知识蒸馏研究综述

黄震华[1),2)] 杨顺志[1)] 林 威[1)] 倪 娟[3)] 孙圣力[4)] 陈运文[5)] 汤 庸[1)]

[1)](华南师范大学计算机学院 广州 510631)

[2)](同济大学电子与信息工程学院 上海 201804)

[3)](华南师范大学哲学与社会发展学院 广州 510631)

[4)](北京大学软件与微电子学院 北京 102600)

[5)](达而观智能（深圳）有限公司研发部 广东 深圳 518063)

摘 要 高性能的深度学习网络通常是计算型和参数密集型的，难以应用于资源受限的边缘设备. 为了能够在低资源设备上运行深度学习模型，需要研发高效的小规模网络. 知识蒸馏是获取高效小规模网络的一种新兴方法，其主要思想是将学习能力强的复杂教师模型中的“知识”迁移到简单的学生模型中. 同时，它通过神经网络的互学习、自学习等优化策略和无标签、跨模态等数据资源对模型的性能增强也具有显著的效果. 基于在模型压缩和模型增强上的优越特性，知识蒸馏已成为深度学习领域的一个研究热点和重点. 本文从基础知识,理论方法和应用等方面对近些年知识蒸馏的研究展开全面的调查，具体包含以下内容：(1)回顾了知识蒸馏的背景知识，包括它的由来和核心思想；(2)解释知识蒸馏的作用机制；(3)归纳知识蒸馏中知识的不同形式，分为输出特征知识、中间特征知识、关系特征知识和结构特征知识；(4)详细分析和对比了知识蒸馏的各种关键方法，包括知识合并、多教师学习、教师助理、跨模态蒸馏、相互蒸馏、终身蒸馏以及自蒸馏；(5)介绍知识蒸馏与其它技术融合的相关方法，包括生成对抗网络、神经架构搜索、强化学习、图卷积、其它压缩技术、自动编码器、集成学习以及联邦学习；(6)对知识蒸馏在多个不同领域下的应用场景进行了详细的阐述；(7)讨论了知识蒸馏存在的挑战和未来的研究方向.

关键词 知识蒸馏；模型压缩；模型增强；知识迁移；深度学习

中图法分类号 TP311 **DOI 号** 10. 11897/SP. J. 1016. 2022.00624

Knowledge Distillation: A Survey

HUANG Zhen-Hua[1),2)] YANG Shun-Zhi[1)] LIN Wei[1)] NI Juan[3)]
SUN Sheng-Li[4)] CHEN Yun-Wen[5)] TANG Yong[1)]

[1)](*School of Computer Science, South China Normal University, Guangzhou* 510631)

[2)](*School of Electronic and Information Engineering, Tongji University, Shanghai* 201804)

[3)](*School of Philosophy and Social Development, South China Normal University, Guangzhou* 510631)

[4)](*School of Software & Microelectronics, Peking University, Beijing* 102600)

[5)](*Research and Development Department, DataGrand Inc, Shenzhen, Guangdong* 518063)

Abstract High-performance deep learning models are usually computationally and parameter-intensive, making it hard to deploy on edge devices with limited resources. In order to run deep learning models on low resource devices, efficient small-scale networks are needed. Knowledge distillation is a new method to obtain efficient small-scale networks. Its main idea is to transfer the “knowledge” from complex teacher

收稿日期：2021-01-17；在线发布日期：2021-08-02.本课题得到国家自然科学基金(61772366，U1811263，61972328)、上海市自然科学基金(17ZR1445900)、广东省科技计划项目(2019B090905005)资助. **黄震华(通信作者)**，博士，教授，博士生导师，主要研究领域为机器学习、数据挖掘、推荐系统. E-mail: huangzhenhua@m.scnu.edu.cn. **杨顺志**，博士研究生，主要研究领域为高效网络模型、知识蒸馏、图像识别. **林 威**，硕士研究生，主要研究领域为推荐系统、知识蒸馏. **倪 娟**，硕士，讲师，主要研究领域为教育大数据、知识图谱. **孙圣力**，博士，副教授，主要研究领域为机器学习、数据挖掘、数据库. **陈运文**，博士，高级工程师，主要研究领域为机器学习、数据挖掘、自然语言处理. **汤 庸**，博士，教授，博士生导师，中国计算机学会（CCF）杰出会员，主要研究领域为教育大数据、数据挖掘、数据库.

networks with a strong learning ability to simple student networks. In knowledge distillation, a student model improves its generalization ability by imitating the “dark knowledge” of the corresponding teacher. At the same time, it can improve the performance of models by exploiting the optimization strategies such as mutual learning and self-learning of neural networks and the data resources such as unlabeled and cross-modal. Therefore, we can obtain an efficient and effective deep learning network model through knowledge distillation. Based on these predominant characteristics in model compression and enhancement, knowledge distillation has become a research hotspot and focus in the field of deep learning. Currently, there are some surveys on knowledge distillation. However, they lack more systematic studies to present global and comprehensive views on knowledge distillation. First of all, previous investigations have ignored the application prospects of knowledge distillation in model enhancement. Second, previous surveys did not pay attention to structure knowledge, which is indispensable in the knowledge structure of a network. In model enhancement and structural feature knowledge, the application prospects of knowledge distillation have become more and more important in improving the performance of student models in the past two years. In order to overcome the shortcomings of previous works, this paper gives a description based on knowledge distillation from different perspectives and provides more detailed knowledge introductions. Specifically, we conduct a more comprehensive investigation of knowledge distillation in recent years from the aspects of basic knowledge, theoretical methods and applications, etc. It is composed of the following contents. (1) Review the background knowledge of knowledge distillation, including its origin and core ideas. (2) The working mechanism of knowledge distillation is introduced in detail, i. e. , provides the reason why knowledge distillation is effective. (3) The different knowledge forms in knowledge distillation are summarized, which are divided into response-based, feature-based, relation-based and structure knowledge. (4) Detailed analysis and comparison of various key methods in knowledge distillation, which emphasizes knowledge transfer ways, including knowledge amalgamation, learning from multiple teachers, teacher assistants, cross modal distillation, mutual distillation, lifelong distillation, and self-distillation. (5) Related methods of knowledge distillation integration with other technologies are introduced, including generative adversarial networks, neural architecture search, reinforcement learning, graph convolution, other compression techniques, autoencoders, ensemble learning, and federated learning. (6) The application scenarios of knowledge distillation in many fields are described in detail, including its application progress in model compression and enhancement. (7) The current challenges and future development trends of knowledge distillation are also discussed at the end of this paper. In a word, this paper surveys the research progress of knowledge distillation in recent years, and summarizes, compares, and analyzes the following aspects for it: origin, mechanism, knowledge forms, key methods, integration with other technologies, application progress, challenges and perspective.

Keywords knowledge distillation; model compression; model enhancement; knowledge transfer; deep learning

1 引 言

深度学习由于对目标多样性变化具有很好的鲁棒性，近年来得到广泛的关注并取得快速的发展. 然而性能越好的深度学习模型往往需要越多的资源，使其在物联网、移动互联网等低资源设备的应用上受到限制. 因此研究人员开始对高效的(Efficient)深度学习模型展开研究，其目的是使具有高性能的模型能够满足低资源设备的低功耗和实时性等要求，同时尽可能地不降低模型的性能. 当前，主要有 5 种方法可以获得高效的深度学习模型：直接手工设计轻量级网络模型、剪枝、量化、基于神经架构搜索(Neural Architecture Search，NAS)[1]的网络自动化设计以及知识蒸馏(Knowledge Distillation，

KD)[2]. 其中，知识蒸馏作为一种新兴的模型压缩方法，目前已成为深度学习领域的一个研究热点和重点. 国内外许多大学和研究机构已经对知识蒸馏展开了深入研究，并且每年在机器学习和数据挖掘的国际顶级会议和知名期刊中都有关于知识蒸馏的文章发表.

知识蒸馏是一种教师-学生(Teacher-Student)训练结构，通常是已训练好的教师模型提供知识，学生模型通过蒸馏训练来获取教师的知识. 它可以以轻微的性能损失为代价将复杂教师模型的知识迁移到简单的学生模型中. 在后续的研究中，学术界和工业界扩展了知识蒸馏的应用范畴，提出了利用知识蒸馏来实现模型性能的增强. 基于此，本文根据应用场景划分出基于知识蒸馏的模型压缩和模型增强这两个技术方向，即获得的网络模型是否为了应用于资源受限的设备. 图 1 给出了这两种技术对比的一个例子，其中的教师模型都是提前训练好的复杂网络. 模型压缩和模型增强都是将教师模型的知识迁移到学生模型中. 所不同的是，模型压缩是教师网络在相同的带标签的数据集上指导学生网络的训练来获得简单而高效的网络模型，如左图的学生是高效的小规模网络. 模型增强则强调利用其它资源(如无标签或跨模态的数据)或知识蒸馏的优化策略(如相互学习和自学习)来提高一个复杂学生模型的性能. 如右图中，一个无标签的样本同时作为教师和学生网络的输入，性能强大的教师网络通常能预测出该样本的标签，然后利用该标签去指导复杂的学生网络训练.

