ECE 580 Project Proposal

Team Members

Yiming Mao(ym241)
Enmu Liu(el296)

Problem Statement

Currently, most generative models excel at generating images based on user-provided text prompts. However, when a user wants to edit an existing image while keeping the subject looking the same, these models often fails. Either the model changes the subject so much that it's no longer recognizable, or it fails to follow the new instructions properly.

Last year, I tried to make a birthday card with my cat on it using models like DALL·E and DeepAI. But, the results were disappointing. Either my cat looked completely different, or the model just didn't understand what I wanted.

To address this, we're proposing a model that can bind a unique identifier to a subject. This would allow the model to generate high-quality images of the subject in different contexts while preserving its original appearance.

This really interests us since it could have a lot of useful applications. For example, in fashion, brands could use it to show how their products— clothes or jewelry —would look on individuals of various heights, ethnicities, and body types without needing to hire multiple models and conduct expensive photoshoots. Furniture companies could use it to let customers see how a piece of furniture fits into their space. Lastly, I am finally able to make a proper birthday card with my cat on it saying "Happy Birthday!".

ML Task(s)

Our project focuses on using machine learning to modify images while keeping the subject's appearance consistent. The key tasks include:

- **Subject Recontextualization**: The model can place a subject into new environments while keeping their original look.
- **Text-Guided View Synthesis**: Users can describe changes in natural language, and the model will adjust the image accordingly. This combines text-to-image generation with multi-view synthesis.
- **Appearance Modification**: The model allows small changes, like switching outfits or adding accessories, without altering key features.
- **Artistic Rendering**: Users can apply different artistic styles (like sketches, paintings, or cartoons) while keeping the subject recognizable.

To achieve this, we will use **diffusion models, transformers, and fine-tuned embeddings**, ensuring that the generated images stay true to the original subject while allowing creative modifications.

Data Involved

Description: We will use the official dataset from Google's DreamBooth paper. This dataset consists of images of various subjects captured under different conditions, environments, and angles to improve model generalization.

Source: The dataset is available on GitHub: https://github.com/google/dreambooth

Features: The dataset includes 30 subjects from 15 different classes. Each subject has a variable number of images (4-6 per subject), ensuring diverse representations of the same entity in different contexts.

Models to Consider

To generate and edit images while keeping the subject's appearance consistent, we may use:

Diffusion Models

- Stable Diffusion
- DeepFloyd IF

Fine-Tuning Methods

- DreamBooth
- LoRA (Low-Rank Adaptation)
- ControlNet

Expected Results

By fine-tuning a model with our approach, we expect to achieve accurate subject re-contextualization while maintaining high fidelity to its key features. Our model will ensures:

- The subject remains **recognizable** across different scenes.
- The images look **natural**, with realistic lighting, perspective, and interactions with their surroundings.
- Users can guide the generation process through **text-based prompts**, making modifications intuitive and flexible.

List at least one relevant paper to the project

[1] Ruiz, Nataniel, et al. "Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation." *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2023.
[2] Cao, Pu, et al. "Controllable generation with text-to-image diffusion models: A survey." arXiv preprint arXiv:2403.04279 (2024).
[3] Zhang, Lvmin, Anyi Rao, and Maneesh Agrawala. "Adding conditional control to text-to-image diffusion models." *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2023.
[4] Chen, Wenhu, et al. "Subject-driven text-to-image generation via apprenticeship learning." *Advances in Neural Information Processing Systems* 36 (2024).