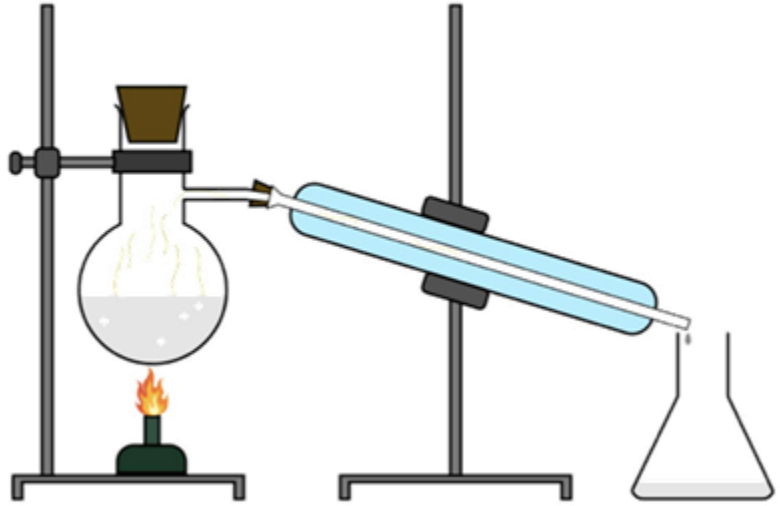
知识蒸馏 Knowledge Distillation

郝飞洋

2024年11月24日



Distilling the Knowledge in a Neural Network

Geoffrey Hinton, Oriol Vinyals, Jeff Dean

模型训练和部署的需求是有差距的！

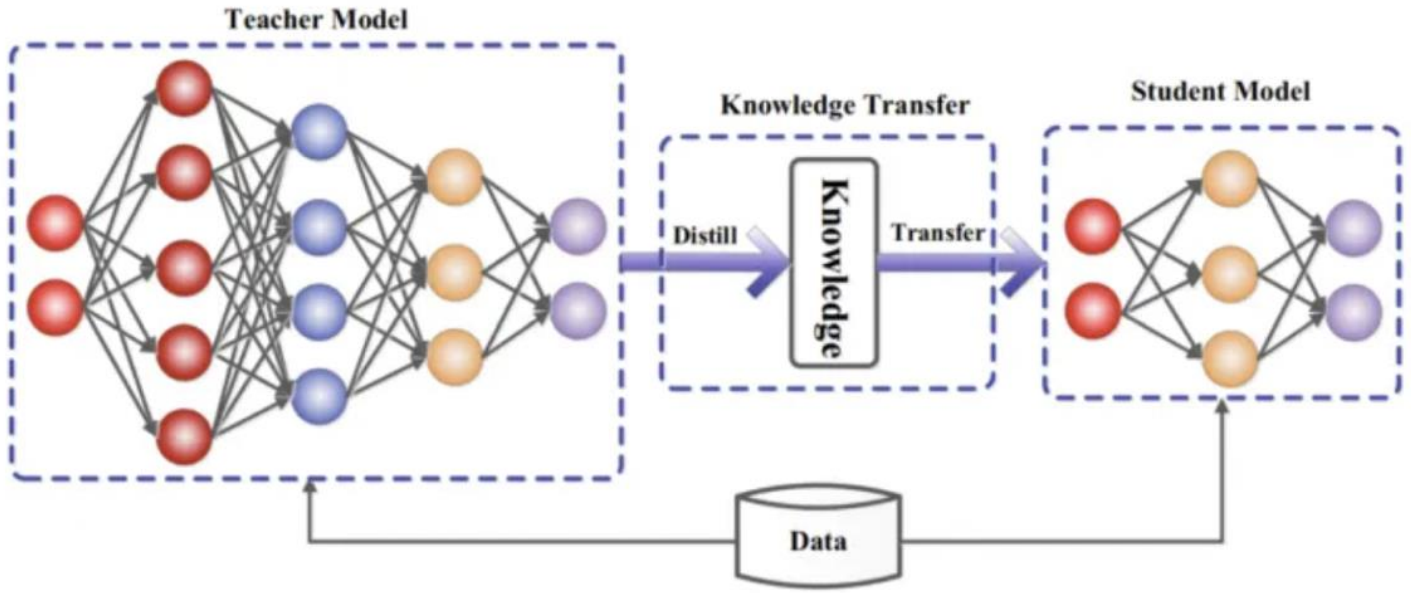
- 学习：十年磨一剑
- 考试：不光要能做出来，还要在规定时间内做出来

- 训练：获取更多的知识，可以用较多的时间和算力
- 部署：快速高效地完成任务

蒸馏->教学

- 两个模型：“师者，所以传道授业解惑也。”
- 两个loss：“吾爱吾师，吾更爱真理。”
  - 平时成绩
  - 考试成绩

$$L = \alpha L^{(soft)} + (1-\alpha) L^{(hard)}$$



