[1] Distilling the Knowledge in a Neural Network

首先提出目标：将一个模型（大规模神经网络）的知识部署到较小的模型中，实现知识迁移，构成更加适用于某一场景的小模型。 介绍中，以昆虫的类比介绍了知识蒸馏的基本理念，将训练好的较为庞大的泛化模型的知识迁移到较小的单一的适于部署的模型中。在训练阶段和部署阶段，尽管任务需求不同，但通常使用类似的模型。大型模型或模型集的训练需要大量计算资源，但部署到大规模用户时资源受限。因此，作者提出先训练一个笨重的模型（例如模型集或带有强正则化的大模型），然后通过蒸馏技术将知识转移到一个较小的、适合部署的模型中。

知识蒸馏：

蒸馏的核心思想是通过将复杂模型的输出作为"软目标"（Soft Target）（通常为预测分布的算数或几何平均值）来训练较小模型。这些软目标并非传统的hard labels（正确标签），而是包含了更加丰富信息的概率分布。通过在较大的泛化模型中使用更高的蒸馏温度T，可以使得输出结果更加平滑（也会导致数据不敏感），从而软化softmax输出，得到软目标，从而保留泛化模型对于错误类别的置信度信息。通过这种方法可以有效的将大型模型的泛化能力迁移到小模型中。这种方法下得到的小模型训练所需的数据量较小，并且学习率更高。 □□□

神经网络在分类任务中通常使用Softmax层来输出各类别的概率分布，公式：

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

qi为类i的概率，zi为类i的logit（即softmax层之前的输出），T为蒸馏温度参数。 常规训练中，T通常设为1，如果增加T，可以使得概率分布更加平滑（熵更高），softmax输出更均匀，也意味着模型对所有类别的预测更加不确定。相反，T降低会使概率分布更加尖锐，结果也更接近hard label，一个概率接近1，其他类别概率接近0.

Logits 是神经网络在输出层之前计算的值，通常是线性变换（权重加偏置）的结果。假设一个分类问题的最后一层神经元有 n 个输出（即 n 个类别），那么每个输出单元的 logit z_i 可以表示为：

$$z_i = W_i^T x + b_i$$

其中：

- W_i 是与类别 i 对应的权重向量；
- x 是前一层的输入；
- b_i 是偏置；
- z_i 是类别 i 对应的 logit 值。

反映了模型对不同分类的信心程度。

蒸馏过程：

通俗的来说，蒸馏就是让较小的模型模仿较大模型的概率输出分布来传递知识，这些概率分布即为软目标。在训练时，通常使用较高的T对大模型训练，在高温情况下的概率分布包括了模型对每个类别得到相对置信度和不正确类别的置信度。然后，在同样的T下，用大模型生成的软目标训练小模型。通过学习这些软目标，能够捕捉到大模型的泛化能力和类别之间的细微差异。在小模型训练完成后，在实际推理时，小模型的T设置为1，即不再需要平滑概率来进行预测。

当训练集的hard labels（正确标签）已知时，论文中提出了一个增强蒸馏方法，即同时使用软目标与硬标签的加权平均值进行训练：

· 目标函数1--交叉熵损失与软目标：首先，用软目标计算交叉熵损失（在较高T下进行），目的是让小模型更好地模仿大模型的预测分布。

· 目标函数2--交叉熵损失与硬标签：然后，用正确的硬标签计算另一个交叉熵损失（在T=1下完成），从而确保小模型在训练过程中能够输出正确的预测。

通常情况下，对于硬标签的损失给予较低的权重，主要关注通过软目标的学习。而软目标的梯度大小与1/T^2成比例，因此需要将梯度乘以T^2来确保不同目标的相对贡献保持平衡。

