Jacky Lin, Matthew Thomas
COE379L
Fall 2025

Initial Proposal

Introduction

One of the most consequential features of deep learning models is that they require an extensive amount of training data to be able to learn patterns and perform accurate inference. This can be a major roadblock to building models that can make predictions on images that are highly limited in nature. This is the motivating factor behind synthetic data generation, which involves artificially creating images that deep learning models can train on to better learn the features of the object of interest when obtaining real images is difficult. For this project, we aim to explore how well the feature representations learned by a CNN from real-world photographic data transfer to images synthesized by generative AI models and vice versa. Synthetic images possess many differences from real images stemming from the prompt used to generate it, hallucinations, random factors, and more. This work aims to quantify the resulting performance gap in classification.

Data Sources

The existing data used for this project will be a subset of CIFAR-10, which is a widely known dataset containing a variety of different labeled images. We decided to use this dataset because it is one of the standard datasets that are used to generate baseline/benchmark performances in computer vision tasks, and it is easily available via the Keras API. The subset we will use will contain only images of the animal classes (bird, deer, dog, cat, horse, frog), but with the full 6,000 images of each class. The second set of data we will be using will be generated by a stabilityai text-to-image diffusion model from Hugging Face. Using basic prompt engineering, we will produce 6,000 images of each animal class. The CIFAR-10 dataset can be found here https://www.cs.toronto.edu/~kriz/cifar.html.

Techniques and Technologies

For this project, we want to chain together a generative AI model and a CNN classifier to evaluate how well they can be used together to produce accurate inference. To first create the synthetic dataset, we will prompt a diffusion model to generate 6,000 images of each animal class in CIFAR-10. Once we have the paired generated and real data, we will run three experiments:

1. Train CNN on synthetic data, test on CIFAR-10 (most applicable to real-world)
2. Train CNN on CIFAR-10, test on synthetic data

3. Train and test CNN on a mixture of the two data types

We will ideally utilize the same CNN architecture for each experiment and assess each model's performance by analyzing accuracy and confusion matrices. This allows for both a visual and numerical understanding of how well the model trained on synthetic data can generalize to real data and vice versa.

Products and Deliverables

For this project, our main deliverable will be a report detailing our discoveries through our exploration of this problem. We expect to return quantitative results on how well generative AI models can be used to produce additional data to feed into image classifiers. Secondly, we will deploy a Flask server that users can use to experiment with generating images and classifying them with one of our models. The entire software will be containerized in Docker and is meant to allow inexperienced users to explore generative AI models and model chaining.