教什么：如何定义“知识”？

- 大脑皮层里面的突触连接是知识？
- 对题目的理解是知识？
- 做题结果是知识？

- 模型的参数是知识？
- 模型的中间计算结果是知识？
- 模型输出的最终结果是知识？

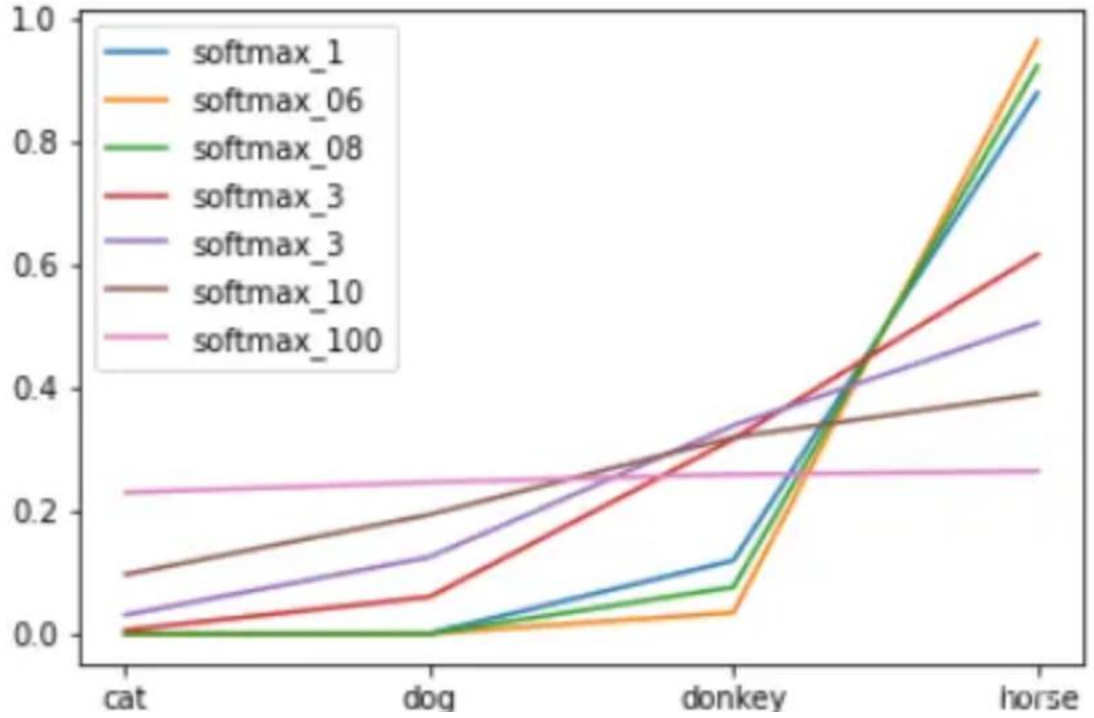
教什么：

- “考试的时候用排除法，下来分析卷子的时候把四个选项都看看。”
- 之前训练信息密度其实是不大的
- 可以用已经训练好的模型提供信息密度更大的数据进行训练

- hard targets: 正确的类别是1，错误的是0
- soft targets：各个类别的概率（包含了更多的信息）

设置一个蒸馏温度！

- 教师模型训练的时候是向正确答案对齐的，但是在教学生的时候要尽可能多展示出自己的“思考过程”(Dark Knowledge)
- 蒸馏温度越大，各个类别的概率差距会越小，“目标越软”



