这篇文章使用了“蒸馏”的方法从一个大的模型中学习出一个效果相当的小模型。

1.引言
在大规模的机器学习中，我们通常在训练阶段和测试阶段使用一样的模型，然而两个阶段的任务要求并不一样：训练阶段需要从大规模的数据中提取特征，因此需要花费很多的时间以及计算资源。而在测试阶段，对于大多数用户来说，需要更加严格的时间以及计算资源的限制。因此，可以在训练阶段采用大模型，训练完之后将大模型所学到的知识迁移到小模型中进行应用。

一种迁移知识的方式是通过让小模型学习大模型的概率分布。因为概率分布通常比 ground-truth 能够提供更多的信息。
For cumbersome models that learn to discriminate between a large number of classes, the normal training objective is to maximize the average log probability of the correct answer, but a side-effect of the learning is that the trained model assigns probabilities to all of the incorrect answers and even when these probabilities are very small, some of them are much larger than others. The relative probabilities of incorrect answers tell us a lot about how the cumbersome model tends to generalize.

An obvious way to transfer the generalization ability of the cumbersome model to a small model is to use the class probabilities produced by the cumbersome model as “soft targets” for training the small model.
When the soft targets have high entropy, they provide much more information per training case than hard targets and much less variance in the gradient between training cases, so the small model can often be trained on much less data than the original cumbersome model and using a much higher learning rate.

比如说，对于数字“2”，可能有 10^-6 的置信度识别为“3”，有 10^9 的置信度识别为“7”，这说明“3”比“7”看起来更像“2”。但由于这些概率值很小，所以很难对损失函数有影响。文章采用了“蒸馏”的方法来解决。

2.蒸馏
置信度的计算公式为：

$$q_i = \frac{exp(z_i/T)}{\sum_j exp(z_j/T)}$$

通过提高 T 的值来使置信度更加 soft（也就是使类别之间的置信度之差更小）

训练过程：
联合大模型的置信度分布(soft targets)和 groud-truth 信息(hard targets)来监督小模型的训练
The first objective function is the cross entropy with the soft targets and this cross entropy is computed using the same high temperature in the softmax of the distilled model as was used for generating the soft targets from the cumbersome model. The second objective function is the cross entropy with the correct labels. This is computed using exactly the same logits in softmax of the distilled model but at a temperature of 1. . We found that the best results were generally obtained by using a condiderably lower

weight on the second objective function.

3.mnist 实验结果

文章将数字“3”从小模型的训练集中删掉，使用“蒸馏”的方法训练小模型，对测试集 1010 个“3”中只识别错了 133 个，说明了小模型能够从大模型的 soft targets 中学习到有关数字“3”的信息。