions can reside in distant locations within a common representation space (CLIP). Emogen tackles this by manipulating CLIP space. It utilizes a diffusion model, which progressively adds noise to an image until it resembles pure noise. The model then learns to reverse this process, guided by emotionally labelled image-text pairs, to generate images that evoke the desired emotion. By effectively reorganizing the emotional landscape within CLIP, Emogen allows for more targeted image generation based on the intended emotion. This holds promise for various creative fields, offering artists and designers a tool to generate concept art infused with specific moods or creating visuals that perfectly complement the emotional tone of a story.

Disadvantages

Limited to generating new images: Emogen might not be suitable for adding specific emotions to existing images. It seems more focused on generating entirely new images that evoke a particular emotion.

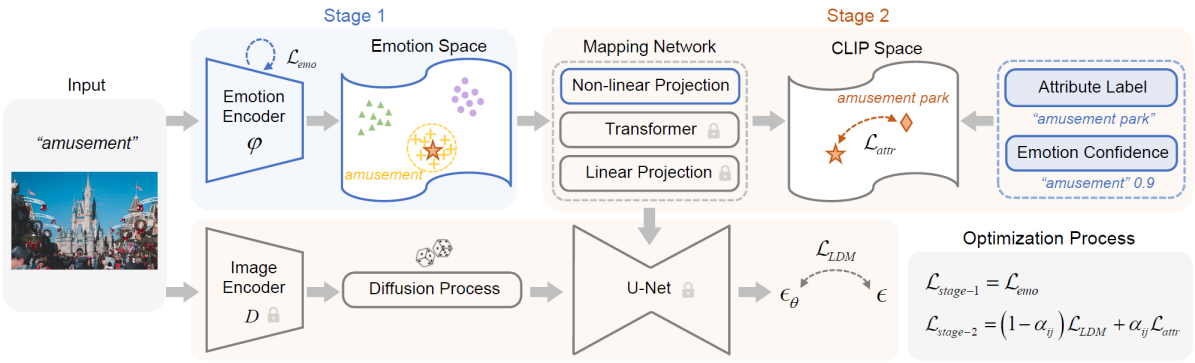

Figure 2.1: Outline of Emogen methodology.

2.2 Textual inversion based methods

Stable diffusion relies on text encoders to translate prompts into numerical representations (embeddings).These embeddings guide the image generation process. We can even swap token embeddings (think of them as unique codes for words) to influence the final image. Textual inversion[**Rinon2016inversion**] builds on this idea. By providing a few example images, we can essentially "teach" the model a new word and its corresponding embedding. This new embedding captures the visual concept represented by the images. We can then use this "pseudo-word" like any other word in the vocabulary, allowing us to create prompts like "a photo of a [learned concept] on the beach" or even combine concepts like "a drawing of concept 1 in the style of concept 2." Remarkably, this is achieved without modifying the underlying image generation model itself.

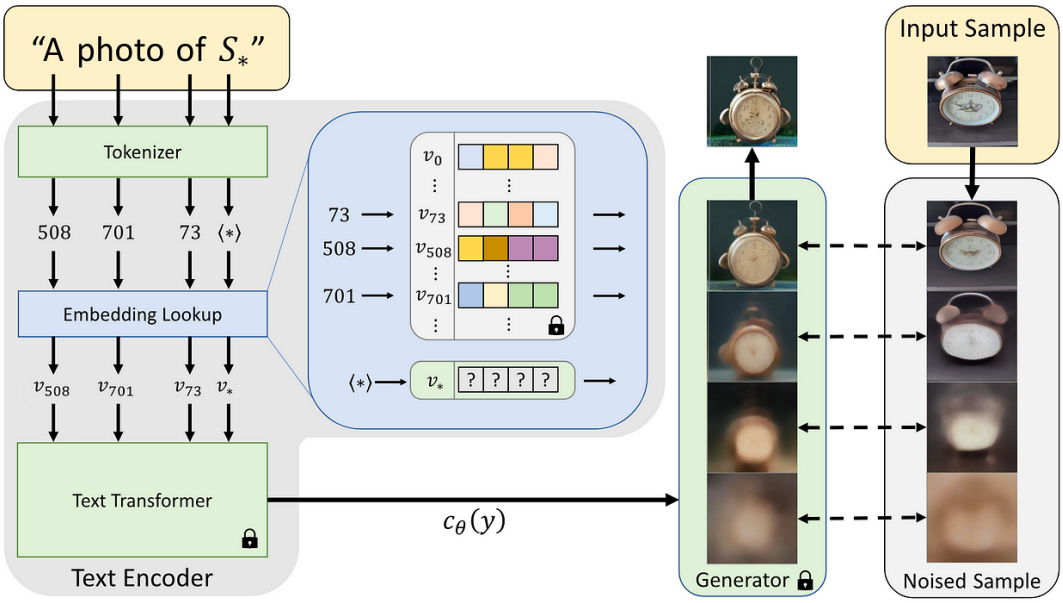

Figure 2.2: Outline of Textual inversion methodology.