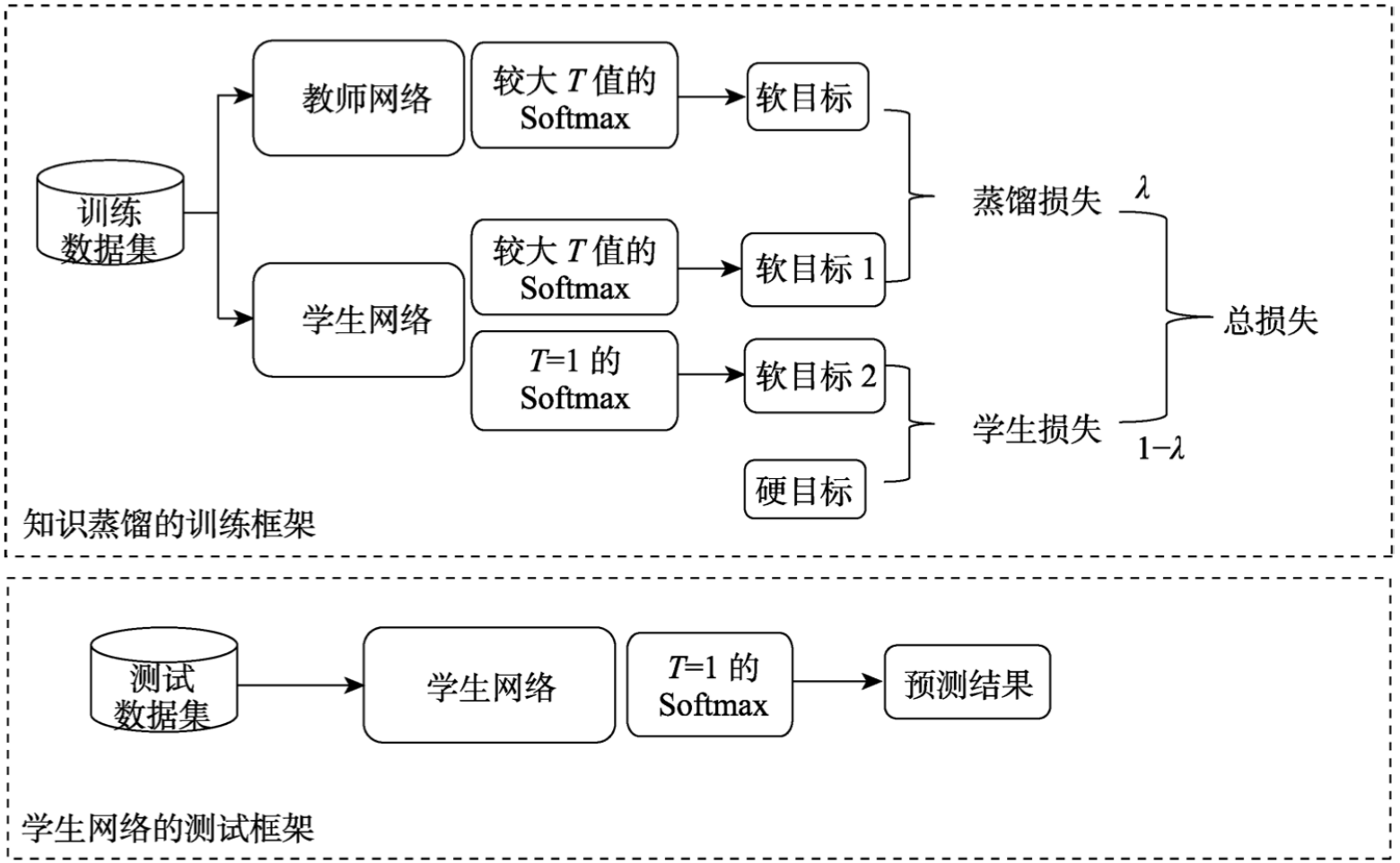
$$q_i = \frac{exp(z_i/T)}{\sum_j exp(z_j/T)}$$

能不能学习logit？

- logit包含很多噪声，效果不算太好
- 当蒸馏温度足够大且假设logit均值为0时，就相当于是在对齐到logit了
- 但是温度高的时候会带进来很多噪声

- 所以需要调整好蒸馏温度！

总览


教师网络
较大 T 值的 Softmax
软目标
训练数据集
蒸馏损失
λ
较大 T 值的 Softmax
软目标 1
学生网络
总损失
T=1 的 Softmax
软目标 2
学生损失
1−λ
硬目标
知识蒸馏的训练框架
测试数据集
学生网络
T=1 的 Softmax
预测结果
学生网络的测试框架

方法的优点和应用

- 学生模型可以从教师模型中学到一些自己从未见过的知识
- 0-shot的分类
- 甚至可以手动调整learned bias，进一步调整训练集和测试集的偏差

在语音识别集成模型的运用

- 小模型的性能超过了单个模型，直逼集成模型

System	Test Frame Accuracy	WER
Baseline	58.9%	10.9%
10xEnsemble	61.1%	10.7%
Distilled Single model	60.8%	10.7%

Training ensembles of specialists on very big datasets

架构设计

- 对于每一个样本，选择generalist 预测中的前k名(k=1)
- 找到specialist中涉及到这个类别的模型来参与预测
- 找一个分布，使得其与generalist 和各个specialist的分布差异最小（最小化KL散度）
- 根据这个分布来获得最后结果

specialist models

- 通过**聚类**找出 generalist model 认为相似度很高的类别，然后设计相应的specialist models 来协助分析
- 用generalist model 初始化，以获得“**通识教育**”
- 在一半目标类别和另一半其他类别的数据集上进行**微调**
- specialist models：小、快、需要的数据少
- 结果表明某一个类设计的specialist越多，效果越好

- 在专门的数据上训练specialist模型容易过拟合，使用soft targets可以缓解过拟合问题

Discussion

Discussion

- 除了压缩，还可以用于模型增强。
- “是故弟子不必不如师，师不必贤于弟子。”翻转课堂
- 讲授什么层次的知识？不同任务有不同的选择！
  - 输出特征知识、中间特征知识、关系特征知识
- “圣人无常师”：多个老师教同一个学生【多教师学习】
- 学生互助自学【相互蒸馏】
- 和其他模型的结合……
- 解决隐私安全问题