Distilling the Knowledge in a Neural Network

**Authors: Geoffrey Hinton, Oriol Vinyals, Jeff Dean**

**Link to the paper:** [**https://arxiv.org/abs/1503.02531**](https://arxiv.org/abs/1503.02531)

In this paper the authors show that the acoustic model can be significantly improved by distilling the knowledge in an ensemble of models into a single model.

A Novel method is presented in the paper for pruning filters from convolution neural networks with the objective of reducing computation for inference. New criterion based on Taylor expansion of the neural network function is proposed for pruning. The authors also have introduced a new type of ensemble having many specialist models and one or more full models which differentiate fine-grained classes that the full models confuse.

The major inference in the paper was

1. Knowledge can be transferred from an ensemble or from a large highly regularized model into a smaller, distilled model by Distilling.
2. Performance of a very large net can be improved by having large number of specialist nets, where each of them learns to differentiate between the classes in a highly confusable cluster.
3. A deep acoustic model, which is an ensemble of deep neural networks can be distilled to a single net of same size.

The paper doesn’t have any graphical representation of any results or data, which if was added would have been easier for reader to visualize the results. It would be difficult to Implementation the paper, as details of Implementation was not well explained. The paper needed more elaboration.