

1. Definition

Project Overview

According to recent statistics on car accidents, one in five are caused by distracted drivers. State Farm¹, an insurance company based in the USA, hope to improve these alarming statistics, and better insure their customers, by testing whether dashboard mounted cameras can automatically detect drivers engaging in distracting behaviours.

State Farm has challenged the Kaggle² community to classify drivers' behaviour based on a dataset of 2D dashboard camera images. The aim is to determine whether the driver is paying attention to the road, distracted by passengers in the car, drinking coffee, or using a mobile phone.

The aim of this Capstone project is to train a machine learning algorithm to classify the behaviour of drivers using this dataset of 2D dashboard camera images.

Problem Statement

The problem that this project aims to solve is to train a machine learning algorithm that can correctly identify the behaviour of drivers based on 2D images captured by a dashboard camera. These images will be used to determine whether a driver spends more time focused and driving attentively, or whether they are distracted by other passengers, by drinking coffee, using a mobile phone or applying makeup.

There are a number of different machine learning approaches that could be taken, but in this study we look into CNN-based image recognition algorithms. CNNs, or Convolutional Neural Networks, are a type of feed-forward artificial neural network in which the connectivity pattern between neurons is inspired by the organisation of an animal visual cortex, whose individual neurons are arranged in such a way that they respond to overlapping regions tiling the visual field³. When used for image recognition, CNNs consist of multiple layers of small neuron collections which process portions of the input image, called receptive fields. The outputs of these collections are then tiled so that their input regions overlap, to obtain a better representation of the original image; this is repeated for every such layer. Tiling allows CNNs to tolerate translation of the input image.

Convolutional neural networks are often used in image recognition systems. They have set the standard in many image classification benchmarks such as MNIST⁴ and ILSVRC⁵. As this is an image classification problem, a CNN based machine learning algorithm will be used. The input of the network will be the 2D image taken from the dashboard camera and the output from the network is the predicted likelihood of what the driver is doing in each image.

The approach is illustrated in the following diagram:


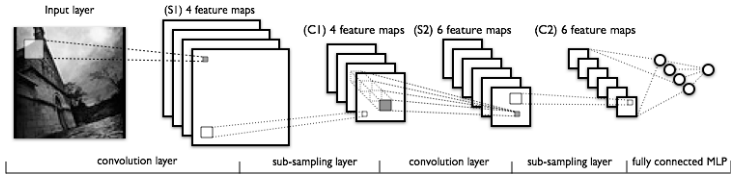
¹ <https://www.statefarm.com/>

² <https://kaggle.com> is a platform for predictive modelling and analytics competitions on which companies and researchers post their data and statisticians and data miners from all over the world compete to produce the best models.

³ https://en.wikipedia.org/wiki/Convolutional_neural_network

⁴ https://en.wikipedia.org/wiki/MNIST_database

⁵ https://en.wikipedia.org/wiki/ImageNet#ImageNet_Challenge

2D image of the driver	Neural Network processes the image	Classification
		Using phone

The following have been defined by State Farm as the list of classes for this exercise:

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

The neural network will be built using Python, Theano⁶ and Keras⁷. Theano is a python library that allows the programmer to define, optimise and evaluate mathematical expressions involving multi-dimensional arrays efficiently. It has the ability to use a GPU if one is available purely by changing configuration options. Keras is a minimalist, highly modular neural networks library written in Python and capable of running on top of Theano (or TensorFlow), and will be the basis for the CNN built to solve this problem.

Metrics

Kaggle's requirement is to predict the likelihood of what the driver is doing in each image across all 10 possible classifications. Performance of the classifier is to be measured using a categorical cross entropy function over validation data.

Log loss is a classification loss function often used as an evaluation metric in Kaggle competitions, and is defined as:

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

where N is the number of samples, M is the number of possible labels/classes, y_{ij} is a binary indicator of whether or not label j is the correct classification for instance i, and p_{ij} is the model

⁶ <http://deeplearning.net/software/theano/>

⁷ <https://keras.io>

probability of assigning label j to instance i . A perfect classifier would have a log loss of precisely zero, and less ideal classifiers have progressively larger values of log loss.

One of the challenges with Log Loss is that it heavily penalises classifiers that are confident about an incorrect classification. This means that it is better to be somewhat wrong than completely wrong, and suggests that smoothing the results set may provide a better overall benchmark.

Although the competition has now closed, Kaggle still permit submissions of test results to give an indication of where you would have ended up had you entered when the competition was still open. Because of this, it is possible to get an accurate indication of the log loss from the test data provided by submitting results back to Kaggle.

