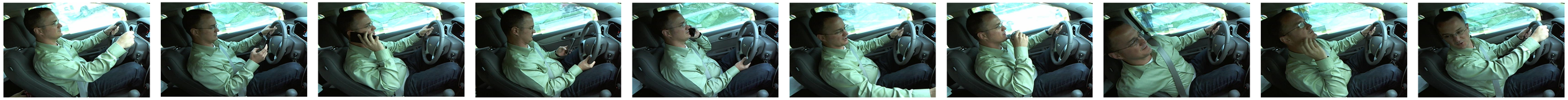
Classification of Driver Distraction

Samuel Colbran, Kaiqi Cen, Danni Luo

samuco@stanford.edu, kaiqi@stanford.edu, danniluo@stanford.edu

Stanford University



Summary

Distracted driving causes more than 3,000 deaths and 424,000 injuries every year. Motivated to reduce these statistics, our project aims to accurately classify what drivers are doing and whether they are distracted based on images of them while driving. We use convolutional neural networks and investigate techniques to improve their performance in an image recognition setting. We ran experiments with different combination involving VGG-16 and AlexNet. The best result so far, achieved with an ensemble, ranks No. 217 (top 15%) on Kaggle with a leaderboard (LB) score of 0.32706.

Data

The data was supplied by StateFarm (an insurance company) for a public Kaggle challenge. It consists of 22400 training and 79727 testing images (640x480 full color) of people either driving safely or doing 9 distracted behaviours:

c0: safe driving, c1: texting - right, c2: talking on the phone - right
c3: texting - left, c4: talking on the phone - left
c5: operating the radio, c6: drinking, c7: reaching behind
c8: hair and makeup, c9: talking to passenger.

An example of each class is shown in the ribbon of images at the top of the poster. The train images come with correct labels and the challenge is to make the best multi-class classifications we can.

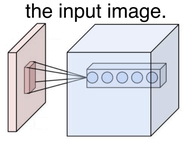
To evaluate the success of models, we split the train images into "train" and "test" sets. The trick is to choose the images of a certain driver in the training set to be the “test images”. Only then can these “test images” be independent from the remaining training images of other drivers and we can avoid a falsely high test accuracy. The provided test images are used only when making a submission to Kaggle, which provides us with a ranking that we can compare with other competitors.

Models

Convolutional networks generally use several common building block layers:

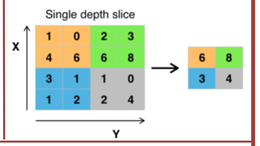
Convolutional

A layer that learns filter functions that activate when a specific type of feature appears at some location in the input image.



Pooling

Used as a form of regularization. Partitions an image into regions and returns the maximum.



ReLU

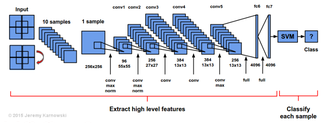
Increases the nonlinear properties of the decision function using the activation function f(x) = max(0, x)

Fully Connected (FC)

Performs high-level reasoning using a large matrix multiplication based on full connections to all activations in the previous layer. Both AlexNet and VGG-16 use Softmax activation on the final FC layer.

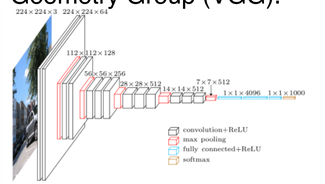
AlexNet

An 8-layer convolutional network developed by Alex Krizhevsky.



VGG-16

A 16-layer convolutional network developed by the University of Oxford Visual Geometry Group (VGG).

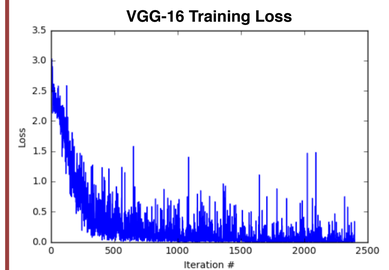


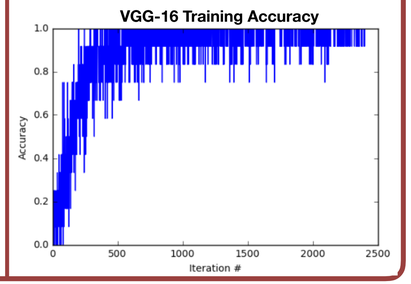
Results

Models	Iterations	Learning Rate	Weight Decay (regularizaion)	LB Score
AlexNet	600	1E-03	5E-04	0.69126
VGG1	2400	1E-03	5E-04	0.49005
VGG2	2400	1E-04	5E-04	0.45361
VGG3	2400	1E-04	1E-03	0.55385
VGG4	5000	1E-03	53-04	0.53085
Ensemble1	CaffeNet + Vgg2			0.44084
Ensemble2	CaffeNet + Vgg1 + Vgg2 + Vgg3 +Vgg4			0.34354
Ensemble3	Vgg1 + Vgg2 + Vgg3 + Vgg4			0.32706

With Ensemble 3, we rank No. 217 (top 15%) among total 1440 participants on Kaggle's LB with a score of 0.32706. The metric used by Kaggle is the following multi-class log loss function:

$$logloss = -\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{M} y_{ij}\log(p_{ij}),$$





Discussion

- **Transfer learning with pre-trained VGG-16 model** gave us a significant boost in terms of speed and performance. We modified the last fully connected layer to output 10 class predictions.
- **Training** - The test set contains images of drivers that **never** appeared in the training set to well-reflect that we randomly picked 3 drivers and all of their images as validation set for training (10% of the whole training set). Due to limitation in memory, our implementation in Caffe uses batch size of 12 images per iteration (relatively small size) and this might be the reason for our fuzzy training curve.
- **Fine tuning parameters** - With **learning rate** from 0.0001 to 0.00001, our VGG model converges well. 2400 **training iterations** (1.5 epoch) generally has the best performance with validation accuracy around 80% - 86%. Higher iterations will likely cause overfit and lower the validation accuracy. Changing the default **regularization parameters** is not helpful in terms of improving. Therefore, we stick with the default value 5e-04 most of the time.
- **Ensemble** - **Overfitting** is the major problem we faced as testing set is way larger than training set. Ensembling different models effectively reduce our results' general error and improved our final score (loss) on the leaderboard. Our best results ensembles top 4 single VGG-16 models with various parameters.

Ongoing and Future Work

- Currently training and debugging ResNet-152 model - aiming to add to the final ensemble
- K-fold cross validation + Ensemble K models
- Data Augmentation (NN: neighbour images highly correlated)

References

[1] Centers for Disease Control and Prevention. Distracted Driving. https://www.cdc.gov/motorvehiclesafety/distracted_driving/
[2] Kaggle. State Farm Distracted Driver Detection. https://www.kaggle.com/c/state-farm-distracted-driver-detection/data
[3] Kaggle. A brief summary. https://www.kaggle.com/c/state-farm-distracted-driver-detection/forums/t/22906/a-brief-summary
[4] BVLC. http://dl.caffe.berkeleyvision.org/
[5] GitHub. kaggle-statefarm. https://github.com/alireza-a/kaggle-statefarm/blob/master/src/fine_tune_caffe_net.ipynb
[6] Cornell University Library. Deep Residual Learning for Image Recognition. ResNet: https://arxiv.org/abs/1512.03385