Disadvantages

Existing code implementations, particularly those from Hugging Face, might only work with pre-defined concept images. Hence Textual inversion excels at representing specific concepts but may struggle with broader ideas. For instance, a textual inversion for "cat" might generate great cat images, but it wouldn't be ideal for capturing the vastness of the concept "animal."

Also, the quality of the generated image heavily relies on the quality and quantity of the training images used. A small set of blurry photos might lead to blurry or inaccurate representations of the concept.

2.3 Semantic Latent Space

Diffusion models are great at generating images, but they lack a key feature: a dedicated space to control the image content. This space, called a semantic latent space, is crucial for fine-tuning the image during generation.ASYRP [KJU23] introduces the concept of a semantic latent space within the diffusion model, called h-space. This h-space has several advantages for editing images: it's consistent, predictable, and works well across different generation steps. On top of that, ASYRP introduces a method for controlling the editing intensity and measuring the quality of the generated image at each stage. This allows for precise control and improvement of the final image. ASYRP is versatile and can be applied to various diffusion model architectures and datasets, making it a powerful tool for image editing and generation.

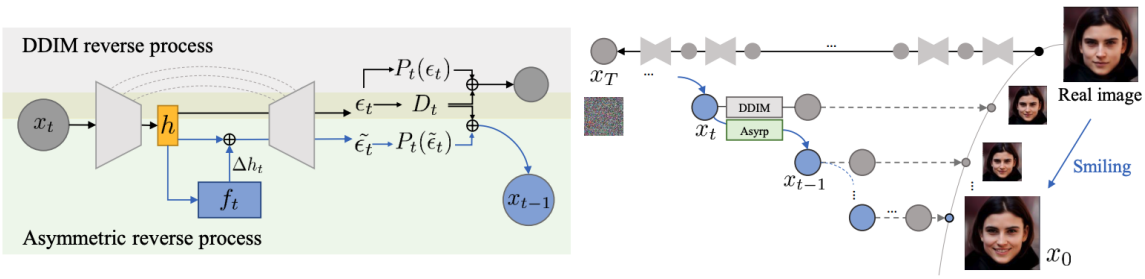

Figure 2.3: Generative process of ASYRP.

In the above figure, the green box on the left illustrates ASYRP which only alters Pt while preserving Dt shared by DDIM. The right describes that ASYRP modifies the original reverse process toward the target attribute reflecting the change in h-space.

2.4 Summary

In this section, we covered different approaches to generating emotion-driven images. The next chapter will explore the proposed approach and the results.

Chapter 3

Investigating Emotion in Images: Experimental Findings

This chapter details a series of experiments designed to evaluate the effectiveness of emotion infusion in images. We tried exploring the capabilities of combining Textual Inversion and Asyrp techniques. Here presenting the results obtained for various emotions on a diverse dataset of images.

3.1 Proposed Framework

We propose a novel approach that integrates Textual Inversion with Asyrp to address this limitation. This combined framework leverages the following aspects:

- **Textual Inversion's Strength:** Capture emotional concepts as custom text embeddings using Textual Inversion.
- **ASYRP's Targeted Editing:** Utilize ASYRP's ability to manipulate the h-space of the diffusion model for targeted edits. But we were unable to combine both the methods.

In our initial approach, we were tried combining Textual Inversion and ASYRP for emotion-driven image synthesis. ASYRP's ability to manipulate the image generation process for targeted edits seemed like a perfect match for Textual Inversion's strength in capturing emotional concepts within image-text relationships. We envisioned a framework that would leverage both techniques: ASYRP for precise control and Textual Inversion for emotional infusion. However, during our experiments, we encountered challenges in integrating these two methods seamlessly within our code. While both ASYRP and Textual Inversion work quite well, results of which are shown in the next chapter. Combining

them proved to be more intricate than initially anticipated. This hurdle prevented us from fully exploring the potential of this combined approach in this project. Despite this setback, the individual results from both ASYRP and Textual Inversion remain almost good, and future advancements in integrating these techniques could lead to significant progress in emotion-driven image generation.

3.2 Textual Inversion

Textual Inversion by learning textual embeddings from example images, it can effectively translate emotional prompts into visual representations. This give good results for pre-trained models. However, its true power lies in the potential for fine-tuning – teaching the model to associate specific emotions with user-provided concepts. Unfortunately, this fine-tuning process often yields less satisfactory results. The model struggles to adapt its learned embeddings to entirely new emotional concepts.

Figure 3.1: Output of Textual inversion with prompt "irritated "

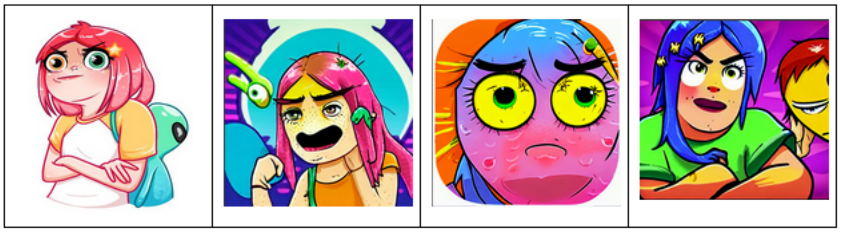

Figure 3.2: Output of Textual inversion with prompt "angry"

3.3 Industry SoTA - Stability AI

Even though our methods were giving decent results, they were nothing in comparison to paid methods which give excellent results. The image quality was extremely high and had many good characteristics from the source image. Below is an example image 3.3. Stability AI provided developer API to transform images which was utilised by us to compare with our methods.

