Title: CarID: A Machine Learning-Based Application for Accurate Car Model, Make, and Year Recognition from User-Uploaded Photos

Team Members: Nuoya Jiang, Harper Wang

Contact Information: nuoyajiang@hms.harvard.edu, xinyi_wang2@hms.harvard.edu

Project Repository: https://github.com/xinyi-wang02/ac2152024_group_X/tree/milestone4

Project Scope: Develop an application that can accurately identify the car model, make, and year from user-uploaded photos. Implement microservices for each step of the application, ensuring scalability and the ability to handle multiple requests concurrently.

Minimum Components for a Good Project
Large Data: Leveraging a dataset of 16,185 car images of 196 different classes
Scalability: Capable of processing multiple user queries simultaneously
Complex Models: Fine-tune machine learning-based image analysis models for accurate prediction of car models, make, and year
Computationally Expensive Inference: Deploy trained model on cloud to reduce latency during prediction

Objectives:
1. Preprocess and augment the original image dataset, then store the processed data in a Google Cloud Platform bucket for easy access and mounting.
2. Compare various computer vision model architectures and fine-tune the best-performing model to fit the car dataset.
3. Deploy the whole machine learning pipeline and the top-performing model to Vertex AI and create an endpoint to enable API access.
4. Design an intuitive and user-friendly frontend for users to upload the photos and receive their results.

Background and Motivation: This project aims to develop a machine learning-based application that can accurately identify car models, makes, and years from user-uploaded images, addressing the rising demand in the automotive industry for automated vehicle recognition. Using a dataset of 16,185 images spanning 196 car classes, the project tackles the challenges posed by the variety of car models and image conditions. By comparing and fine-tuning different deep learning models, the best-performing structure will be deployed to the cloud for scalability and reduced latency. The end goal is to provide users with a fast, intuitive tool capable of handling multiple requests efficiently.

Source of Data: The data source is from Stanford Car Dataset by classes folder on Kaggle. The data can be accessed here:
https://www.kaggle.com/datasets/jutrera/stanford-car-dataset-by-classes-folder/data

Description of Dataset: The dataset used in this project is the Stanford Cars dataset, which comprises 16,185 images of cars across 196 classes. The dataset is divided into 8,144 training images and 8,041 testing images, with each class representing a specific car make, model, and year (e.g., 2012 Tesla Model S or 2012 BMW M3 coupe) and split roughly 50-50 between training and testing sets. The dataset is organized in a folder structure that separates the images into “train” and “test” directories, each containing subfolders labeled by car class.

Data Quality Concerns: The images are of high quality; however, there is a class imbalance among some categories. For instance, the dataset contains a disproportionately large number of Chevrolet cars and cars produced in 2012.