- Write a summary of a research paper of your choice.

Distilling the knowledge in a neural network

In [1], the authors propose a compression method of transferring the knowledge of a large model or ensemble of models to a smaller one with almost the same performance. The process involves transferring the knowledge from the teacher to the student model by training a large model on data or using a pre-trained model. Using the logits, which are the inputs to the last layer, the objective of the small model is composed of two parts, minimizing the loss on data prediction and minimizing the squared difference between its logits and the logits of the teacher. The authors tested the performance on image classification and speech recognition, with the student models almost half the neurons of the teacher, it was able to achieve similar performance with negligible loss in accuracy. The authors also introduced a new type of ensemble that is composed of one or more full models and many specialized models. The idea is to train one model or more on the whole data and use its weights to initialize the specialists. The specialists focus on the confused classes by the generalist model. The data used to train specialists are the samples of the targets that are harder to distinguish and random samples of all the other targets grouped in one class called the dustbin. According to their results, adding more specialized models improves the performance until a certain point.

- Share with us why you selected this paper.

Usually, ensemble models provide better performance than a single one. Most winning models on Kaggle use ensemble techniques. However, deploying these models can be cumbersome, and the inference process can be slower. This technique can effectively reduce the required storage without affecting the performance.

- Share with us the limitations of the authors' work.

The limitation of this study is the training process of the student. If there is no available pre-trained model, we have to do the training twice, once for the teacher and once for the student. The second limitation is the method focuses on the features of the last layer only, ignoring the intermediate layers that can hold valuable information.

- Comment on how one can benefit from the authors' work.

When working with a large number of classes with similar features that are hard to differentiate, we can train an ensemble of specialized models to improve the performance and then use knowledge distillation to compress it in one model. Also, to compress a large model that is gigabytes of size, that to be deployed on IoT devices since they have limited resources.

References:

[1] G. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," arXiv.org, https://arxiv.org/abs/1503.02531.