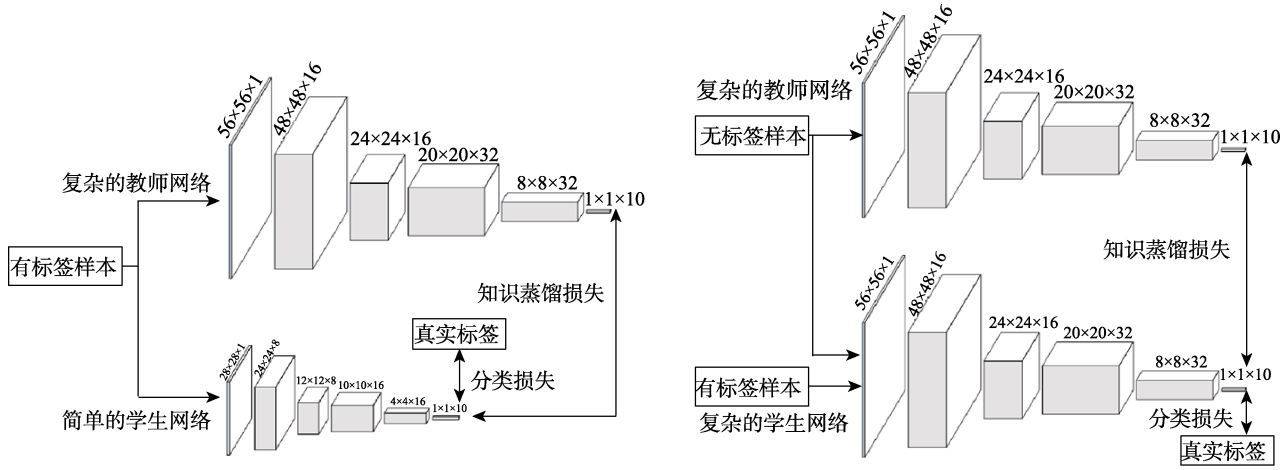

图 1 知识蒸馏的两个技术方向：模型压缩(左)和模型增强(右)

本文重点收集了近些年在人工智能、机器学习以及数据挖掘等领域的国际顶级会议(如 ICCV，ICML，EMNLP，KDD)与重要学术期刊(如 PAMI，TOIS，TKDE，TMM)上有关知识蒸馏的论文并加以整理、归纳和分析. 据我们所知，目前国内没有知识蒸馏相关的中文综述，而先前两篇英文综述[3,4]和我们工作相似，但本文进一步完善了知识蒸馏的综述. 具体地，本文与先前的英文综述[3,4]至少有以下三点的不同：

（1）先前的研究都忽略了知识蒸馏在模型增强上的应用前景. 在本文的研究调查中，知识蒸馏不仅可以用于模型压缩，它还能通过互学习和自学习等优化策略来提高一个复杂模型的性能. 同时，知识蒸馏可以利用无标签和跨模态等数据的特征，对模型增强也具有显著的提升效果.

（2）先前的研究都没有关注到结构化特征知识，而它在知识架构中又是不可或缺的. 某个结构上的知识往往不是单一的，它们是有关联的、多个知识形式组合. 充分利用教师网络中的结构化特征知识对学生模型的性能提升是有利的，因此它在近两年的工作中越发重要[5,6].

（3）本文从不同视角给出了基于知识蒸馏的描述，并提供了更多的知识介绍. 在知识蒸馏的方法上，本文增加了知识合并和教师助理的介绍；在技术融合的小节，本文增加了知识蒸馏与自动编码器、集成学习和联邦学习的技术融合；在知识蒸馏的应用进展中，本文分别介绍了知识蒸馏在模型压缩和模型增强的应用，并增加了多模态数据和金融证券的应用进展；在知识蒸馏的研究趋势展望中，本文给出了更多的研究趋势，特别是介绍了模型增强的应用前景.

总的来说，本文在文献[3,4]基础上，以不同的视角，提供更加全面的综述，以便为后续学者了解或研究知识蒸馏提供参考指导.

本文组织结构如图 2 所示. 第 2 节回顾了知识蒸馏的背景知识，包括它的由来；第 3 节解释知识蒸馏的作用机制，即为什么知识蒸馏是有效的；第 4 节归纳知识蒸馏中知识的不同形式；第 5 节详细分析了知识蒸馏的各种方法，其强调的是知识迁移的方式；第 6 节介绍知识蒸馏与其它技术融合的相关方法；第 7 节归纳知识蒸馏的应用进展；第 8 节给出了知识蒸馏的研究趋势展望. 最后，第 9 节对本文工作进行总结.

2　知识蒸馏的提出

知识蒸馏与较早提出的并被广泛应用的一种机器学习方法的思想较为相似，即迁移学习[7]. 知识蒸馏与迁移学习都涉及到知识的迁移，然而它们有以下四点的不同：

(1)数据域不同. 知识蒸馏中的知识通常是在同一个目标数据集上进行迁移，而迁移学习中的知识往往是在不同目标的数据集上进行转移.

(2)网络结构不同. 知识蒸馏的两个网络可以是同构或者异构的，而迁移学习通常是在单个网络上利用其它领域的数据知识.

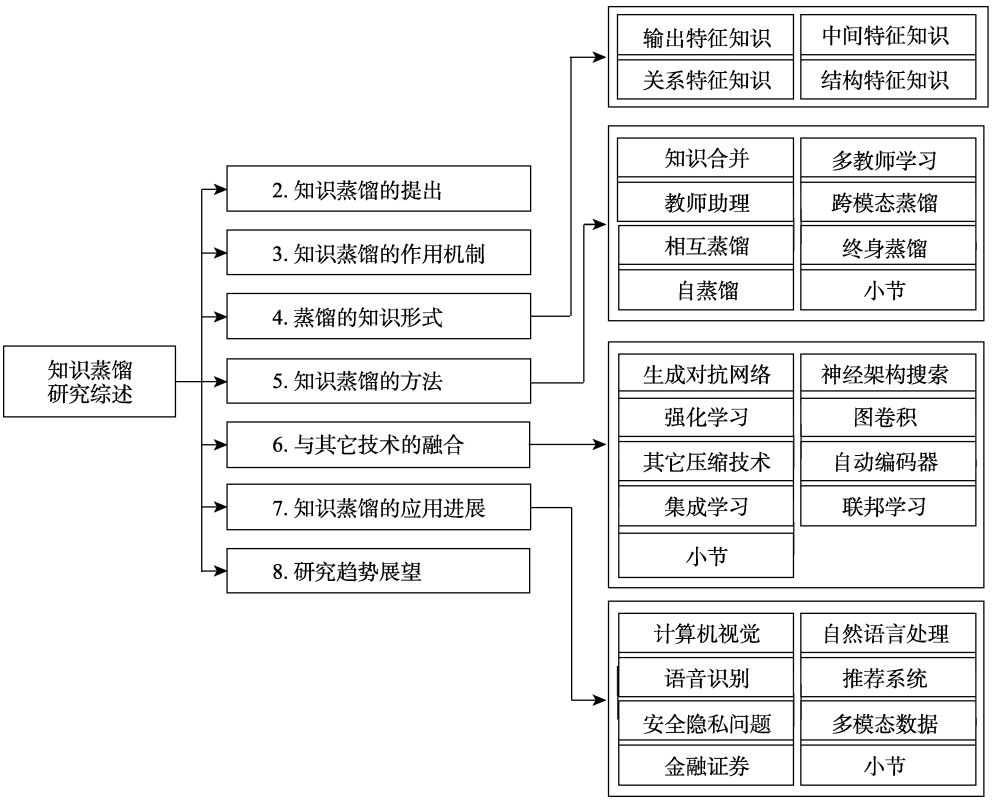

图 2　本文的组织结构图

(3)学习方式不同. 迁移学习使用其它领域的丰富数据的权重来帮助目标数据的学习，而知识蒸馏不会直接使用学到的权重.

(4)目的不同. 知识蒸馏通常是训练一个轻量级的网络来逼近复杂网络的性能，而迁移学习是将已经学习到相关任务模型的权重来解决目标数据集的样本不足问题.

因此，知识蒸馏更强调的是知识的迁移，而非权重的迁移. 与知识蒸馏思想最为接近的工作是 Bucilu 等人[8]在 2006 年提出的“模型压缩”(Model Compression)，它通过学习较大但性能更好模型的近似特征来获得轻量级的网络模型. 在 Bucilu 等人[8]“模型压缩”思想的启发下，Liang 等人[9]使用未标记的数据将条件随机场模型的预测能力转移到计算简单的独立逻辑回归模型中. Ba 和 Caruana[10]通过 L2 损失函数学习大网络输出的逻辑单元(Logits)来训练小网络. 另外，Li 等人[11]则是利用深度神经网络输出分布特性，最小化小网络和大网络输出分布之间的 KL (Kullback- Leibler)散度.

