论文题目

Distilling the Knowledge in a Neural Network [PDF]

简介

Knowledge Distilling 方法受到了ensemble的启发，利用训练好的cumbersome net来作为soft target监督小型网络的训练。并且本文还提出了一种新的ensemble方法。

主要贡献

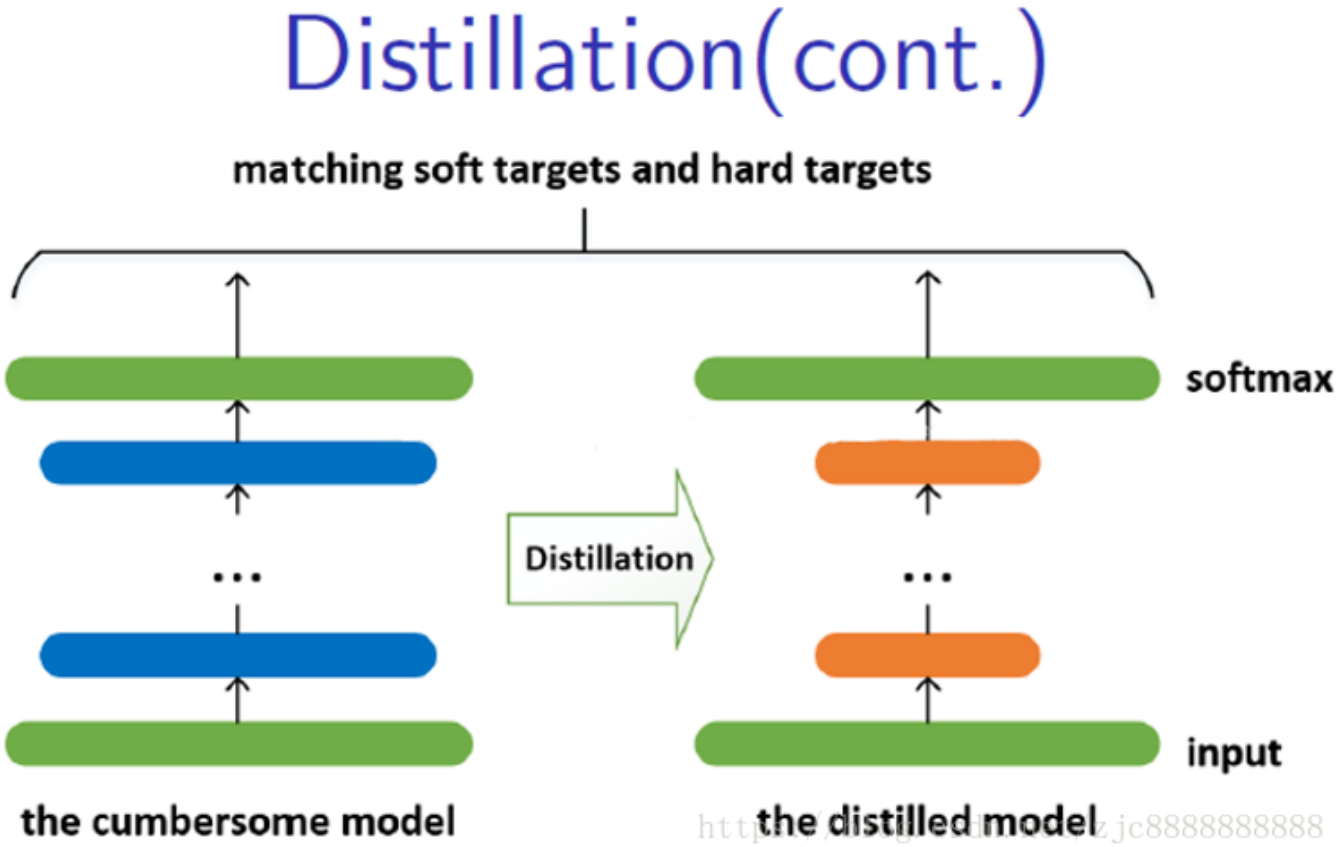
1. 提出一种 知识蒸馏（Knowledge Distillation）方法，从大模型所学习到的知识中学习有用信息来训练小模型，在保证性能差不多的情况下进行模型压缩

2. 提出一种新的 集成模型（Ensembles of Models）方法，包括一个通用模型（Generalist Model）和多个专用模型（Specialist Models），其中，专用模型用来对那些通用模型无法区分的细粒度（Fine-grained）类别的图像进行区分

Knowledge Distillation

知识蒸馏的整体框架如图所示，

- cumbersome model：复杂的大模型
- distilled model：经过knowledge distillation后学习得到的小模型
- hard targets：输入数据所对应的label 例：[0,0,1,0]
- soft targets：输入数据通过大模型（cumbersome model）所得到的softmax层的输出 例：[0.01,0.02,0.98,0.17] soft targets 在训练过程中可以提供更大的信息熵，将已训练模型的知识更好地传递给新模型



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（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（保证小模型的结果和实际类别标签尽可能一致）

Training ensembles of model

当数据集非常巨大以及模型非常复杂时，训练多个模型所需要的资源是难以想象的，因此提出一种新的集成模型方法，包括：

- 一个 Generalist model ：使用全部数据进行训练
- 多个 Specialist models ：对某些易混淆的类别进行专门训练的专有模型

在该方法中，只有 generalist model 耗时较长，剩余的 specialist model 由于训练数据较少，且相互独立，可以并行训练，因此整体运算量少了非常多。

但是，specialist model由于只使用特定类别的数据进行训练，因此模型对别的类别的判断能力几乎为0，导致非常容易过拟合，我们可以采用如下方法来解决：

- 当 specialist model 通过 hard targets 训练完成后，再使用由 generalist model 生成的 soft targets 进行 finetune，这样做是因为 soft targets 保留了一些对于其他类别数据的信息，因此模型可以在原来基础上学到更多知识，有效避免了过拟合

测试方法

1. 通过 generalist model 生成预测概率
2. 由预测概率选择相关的 specialist model 进行再次预测
3. 对相关 specialist model 的输出进行加权组合运算，并作为最终的预测结果