匹配logits的特殊情况

通过计算交叉熵损失函数来评估小模型输出分布是否接近大模型的输出分布：

$$\frac{\partial C}{\partial z_i} = \frac{1}{T}(q_i - p_i)$$

C为损失函数，zi是小模型在类i上的logit，qi是小模型的输出概率，pi是大模型的输出概率，T为蒸馏温度。即计算了交叉熵损失对小模型logits zi的梯度

T较高时，softmax公式：

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

$$p_i = \frac{\exp(v_i/T)}{\sum_j \exp(v_j/T)}$$

当T足够高时，exp(zi/T) 和 exp(vi/T) 可以近似表示为线性函数，

$$\frac{\partial C}{\partial z_i} \approx \frac{1}{NT^2}(z_i - v_i)$$

zi为小模型在类i上的logits，vi为大模型在类i上的logits，则蒸馏过程等价于最小化两个模型logits均方差：$\frac{1}{2}(z_i - v_i)^2$ 。

但是过度关注较大负对数可能会引入噪声，因此需要选择合适的T，适当忽略无用的logits信息，从而帮助小模型更好的学习。

负logits处理

负logits表示非常接近于0的概率，通常不会对分类结果产生较大影响，所以在T较小的情况下可以忽略。小模型只需要专注于对分类结果影响更大的logits。

所以当选择中等T时，既忽略了无关紧要的logits，同时也保留了足够的知识用于分类任务。

以MNIST数据集为例：通过蒸馏将一个大型神经网络（测试误差67）的软目标传递给小模型（测试误差146），后者的错误率降至74，显著提高了性能；

语音识别实验：采用了8层隐藏层的深度神经网络（DNN）作为声学模型，单独模型的词错误率（WER）为10.9%。通过训练10个不同的模型组成模型

集，WER降至10.7%。然后通过蒸馏技术，将10个模型的知识提取到一个单模型中，该单模型的WER同样为10.7%，接近模型集的性能。

为了进一步减少计算开销，作者提出了一种由全局模型（generalist model）和多个专家模型（specialist models）组成的模型集。每个专家模型专注于区分特定类别中的细粒度差异。通过对这些专家模型进行训练，可以显著提高整体模型的性能，同时通过蒸馏技术，可以避免过度的计算负担。

论文最后总结，蒸馏技术可以有效地将大型模型或模型集的知识压缩到较小的模型中，使其更适合部署。此外，在非常大的数据集上，使用专家模型集可以显著提高性能，而蒸馏技术则可以减少计算成本。

CONTRASTIVE REPRESENTATION DISTILLATION

主要提出了一种基于对比学习的知识蒸馏方法，能够捕获更多的结构化知识。

首先指出了现有方法的不足：主要通过最小化教师模型和学生模型输出分布的差异，即KL散度来优化学生模型，使其输出结果尽可能接近教师模型，但是KL散度只考虑了独立的输出维度之间的差异，忽视了可能存在的相互依赖的关系（结构性知识）；同时，当处理表示层进行蒸馏时，KL散度并不总是适用。

而通过引入对比学习，最大化正样本对（相似样本在表示空间中的相似性）和最小化负样本对（不同样本之间的差异）来学习教师模型中的表示，从而在跨模态转移（图像-声音）和模型集成蒸馏等任务上展现了优异的性能。【对比表示蒸馏CRD】

方法：

对比损失（Contrastive Loss）：对于两个深度神经网络：学生模型fs和教师模型ft，他们对输入x进行处理，得到表示ft(x)和fs(x)。目的是让fs(xi)尽可能接近ft(xi)，同时让fs(xi)和ft(xj)尽可能远离。定义具有潜在变量C的分布q，决定给定的样本对（S,T）是否从联合分布中抽取（同一个输入产生的表示），还是从独立的边缘分布中抽取（不同输入的表示）。通过最大化联合分布与边缘分布的KL散度，最大化学生和教师表示之间的互信息。公式：

$$
\begin{aligned}
q(C=1|T,S) &= \frac{q(T,S|C=1)q(C=1)}{q(T,S|C=0)q(C=0)+q(x,y|C=1)q(C=1)} \\
&= \frac{p(T,S)}{p(T,S)+Np(T)p(S)}
\end{aligned}
$$

