Contribution ID: **2** Type: **not specified**

Distilling the Knowledge in a Neural Network

A very simple way to improve the performance of almost any mac
hine learning algorithm is to train many different models on the same data a
nd then to average their predictions [3]. Unfortunately, making predictions
using a whole ensemble of models is cumbersome and may be too computationally expen sive to allow deployment to a large number of users, especially if the indivi dual models are large neural nets. Caruana and his collaborators [1] have shown that it is possible to compress the knowledge in an ensemble into a single model which is much easier to deploy and we develop this approach further using a different compression technique. We achieve some surprising results on MNIST and w e show that we can significantly improve the acoustic model of a heavily used commercial systemby distilling the knowledge in an ensemble of models into a single model.

Authors: Mr VINYALS, Oriol (Google); Mr HINTON, Geoffrey (Google)

Presenters: Mr VINYALS, Oriol (Google); Mr HINTON, Geoffrey (Google)

Track Classification: Machine Learning