Fine-Grained Image Classification for Vehicle Makes & Models

Utilizing Convolutional Neural Networks (CNNs)

Nicholas Benavides & Christian Tae - {nbenav, ctae}@stanford.edu

Overview

Fine-grained classification is a challenging problem because differences between classes tend to be much more nuanced compared to class differences in tasks such as object detection. Specifically for vehicle classification, humans perform well due to their ability to identify small details like a logo or lettering, but the complexity of a car has made this task difficult for computers. In this project, we applied transfer learning from ImageNet on VGG16 to discern a vehicle's make and model from an image.

After tuning our modified VGG16 network, we achieved a test accuracy of 85%, top-3 test accuracy of 93.9%, and top-5 test accuracy of 96.3%, surpassing state-of-the-art results.

Data

The Stanford Cars dataset[1] contains 16,185 image classification pairs of 196 different classes, where a class is defined a vehicle's make, model, and year.

Figure 1. Sample images of cars dataset.

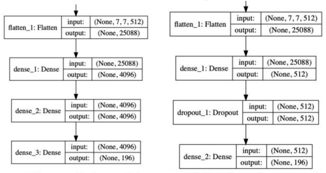
As shown above, each image contains a car in the foreground against various backgrounds viewed at different angles. For each image, coordinates of the bounding boxes that enclose the car are given for preprocessing.

Model

We applied transfer learning to a VGG-16 model. For our baseline, we changed the output layer to have 196 classes and fine-tuned the dense layers. For our best model, we eliminated the first dense layer, changed the dimension of the second dense layer to 512, and added a Dropout layer. Both models used the categorical cross-entropy loss, defined below.

$$-\sum_{c=1}^{196} y_{o,c} \log(p_{o,c})$$

In the figure below, you can see how our best model differs from the baseline model. The only differences come in the last 4 layers of the network.



flatten_1: Flatten | input: (None, 7, 7, 512) | output: (None, 25088)

dense_1: Dense | input: (None, 25088) | output: (None, 4096)

dense_2: Dense | input: (None, 4096) | output: (None, 4096)

dense_3: Dense | input: (None, 4096) | output: (None, 196)

flatten_1: Flatten | input: (None, 7, 7, 512) | output: (None, 25088)

dense_1: Dense | input: (None, 25088) | output: (None, 512)

dropout_1: Dropout | input: (None, 512) | output: (None, 512)

dense_2: Dense | input: (None, 512) | output: (None, 196)

Figure 2. Last 4 Layers of the Baseline model (left) vs Modified Vgg-16 model (right).

Methods

Each vehicle image is represented as a 224 x 224 x 3 matrix, which is the input size for VGG16. To process the images, we first cropped each image using the bounding box provided, normalized the image pixel values, and then ran a random stratified train-test split to divide the dataset. For the training images, we augmented them by 1) flipping images horizontally, 2) randomly rotating images, and 3) applying random horizontal and vertical shifts.

VGG Network Architecture

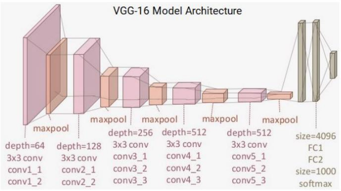


Figure 3. VGG-16 Model Architecture

Results

Table 1. Vehicle Classification Results. For each image, the model assigns a probability to each of the 196 classes. In the table, accuracy corresponds to to when the true class is assigned the highest probability, and top-k accuracy corresponds to when the true class is assigned one the k highest probabilities.

Model	Training Acc.	Training Top-3 Acc.	Training Top-5 Acc.	Test Acc.	Test Top-3 Acc.	Test Top-5 Acc.
VGG16 Baseline	85.8%	96.6%	98.2%	52.1%	75.4%	83.3%
Modified VGG16 Network	84.1%	95.6%	97.9%	84.0%	93.9%	96.3%

Training Set Size: 11,303 images
Test Set Size: 2,426 images

Discussion

For error analysis, we examined 100 images that were not properly classified to gain insights into how the model works. From this, we identified **three main groups of misclassifications:** one where the model correctly predicts the make but not the model of the car, one where the model predicts a similar style of car from a different make, and rearview images. Examples of these misclassifications can be seen below.

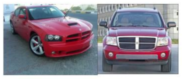


Figure 4. Input = Dodge Charger Sedan 2012 (left), Model Prediction = Dodge Durango SUV 2007.

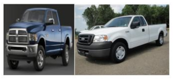


Figure 5. Input = Dodge Ram Pickup 3500 Crew Cab 2010 (left), Prediction = Ford F-150 Regular Cab 2007.

Future Work

To improve our model's test performance on classifying rear-view images, we would obtain more of those image types for our training set. In addition, we would want to experiment with other model architectures as well as conduct more visualizations of the network to better understand how the model distinguishes between classes.

References

[1] J. Kraus et al., "Collecting a Large-Scale Dataset of Fine-Grained Cars," Department of Computer Science, Stanford University. Available: http://ai.stanford.edu/~jkrause/papers/fgvc13.pdf

[2] Simonyan, Karen, and Andrew Zisserman. "Very deep convolutional networks for large-scale image recognition." *arXiv preprint arXiv:1409.1556*(2014).