Project Vision Document

Project Title and Name: Visual Layer

Team Members:
- Alec Song: alecsong@ucsb.edu
- Bhavya Ranjan: bhavyaranjan@ucsb.edu
- Karen Yuan: karenyuan@ucsb.edu
- Rushil Gupta: rushil@ucsb.edu
- Saeed Arellano: saeedarellano@ucsb.edu

Team Lead: Alec Song

Company Overview:

Visual Layer is a Tel Aviv-based startup founded in 2022 that builds tools to manage, clean, and enrich massive visual datasets used for training AI and computer vision models. Their platform, built on the open-source project fastdup, helps organizations detect duplicates, mislabels, corrupt files, and other quality issues at scale, ensuring models are trained on the most reliable data. Visual Layer's mission is to make large-scale visual data curation efficient, scalable, and accessible so teams can focus on building better-performing AI systems.

Problem Statement:

Despite the rapid advancement of computer vision models, dataset quality remains one of the most under-quantified determinants of model performance. Prior studies have shown that even modest data degradation can have measurable effects. For example, image blur alone can reduce classification accuracy by 20-40%, and removing just 5-10% of low-quality or mislabeled samples can improve accuracy by 3-7%. However, most large-scale pretraining datasets like ImageNet and COCO contain 10-25% noisy or low-quality samples, ranging from mislabeled images to compression artifacts and poor lighting conditions. We aim to empirically evaluate how dataset quality affects deep learning performance in computer vision. Models will be pretrained on both raw and curated datasets to measure differences in accuracy, robustness, and generalization. Our goal is to quantify the direct impact of data cleanliness and enrichment on model outcomes.

Project Overview:

We will be investigating how the quality of pretraining data directly influences model accuracy, robustness, and generalizability, and how data leakage occurs in training large language models for computer vision depending on the pretraining data. We aim to develop new methods to

prevent/fix data leakage in these visual dataset through systematic experiments. Hence, it is important to focus on finding a generalized method of how the quality of the pretraining data affects the quality of the machine learning model performance and which type of datasets gives the best results for training machine learning models in computer vision.

Project Goals:

- Investigate how dataset quality impacts model performance
    - Evaluate how cleaning and enriching datasets affect the accuracy, robustness, and generalizability of deep learning models
    - Identify and address biases, inconsistencies, and quality issues common in public datasets
    - Use Visual Layer's tools to enhance datasets and evaluate how each improvement influences model performance and outcomes
- Conduct controlled deep learning experiments
    - Train and benchmark computer vision models on a variety of researched datasets, on both the original and enriched datasets
    - Ensure rigorous experimental design and statistical validity
    - Maintain consistent training parameters to isolate the effects of data quality
- Contribute open-source resources
    - Release optimized datasets and code on HuggingFace
    - Publish findings via a technical research paper suitable for submission
        - Summarizing methods and sharing experiments and results

Milestones:

- Compile list of open-source, public training data to run experiments on.
- Identify and address biases, inconsistencies, and quality issues inherent in commonly used public datasets.
- Get acquainted with and learn how to use Visual Layer's cutting-edge tools to systematically enhance these datasets.
- Gather performance data from models trained on cleaned versus uncleaned datasets.
- Compile data into a technical paper suitable for submission to a conference or journal.

Technical Approach:

1. Visual Datasets Research and Selection
    a. Conduct review of existing public computer vision datasets of 1-4 million images (e.g. Imagenet, Coco, Pixmo) to identify potential datasets and compile a comprehensive list of candidates
    b. Evaluate list of datasets based on accuracy, diversity and usefulness
    c. Select ~10 datasets that will serve as the foundation of research project

2. Model Training
    a. Train deep learning models and architectures on original and curated datasets under identical factors and conditions to ensure equal comparison
    b. Track model performance across key metrics such as accuracy, robustness to noise, etc.
3. Experimental Evaluation and Statistical Analysis
    a. Conduct statistical significance tests to rigorously assess the performance differences between datasets trained in models
    b. Visualize results through error heatmaps and graphs
4. Documentation and Publication
    a. Prepare a comprehensive technical report explaining methodologies, results, experiments, and more for potential publication in machine learning and computer vision journals or conferences.

Conclusion:

This project aims to prove that better, cleaned data translates directly to improved model performance and quality. By comparing datasets curated by Visual Layer against unfiltered, uncleaned datasets, we will compile and present relevant data that proves this hypothesis, and to further validate the value and relevance of clean training data going forwards.