2. Analysis

Data Exploration

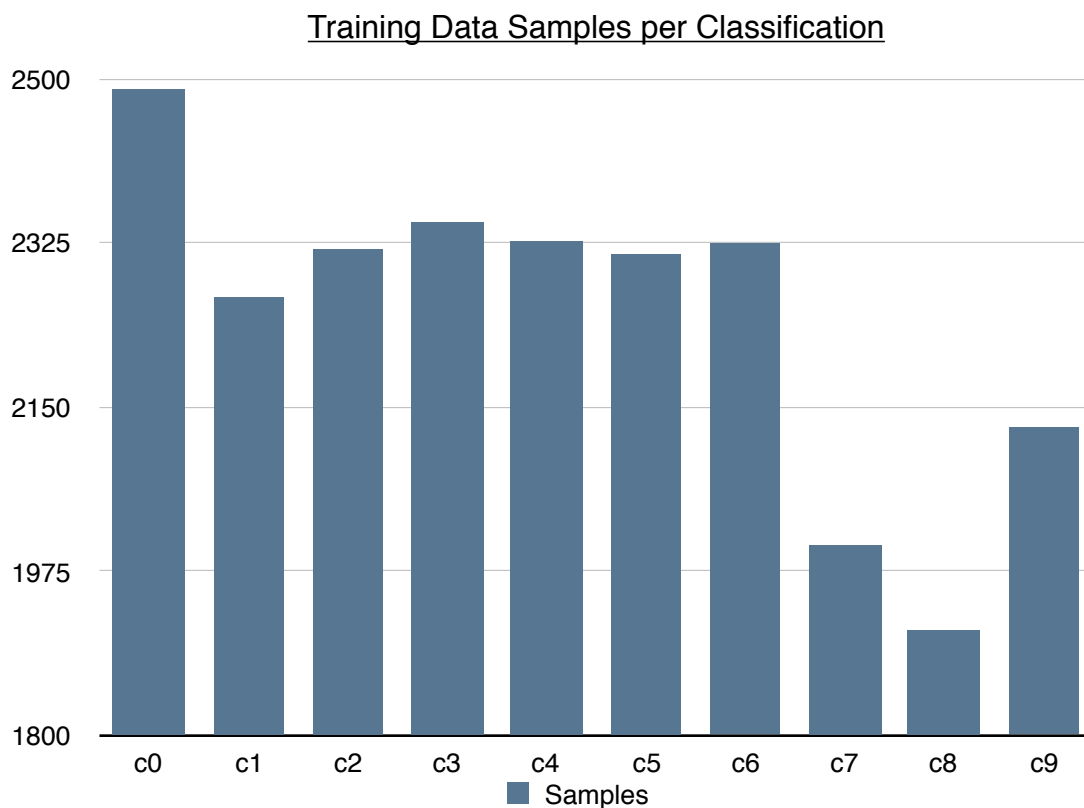
For this particular Kaggle competition, the data has been provided by State Farm in the form of 2D JPEG images. Each image is sized 640x480 pixels and are in colour.

Each of the images show the inside of the car and a number of different drivers performing various actions.

The following key points have been identified:

Training Data

There are 22,424 training images split across each of the 10 classifications, although this split is not equal. Therefore some classes have more training samples than others as can be seen in this graph.



What this means is that the learning algorithm is likely to find it harder to classify images from c7 (reaching behind) and c8 (applying hair and make-up) based strictly on the number of images it has to learn from. In theory, the algorithm should find it easier to predict normal driving, and therefore might actually predict that this is the case for any borderline samples (e.g. the network may prefer to predict normal driving). In terms of the outcome of the exercise, predicting normal driving when the driver is distracted is not an ideal outcome.

Additionally there are only 26 drivers in the complete training data set, which means there is significant risk of overfitting during training. This can be highlighted when comparing similar images that are from different classes:



If the machine learning algorithm is not structured correctly, then it is likely to take features that identify the driver or the vehicle rather than from the action that the driver is taking. This may make it inherently hard to train a model that has a good chance of predicting the driver's behaviour.

Validation Data

The dataset does not provide pre-selected validation images, so these will need to be selected from the training images. Due to the limited number of drivers, it is envisaged that the validation set contains drivers that are not used to train the model.

