



Driven Data

Intelligent Vehicle Identification for Market Data Generation

Chase Kent-Dotson

April 4, 2022

Overview


Car manufacturers face an increasingly cutthroat business environment, and in modern times the company that has the data has the advantage. Through channels such as dealership data, customer address data, and vehicle GPS, companies can obtain varying levels of market data on the location of cars they have sold, but what about such information on their competitor's vehicles? What about information on the distribution of makes and models within a target market? This data is difficult to even estimate via most traditional methods such as surveys, which presents the perfect opportunity to apply machine vision techniques at the local level combined with a big data backbone at the national level.

Solution Details

My proposed solution is as follows: use deep learning to identify vehicles using on-car and stationary cameras that are already in use, recording location data, make, model, year, and other machine vision identifiable features, then use that data to populate a national database containing information about cars all over the country. This database could be stored using commercial cloud computing services, then data engineers could build pipelines for data scientists to access the data, who in turn could build further models, dashboards, and tests to deliver insights to marketing teams and executives which provide a competitive advantage through data informed decision making.

Existing Work

Much work has been done on this problem already, and for this project I will focus on the identification of vehicle identities by machine vision, that is to say: telling the difference between one vehicle and another. A large component of the problem and an area in which much work has been done already is the identification of vehicle locations within an image, that is to say: telling the difference between a vehicle and the road or the surrounding environment. The article [here](#) discusses this part of the problem in detail. It is an essential component of a complete solution which goes hand in hand with the model at the heart of this project (determining vehicle identities) used to generate the market data from vehicle images.



Determining vehicle makes, models, and years is something that is currently being tackled by state of the art tools. The golden standard frameworks use “bootstrapping” methods to train an initial model, then use the features extracted from that model to reduce the labor required to label large amounts of additional data, which can then be used to further improve the model. An example can be seen [here](#). My project uses the same deep learning methods that would be found at the heart of such a bootstrapping model.

The Dataset

To train a model in classifying cars, the first thing necessary is a dataset of car images. I used the Stanford AI Car Dataset, found [here](#). It is a collection of roughly 16,000 vehicle images from 196 classes, with a class format of “Make Model Type Year.” The images are contained in directories divided and labeled by class, and the entire dataset is split roughly in half for training and testing use. CSV files are also included with class names and numbers as well as bounding box information for the location of the vehicle within each image.

The source website and files do not list the acquisition methods, though it does appear to be human labeled or at least human reviewed as I found zero label mistakes throughout all stages of my entire project. The data is somewhat lacking in that it has not been updated with additional current vehicles and will not be updated again. The problem becomes easier with greater amounts of car images which I would like to have, but without building more advanced automated systems to label them that is difficult to achieve.

Data Exploration and Preparation

I started the process of getting my data ready for modeling by diving into the image files and the support files containing the class information. I verified that the class distribution was appropriate for successful model training, which it is as the images are distributed rather evenly among all classes which are fully represented in both the training and test data.

The provided files contain bounding box information for each image, which would be used as the target by a model attempting to differentiate cars from not cars, or find a car within an image as I explained above. For the purposes of my model, determining the type of car, I decided to use the bounding box information to crop the images such that each image contains as little environment information and as much vehicle information as possible to allow the model to focus on vehicle features. With the images cropped, I applied scaling and normalization. Scaling sets the image values within the range expected by the model, and normalization helps the model's parameters apply evenly to all images. Random augmentations were also applied to help prevent overfitting.

Modeling

Initially, CNNs were built and trained from scratch. Though initial attempts lacked predictive power, improvements that minimized overfitting and maximized learning capability improved these models dramatically, but not to a level sufficient for generating quality market data. I began training and optimizing transfer learning models with pre-trained weights to better identify the vehicles, and achieved a final accuracy of 77% using my best model to predict the vehicle classes in the test data.

Model Evaluation

Generally speaking, the model is very good at identifying boxy-shaped vehicles with unique features, such as Lamborghinis, Jeeps, and Hummers. The model falters with inversely described classes, those with rounded shapes and generic features. It seems color may play a role as well in that some classes with high percentages of images containing an unorthodox paint color were also predicted with high precision.

Digging into the class-by-class evaluation metrics revealed that false positives were being generated disproportionately among some vehicle classes. While performing this evaluation, I discovered I could re-evaluate the model's capabilities by identifying vehicles which were predicted as the wrong class, but the correct make. I re-scored the model using makes as the classes and achieved an accuracy of 86%. In further evaluation, I again identified false positives predicted as the wrong make or model but correct body type. Re-scoring the model by body type yielded an accuracy of 90%.

Conclusion and Next Steps

I met my original goal of creating a model capable of predicting the make, model, and year of vehicles with an accuracy of greater than 70%. The value of this model as it stands is good, and could be made far better in a production system with automated labeling to obtain additional data as I mentioned earlier, which would be the next step. After that, the model could be coupled with a "vehicle or not vehicle" model to identify where cars are in images, deployed to on-car and stationary cameras, then start collecting and feeding data to a large database. The company deploying this model could then use the data to determine what types of cars are most popular in certain areas, which cars are most common for different uses (commuting, road trips, city driving, dirt roads, etc.), or even trends among body types—which in turn can be used to make smart, data-informed decisions about marketing, production, and design.