推理引擎-模型压缩

知识蒸馏

ZOMI

Talk Overview

1. **推理系统介绍**
   - 推理系统与推理引擎区别
   - 推理工作流程
   - 推理系统介绍
   - 推理引擎介绍
2. **模型小型化**
   - 基础参数概念
   - CNN小型化结构
   - Transform小型化结构
3. **离线优化压缩**
   - 低比特量化
   - 模型剪枝
   - 知识蒸馏
   - 二值化网络
4. **部署和运行优化**
   - 图转换优化（算子融合/重排/替换）
   - 并发执行与内存分配
   - 动态batch与bin Packing

Talk Overview

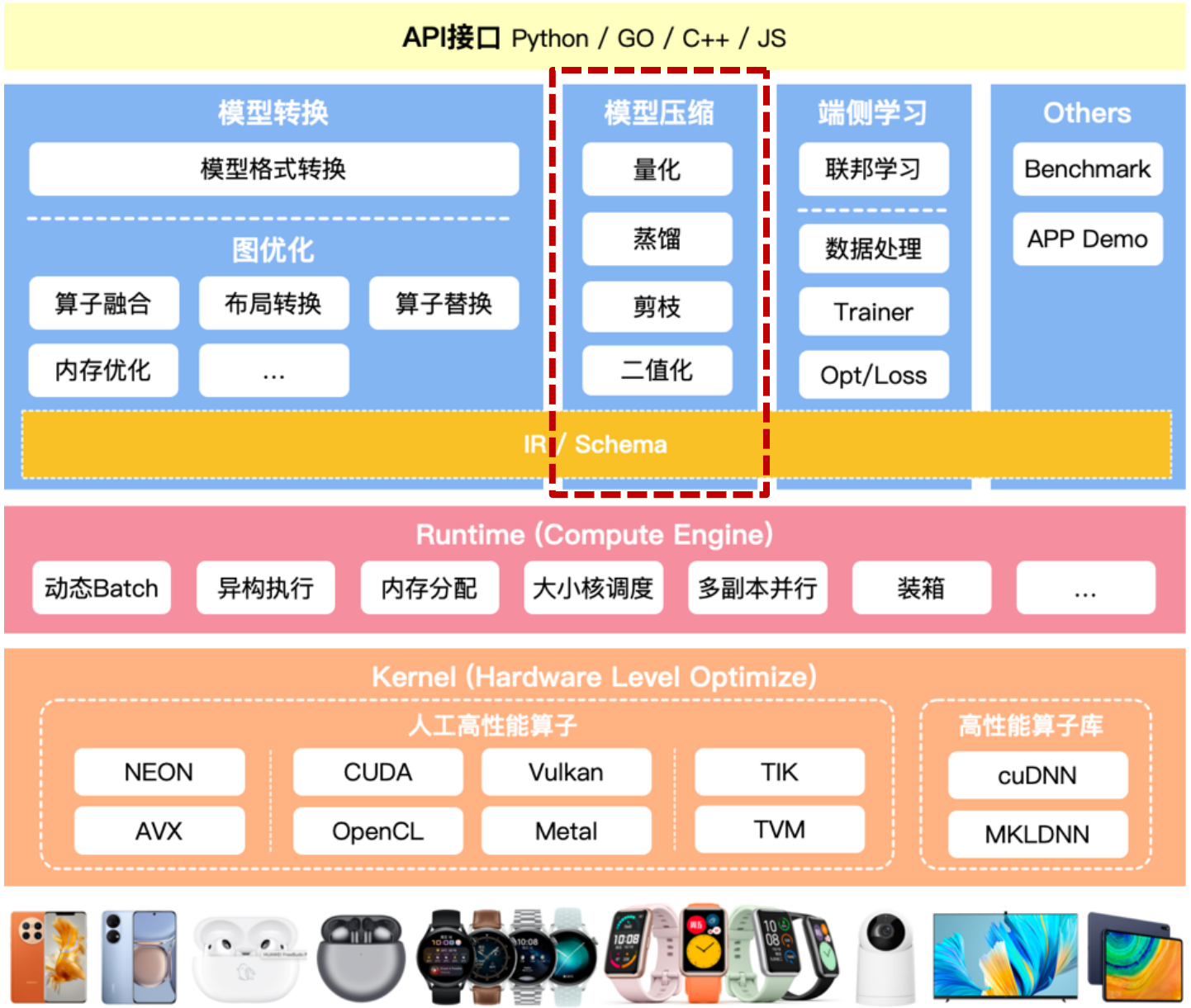
I. 知识蒸馏

- Background of KD - 知识蒸馏的背景
- Knowledge Format - 蒸馏的知识形式
- Distillation Schemes – 具体方法
- Hinton 经典蒸馏算法解读