这些早期的工作都涉及到小网络利用大网络的输出知识，如逻辑单元[8,10]和类概率(Class Probabilities)[9,11]. 逻辑单元是 Softmax 激活的前一层，而类

概率是逻辑单元通过 Softmax 激活函数转化而来：

$$p_i(z_i)=\frac{\exp(z_i)}{\sum_{j=0}^{k}\exp(z_j)} \tag{1}$$

其中，z_i是第 i 类的逻辑单元值，p_i是第 i 类的类概率以及 k 表示类别的数量. 网络输出逻辑单元和类概率的关系如图 3 所示.

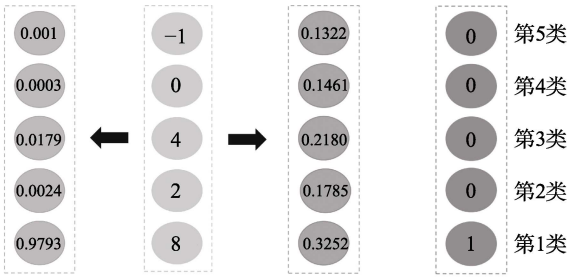

图 3 软目标、逻辑单元和硬目标的关系

上述工作在训练和测试阶段使用的知识都是一致的：要么使用逻辑单元[8,10]，要么使用类概率[9,11]. 但是，如果使用逻辑单元来表示知识，这些不受约束的值在测试的时候可能会包含噪声信息. 如图 3 所示，在网络测试时逻辑单元层中第 3 个类别的输出值是噪声. 如果将其包含噪声信息的逻辑单元作为学生的监督信号，则会因过拟合而限制了学生的泛化能力. 另一方面，如果使用类概率来表示知识，那么负标签的概率被 Softmax 压扁后将会接近零而导致在网络训练时这部分信息丢失. 例如，在网络训练时，类概率层(如图 3 中 T=1 的软目标)的负标签输出的信息基本已经丢失. 将该类概率作为学生的监督信号，相当于让学生学习硬目标知识.

为解决上述问题，Hinton 等人[2]在 2015 年引入软目标(即带有参数 T 的类概率)并提出知识蒸馏概念：

$$p_i(z_i,T)=\frac{\exp(z_i/T)}{\sum_{j=0}^{k}\exp(z_i/T)} \tag{2}$$

其中，T 为温度系数，用来控制输出概率的软化程度. 不难看出，当 T=1 时，公式(2)表示网络输出 Softmax 的类概率. 当 T 为正无穷大时，Hinton 等人[2]在论文中证明了公式（2）此时表示网络输出的逻辑单元.

具体来说，教师模型和学生模型在逻辑单元匹配时的蒸馏损失定义为

$$L_{KD}\big(p(u,T),p(z,T)\big)=\sum_{i=0}^{k}-p_i(u_i,T)\log\big(p_i(z_i,T)\big) \tag{3}$$

其中，u 和 z 分别为教师和学生模型输出的逻辑单元. 当 T 为正无穷大时，反向传播时的蒸馏损失优化为

$$\begin{aligned}\frac{\partial L_{KD}\big(p(u,T),p(z,T)\big)}{\partial z_i}&=\frac{p_i(z_i,T)-p_i(u_i,T)}{T}\\&\approx\frac{1}{T}\left(\frac{1+\left(\frac{z_i}{T}\right)}{k+\sum_{j=0}^{k}\frac{z_i}{T}}-\frac{1+\left(\frac{u_i}{T}\right)}{k+\sum_{j=0}^{k}\frac{u_i}{T}}\right)\end{aligned} \tag{4}$$

假定 z_i 是 0 均值的，则有

$$\frac{\partial L_{KD}\big(p(u,T),p(z,T)\big)}{\partial z_i}\approx\frac{1}{kT^2}(z_i-u_i) \tag{5}$$

优化逻辑单元匹配的蒸馏损失为标签匹配的均方误差$(z_i-u_i)/(kT^2)$，这表明逻辑单元实际上是知识蒸馏的一个特例. 因此，可以通过调节 T 来解决把逻辑单元和类概率作为知识时产生的问题. 通常，知识蒸馏在测试的时候令 T=1，在训练的时候则使用较大的 T 值. 知识蒸馏在测试阶段令 T=1 时，不同逻辑单元值的软目标的差异很大，所以在测试时能够较好地区分开正确的类和错误的类. 而在训练时，较大 T 值的软目标差异比 T=1 时的差异小，模型训练时会对较小的逻辑单元给予更多的关注，从而使学生模型学习到这些负样本和正样本之间的关系信息. Hinton 等人[2]将这样的蕴含在教师模型中的关系信息称之为“暗知识”(Dark Knowledge)，而知识蒸馏就是在训练过程中将教师模型的“暗知识”传递到学生模型中. 因此，知识蒸馏通过调节 T 来解决了先前工作[8-11]存在的问题，从而提高学生模型的预测性能.

除了利用教师模型输出的软目标外，Hinton 等人[2]还发现了在训练过程加上正确的数据标签(即硬目标)会使学习效果更好. 具体是对两个不同的目标函数进行权重平均. 第一个目标函数是具有较高 T 值教师模型和具有较高 T 值学生模型的交叉熵损失(称为蒸馏损失)，第二个目标函数是 T=1 的学生和硬目标的交叉熵损失(称为学生损失). 因此，知识蒸馏的总损失可以表示为

$$\begin{aligned}L_{total}=&\lambda\cdot L_{KD}\big(p(u,T),p(z,T)\big)\\&+(1-\lambda)\cdot L_S\big(y,p(z,1)\big)\end{aligned} \tag{6}$$

其中，λ 是超参数，$L_S(y, p(z, T))$是学生损失. λ 的值通常是经验调参的固定值，也可以动态地调整. 学生损失表示为

$$L_S\big(\boldsymbol{y},p(z,1)\big)=\sum_{i=0}^{k}-y_i\log\big(p_i(z_i,1)\big) \tag{7}$$

其中$\boldsymbol{y}$是硬标签的向量. 知识蒸馏的框架如图 4 所示.

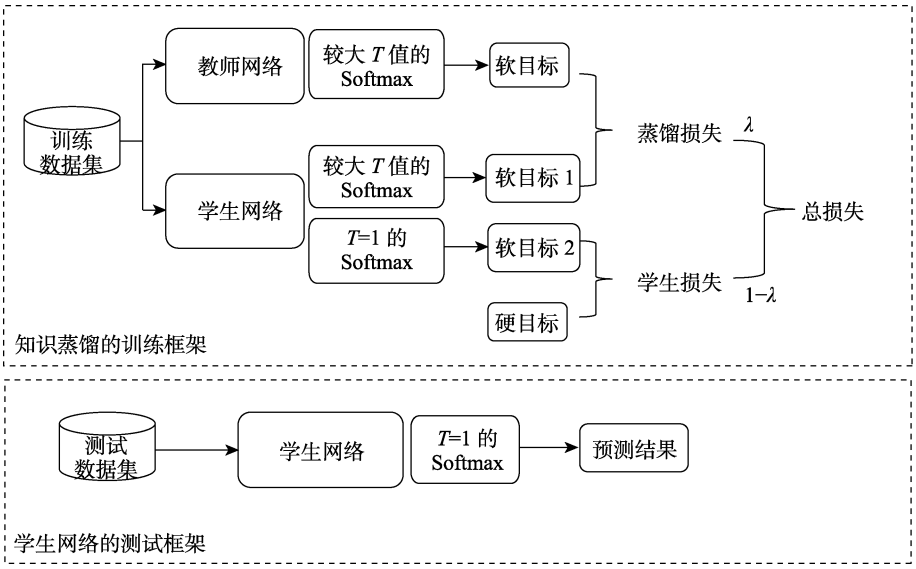

图 4　Hinton 等人[2]的知识蒸馏框架

3　知识蒸馏的作用机制

Hinton 等人[2]认为，学生模型在知识蒸馏的过程中通过模仿教师模型输出类间相似性的“暗知识”来提高泛化能力. 软目标携带着比硬目标更多的泛化信息来防止学生模型过拟合. 虽然知识蒸馏已经获得了广泛的应用,但是学生模型的性能通常是仅接近于教师模型. 特别地，给定学生和教师模型相同的大小却能够让学生模型的性能超越教师模型[12]，性能越差的教师模型反倒教出了更好的学生模型[13]. 为了更好地理解知识蒸馏的作用，一些工作从数学或实验上对知识蒸馏的作用机制进行了证明和解释. 本文归纳为以下几类：

(1) 软目标为学生模型提供正则化约束. 这一结论最早可以追溯到通过贝叶斯优化来控制网络超参数的对比试验[14]，其表明了教师模型的软目标为学生模型提供了显著的正则化. 软目标正则化的作用是双向的，即还能将知识从较弱的教师模型迁移到能力更强大的学生模型中[15,16]. 一方面，软目标通过标签平滑训练提供了正则化[15,16]，标签平滑是通过避免了过分相信训练样本的真实标签来防止训练的过拟合[15]. 另一方面，软目标通过置信度惩罚提供了正则化[12]，置信度惩罚让学生模型获得更好的泛化能力，其主要依赖于教师模型对正确预测的信心. 这两种正则化的形式已经在数学上得到了证明. 总的来说，软目标通过提供标签平滑和置信度惩罚来对学生模型施加正则化训练. 因此，即使没有强大的教师模型，学生模型仍然可以通过自己训练或手动设计的正则化项得到增强[16].