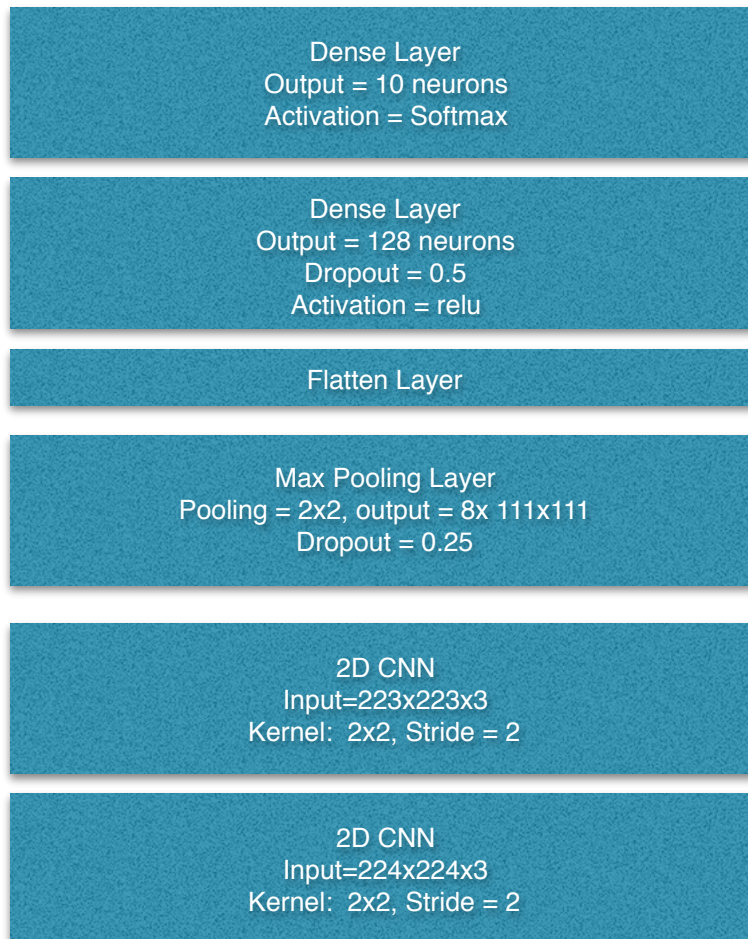
Testing Data

There are 79,726 testing images, significantly higher than the number of training images. None of the drivers in the test data set are in the training data set.

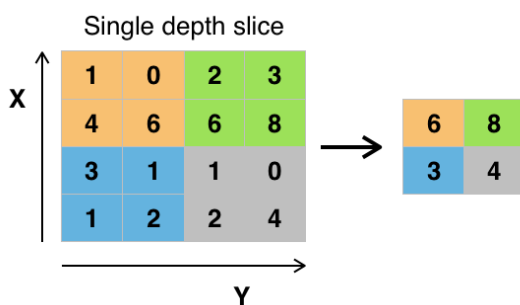
Algorithm and Techniques

The overall aim of this architecture is to take the input image (in this case resized) and gradually identify features within the image, reduce its dimensionality, and then reason as to which features provide the best indication of the driver's behaviour.

As discussed previously, the algorithm to be used is a deep convolutional neural network. The proposed network architecture is:



As discussed previously, CNN-based deep-learning architectures are ideally suited to image recognition tasks. In this architecture, the bottom 2 layers are used to identify features in each image. The bottom layer is likely to identify such things as edges or textures, and the 2nd layer aims to identify more complex structures such as an arm, a face or a phone.



A max-pooling⁸ layer is a common approach to perform non-linear down-sampling of the identified features. Max-pooling partitions the image into non-overlapping sub-regions, and for each sub-region output the maximum. The intuition is that once a feature is found, it doesn't matter where in the image it is. The max-pooling layer also reduces the size of the representation of the image, and therefore also reduces the amount of parameters and computation in

⁸ http://ais.uni-bonn.de/papers/icann2010_maxpool.pdf

the network. It also is a way to control over-fitting, and also provides a form of translation invariance.

The flatten layer takes the 2D representation of the image and converts it into a 1D representation suitable for feeding into a normal dense layer of neurons.

The 2 dense layers, also known as fully-connected layers, are designed to translate the identified features into a prediction of the driver's action. This is the higher-level reasoning of the network.

The final layer is the loss layer and determines how the network penalises the deviation between the predicted labels and the true labels. In this case, there is a single mutually-exclusive class to be predicted, so Softmax⁹ is an ideal loss function.