推理引擎架构

对模型进行压缩

- 减少模型大小
- 加快训练速度
- 保持相同精度

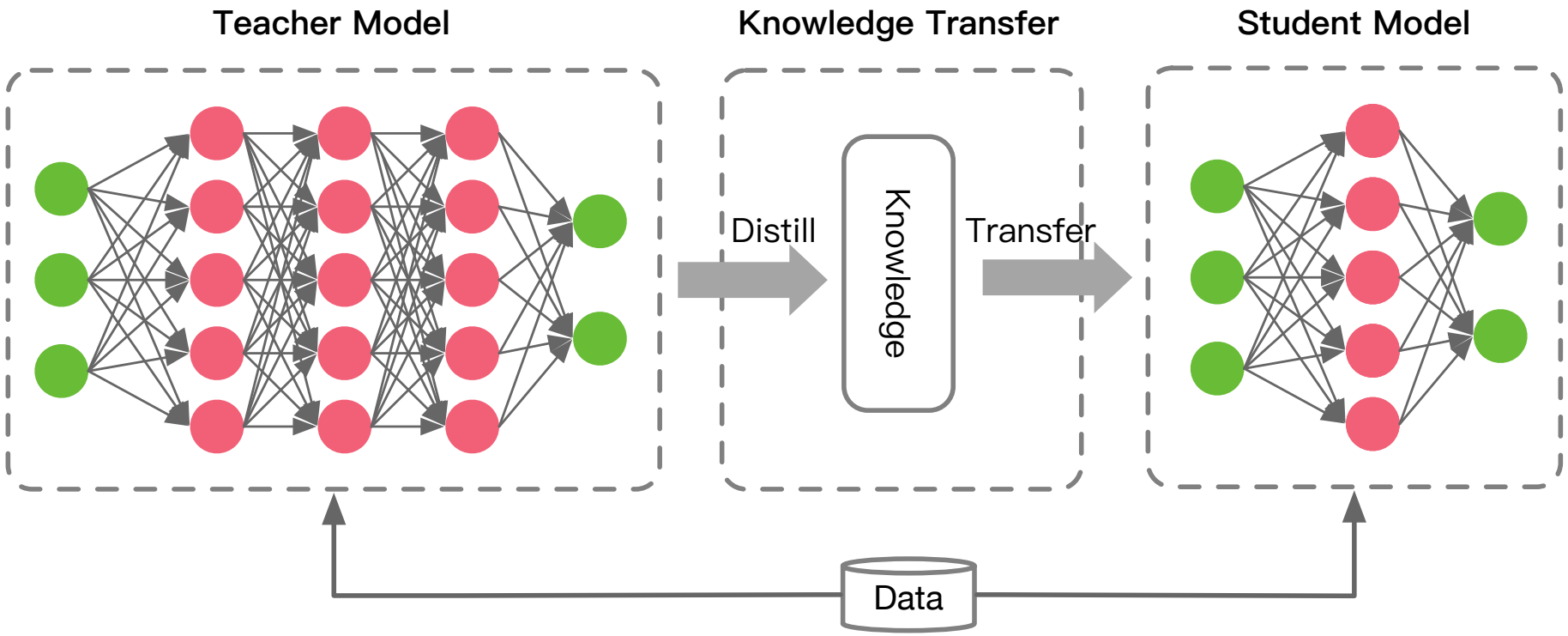


知识蒸馏背景

- Knowledge Distillation（KD）最初是 Hinton 在 “Distilling the Knowledge in a Neural Network”提出，与 Label smoothing 动机类似，但是 KD 生成 soft label 方式通过教师网络得到。
- KD 可以视为将教师网络学到的知识压缩到学生网络中，另外一些工作 “Circumventing outlier of auto augment with knowledge distillation”则将 KD 视为数据增强方法的一种。