(2) 软目标为学生模型提供了“特权信息”(Privileged Information). “特权信息”指教师模型提供的解释、评论和比较等信息[17]. 教师模型在训练的过程中将软目标的“暗知识”迁移到学生模型中，而学生模型在测试的过程中并不能使用“暗知识”. 从这个角度看，知识蒸馏是通过软目标来为学生模型传递“特权信息”.

(3) 软目标引导了学生模型优化的方向. Phuong 等人[18]从模型训练的角度证明了软目标能引导学生模型的优化方向. 同时，Cheng 等人[19]从数学上验证了软目标使学生模型比从原始数据中进行优化学习具有更高的学习速度和更好的性能.

4　蒸馏的知识形式

原始知识蒸馏(Vanilla Knowledge Distillation)[2]仅仅是从教师模型输出的软目标中学习出轻量级的学生模型. 然而，当教师模型变得更深时，仅仅学习软目标是不够的. 因此，我们不仅需要获取教师模型输出的知识，还需要学习隐含在教师模型中的其它知识，比如中间特征知识. 本节总结了可以使用的知识形式有输出特征知识、中间特征知识、关系特征知识和结构特征知识. 知识蒸馏的 4 种知识形式的关系如图 5 所示. 从学生解题的角度，这 4 种知识形式可以形象比喻为：输出特征知识提供了解题的答案，中间特征知识提供了解题的过程，关系特征知识提供了解题的方法，结构特征知识则提供了完整的知识体系.

4.1 输出特征知识

输出特征知识通常指的是教师模型的最后一层特征，主要包括逻辑单元和软目标的知识. 输出特征知识蒸馏的主要思想是促使学生能够学习到教师模型的最终预测，以达到和教师模型一样的预测性能. 原始知识蒸馏是针对分类任务来提出的仅包含类间相似性的软目标知识，然而其它任务(如目标检测)网络最后一层特征输出中还可能包含有目标定位的信息. 换句话说，不同任务教师模型的最后一层输出特征是不一样的. 因此，本文根据任务的不同对输出特征知识分别进行归纳和分析，如表 1 所示.

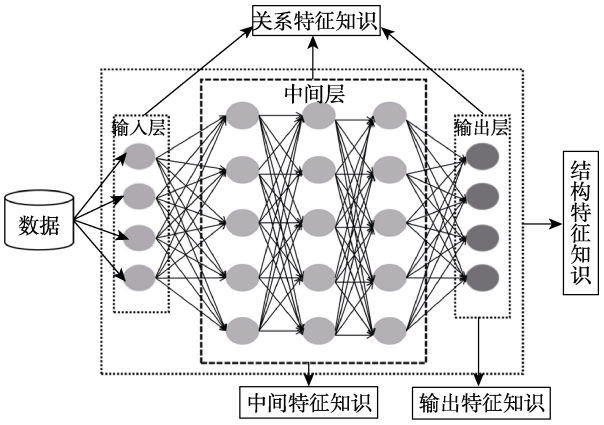

图 5 蒸馏的知识形式

表 1 输出的特征知识

任务名称	知识的类型	相关工作	描述
目标分类	软目标知识	文献[2,13]	分类任务重点是学生如何模仿到教师的软目标. 与硬目标知识相比，教师的软目标提供给学生模型更多的类间知识. 目标分类任务的知识蒸馏就是学习教师模型输出的软目标知识.
目标检测	边界框回归知识和软目标知识	文献[20,21]	目标检测网络的最后输出层包含区域建议网络(Region Proposal Network，RPN)的边界框回归和区域分类网络(Region Classification Network，RCN)的软目标知识. 前者的知识用于定位，后者用于分类.
目标分割	像素级软目标知识和空间上下文结构知识	文献[22,23]	目标分割需要分割出具有抽象语义的目标，即要对目标的每一个像素都要分类. 同时由于教师和学生模型输出的特征尺寸并不总是能直接进行匹配，因此要求学生模型在特征映射时学习教师模型的空间上下文结构知识.
序列特征	序列级输出概率分布	文献[24,25]	序列特征任务在数据对齐的情况下都可以使用帧级的知识蒸馏，其主要思想是匹配教师和学生间输出离散化的软目标. 然而在数据没有对齐的情况下，帧级的知识蒸馏实时地使用教师模型输出的每一种可能的序列分布来指导学生模型的训练. 序列级的知识蒸馏则在一个样本中，只将教师模型得分最高的输出序列分布作为学生模型的监督信号，最优的序列分布代表了序列级的知识特性.

4.2 中间特征知识

Gotmare 等人[26]的研究表明：教师的软目标主要是指导学生在深层次的网络层训练，而在学生网络的特征提取层的指导较少. 换句话说，如果网络较深的话，单单学习教师的输出特征知识是不够的. 复杂教师和简单学生模型在中间的隐含层之间存在着显著的容量差异，这导致它们不同的特征表达能力. 教师的中间特征状态知识可以用于解决教师和学生模型在容量之间存在的“代沟”(Gap)问题，其主要思想是从教师中间的网络层中提取特征来充当学生模型中间层输出的提示(Hint). 这一过程称之为中间特征的知识蒸馏，它不仅需要利用教师模型的输出特征知识，还需要使用教师模型隐含层中的特征图知识. 最早使用教师模型中间特征知识的是 FitNets[27]，其主要思想是促使学生的隐含层能预测出与教师隐含层相近的输出. 学生模型学习教师中间隐含层的损失定义为

$$L_{\text{Hint}}\left(W_{Guided},W_r\right)=\frac{1}{2}\left\|u_h\left(x;W_{\text{Hint}}\right)-r\left(u_g\left(x;W_{Guided}\right);W_r\right)\right\|^2 \tag{8}$$

W_{Hint} 是教师前 h 层的权重，W_{Guided} 是学生前 g 层的权重，而 u 为特征的输出. r 是针对师生间的隐含层尺寸不一致而设计的回归函数，期望学生的前 g 层网络经过这个回归函数后，中间隐含层的输出特征能与教师的前 h 层相近. 学生模型学习到教师模型中间特征之后，FitNets 再结合原始的知识蒸馏损失 L_{total} 对整个网络做知识蒸馏. FitNets 的完整训练过程如图 6 所示.

FitNets 使学生模型模仿到教师模型全局的中间特征表达能力，而部分后续工作仅模仿教师模型中的重要特征[28,29]. 其中有代表性的是 Li 等人[29]通过监督学习来筛选出重要特征，其第 j 个通道特征的重要性定义为

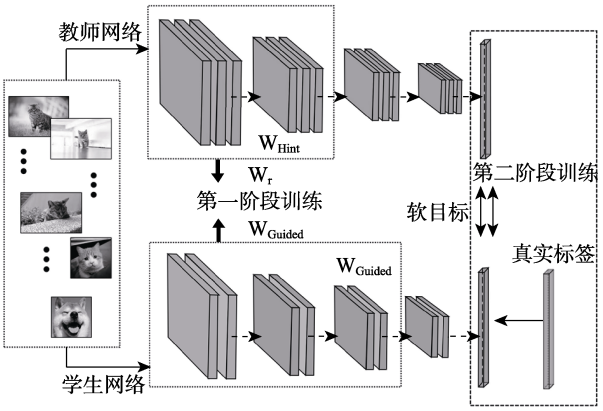

图 6　FitNets 的中间特征知识蒸馏

$$W_j\left(z,w^*,\mathrm{x}_i,y_i\right)= \\ \mathrm{soft\,max}\left(L\left(z\left(\mathrm{x}_i,w^{*\backslash \mathrm{j}}\right),y_i\right)-L\left(z\left(\mathrm{x}_i,w^*\right),y_i\right)\right) \tag{9}$$

其中，$L(z(x_i,\ w^*),\ y_i)$表示数据拟合的经验损失，$L(z(x_i,\ w^{*\backslash j}),\ y_i)$指删除第 j 个通道特征后的经验损失，softmax 函数对结果进行归一化. 通过筛选出更具识别能力的特征，促使学生模型更多关注该特征来获得更高的性能增益.

中间特征的知识蒸馏要求将教师模型的特征提取能力迁移到学生模型中. 在网络层的迁移点上，可以隔层，逐层和逐块地将教师的中间特征知识转移到学生模型中，或者仅迁移教师模型较高的隐含层和最后一个卷积层的特征知识. 在网络层的迁移手段上，可以借助于构造的网络块使学生模型通过模仿学习来获得教师模型的中间层特征，如可学习的投影矩阵和定义的注意力映射图[28]. 不同于将教师模型的中间特征知识迁移或投射到学生模型，一些工作通过共享网络的网络层直接利用教师的中间特征. 不难看出，中间特征的知识蒸馏是要最小化教师与学生之间的中间特征映射距离，这一目标和度量学习的思想很相似. 知识蒸馏中应用最广的度量学习算法是 KL 散度，如用于最小化教师与学生模型输出的相对概率分布[30].

