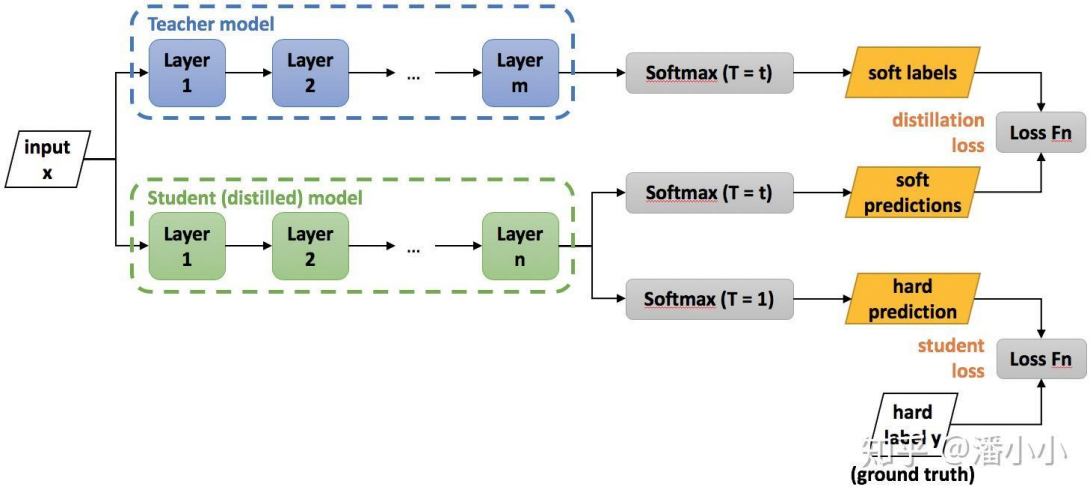
知识蒸馏的方法

第一步训练 Net-T；第二步在高温 T 下蒸馏 Net-T 的知识得到 Net-S



softmax 层的输出，除了正例之外，负标签也带有大量的信息，比如某些负标签对应的概率远远大于其他负标签。而在传统的训练过程(hard target)中，所有负标签都被统一对待。也就是说，只是蒸馏的训练方式使得每个样本给 Net-S 带来的信息量大于传统的训练方式。
但要是直接使用 softmax 层的输出值作为 soft target，这又会带来一个问题：当 softmax 输出的概率分布熵相对较小时，负标签的值都很接近 0，对损失函数的贡献非常小，小到可以忽略不计。因此"温度"这个变量就派上了用场。
下面的公式时加了温度这个变量之后的 softmax 函数:

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

这里的 T 就是温度。
原来的 softmax 函数是 T = 1 的特例。 T 越高，softmax 的 output probability distribution 越趋于平滑，其分布的熵越大，负标签携带的信息会被相对地放大，模型训练将更加关注负标签。

高温蒸馏的过程

（注：**logits**: 未归一化的概率， 一般也就是 softmax 层的输入。所以 logits 和 lables 的 shape 一样，也可以做为 sigmoid 的输入）

高温蒸馏过程的目标函数由 distill loss(对应 soft target)和 student loss(对应 hard target)加权得到，目标函数如下。

$$L = \alpha L_{soft} + \beta L_{hard}$$

- v_i: Net-T的logits
- z_i: Net-S的logits
- p_i^T: Net-T的在温度=T下的softmax输出在第i类上的值
- q_i^T: Net-S的在温度=T下的softmax输出在第i类上的值
- c_i: 在第i类上的ground truth值, $c_i \in \{0, 1\}$, 正标签取1，负标签取0.
- N: 总标签数量
- Net-T 和 Net-S同时输入 transfer set (这里可以直接复用训练Net-T用到的training set), 用Net-T产生的softmax distribution (with high temperature) 来作为soft target，Net-S在相同温度T条件下的softmax输出和soft target的cross entropy就是**Loss函数的第一部分** L_{soft}

$$L_{soft} = -\sum_j^N p_j^T \log(q_j^T)\text{，其中 } p_i^T = \frac{\exp(v_i/T)}{\sum_k^N \exp(v_k/T)}\text{，} q_i^T = \frac{\exp(z_i/T)}{\sum_k^N \exp(z_k/T)}$$

- Net-S在T=1的条件下的softmax输出和ground truth的cross entropy就是**Loss函数的第二部分** L_{hard} 。

$$L_{hard} = -\sum_j^N c_j \log(q_j^1)\text{，其中 } q_i^1 = \frac{\exp(z_i)}{\sum_k^N \exp(z_k)}$$

知识蒸馏的种类

1、 离线蒸馏
2、 半监督蒸馏
3、 自监督蒸馏