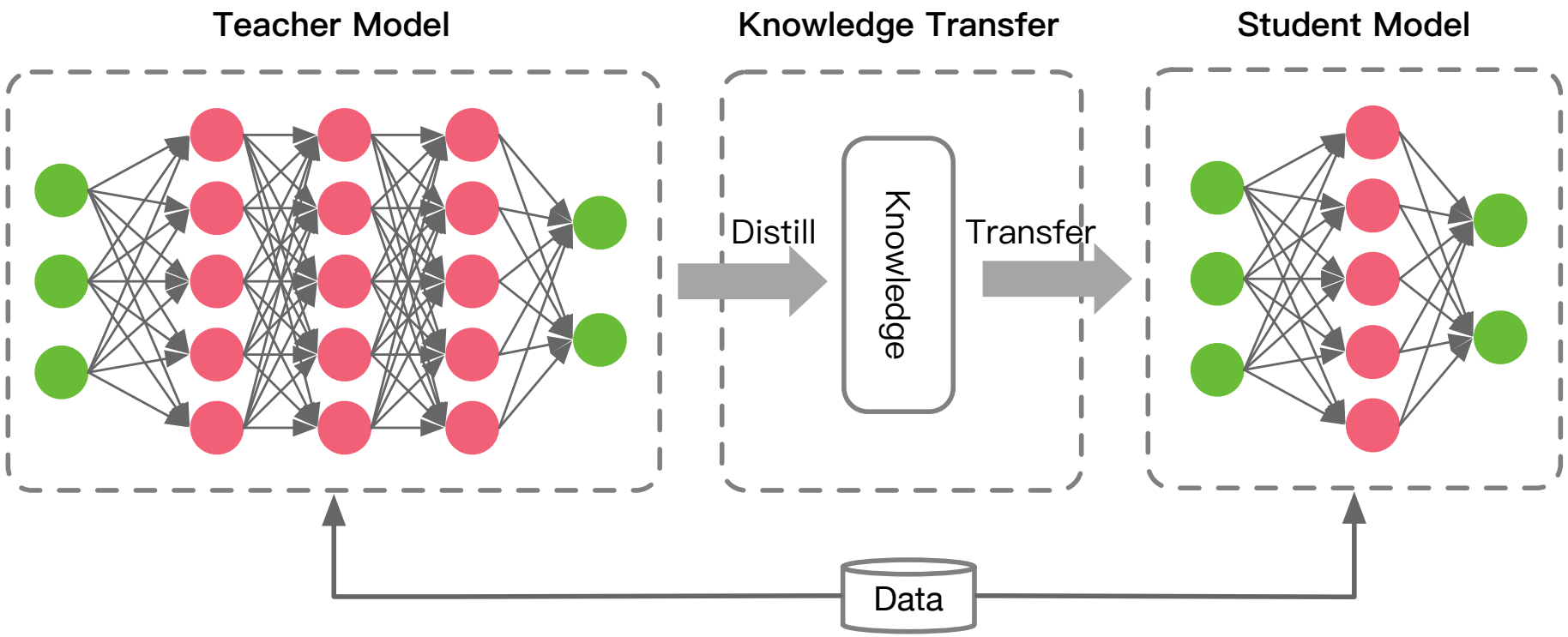
知识蒸馏主要思想

- Student Model 学生模型模仿 Teacher Model 教师模型，二者相互竞争，直到学生模型可以与教师模型持平甚至卓越的表现：



知识蒸馏组成

- 知识蒸馏的算法，主要由：1）知识 Knowledge、2）蒸馏算法 Distillate、3）师生架构三个关键部分组成：

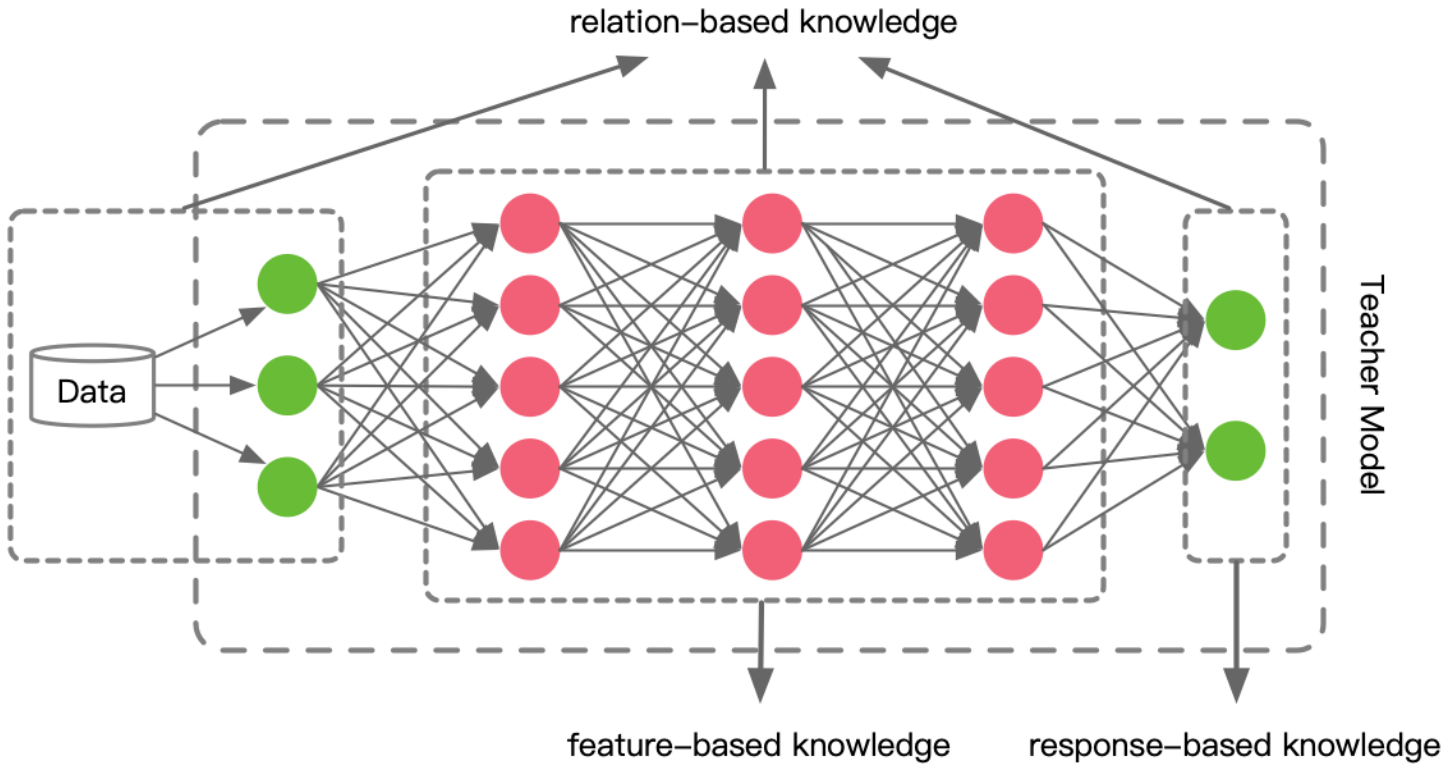


蒸馏知识的方式

Knowledge Distillation: A Survey

Knowledge 知识的方式

1. response-based knowledge
2. feature-based knowledge
3. relation-based knowledge
4. Architecture-base knowledge



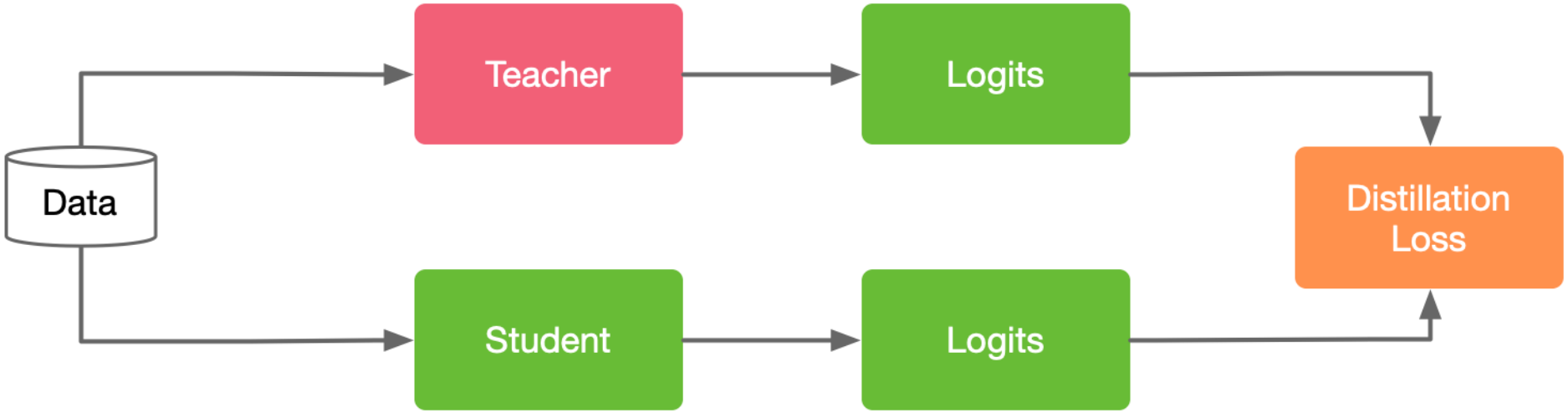
Response-Based Knowledge

- 主要指Teacher Model 教师模型输出层的特征。主要思想是让 Student Model 学生模型直接学习教师模式的预测结果（ Knowledge ）。
- 假设张量 z_t 为教师模型输出，张量 z_s 为学生模型输出，Response-based knowledge 蒸馏形式可以被描述为：

$$L_{Res\ D}(z_t, z_s) = \mathcal{L}_R(z_t, z_s)$$

Response-Based Knowledge

- Response-based knowledge 主要指 Teacher Model 教师模型最后一层 —— 输出层的特征。其主要思想是让学生模型直接模仿教师模式的最终预测：