最终的对比损失表达式类似InfoNCE损失，通过最大化互信息的下界来学习教师的表达。

从而得到了互信息的联系：

$$\begin{aligned}\log q(C=1|T,S) &= \log \frac{p(T,S)}{p(T,S)+Np(T)p(S)} \\ &= -\log(1+N\frac{p(T)p(S)}{p(T,S)}) \leq -\log(N)+\log\frac{p(T,S)}{p(T)p(S)}\end{aligned}$$

接下来的公式推导，通过修改InofNCE损失，通过最大化Lcritic，将问题转化为了学生模型fs和判别器h（用于区分正负样本对）的优化，判别器h的公式：

$$h(T,S)=\frac{e^{g^T(T)'g^S(S)/\tau}}{e^{g^T(T)'g^S(S)/\tau}+\frac{N}{M}}$$

N为负样本数量，M为数据集大小

知识蒸馏目标

除了对比损失，论文结合Hinton的经典知识蒸馏损失：①学生模型输出与真实标签之间的交叉熵损失 ②学生模型输出与教师模型输出之间的交叉熵损失，通过调节温度参数ρ，使输出概率分布平滑。

$$\mathcal{L}_{KD}=(1-\alpha)H(y,y^S)+\alpha\rho^2H(\sigma(z^T/\rho),\sigma(z^S/\rho))$$

α为平衡参数，σ为softmax函数。

跨模态迁移

在跨模态任务中，教师模型在源模态x上训练，并将知识转移到学生模型的目标模态y，使用对比损失来匹配学生和教师的特征表示。可以结合其他蒸馏目标，例如注意力转移、FitNet进一步提升迁移效果。这种方法在无监督场景下也适用。

集成蒸馏损失

对于集成蒸馏任务，假设有M（>1）个教师网络，fTi和一个学生网络fS，定义每个教师网络特征之间的多个成对对比损失Ti和学生网络fS，将损失相加，得到最终损失：

$$\mathcal{L}_{CRD-EN} = H(y, y^S) - \beta \sum_i \mathcal{L}_{critic}(T_i, S)$$

H(y,ys)表示学生模型输出ys和真实标签y之间的交叉熵损失。β为权重超参数，用于调节对比损失的影响力。后面将每个教师模型与学生模型之间的对比损失进行加权求和，使得学生模型能够同时学习多个教师模型的知识，提高模型性能和泛化能力。

实验

文章主要将CRD用于三个蒸馏任务场景进行测试：①将大模型压缩为小模型 ②跨模态知识迁移 ③多教师模型到单一学生模型的集成蒸馏

数据集：

1. CIFAR-100：该数据集包含60000张彩色图像，分为100个类别，每个类别有600张图像，选择ResNet-56作为教师模型，并转移到ResNet-20作为学生模型。
2. ImageNet：包含1000个类别和超过120万张图像，采用ResNet-34为教师模型，ResNet-18作为学生模型。
3. STL-10：10分类的5000张标注图像和100000张为标注图像，训练集为8000张图像
4. TinyImageNet：200类别，每个有500训练图像和50验证图像
5. NYU-Depth V2：1449室内图像，标注密度深度和语义地图 【RGB到深度图像跨模态知识迁移】

结果显示，CRD在CIFAR-100和ImageNet上显著优于其他方法，尤其在学生模型较小时，能够有效提升模型性能。

表明CRD能够捕获类别间相关性：

教师模型和学生模型的logit相关性矩阵差异最小，从而拥有较低的错误率

ImageNet实验结果：

为了和Zagoruko(2016)的工作进行比较，使用相同模型结构(ResNet-34作为教师模型，ResNet-18作为学生模型)。结果显示，CRD方法使学生模型和教师模型之间的准确率差距减少了1.42%，相对提高了50%，证明了CRD在大规模数据集上的可扩展性和优越性。

表示学习的可迁移性：

测试了蒸馏得到的表示能否较好的迁移到其他任务和数据集。实验中，使用WRN-16-2作为学生模型，①使用WRN-40-2教师模型蒸馏 ②从头开始在CIFAR-100上训练学生模型。然后使用学生模型作为表示提取器（logit前一层表示）来处理STL-10和TinyImageNet数据集的图像。

结论：CRD方法能够捕捉类别间相关性，迁移效果显著优于传统KD方法。

BIG TRANSER (BIT)