'ReLU', or Rectified Linear Units are used across the network as they provide a number of distinct advantages of other types of neurons. Although they help mitigate the risk of vanishing gradients, in this case they have been chosen as they can be trained effectively¹⁰ without pre-training the network on other data.

Dropout¹¹ has been used in the architecture as this strengthens the network by forcing it to learn a number of different representations of the data, and additionally it also reduces overfitting. As described earlier, there is a risk of overfitting due to the limited number of drivers in the training data set.

The following parameters can be tuned to optimise the classifier:

- Training parameters
 - Training length (number of epochs)
 - Batch size (how many images to train with in a single training step)
 - Weight decay and momentum
 - Learning rate
 - Size of the training vs validation sets
- Neural network architecture
 - Number and type of layers
 - Network parameters, such as drop outs, stride, kernel size
- Pre-processing of the images
 - Image size
 - Greyscale vs colour
 - Cropping
 - Random ordering of training images

During training, both training and validation sets are loaded into RAM. After that, batches are selected and to be loaded into the GPU memory for processing, and training is done using standard gradient descent without momentum.

Benchmark

To create an initial benchmark for the classifier, I aimed to achieve a Kaggle score in the top 500 which represented a log loss score of 0.88952¹² (noting that this is my first Kaggle competition). Additionally, with an end-user market in mind, I wanted to ensure that the trained neural network could predict the behaviour of a driver in under 2 seconds using a typical laptop (in my case a

⁹ https://en.wikipedia.org/wiki/Softmax_function

¹⁰ <https://www.quora.com/What-is-special-about-rectifier-neural-units-used-in-NN-learning>

¹¹ <https://www.cs.toronto.edu/~hinton/absps/JMLRdropout.pdf>

¹² <https://www.kaggle.com/c/state-farm-distracted-driver-detection/leaderboard>

Macbook Pro) without GPU support¹³ in order to represent a low-power, non-GPU based device within the car. I assumed that 2 seconds on a Mac would translate to being able to predict driver behaviour every 10 seconds in-car.

These values were determined based on my estimations of the use case described by State Farm, no indication has been given by them or Kaggle as to how often they wanted to predict the driver's behaviour.

¹³ A GPU would be used during training in order to improve performance and reduce the time to train and analyse different network architectures.

3. Methodology

Data Preprocessing

It would be possible for a CNN based deep network to learn directly from the images provided by State Farm. However, in order to reduce the size of the network and accordingly the number of parameters, the images are resized from full colour 640x480 pixels to grey scale 224x224 pixels.

The training set is also split into training and validation sets. The number of drivers in the training set vs the number in the validation set is a tuneable parameter, currently set to 95% (this means 2 drivers are allocated to the validation set).

Pre-processing is done entirely before the neural net is trained. Here are examples of pre-processed images:



Due to the risk of overfitting, it would be ideal to increase the size of training images by rotation, skewing or other image manipulation techniques. This will be addressed in the section on “Improvements”.

Implementation

The following steps have been taken to implement the pre-processing, training and testing:

Pre-processing:

A complete list of images and their labels is provided in the file “driver_imgs_list.csv”. The format of this csv file is “subject, classname, img”. This file is read in for future use. There is no list of testing images provided directly, so this list is created by reading the directory of testing images.

Next, the 2 arrays are created, these are an array of image names and an array of the classes (or labels). As the pre-processing of images takes a long period of time, the images are processed and then stored in a separate directory (defined by the variables `train_images_dir` and `test_images_dir`). The option to create pre-processed images is handled by setting the variable `create_repository` to true (to create a new set of pre-processed images) or false (to use an existing set).

Next, a set of district drivers is collated. This is to ensure that the split between training and validation sets are split by driver in an aim to reduce overfitting, and to provide better metrics in training and validation.

The training images can then be split into training and validation sets. This could be done at training time using the “validation_split” parameter in the Keras `model.fit` command, but as mentioned previously, the intuition is that the network would learn drivers as opposed to drivers behaviour, and therefore the train/validation split is performed separately and is determined from the list of drivers and then their associated images.

Defining the Neural Network

The next step is to create a network using Keras. Keras has been chosen due its portability (it works on any Python platform with Theano or Tensorflow installed) and has a simple approach for defining and training neural networks. Caffe was originally considered for this but there was a higher level of complexity around creating training, validation and test data sets.

The neural network is based on the LeNet model as detailed previously. Categorical cross-entropy is used as a loss function in line with recommendations for a multi-class classifier implemented in Keras.