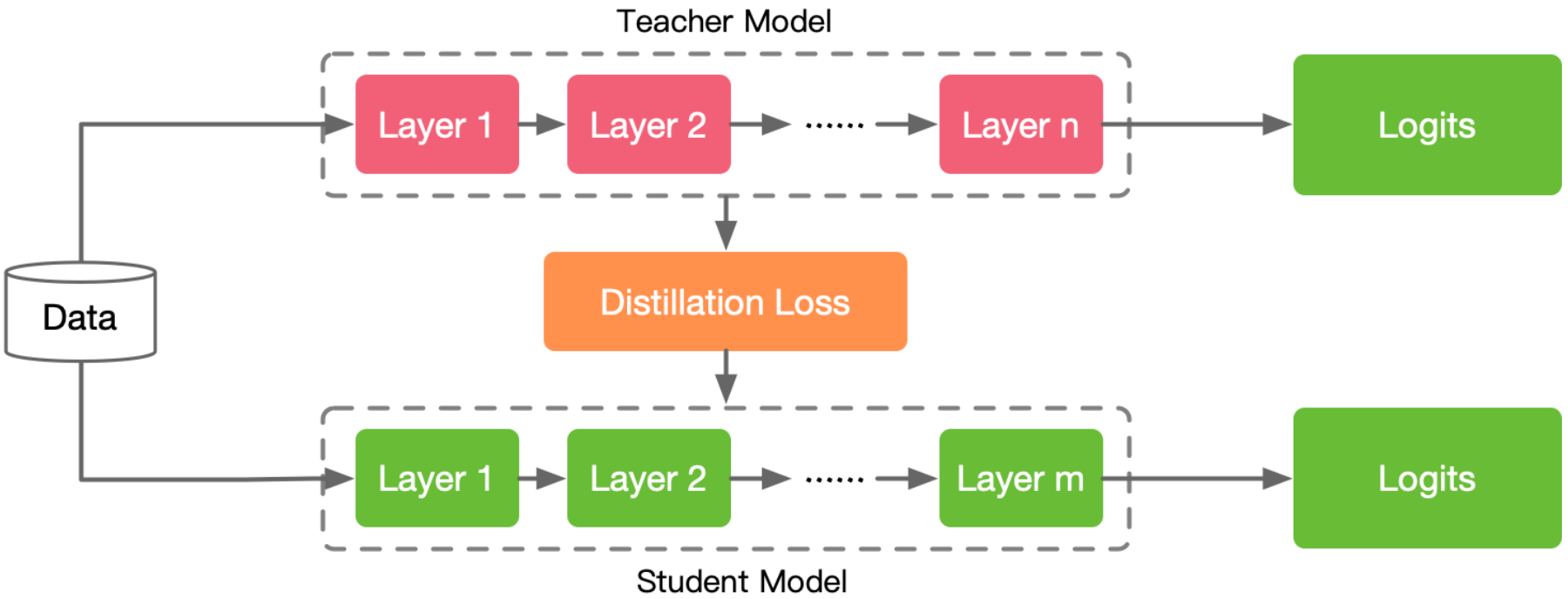
Feature-Based Knowledge

- 深度神经网络善于学习到不同层级的表征，因此中间层和输出层的都可以被用作知识来训练学生模型，中间层学习知识的 Feature-Based Knowledge 对于 Response-Based Knowledge是一个很好的补充，其主要思想是将教师和学生的特征激活进行关联起来。Feature-Based Knowledge 知识转移的蒸馏损失可表示为：

$$L_{Fea\ D}(f_t(x), f_s(x)) = \mathcal{L}_F(\phi_t(f_t(x)), \phi_t(f_s(x)))$$

Feature-Based Knowledge

- 虽然基于特征的知识迁移为学生模型的学习提供了良好的信息，但如何有效地从教师模型中选择提示层，从学生模型中选择引导层，仍有待进一步研究。由于提示层和引导层的大小存在显著差异，如何正确匹配教师和学生的特征表示也需要探讨。



