Summary of “Distilling the Knowledge in a Neural Network”, Hinton et al.

We use neural networks and deep learning for several purposes. One of which is prediction. Prediction of new data by learning from historical data. This historical data can be immense with data points in the range of millions and billions. For such data and improved performance, a simple way is to train the data using numerous models and then average out the accuracies. This usage of an ensemble of models proves to be computationally expensive in terms of memory and time. In these situations, it helps to “compress” the knowledge obtained from an ensemble into a single model that generalises like the ensemble but is also computationally cheaper.

Distillation is defined as the transfer of the knowledge learnt by a cumbersome model to a smaller model that helps in making the training easier to deploy.

The conceptual hurdle in model compression as proposed by Caruana and his collaborators - keeping the knowledge intact from model to model is tough because knowledge is usually identified as the learned parameter values which normally differ from model to model. For cumbersome models, the primary objective during training is maximising the average log probability of the correct value so that predictions are closer to them. By this intuition, even the incorrect values are assigned some probabilities which causes the models to optimize performance over the existing training data rather than generalising to new data.

A method to make the transferring of knowledge compatible with similar generalization is to use class probabilities generated by the cumbersome models as soft targets, which are the arithmetic or geometric mean of respective predictive distributions.

According to the paper, distillation is carried out by raising the temperature (or entropy) of the final softmax being used in the neural network activations until the cumbersome model produces a suitably soft set of targets. The same high temperature should be used to train the smaller model to match the soft targets.

using a higher value of T produces a softer probability distribution over classes. We use two objective functions. The first calculates the cross entropy with the soft targets using the high temperature used in the distilled model's softmax. The second objective function computes the cross entropy with the correct labels.

When the distilled model is too small and is not able to interpret full knowledge from the cumbersome model, it was found that intermediate temperatures worked well. This causes the inference that ignoring the larger logits i.e., inputs of the final softmax is helpful for distillation.

* MNIST

This dataset consists of a single large NN with 2 hidden layers of 1200 ReLu hidden units. Has been strongly regularised using dropout and weights are constrained. Jittering is used which causes test errors. As the number of units in the hidden layer reduces, effective distillation and performance are achieved with decreasing temperature. Increasing the bias for mythical characters and decreasing bias for repeating characters reduces test error.

* SPEECH RECOGNITION

SOTA Automatic Speech Recognition systems currently use Deep NNs to map context of features temporally derived from the speech waveform to a probability distribution. Results of distilling the ensemble of 10 ASR models into a single model show minimal change in test frame accuracy and no change in the word error rate. distillation of an ensemble of models takes advantage of the parallel computation capacity of the ensemble as well as reduces computation at test time.

If the individual models are very large NNs or/and the data is immense, it helps to make use of specialist models which each focus on different subsets of the data and hence reduce computation.

* JFT dataset

100 million labeled images with 15k labels - very large data

Here the authors used an ensemble that contains one generalist model trained on all of the data and many ‘specialist models’ which are individually trained on example enriched subsets of the data. Results show that the baseline of CNN with 6 months training using asynchronous stochastic gradient descent gives a conditional test accuracy of 43.1% (conditional test accuracy is the accuracy when only examples belonging to the specialist classes are considered). Using an ensemble of 61 specialist models improves the CTA to 45.9, and a significant improvement is also observed in general test accuracy.

Training independent specialist models is very easy to parallelize hence reducing computation. This improves accuracy when more specialists cover the same class (overlapping). Using soft targets as regularizers on smaller data subsets helps reduce overfitting and improve generalisation.