介绍：为了减少深度学习任务中为了达到较好表现所需要的计算资源，转移学习通过预训练阶段在大型通用数据集上训练网络，并将其权重用于初始化后续任务，从而大大提高了样本效率并简化了超参数调优。

BIT的核心思想是通过在大型监督数据集上进行预训练，然后对目标任务进行微调，从而实现最小化资源达到较好效果。

在和Generalist SOTA和Baseline (ILSVRC-2012)的训练结果对比中，准确率最高。

（1）上游（预训练）组件

--1 训练规模

BIT方法基于ResNet架构，通过在三个不同规模的数据集上进行预训练，验证数据集和模型规模对预训练表现和迁移学习效果的影响：

· BIT-L：在JFT-300M数据集上训练，包含3亿张图像和18291个类别

· Bit-M：在ImageNet-21k数据集上训练，包含1420万图像和2.1万个类别

· Bit-S：在ILSVRC-2012数据集上训练，包含约128万张图像和1000个类别

--2 归一化和权重标准化

组件归一化（Group Normalization，GN）和权重标准化（Weight Standardization，WS）：

传统视觉模型使用批归一化（Batch Normalization）来稳定模型，但是对于大规模迁移学习并不理想，原因：

① 在训练大型模型时，BN在设备上使用的小批量下表现较差，或者需要进行设备间的同步，增加了训练成本

② BN需要更新运行时统计信息，对于迁移任务不利

作者采用GN和WS组合来替代BN，在实验中显示出了这种组合对于小批次训练任务有所提升。

（2）下游（微调）任务组件

--1 微调协议

Bit提出了一个简单的超参数设置规则，称为Bit-HyperRule，避免了对每个任务进行昂贵的超参数搜索。该规则根据任务的图像分辨率和数据集规模设置训练调度长度、分辨率以及是否使用MixUp正则化。

主要选择最重要的超参数，包括了任务的固有图像分辨率和数据点数量的函数，对于每个任务，作者主要调整了以下超参数：① 训练调度长度 ② 图像分辨率 ③ 是否使用MixUp正则化

该规则被应用于超过20个不同的任务，从每类只有一个样本的小样本任务到拥有超过100万张样本的大数据集。

--2 数据预处理和增强

在微调过程中，使用标准的数据处理和增强技术：

· 将图像调整为正方形

· 随机裁剪出一个小正方形

· 在训练时随机水平翻转图像

· 在测试时仅将图像调整为固定大小

--3 测试时的分辨率调整

研究表明，现有的数据增强方法在训练和测试分辨率之间引入了不一致性。因此，通常在测试时将分辨率放大一个小的比例，或在微调步骤中添加分辨率变化。对于迁移学习，调整分辨率更加适用。

--4 MixUp正则化应用

在预训练中，MixUp正则化的效果有限（可能是因为数据量充足）。而在中等规模的数据集的迁移任务中，MixUp较为有效

--5 正则化和权重衰减

作者在实验中并没有使用常见的正则化形式（权重衰减至0，权重衰减至初参数，dropout），BiT-L模型有9.28亿个参数，但是即使没有使用正则化技术，将其迁移到小数据集时仍表现良好。即为较大的数据集设定合适的训练调度长度可以提供足够的正则化效果。

1. 知识蒸馏的发展进步：时间顺序分析

(1) Hinton等人的经典知识蒸馏 (2015)

- **背景**：知识蒸馏由Hinton等人在2015年首次提出，主要解决如何将复杂的、计算开销大的大模型（教师模型）的知识压缩到较小的模型（学生模型）中。
- **核心思想**：其方法通过最小化教师模型和学生模型的输出分布之间的KL散度，使学生模型尽可能接近教师模型的表现。该技术使得模型的推理速度大大提高，尤其适用于资源受限的设备。
- **贡献**：这项工作开启了知识蒸馏研究的一个重要方向，即如何在不显著降低性能的前提下，利用较小的模型进行高效推理。这一框架也成为之后蒸馏技术改进的基础。

(2) Contrastive Representation Distillation (CRD, 2019)