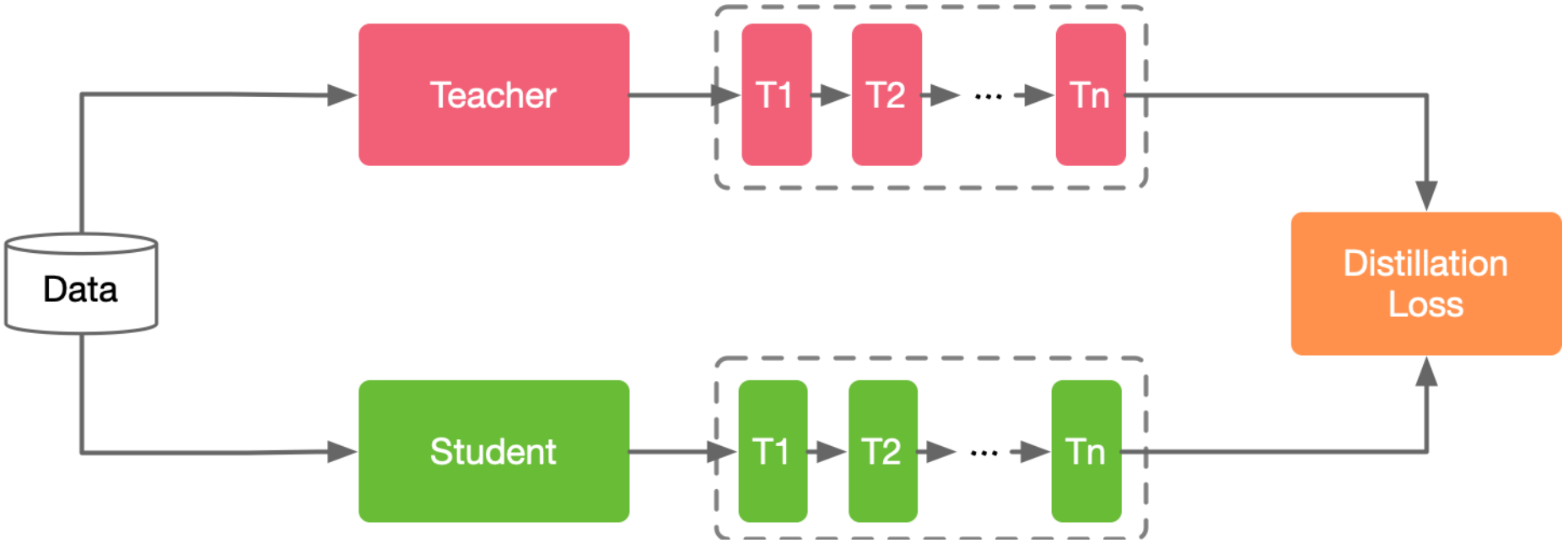
Relation-Based Knowledge

- 基于 Feature-Based Knowledge 和 Response-Based Knowledge 知识都使用了教师模型中特定层中特征的输出。基于关系的知识进一步探索了不同层或数据样本之间的关系。一般情况下，基于特征图关系的关系知识的蒸馏损失可以表示为：

$$L_{Rel\ D}(f_t(x), f_s(x)) = \mathcal{L}_R\left(\psi_t(\hat{f}_t, \check{f}_t), \psi_t(\hat{f}_s, \check{f}_s)\right)$$

Relation-Based Knowledge

- 传统的知识转移方法往往涉及到个体知识的提炼。教师的个人软标签 Soft Label 被直接提炼到学生中，实际上经过提炼的知识不仅包含了特征信息，还包含了数据样本之间的相互关系。