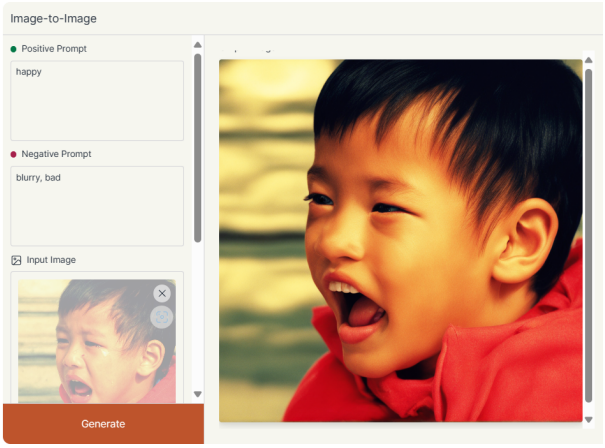

Figure 3.3: Stability AI Sample output

3.4 ASYRP

While ASYRP offers a powerful method for targeted semantic edits, it doesn't directly utilize the image-text mapping capabilities of CLIP. This limits its capacity to leverage pre-trained emotional concepts.

Figure 3.4: Output of ASYRP model

3.4.1 StyleGAN

StyleGAN is a powerful Generative Adversarial Network (GAN) framework known for generating high-quality, realistic images. A key component of this network is W space, which encodes the image's style in a code format where changes have more interpretable effects. StyleGAN used this code to modulate different layers in its generator network, to control the image's appearance. By manipulating the code in W space, we can target the specific features of an image. Since this W space has more semantic structure, modifying the code and so the changes in expressions can be more meaningful.

Figure 3.5: Output of StyleGAN model for different emotions

3.4.1.1 Observations

While it excels on familiar data, new images suffer from distortions like Closed eyes, Corrupted background, Jumbled facial features etc.

3.4.2 VAE based inducer

https://github.com/leon-schi/face-flex[leo]

We observe that the system generates smiling images from inputs, but often with distortion or blurring. Additionally, it seems limited to producing smiles. Our proposed solution to address this issue involves expanding the range of emotions portrayed while eliminating the distortion effect applied to the original image. [Jin24]

Figure 3.6: Output of VAE-based inducer for smiling as emotion

3.5 Challenges

Diffusion models have randomness built in, so controlling them with an image isn't perfect. Edits can blur the image, making details like eyes look worse. The model might only work well for images similar to its training data (e.g., portraits). Complex backgrounds or non-portrait images are difficult to control effectively.

Chapter 4

Conclusion

This section outlines the next steps in our project and areas for further exploration.

- **Codebase Exploration:** We have familiarized ourselves with the codebases of Textual inversion, existing Generative Adversarial Networks (GANs) and Variational Autoencoders (VAEs) designed for emotion induction.
- **Hugging Face Integration:** We have explored relevant Hugging Face models, including CLIP (for image-text mapping), Stable Diffusion (for image generation), and text-to-image pipelines. We have also conducted preliminary experiments.
- **ASYRP Adaptation:** We can work on modifying the baseline ASYRP method to achieve better results with custom datasets.

4.1 Summary and Conclusion

Our literature survey examined various approaches, including EmoGen for enhanced emotional control in new image generation, Textual Inversion for adding emotions to existing images, and ASYRP for targeted emotional edits within diffusion models.

To address the limitations of these techniques, we proposed a novel framework that combines Textual Inversion's strength in capturing emotional concepts with ASYRP's ability for targeted editing in the latent space of diffusion models. This approach has the potential to achieve both enhanced emotional control through pre-trained CLIP concepts and preserved image quality by leveraging ASYRP's core editing process.

Textual inversion based image editing was successfully explored and experimented in this project and the overall learning experience was really smooth.

Bibliography

[Gal+22] Rinon Gal et al. *An Image is Worth One Word: Personalizing Text-to-Image Generation using Textual Inversion*. 2022. arXiv: 2208.01618 [cs.CV].

[KJU23] Mingi Kwon, Jaeseok Jeong, and Youngjung Uh. "Diffusion Models already have a Semantic Latent Space". In: *arXiv preprint arXiv:2210.10960* (2023).

[Jin24] Hui Huang Jingyuan Yang Jiawei Feng. "EmoGen: Emotional Image Content Generation with Text-to-Image Diffusion Models". In: *arXiv preprint arXiv:2401.04608* (2024).

[leo] leon-schi. *face-flex*. URL: https://github.com/leon-schi/face-flex.