- **背景**：随着研究的深入，研究者们发现，传统的知识蒸馏方法（如KD）只关注教师模型和学生模型的输出层信息，而忽略了中间层中的结构化知识。这种不足导致学生模型在某些任务上的泛化能力较弱。
- **核心贡献**：CRD通过引入**对比学习**，弥补了KD的这一不足。CRD不仅最小化输出分布的差异，还通过对比损失捕捉教师模型和学生模型的中间层表示（representation）之间的相似性。
- **进步**：CRD的显著贡献在于捕获了更深层次的结构信息，使学生模型能够学习到类别间的相关性。这极大地提升了学生模型的泛化能力，尤其是在需要表征学习的任务上，如表示迁移和跨模态任务。

(3) Big Transfer (BiT, 2019)

- **背景**：随着模型规模的扩大和数据集的增加，研究者们开始探讨如何通过大规模的预训练模型来实现高效的迁移学习。BiT 方法旨在解决如何将

大模型的知识转移到各种不同任务和数据集上，特别是那些数据稀少的任务。

- **核心贡献**：BiT 的核心在于通过在超大规模数据集上预训练（如 JFT-300M），然后针对特定任务进行微调。这种做法允许模型在少样本甚至单样本学习的任务中表现出色。与传统的知识蒸馏方法不同，BiT 强调的是从大数据中提取到丰富的通用知识，适用于各种视觉任务。
- **进步**：BiT 的影响力不仅在于其简单有效的迁移学习框架，还在于它展示了**大规模预训练模型**在少样本场景下的卓越表现。这一成果对迁移学习、预训练模型和视觉任务等领域产生了深远影响，推动了之后的大规模预训练模型的研究浪潮。

2. CRD 和 Big Transfer 影响力的分析

(1) CRD 的影响力

CRD 作为对传统知识蒸馏方法的改进，影响力来源于以下几个关键方面：

- **创新性方法**：通过引入对比学习，CRD 不仅仅关注输出层的概率分布，还深入捕捉了教师模型的内部结构和类别之间的相似性。这种创新性的对比损失框架有效地提高了学生模型的性能，尤其在表示迁移任务和细粒度分类任务中表现突出。
- **广泛适用性**：CRD 可以应用于多种知识转移场景中，如模型压缩、跨模态迁移、模型集成等。这种方法具有较强的普适性，适用于多种任务和架构。
- **性能提升**：实验结果表明，CRD 在多个任务上显著超越了传统的KD方法，展示了对比学习的强大能力，特别是在处理复杂结构化数据时。

(2) Big Transfer (BiT) 的影响力

BiT 的巨大影响力主要来自以下几点：

- **大规模预训练模型的引领者**：BiT 展示了大规模预训练模型在迁移学习中的强大能力，特别是在处理小样本任务时的表现极为突出。通过在超大数据集上进行预训练，BiT 成功将这些知识转移到小数据集或特定任务中，从而显著提升了模型性能。
- **易于迁移的框架**：BiT-HyperRule 简化了任务的微调过程，不需要对每个新任务进行复杂的超参数搜索。这种简化使得迁移学习过程更加高效，降低了应用的复杂性，进一步提高了其实际应用价值。

- **影响后续大模型发展**：BiT 的成功证明了预训练和微调策略在各类视觉任务中的巨大潜力，推动了大模型和少样本学习的研究。许多后续的预训练模型（如 GPT-3 和 CLIP）受到了 BiT 的启发，进一步发展了大规模预训练的理念。

总结

从2015年的经典知识蒸馏到2019年推出的CRD和BiT，知识蒸馏领域取得了显著的进展。从最初的单纯输出层知识传递，到CRD引入对比学习捕捉深层表示，再到BiT通过大规模预训练增强迁移学习能力，知识蒸馏逐渐变得更加复杂和强大。这些进步使得知识蒸馏方法在处理复杂任务、跨模态迁移以及少样本学习中，展示出更高的灵活性和表现力。CRD 和 BiT 的成功也为后续的大模型预训练和高效知识转移提供了新的方向和思路。