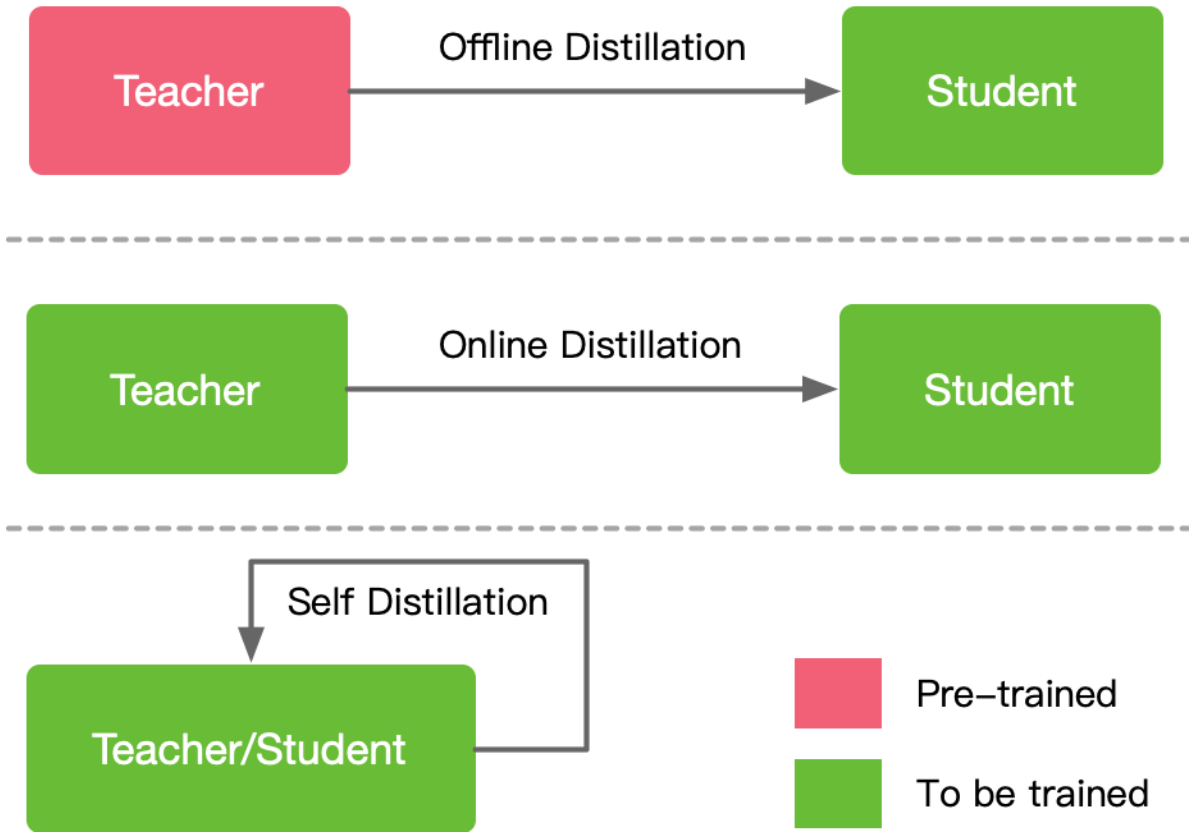


知识蒸馏方法

Knowledge Distillation: A Survey

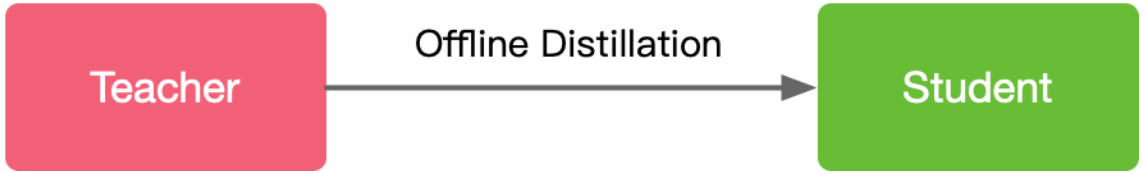
蒸馏方法

- 知识蒸馏可以划分为1）offline distillation, 2）online distillation，3）self-distillation



Offline Distillation

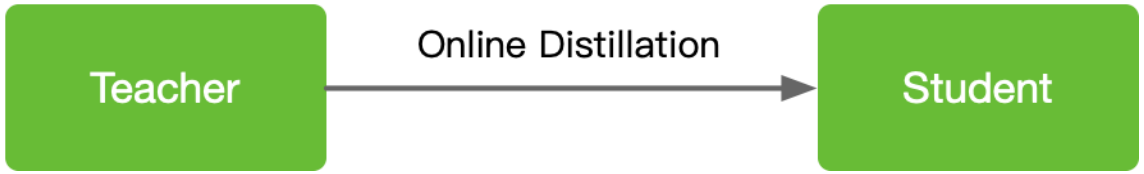
- 大多数蒸馏采用 Offline Distillation，蒸馏过程被分为两个阶段：1）蒸馏前教师模型预训练；2）蒸馏算法迁移知识。因此 Offline Distillation 主要侧重于知识迁移部分。
- 通常采用单向知识转移和两阶段训练过程。在步骤1）中需要教师模型参数量比较大，训练时间比较长，这种方式对学生模型的蒸馏比较高效。



- **Cons**：这种训练模式下的学生模型往往过度依赖于教师模型

Online Distillation

- Online Distillation 主要针对参数量大、精度性能好的教师模式不可获得的情况。教师模型和学生模型同时更新，整个知识蒸馏算法是一种有效的端到端可训练方案。



- **Cons**：现有的 Online Distillation 往往难以获得在线环境下参数量大、精度性能好的教师模型。

Self-Distillation

- 教师模型和学生模型使用相同的网络结构，同样采样端到端可训练方案，属于 Online Distillation 的一种特例。

蒸馏方法

1）offline distillation, 2）online distillation，3）self-distillation 三种蒸馏方法可以看做是学习过程：

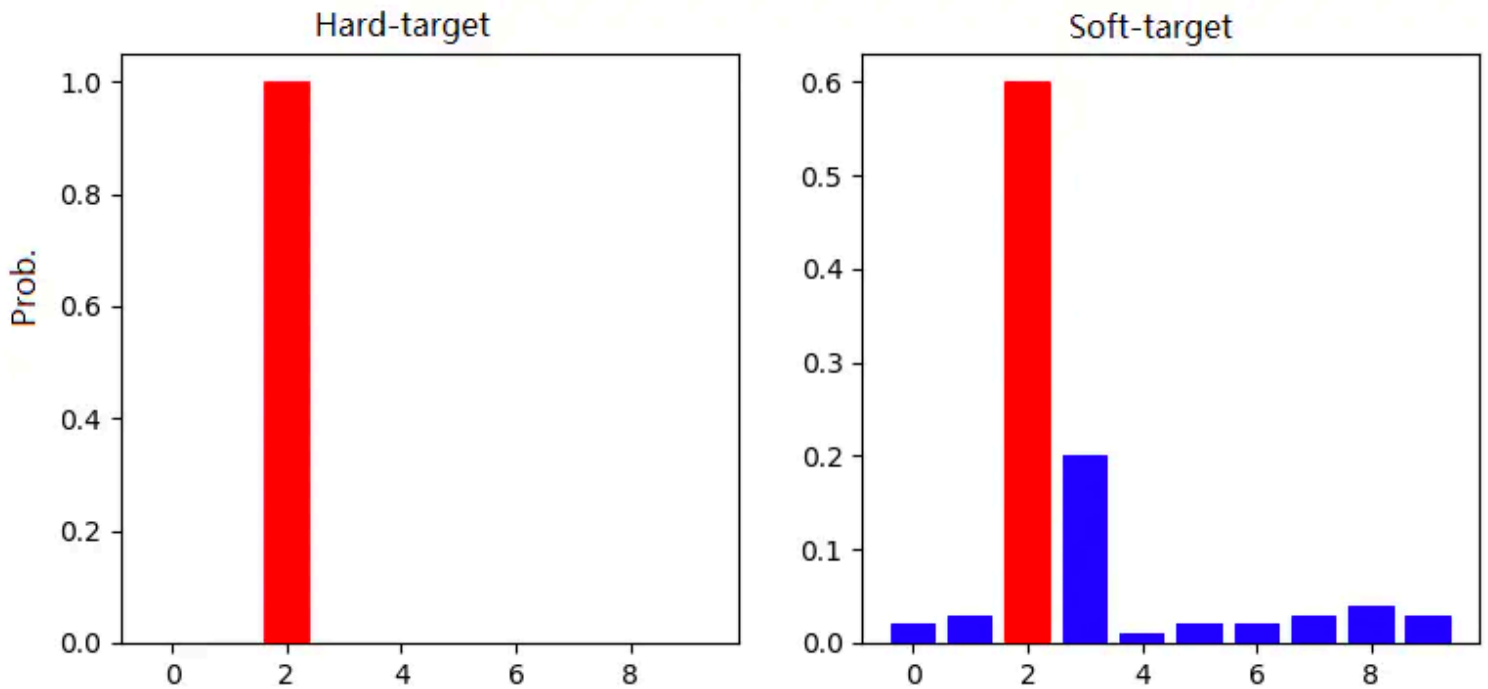
1. Offline Distillation 指知识渊博教师向传授学生知识；
2. Online Distillation 是指教师和学生共同学习；
3. Self-Distillation 是指学生自己学习知识。

Hinton经典蒸馏算法解读

Distilling the Knowledge in a Neural Network

Hard-target 和 Soft-target

- 传统的神经网络训练方法是定义一个损失函数，目标是使预测值尽可能接近于真实值（Hard- target），损失函数就是使神经网络的损失值和尽可能小。这种训练过程是对ground truth求极大似然。在知识蒸馏中，是使用大模型的类别概率作为 Soft-target 的训练过程。