4.3　关系特征知识

关系特征指的是教师模型不同层和不同数据样本之间的关系知识. 关系特征知识蒸馏认为学习的本质不是特征输出的结果，而是层与层之间和样本数据之间的关系. 它的重点是提供一个恒等的关系映射使得学生模型能够更好的学习教师模型的关系知识. 最早的关系特征知识蒸馏可以追溯到 Yim 等人[31]的“Flow of Solution Procedure”(FSP)矩阵，其中通过模仿教师生成的 FSP 矩阵来实施对学生模型训练的指导. FSP 矩阵的知识蒸馏可视化结构如图 7 所示. 其中，F 表示一组特征图，h 和 w 分别表示特征图的高和宽，m 和 n 是选择特征的数量. FSP 矩阵 $\boldsymbol{G}\in\mathbb{R}^{m\times n}$ 的计算公式为

$$\boldsymbol{G}_{i,j}\left(x,W\right)=\sum_{s=1}^{h}\sum_{t=1}^{w}\frac{F_{s,t,i}^{1}\left(x,W\right)\times F_{s,t,j}^{2}\left(x,W\right)}{h\times w}. \tag{10}$$

其中，x 和 W 分别表示输入的图片和权重. $F_{s,t,i}^{1}$ 表示前一组特征中第 i 个特征图，$F_{s,t,j}^{2}$ 表示后一组特征中第 j 个特征图. 然后使用 L_2 范数最小化教师与学生的 FSP 矩阵距离.

$$L_{FSP}\left(W_t,W_s\right)= \\ \frac{1}{N}\sum_{x}\sum_{i=1}^{n}\lambda_i\times\|\boldsymbol{G}_i^T\left(x,W_t\right)-\boldsymbol{G}_i^S\left(x,W_s\right)\|_2^2. \tag{11}$$

其中，N 表示样本数量，λ_i 表示师生间每一个特征对损失函数的权重，而 T 和 S 分别代表教师和学生模型.

特别，Yim 等人[31]的工作分为两阶段训练. 第一阶段最小化师生间的 FSP 矩阵距离，以使学生能学习到教师模型层间的关系知识. 第二阶段是使用正常的分类损失来优化学生模型. FSP 矩阵是测量网络间的关系特征，而后续工作更强调样本的关系知识. 例如，Park 等人[32]提出了基于样本的角度关系和距离关系蒸馏. 其中的角度关系蒸馏用来测量三个样本角度关系.

$$\begin{cases}\psi_{\mathrm{A}}\left(t_i,t_j,t_k\right)=\cos\angle t_it_jt_k=\left\langle \mathrm{e}^{ij},\mathrm{e}^{kj}\right\rangle \\ \mathrm{e}^{ij}=\dfrac{t_i-t_j}{\left\|t_i-t_j\right\|_2},\ \mathrm{e}^{kj}=\dfrac{t_k-t_j}{\left\|t_k-t_j\right\|_2}\end{cases} \tag{12}$$

其中 t_i 表示第 i 个样本的函数. 而距离关系蒸馏用来测量一对样本的距离关系，定义为

$$\begin{cases}\psi_D\left(t_i,t_j\right)=\dfrac{1}{\mu}\left\|t_i-t_j\right\|_2 \\ \mu=\dfrac{1}{\left|\chi^2\right|}\displaystyle\sum_{\left(x_i,x_j\right)\in\chi^2}\left\|t_i-t_j\right\|_2\end{cases} \tag{13}$$

其中，χ^2 表示一对具有区分性样本的集合，μ 是距离的归一化因子，其值为一个训练批次中样本对的平均距离. 通过上述工作[31,32]可知，关系特征知识蒸馏不仅可以利用不同网络层的关系，还可以使用数据样本间的关系. 我们将基于关系的特征知识蒸馏分为三类：基于网络层的关系、基于样本间的关系和相关任务的关系.

基于网络层的关系特征知识蒸馏中，FSP 矩阵是最早的工作. 但是基于 FSP 矩阵的方法要求网络的中间层具有相同大小和数量的过滤器，当师生网络层的维度不同时，该方法并不能用于表示学习[30].

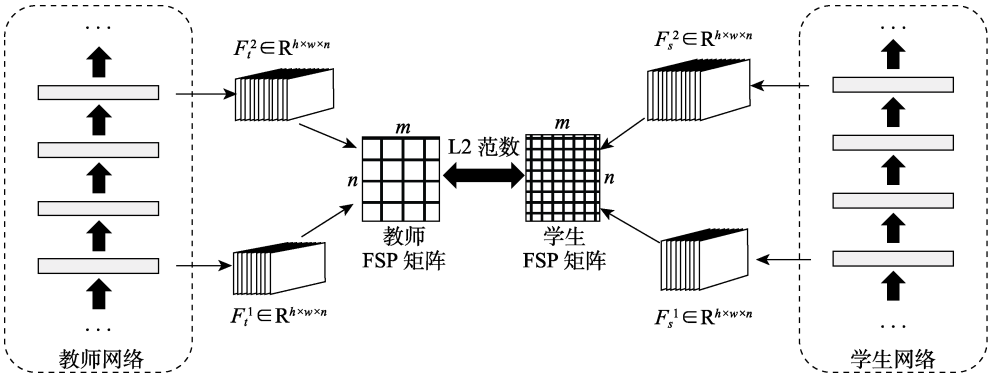

图 7 FSP 矩阵的关系特征知识蒸馏[31]

因此，探索不同架构网络层的内部关系特征成为了必然，如捕获网络层映射相似性的雅可比矩阵[33]和使用径向基函数计算层间的相关性[34]. 这些方法由于不受师生网络结构的限制，特别适用于学生模型的模型压缩.

基于网络层的关系蒸馏仅关注每个样本在网络层间的关系知识，忽略了不同样本之间的关系知识，而该知识也存在于教师模型的空间结构中[35]. 基于样本间的关系特征知识蒸馏是额外利用了不同样本之间的关系知识，即把教师模型捕捉到的数据内部关系迁移到学生模型中. “学习排名”(Learning to Rank)算法[35]是该方法中较早的工作，它将知识蒸馏形式化为师生网络之间样本相似性的排列匹配问题，提出利用不同样本之间的关系，并传递交叉样本的相似性知识来改善学生模型. 除样本间相似性知识之外，关系蒸馏还可以利用相互关系知识[32]和相关性知识[36]. 基于样本的关系知识蒸馏不仅传递了单个样本的信息，而且传输多个样本间的关系知识，使学生模型形成与教师相同的关系. 另外样本的关系知识还能借助于辅助技术，如通过图描述数据内部关系来实现样本关系的知识迁移[37].

在一些特殊任务中，还可以使用相关任务的关系知识. 例如，Bajestani 等人[38]根据人类视觉的原理将教师模型的时间依赖关系知识迁移到学生的目标检测模型中. Deng 等人[39]通过度量长距离跨度视频在外观和几何信息上的关系特征来执行视频目标检测. Wang 等人[40]利用编码器和解码器间的协作关系知识实现风格转移.

4.4 结构特征知识

结构特征知识是教师模型的完整知识体系，不仅包括教师的输出特征知识，中间特征知识和关系特征知识，还包括教师模型的区域特征分布等知识. 网络性能不仅取决于网络的参数或关系，而且还取决于它的体系结构. 结构特征知识蒸馏是以互补的形式利用多种知识来促使学生的预测能包含和教师一样丰富的结构知识. 不同工作构成结构化特征知识的成份是不同的，比如结合样本特征、样本间关系和特征空间变换作为结构化的知识[6]，将成对像素的关系和像素间的整体知识作为结构化知识[41]，还有由输出特征、中间特征和全局预测特征组成的结构化知识[42]. 结构特征知识可以借助于其它的技术来获得，如对比学习[5]和生成对抗网络(Generative Adversarial Networks，GAN)[43]. Tian 等人[5]是利用对比目标函数来捕获结构化特征知识的相关性和高阶输出间的依赖关系来获取教师模型的结构知识，Xu 等人[42]则采用对抗性学习来调整师生网络结构一致的全局预测.

知识蒸馏的首要问题是要明确迁移教师网络中的哪些知识，其知识应当是合适且足够的. 在实际中，教师网络在某个结构上的知识往往不是单一的，它们是有关联的多个知识形式的组合. 例如，中间特征是一种重要的知识形式[27]，其中的各个特征间的关系也是一种重要的知识形式[31]，并且它们都处于同一个网络结构. 这些结构化的特征知识应当被充分利用来提高学生网络和教师模型的全局结构一致性. 同时，结构特征知识包含的多样性知识能提供给学生多个不同视角下的信息，因而是一个更高效的教学范式.

5 知识蒸馏的方法

本节从知识利用的方式，归纳和分析知识蒸馏的主要方法，包括知识合并、多教师学习、教师助理、跨模态蒸馏、相互蒸馏、终身蒸馏以及自蒸馏.