Standard Gradient Descent is used as the optimiser and metrics are captured for accuracy.

Training

Each training run has 10 epochs, and each one uses the same training and validation data. A Keras callback is created in order to display the loss history and display a graph at the end of training.

Testing

The final step is to predict the classes for the test data provided by Kaggle. This uses the `model.predict` function within Keras. Once the predictions have been made, the predicted classes are converted to CSV and saved to the local filesystem for upload to Kaggle's site.

Refinement

The following approaches were taken to improve the algorithm:

- 1) L2 regularisers were implemented for both weights and biases at each layer of the network.
- 2) L1 regularisers were tested but led to a greater loss for both training and validation sets.
- 3) Different learning rates were used in order to determine an optimal rate. The learning rates trained against are 0.001, 0.003, 0.01, 0.03 and 0.1. This is based on intuition from Andrew Ng in the Coursera Machine Learning course.
- 4) The SGD parameters were turned, but the initial values of `decay=0`, `momentum=0` and `nester= False` were deemed to be optimal
- 5) An pre-trained implementation of the VGG-16 network was also tested but this did not converge with the available data so was discarded.

- 6) Varying the size of the training and validation data sets, particularly by increasing the amount of training data

4. Results

Model Evaluation and Validation

The final model is deemed to be reasonable but does not fully align with the expectations set out in the introduction. It typically results in a leaderboard position around 1300th, and not in the top 500 as originally targeted.

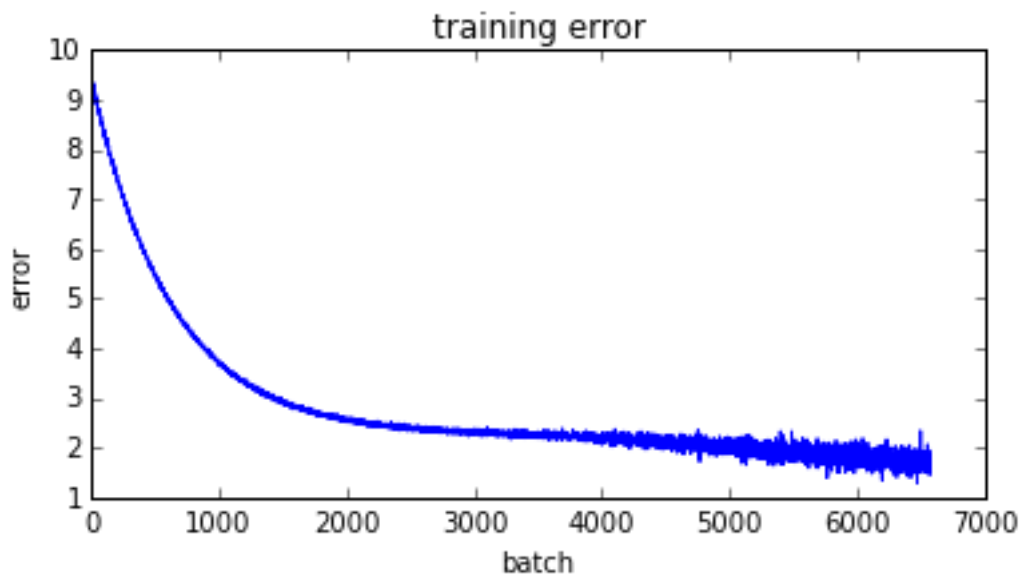
The final model was derived using the following approaches:

- 1) Initially, small training runs were performed with limited numbers of epochs (perhaps 2-5 epochs) to understand the model's behaviour.
- 2) I observed the training and validation metrics of the model to determine whether it was converging when trained and how it performed on validation data. Results of each training run were determined based on data such as:

Starting training iteration 1 with learning rate 0.001

Train on 20996 samples, validate on 1428 samples

```
Epoch 1/10
20996/20996 [=====] - 453s - loss: 6.7022 - acc: 0.1243
- val_loss: 2.2878 - val_acc: 0.1828
Epoch 2/10
20996/20996 [=====] - 379s - loss: 3.8207 - acc: 0.1553
- val_loss: 2.2800 - val_acc: 0.1576
Epoch 3/10
20996/20996 [=====] - 373s - loss: 2.8127 - acc: 0.1752
- val_loss: 2.2702 - val_acc: 0.1352
Epoch 4/10
20996/20996 [=====] - 383s - loss: 2.4550 - acc: 0.1927
- val_loss: 2.2590 - val_acc: 0.1632
Epoch 5/10
20996/20996 [=====] - 375s - loss: 2.3196 - acc: 0.2162
- val_loss: 2.2325 - val_acc: 0.1190
Epoch 6/10
20996/20996 [=====] - 372s - loss: 2.2443 - acc: 0.2453
- val_loss: 2.1801 - val_acc: 0.1471
Epoch 7/10
20996/20996 [=====] - 370s - loss: 2.1485 - acc: 0.2717
- val_loss: 2.0825 - val_acc: 0.2199
Epoch 8/10
20996/20996 [=====] - 370s - loss: 2.0254 - acc: 0.3135
- val_loss: 1.9216 - val_acc: 0.2780
Epoch 9/10
20996/20996 [=====] - 369s - loss: 1.8955 - acc: 0.3611
- val_loss: 1.8931 - val_acc: 0.2185
Epoch 10/10
20996/20996 [=====] - 369s - loss: 1.7738 - acc: 0.4002
- val_loss: 1.6683 - val_acc: 0.3410
```



- 3) When the model didn't behave as expected, I investigated best practice for deep CNNs and made modifications accordingly. These changes included:
 - i. regularisers such as trying L1 and L2.
 - ii. Modifying the kernel size (from 2x2 to 3x3)
 - iii. Changing the number of filters in the CNN layers
 - iv. Running for different number of epochs
- 4) I made modifications to the training parameters, such as the values for L2, the learning rate, and the SGD values. Some of these values were taken from pre-existing models that I researched (for example, I tried SGD values from a VGG-16 model), and others were determined by training using different values.

The following observations have been made with the final model:

- 1) It tends to converge to similar training and validation loss metrics irrespective of which samples are in the train and validation data sets. This can be tested by created new training and validation data sets and re-running the model.
- 2) It generalises reasonably well to unseen data, although this is measured based on the position on the Kaggle leaderboard and the validation data sets.
- 3) It is possible to over-train the model if too many training epochs are used. This is observed by the fact that the training loss continues to decrease, but the validation loss tends to increase over time.
- 4) If the initial learning rate is too large, then there are occasional and dramatic increases in the validation loss.
- 5) Implementation of the L2 regulariser has reduced the smoothness of the loss over the training runs, but has improved the validation metrics.
- 6) With the implementation of the L2 regulariser, a fixed learning rate degrades the performance of the algorithm after 3000 iterations.

Justification

The model is able to predict on the 77k testing samples in approximately 10 minutes on a Macbook Pro, this equates to circa 130 samples per second which is well within the benchmark defined earlier. However, the predictive performance of the model is not deemed to be suitably reliable for

real world performance as its log loss metric on the test data is 2.5x larger than that of the targeted 500th place benchmark.

To become a suitable model, it would be necessary to implement some of the improvements mentioned in section 5 of this document.

5. Conclusion

Free-form Visualisation

To be completed...

Reflection

The process used for this project can be summarised using the following steps:

- 1) An initial problem was found on Kaggle that was deemed to be interesting and provided a way to explore computer vision and deep learning.
- 2) The data was provided and downloaded
- 3) The images were pre-processed to reduce their size
- 4) A benchmark was created for the classifier
- 5) The available training data was split into training and validation data sets
- 6) The classifier was trained on the available data over a number of epochs
- 7) After each training run, the parameters for the network were tuned, and additional controls such as regularisers were added to the network in order to determine an optimal network configuration
- 8) Once a suitable model was created, predictions were made against the test data and the results were uploaded to Kaggle. Typically the predictions rank the algorithm in the 1300's.

Improvements

The following improvements are suggested in order to make this a more usable model:

- 1) Reduce learning rate after 3000 iterations, or as it becomes evident that the loss is becoming erratic (symptomatic that the learning rate is too high). It is not clear how to do this in Keras as the learning rate is part of the model.compile function.
- 2) Increase the volume of training data by rotating and skewing each of the training images
- 3) Trying alternative deep networks, such as a VGG16 model.
- 4) Using a pre-trained network. A pre-trained VGG16 model was used but it did not converge with the existing training data set. This should be tried in conjunction with increasing the amount of training data as described above
- 5) Trying alternative optimisers such as Adam
- 6) Labelling the images in more detail, for example building a network that learnt purely from the arm position and face positions (perhaps either by using a labelling tool, or by training a network on individual parts of an image)
- 7) Instead of randomly picking training and validation data, find a set of training and validation images that perform the best in terms of log loss (and submission to Kaggle).