Softmax with Temperature

- softmax函数：

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

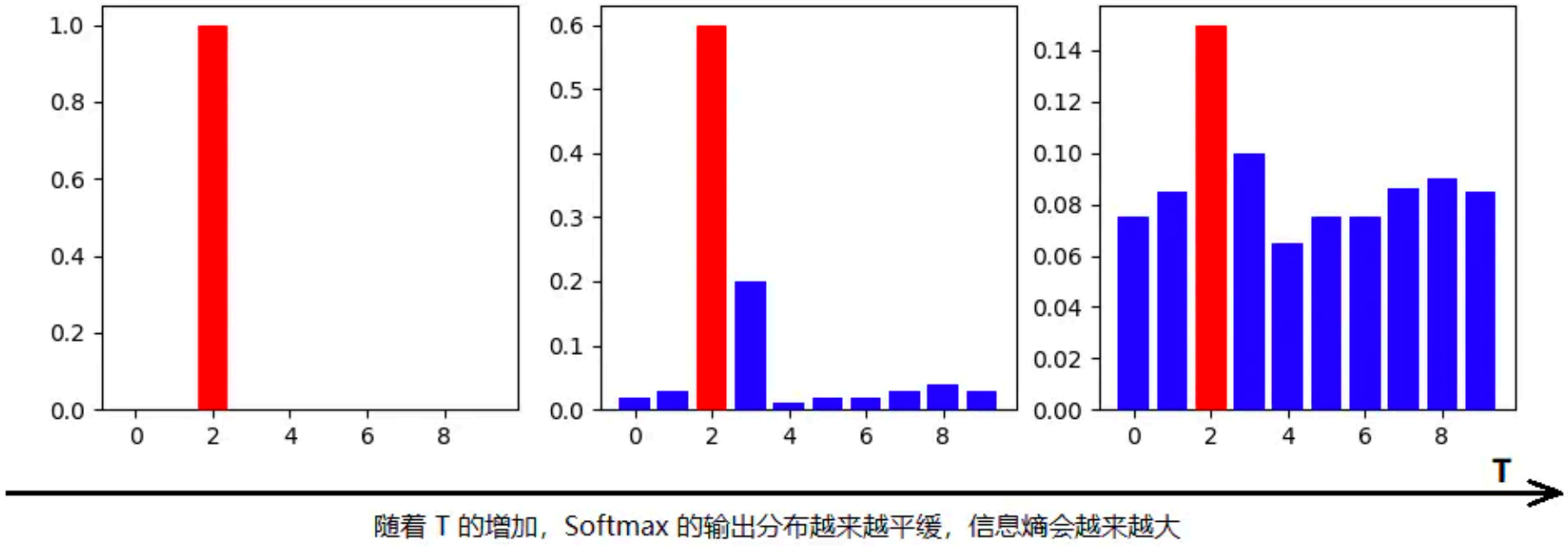
- 使用软标签就是修改了softmax函数，增加温度系数T：

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

- 其中 q_i 是每个类别输出的概率，z_i 是每个类别输出的 logits，T 是温度。温度 T=I 时，为标准 Softmax。T越高，softmax 的output probability distribution越趋平滑，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。

Softmax with Temperature

- 其中 q_i 是每个类别输出的概率，z_i 是每个类别输出的 logits，T 是温度。温度 T=1 时，为标准 Softmax。T越高，softmax 的output probability distribution越趋平滑，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。



如何选择 T？

负标签中包含一定的信息，尤其是那些负标签概率值显著高于平均值的负标签。但由于Teacher模型的训练过程决定了负标签部分概率值都比较小，并且负标签的值越低，其信息就越不可靠。因此温度的选取需要进行实际实验的比较，本质上就是在下面两种情况之中取舍：

- 当想从负标签中学到一些信息量的时候，温度T应调高一些；
- 当想减少负标签的干扰的时候，温度T应调低一些；

与传统训练流程的区别

其中KD的训练过程和传统的训练过程的对比：

- 传统training过程 Hard Targets: 对 ground truth 求极大似然 Softmax 值。
- KD的training过程 Soft Targets: 用 Teacher 模型的 class probabilities作为soft targets。

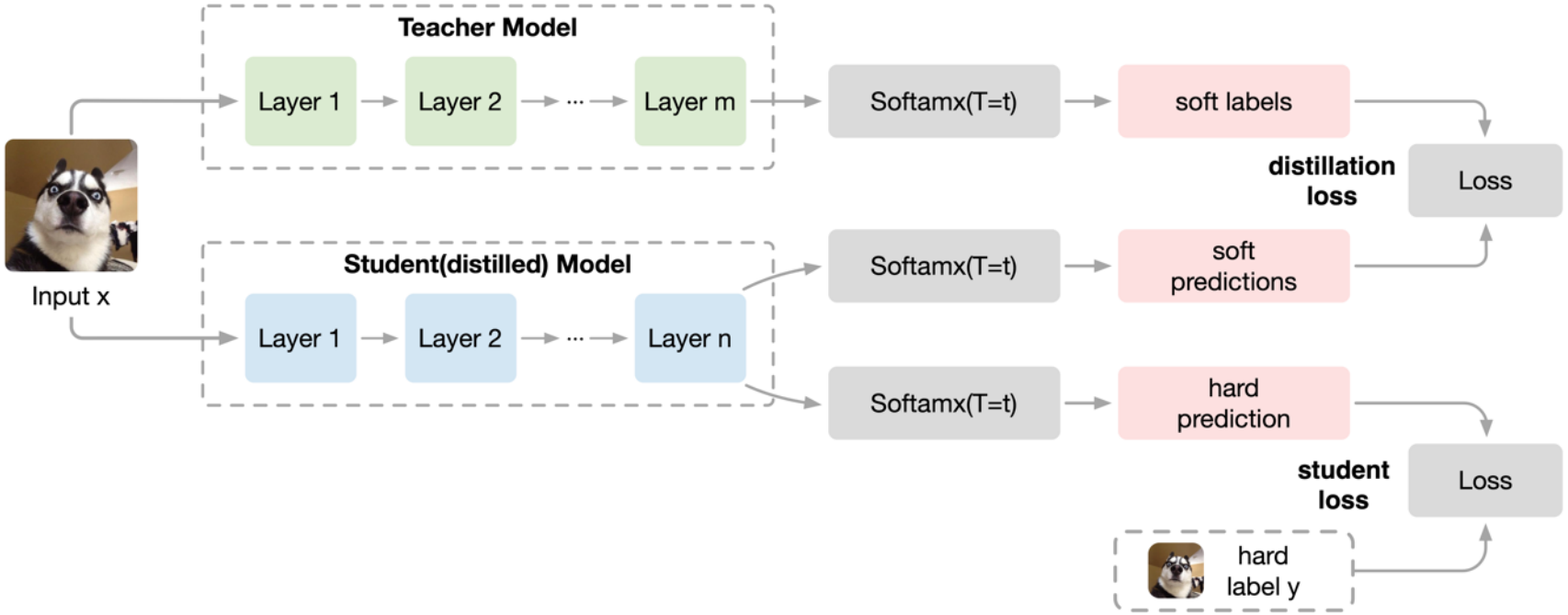
Distilling the Knowledge in a Neural Network