5.1 知识合并

知识合并(Knowledge Amalgamation，KA)是将多个教师或多个任务的知识迁移到单个学生模型

中，从而使其可以同时处理多个任务. 知识合并的重点是学生应该如何将多个教师的知识用于更新单个学生模型参数，并且训练结束的学生模型能处理多个教师模型原先的任务.

目前，一种方法是将多个教师模型的特征知识进行融合，然后将所获得的融合特征作为学生模型学习参数的指导[44,45]. 在融合特征的获得方式上，Shen 等人[44]是将多个教师的特征压缩到紧凑且有区别的特征集，而 Xu 等人[45]是使用辅助模块来提取不同任务所需要的特征. 为了便于理解，这里以 Shen 等人[44]的工作为例显示两阶段知识合并的过程，如图 8 所示.

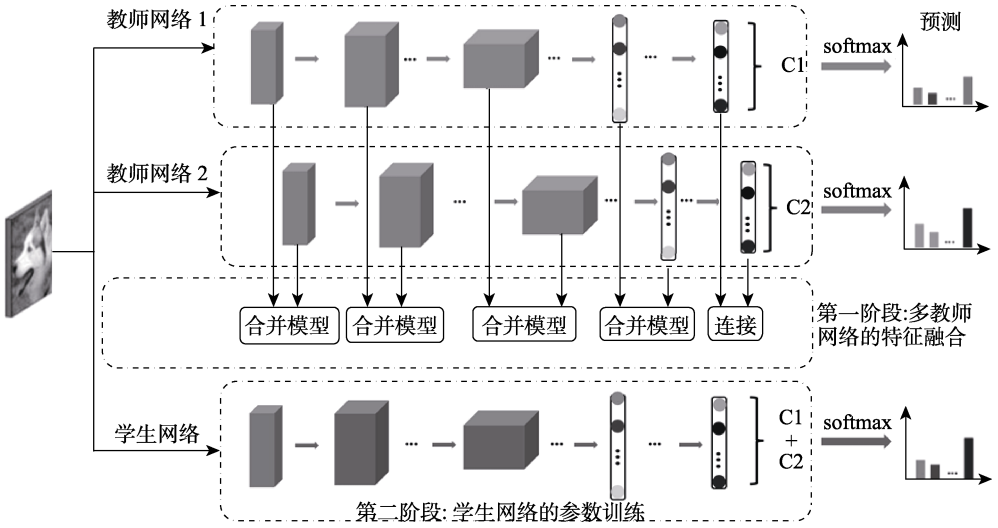

图 8　知识合并示意图[44]

另一种方法是学生模型同时向多个教师模型学习多个任务的特征. 当然，并非所有的教师模型对多任务表示学习都能产生有利的影响. 为了解决这个问题，Shen 等人[46]引入了选择性学习，将多个教师模型中给出置信度最高的预测作为学生模型的学习目标，以降低错误的监督信息对学生模型的误导. 另外，我们可以通过共享网络层直接学习多教师的特征来实现多任务知识的合并[47]，即将教师中的相应层替换为学生中要学习的层，使学生的网络块与相应的教师一起学习. 特别地，Ye 等人[48]通过将学生学习到不同领域的合并知识“投影”(project)到每个教师的专业知识领域，并以计算损失的方式更新学生模型的参数.

5.2　多教师学习

知识合并和多教师学习(Learning from Multiple Teachers)都属于“多教师-单学生”的网络训练结构. 它们的相同点是，知识合并和多教师学习都是学习多个教师模型的知识，但是它们的目标却是不一样的. 知识合并是要促使学生模型能同时处理多个教师模型原先的任务，而多教师学习是提高学生模型在单个任务上的性能. 如图 9 所示，You 等人[49]利用投票策略从多个教师网络中筛选出相对不同的中间特征知识来强化学生模型的性能，其多个教师和单个学生模型的预测任务是相同的.

多个教师模型通过提供多个信息流对学生模型的任务提供了多种解释，学生模型可以利用教师模型对目标任务的“看法”(Views)来提高模型的性能. 因此，多个神经网络预测的集合通常会比单个网络产生更好的性能. 一种简单的方法是仅使用多个教师的软目标来提高单个学生模型的性能，例如随机选择一位教师模型的软目标[50]和通过动态权重选择较高效教师模型的软目标[51]. 动态分配软目标的权重甚至能使学生模型从简单到难地自适应学习多个教师模型中的知识[52]. 但是，单单学习软目标是不够的，充分利用多个教师模型的中间特征知识很有利于学生模型的学习. 目前有投票策略，平均权重和非线性变换等方法能够从多个网络中获得单个网络的中间特征表示.

上述方法均是在教师模型已定的情况下，通过利用多个教师模型的各种知识形式来改善学生模型的预测性能. 然而一些工作特别强调各个教师模型间知识的互补性，即它们自行选择具有互补性知识的教师模型. 例如，Jiang 等人[53]提出同时向能提供稳定信号的长期教师和高质量训练更新的短期教师学习来改善学生模型. 长期的教师信号提供了稳定的教师信息，保证了师生的差异，而短期的教师信

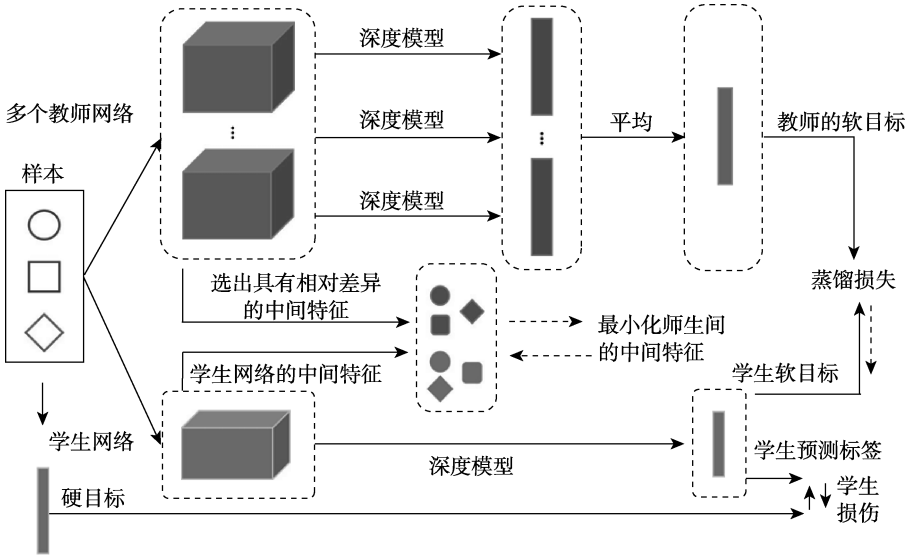

图 9 多教师学习框架[49]

号则保证了高质量的教学. 在实际应用中，向单个教师模型输入具有互补性特征的样本也能充当多教师模型，如将同一个目标的不同方面知识用于指导该目标的学习[54]. 多教师模型除了自行设计之外，还可以直接利用强大相关但不同任务的异构网络，将这些网络应用于目标数据可以生成面向最终预测的高质量特征[55].

5.3 教师助理

教师和学生模型由于容量差异大导致它们存在着“代沟”. “代沟”既可以通过传递教师的特征知识去缓解，也可以使用教师助理(Teacher Assistant)网络去协助学生模型学习. 较早的一个工作是 Wang 等人[56]使用 GAN 的判别器充当教师助理，其工作原理如图 10 所示. 该工作将学生模型当做生成器，判别器促使学生模型对输入数据生成和教师模型同样的特征分布. 随后，Mirzadeh 等人[57]使用一个中等规模的网络(介于教师和学生模型之间)充当教师助理来弥合学生和教师之间的差距，即教师助理先从教师模型中学习到知识后，再传递到学生模型中. 上述工作仅使用了教师模型的软目标,均没利用中间特征. 为了利用异构教师模型的中间特征，Passalis 等人[58]发现教师助理能够允许教师和学生在网络层之间进行直接而有效的一对一匹配来减少“代沟”问题. 另外,Gao 等人[59]让学生模型以及教师助理以互补的方式从教师模型那里获得知识，其中教师助理主要学习教师和学生的残差错误.

5.4 跨模态蒸馏

在许多实际应用中，数据通常以多种模态存在，一些不同模态的数据均是描述同一个事物或事件，我们可以利用同步的模态信息实现跨模态蒸馏(Cross Modal Distillation). 其中有代表性的是 Albanie 等人[60]提出的的跨模态情感识别方法，如图 11 所示. 人在说话时脸部的情感和语音情感是一致的，利用这种同步对齐的模态信息将无标签的视频作为输入数据进行训练，视频中的图片进入预训练的人脸教师模型中产生软目标来指导学生的语音模型训练.

