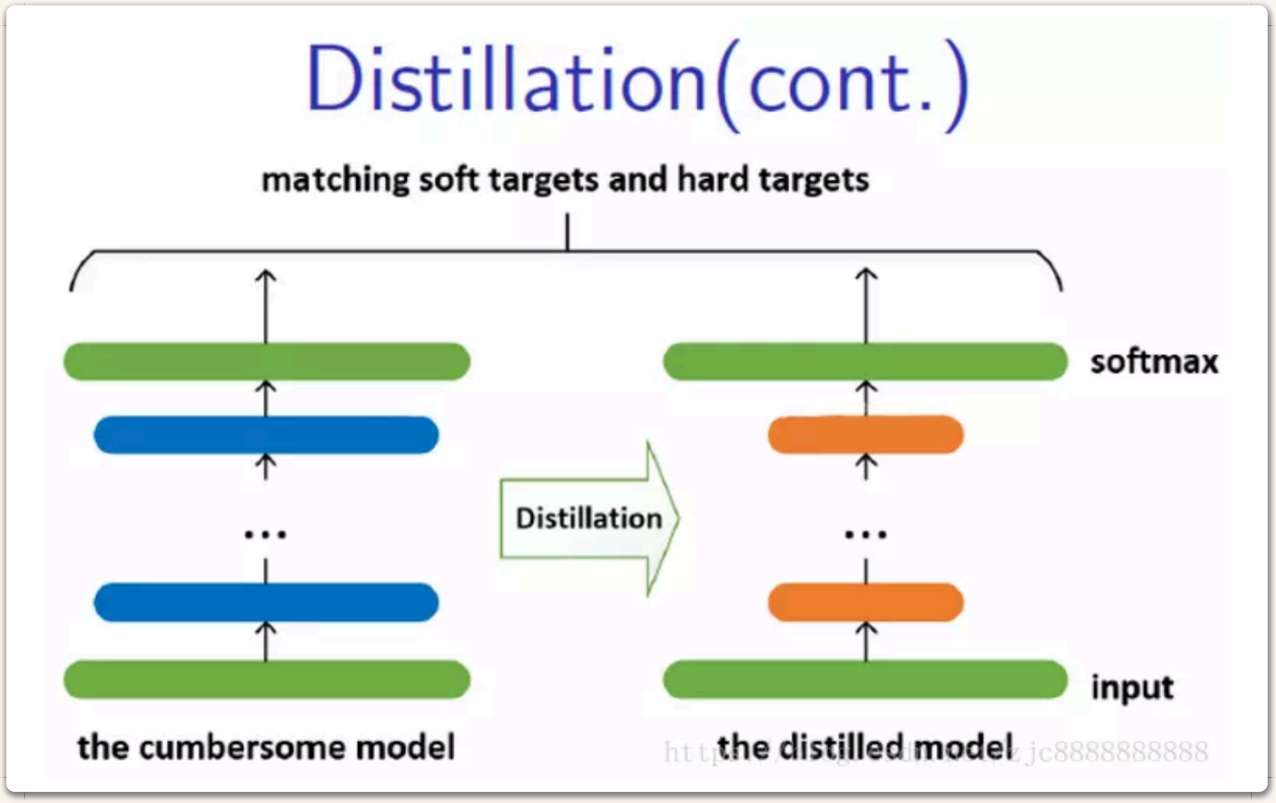
Distill the Knowledge in a Neural Network

贡献

1. 提出 **知识蒸馏 （Knowledge Distillation）** 方法，从大模型中学习到的知识中学习有用信息来训练小模型，在保证性能的前提下进行 **模型压缩**
2. 提出一种新的 **集成模型** 方法，包括通用模型和多个专用模型，其中，专用模型用来对通用模型无法区分的细粒度（Fine-grained）类别图像进行分类

Knowledge Distilling



- cumbersome model：复杂的大模型
- distilled model：蒸馏得到的小模型
- hard target：输入数据所对应的label [0,0,1,0]

- soft target：softmax层得到的输出 [0.01,0.02,0.98,0.17]
- Softmax in distillation:

$$q_i = \frac{e^{z_i/T}}{\sum_j e^{z_j/T}}$$

其中 **温度系数 T** 表示输出概率的soft程度

实验流程

1. 使用一个较大的 **T （例如 T=1）** 和 **Hard target** 训练一个大模型，生产 **Soft target**
2. 使用 **Soft target** 训练一个简单的小模型（distilled model）
3. Distilled model 的Cost Function由以下两项加权平均组成：
    - Soft target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一致性）
    - Hard target和大模型的输出数据的交叉熵 （保证小模型的结果与实际类别标签一致性）

$$\frac{\delta C}{\delta z_i} = \frac{1}{T}(q_i - p_i) = \frac{1}{T}(\frac{e^{z_i/T}}{\sum_j e^{z_j/T}} - \frac{e^{v_i/T}}{\sum_j e^{v_j/T}})$$

Training Ensemble 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

Specialist models 的训练集中，一半是初始训练集中某些特定类别的子集（special subset），另一半由剩余初始训练集中随机采样组成。

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 fine-tune，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

$$KL(p^g, q) = \sum_{m \in A_k} KL(p^m, q)$$

实验流程

1. 训练一个复杂的网络 `N1`
2. 使用数据train `N1`网络并得到 `M1`
3. 根据复杂网络设计一个简单网络 `N0`
4. 将M1 softmax 设 `T=20` 预测数据得到 soft target
5. soft target 和 hard target加权得出Target 推荐`0.1:0.9`
6. 使用 `label = Target` 的数据集训练 `N0 T=20`得到 `M0`
7. 设 `T=1`，`M0` 模型为我们得到的训练好的精简模型