知识蒸馏使用offline distillation 的方式，采用经典的Teacher-Student 结构，其中 Teacher 是“知识”的输出者，Student 是“知识”的接受者。知识蒸馏的过程分为2个阶段：

1. **教师模型训练**：训练 Teacher 模型, 简称为 Net-T，特点是模型相对复杂，精度较高。对 Teacher模型不作任何关于模型架构、参数量等方面限制。唯一的要求就是，对于输入 X , 其都能输出 Y，其中 Y 经过 softmax 映射，输出值对应相应类别的概率值。
2. **学生模型蒸馏**：训练Student模型, 简称为 Net-S，它是参数量较小、模型结构相对简单的模型。同样对于输入 X，其都能输出 Y，Y经过 softmax 映射后能输出与 Net-T 对应类别概率值。

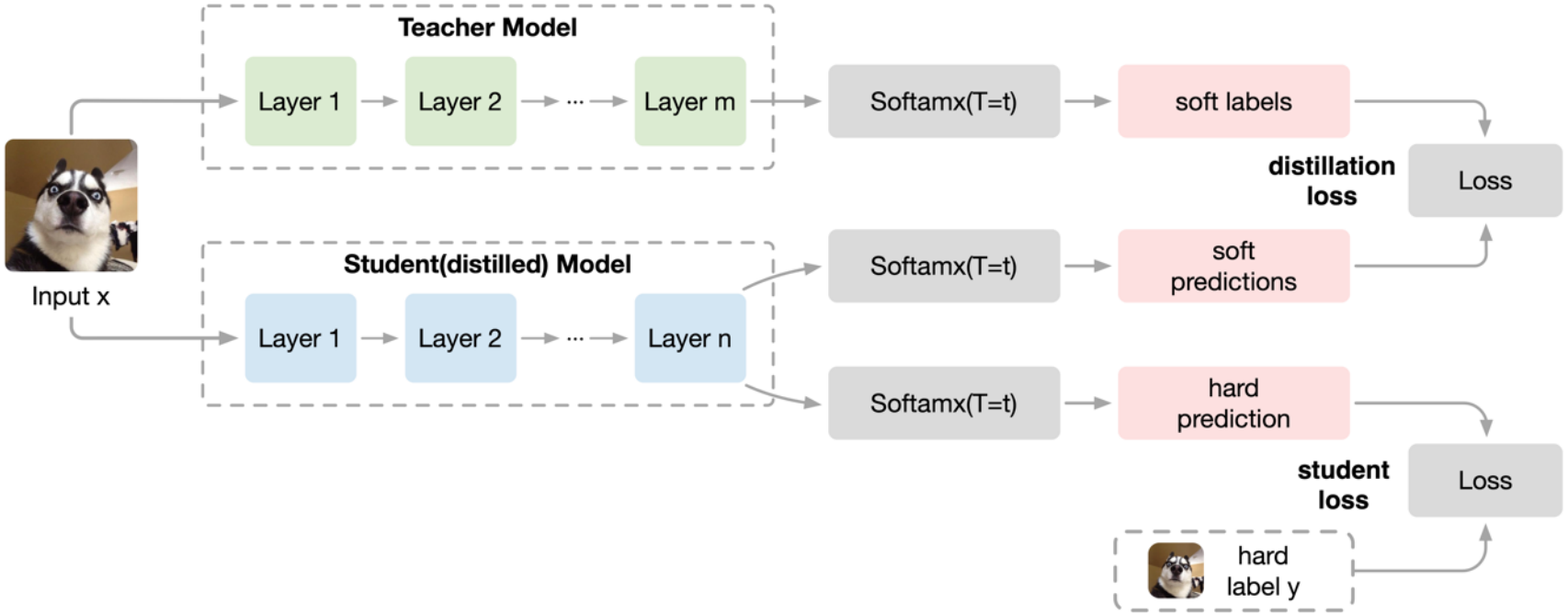
Distilling the Knowledge in a Neural Network

- 论文中，Hinton 将问题限定在分类问题下，或者属于分类的问题，共同点是模型最后有 soft max 层，其输出值对应类别的概率值。知识蒸馏时，由于已经有一个泛化能力较强的 Net-T，在利用 Net-T 来蒸馏 Net-S 时，可以直接让 Net-S 去学习 Net-T 的泛化能力。



Distilling the Knowledge in a Neural Network

1. 训练 Teacher Model；
2. 利用高温 T_{high} 产生 soft target；
3. 使用 $\{\text{soft target}, T_{high}\}$ 与 $\{\text{hard target}, T_{high}\}$ 同时训练 Student Model；
4. 设置温度 $\mathrm{T} = 1$，Student Model用于线上推理；



Distilling the Knowledge in a Neural Network

- 训练 Net-T 的过程中，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：

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

$$L_{soft} = -\sum_{j}^{N} p_j^T \log(q_j^T)$$

$$L_{hard} = -\sum_{j}^{N} c_j \log(q_j)$$

参考文献

1. Knowledge Distillation: A Survey
2. Distilling the Knowledge in a Neural Network
3. Circumventing outlier of autoaugment with knowledge distillation
4. 模型压缩（上）—— 知识蒸馏(Distilling Knowledge) https://www.jianshu.com/p/a6d87b338bcf
5. DeiT：注意力也能蒸馏 https://www.cnblogs.com/ZOMI/p/16496326.html