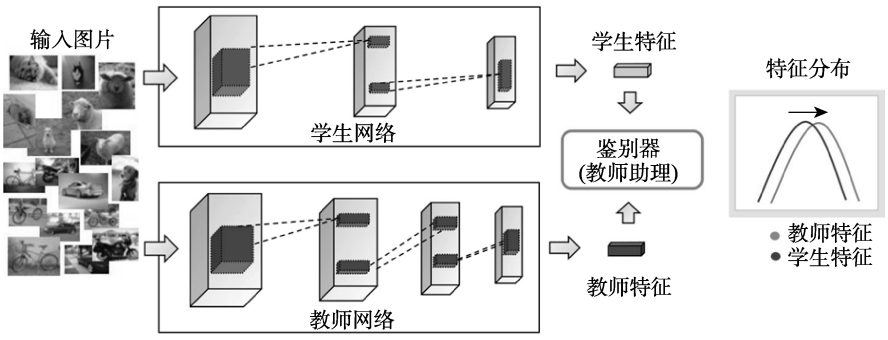

图 10 教师助理学习框架[56]

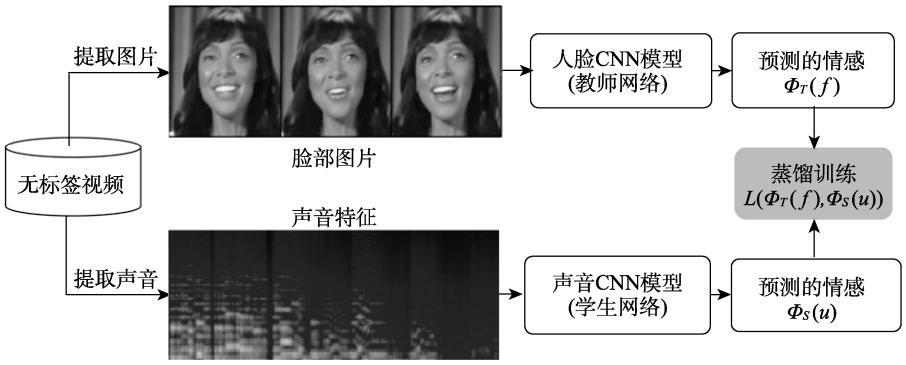

图 11　用于情感识别的跨模态知识蒸馏[60]

跨模态蒸馏还特别适用于由于成本或隐私等原因，无法得到数据多种模态信息的情况. 针对在实际应用中采集不到高分辨率图片，Su 等人[61]使用带标签的高分辨率图像的教师模型来训练识别低分辨率图像. 该工作的主要思想是把教师模型中隐含的“特权信息”传递到学生模型中，如高分辨率图像的颜色信息. 通常，许多带标签的大规模数据集都能在网络上获取，我们可以通过跨模态知识蒸馏将该数据集的信息迁移到不同数据集的学生模型中. 教师同步对齐的模态信息可用于弥补学生网络原本没有的信息流，并通过知识蒸馏的训练来继续增强学生网络的性能，这是知识蒸馏在模型增强中的一个重要的应用方向. 为了便于读者查阅，我们给出了跨模态知识蒸馏用于模型增强的主流工作，如表 2 所示.

5.5　相互蒸馏

相互蒸馏(Mutual Distillation)是让一组未经训练的学生模型同时开始学习，并共同解决任务. 它是一种在线的知识蒸馏，即教师和学生模型是同时训练并更新的. 相互蒸馏的思想由 Zhang 等人[81]于 2017 年提出，其意义在于没有强大教师的情况下，学生模型可以通过相互学习的集成预测来提高性能. 该工作提出的深度互学习(Deep Mutual Learning，DML)模型如图 12 所示.

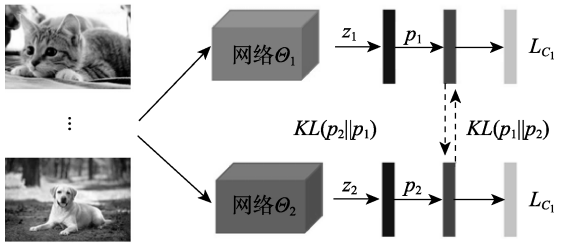

图 12　深度相互蒸馏示意图[81]

DML 方法包含两个网络，输入 M 个类别的 N 个样本，样本集合表示为 $X=\{X_i\}_{i=1}^{N}$，对应的标签表示为 $Y=\{Y_i\}_{i=1}^{N}$，其中 $Y_i=\{1,2,...,M\}$. 样本 X_i 进入神经网络 Θ_1 产生的软目标计算公式为：

$$p_1^m(X_i)=\frac{\exp(z_1^m)}{\sum_{m=1}^{M}\exp(z_1^m)} \tag{14}$$

公式 14 中，z^m 是 softmax 层的输出. 监督损失使用的是交叉熵损失：

$$\begin{aligned} L_{C_1}&=-\sum_{i=1}^{N}\sum_{m=1}^{M}I(y_i,m)\log\left(p_1^m(X_i)\right)\\ I(y_i,m)&=\begin{cases}1 & y_i=m\\ 0 & y_i\neq m\end{cases}\end{aligned} \tag{15}$$

模仿损失使用的是 KL 散度来促使两个分类器的概率分布能够接近，p_1 到 p_2 的 KL 散度为

$$D_{KL}(p_2\,\|\,p_1)=\sum_{i=1}^{N}\sum_{m=1}^{M}p_2^m(X_i)\log\frac{p_2^m(X_i)}{p_1^m(X_i)} \tag{16}$$

在 $D_{KL}(p_2||p_1)$中，p_1 把 p_2 当做数据的真实概率分布，目的是使用 KL 散度使 p_1 的分布近似 p_2 的分布. 从而，神经网络 Θ_1 和 Θ_2 总的损失函数分别定义为

$$\begin{cases}L_{\Theta_1}=L_{C_1}+D_{KL(p_2\|p_1)}\\ L_{\Theta_2}=L_{C_2}+D_{KL(p_1\|p_2)}\end{cases} \tag{17}$$

相互蒸馏避免了对强大教师模型的依赖，同时学生模型能通过相互学习受益. 通过相互蒸馏，各种模型的原始组合能够演变为具有更好性能的新组合[82]. 在实际应用中，相互蒸馏并不一定要使用多个网络进行相互学习，也可以在同一个网络上以相互学习的方式同时训练多个分支点的特征提取器或分类器，各个分支通过补充多样化的信息来生成推理能力更强的模型[83]. 特别，Anil 等人[84]将相互蒸馏应用于分布式训练算法，他们在分布式训练中使用相互蒸馏加快训练速度的同时也提高模型的精度.

表 2 跨模态数据用于模型增强的跨模态知识蒸馏

论文	第一单位与出处	描述
Aytar 等人[62]	麻省理工学院 NIPS	将大规模无标签视频中的图片预测软标签对视频中的声音识别进行指导学习.
Girdhar 等人[63]	卡内基梅隆大学 ICCV	收集和标记大型和干净的静态图像数据集作为教师训练和引导无标签的视频学习丰富的知识表示.
Liu 等人[64]	中国科学院 ACM Multimedia	将现有弱监督检测模型中的局部语义区域和类相关性蒸馏到无标注的多标签图像分类任务中.
Gupta 等人[65]	加利福尼亚大学 CVPR	将从标注过大样本的 RGB 图像模态中学习得到的特征作为监督信号，用于无标注样本深度和光流图像模态的特征学习.
Wang 等人[66]	卡内基梅隆大学 ICCV	使用知识蒸馏纠正 2D 地标注释重建 3D 姿势估计所犯的一些错误，从而避免对 3D 标签的依赖.
Luo 等人[67]	斯坦福大学 ECCV	使用图蒸馏动态学习蒸馏方向和权重来更好利用大规模多模态数据集的“特权信息”.
Liu 等人[68]	约翰霍普金斯大学 ICCV	将 ImageNet 的语义信息迁移到草图检索任务中.
Hoffman 等人[69]	加州大学伯克利分校 CVPR	提出了一个幻觉网络来学习深度图片的特权信息来提高 RGB 目标检测的性能，它通过在测试时学习教师模型的中间层特征来弥补丢失的信息流.
Aditya 等人[70]	Adobe 公司 WACV	将基于问题和场景图的附加信息以空间知识的形式用于视觉推理问题中.
Ye 等人[71]	南京大学 CVPR	通过构造实例与实例之间的关系将不同领域教师模型的知识迁移到新领域的学生模型中.
Yuan 等人[72]	北京大学 IEEE Trans. Multim.	将图像语义理解的知识迁移到文本到图像合成的任务中.
Li 等人[73]	香港中文大学 AAAI	将磁共振成像的知识以相互学习的方式迁移到计算机断层扫描的分割任务中.
Piao 等人[74]	大连理工大学 CVPR	将深度的“特权信息”迁移到 RGB 流中.
Garcia 等人[75]	意大利技术研究院 ECCV	通过时空表示的乘法连接，利用软目标和硬目标以及特征图之间的距离来学习视频中深度特征的“特权信息”.
Tavakolian 等人[76]	奥卢大学 ICCV	通过帧自适应加权将视频中每一帧的时空信息蒸馏到图像中.
Gu 等人[77]	中国科学院 ICCV	通过强制匹配共享特征空间的方式将视频表示网络所学习的时间知识传播到图像表示网络中.
Zhao 等人[78]	罗格斯大学 CVPR	将从包含两个模态的配对样本的源数据集中学习提炼的交叉模式知识推广到目标数据集.
Hu 等人[79]	合肥工业大学 CVPR	通过利用无监督教师模型输出中获得的广泛的相关性信息来监督学生模型的训练.
Li 等人[80]	澳大利亚国立大学 CVPR	传递字幕新闻符号的知识以提高单词级手语识别性能.

