Distilling the Knowledge in a Neural Network

笔记本:	网址		
创建时间:	2018/12/13 14:53	**更新时间:**	2018/12/14 10:09
作者:	13718166945@163.com		

Distilling the Knowledge in a Neural Network(知识蒸馏)

paper链接：https://arxiv.org/pdf/1503.02531.pdf

knowledge含义:

A more abstract view of the knowledge is that it is a learned mapping from input vectors to output vectors

Why：

- A very simple way to improve the performance of almost any machine learning algorithm is to train many different models on the same data and then to average their predictions. Unfortunately, making predictions using a whole ensemble of models is cumbersome and may be too computationally expensive to allow deployment to a large number of users, especially if the individual models are large neural nets.
  （大模型计算复杂高及资源耗费大，想通过小模型去模拟得到大模型的使用效果(提升小模型的效果))

- It is generally accepted that the objective function used for training should reflect the true objective of the user as closely as possible， It would clearly be better to train models to generalize well , but this requires information about the correct way to generalize and this information is not normally available. When we are distilling the knowledge from a large model into a small one, however, we can train the small model to generalize in the same way as the large model.
  （训练的目标函数应该尽量接近用户的使用场景，要求训练出的模型应该具有较好的泛化能力，但是模型获得这种泛化能力的方法通常是难以直接被利用的，当一个大的模型拥有较好的泛化能力的时候，我们可以通过知识蒸馏让小的模型走大模型训练走过的路去模拟这种泛化能力）

蒸馏的含义：

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

1. 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.
   (使用大模型softmax之后的类概率作为软目标soft target，以原始label作为hard target, distilled model 的目标函数由以下两项的加权平均组成：
   - **soft targets 和小模型的输出数据的交叉熵（T较大）（保证小模型和大模型的结果尽可能一致）**
   - **hard targets 和小模型的输出数据的交叉熵（T=1）（保证小模型的结果和实际类别标签尽可能一致）**

)

2. 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. Using a higher value for T produces a softer probability distribution over classes**.**
   (更高的熵意味着更大的混乱程度、不确定性、更多的信息，因此，再大模型上使用较高的T值，小模型也使用同样的T做训练，最终使用时，T恢复为1)

实践测试：

https://github.com/JayjieL/distilliing

扩展：

1 disitilling的延伸

FitNets: Hints for Thin Deep Nets 链接： https://arxiv.org/pdf/1412.6550.pdf

2网络压缩方面的最新进展:

Recent Advances in Efficient Computation of Deep Convolutional Neural Networks 链接: https://arxiv.org/pdf/1802.00939v2.pdf