5.6 终身蒸馏

深度学习网络在学习新任务时，对旧任务的性能就会急剧下降，这个现象被称为灾难性遗忘[85]. 这就需要使用终身学习来减轻这种影响，终身学习也称为持续学习或增量学习.

目前，有些工作使用知识蒸馏方法来实现终身学习，称之为终身蒸馏(Lifelong Distillation). 终身蒸馏就是通过知识蒸馏来保持旧任务和适应新任务的性能，其重点是训练新数据时如何保持旧任务的性能来减轻灾难性遗忘. 知识蒸馏能够较好地解决这一问题，即能最大程度地减少新旧网络对旧类响应之间的差异[85]. 这方面最早的工作可以追溯到 Li 等人[85]通过引入知识蒸馏来保持旧任务性能的“学习而不遗忘”(Learning without Forgetting)算法. 如图 13 所示，θ_s 为网络的共享参数，θ_o 为旧任务参数，目的是增加一个新任务参数 θ_n，并让 θ_n 在新旧任务

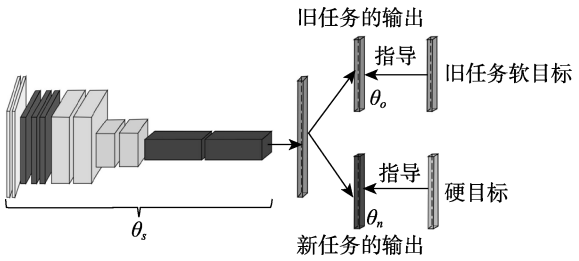

图 13 “学习而不遗忘”算法示意图

上都能获得高性能. 另外，通过缓存一小部分旧任务数据[86]和产生旧任务相似的输出值或视觉模式[87]都能使网络在学习新任务的同时保持旧任务性能.

5.7 自蒸馏

自蒸馏(Self-Distillation)是单个网络被同时用作教师和学生模型，让单个网络模型在自我学习的过程中通过知识蒸馏去提升性能. 它也是一种在线的知识蒸馏. 自蒸馏在作用机制的理论分析上，

Mobahi 等人[88]认为自蒸馏是通过逐渐减少代表解的基函数数量来不断修改正则化. Zhang 等人[89]认为自蒸馏先前的迭代都能为后续的迭代充当教师作用，从而通过多次的迭代之后能学习到多样性的知识.

自蒸馏主要分为两类，如图 14 所示. 第一类是使用不同样本信息进行相互蒸馏. 其它样本的软标签可以避免网络过度自信的预测，甚至能通过最小化不同样本间的预测分布来减少类内距离[90]. 另外一些工作使用增强样本的信息，如利用数据在不同失真状态下的特征一致性来促进类内鲁棒性学习[91]. 另一类是单个网络的网络层间进行自蒸馏. 最通常的做法是使用深层网络的特征去指导浅层网络的学习[92]，其中深层网络的特征包括了网络输出的软目标. 在序列特征的任务中，则是将先前帧中的知识传递给后续帧进行学习[93]. 单个网络的各个网络块学习也可以是双向的，每一个块间可以进行协作学习，并在整个训练过程中互相指导学习.

5.8　小　节

本节介绍了 7 种知识蒸馏的方法，它们的优缺点如表 3 所示.

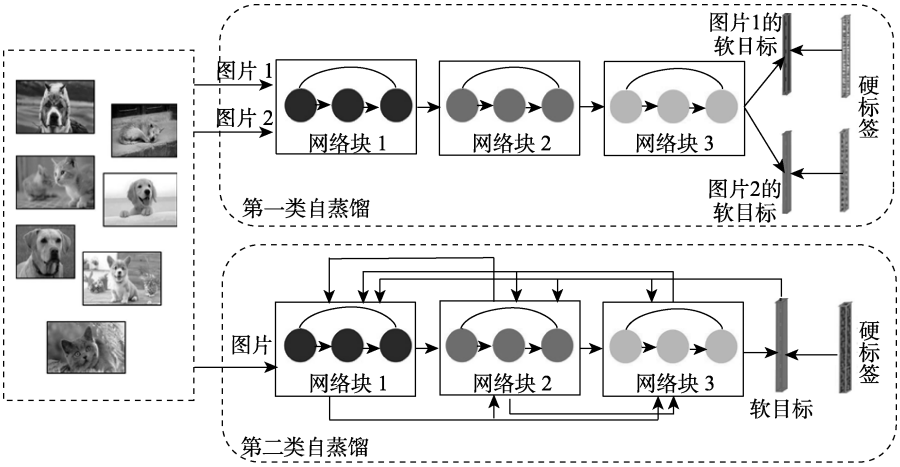

图 14　自蒸馏的示意图

表 3　各种知识蒸馏方法的优缺点

知识蒸馏的方法	优点	缺点
知识合并	多用途的学生网络能减少部署的成本和提高模型的利用率.	相较于单网络预测单个目标任务，多用途学生网络的性能可能会下降.
多教师学习	相比于单个老师，多个老师通常能让学生网络学习到更好的、更丰富的知识.	难以高效地融合各个教师网络的知识.
教师助理	减少师生网络存在的“代沟”，有助于学生网络更好地训练.	通常会增加训练的成本.
跨模态蒸馏	通过跨模态的数据集来实现少样本学习或半监督学习，减少对带标签数据的依赖.	获取同步的跨模态数据并不是一个很容易的任务.
相互蒸馏	节省训练教师网络的时间，提高了训练的效率.	可能会陷入“瞎子带领瞎子”的局面.
终身蒸馏	能保护旧任务的数据隐私，并提高训练的效率.	难以维持旧任务的性能.
自蒸馏	节省训练教师网络的时间，提高了训练的效率.	缺少丰富的外部知识.

6　知识蒸馏与其它技术的融合

近几年，研究人员发现知识蒸馏结合其它主流技术通常能够提高其性能. 目前这些主流技术主要有：生成对抗网络、神经架构搜索、强化学习、图卷积、其它压缩技术、自动编码器、集成学习以及联邦学习.

6.1　生成对抗网络

知识蒸馏和 GAN 的结合就是要在知识蒸馏的学习中引入对抗性学习策略. GAN 是通过对抗性学习来生成新图像，即学习生成无法被判别器网络区分的图像. 它的主要结构包括一个生成器和一个判别器，生成器用来生成尽可能接近真实数据的样本来使判别器无法辨别. 在知识蒸馏中，GAN 判别器的主要任务是区分不同网络的特征映射分布或数据分布，而生成器的主要任务是生成给定实例的相关特征映射分布或数据分布. 生成器一般是学生模型[94]或师生模型共同担任[95].

具体来说,基于 GAN 的知识蒸馏使用判别器来实现教师和学生模型的知识趋于一致，从而使判别器无法辨别知识是来自教师还是学生模型. Haidar 等人[94]将教师模型输出的特征视为真实标签，学生模型充当生成器，在对抗性学习的推动下，生成器的输出近似于真实标签而使得判别器无法辨别. 教师和学生模型的中间特征可以分别作为真样本和假样本，GAN 的判别器对多个网络进行对抗性训练来调整它们的中间特征表示趋向于相同[96].

上述方法均是将学生模型充当生成器，目的是通过对抗性学习策略使学生模型的预测输出或中间特征能够近似于教师模型. 在实际应用中，教师和学生模型能共同担任生成器去欺骗判别器. 例如，它们在知识蒸馏的过程中生成伪标签去欺骗拥有真实标签的辨别器[95]，即促使学生模型能够学习到教师模型所生成的伪标签. 另外，GAN 可以通过传递样本的知识来提高知识蒸馏的性能，如决策边界相关知识[97].

6.2 神经架构搜索

NAS 在给定的搜索空间下，使用某种策略搜索出最优网络结构. 与普通的 NAS 相比，基于知识蒸馏的 NAS 产生的软目标包含着更多的信息. 通过利用软目标的附加信息，能够加快 NAS 搜索网络结构的速度. 例如，Bashivan 等人[98]通过评估候选网络与已知高性能教师模型内部激活特征的相似性来提高 NAS 搜索网络架构的速度. Li 等人[99]从教师模型中提取神经结构的分块知识作为分块结构搜索的指导来提高搜索速度.

知识蒸馏产生的附加信息也能指导 NAS 搜索到最佳的学生模型框架，即提高学生模型的性能. 网络结构是一种很重要的知识形式. Liu 等人[100]的研究表明，任何教师模型结构都有一个最强的学生模型. 因此，他们在给定的教师模型前提下，使用 NAS 找到最佳的学生模型[100]. 在实际的应用中，神经网络的运行速度性能不仅取决于计算量等网络本身的消耗，还需要根据具体的部署环境[101]. 因此一些工作在设定的资源条件约束和知识蒸馏的指导下，通过动态调整的 NAS 去挑选出最好的学生模型[102].

6.3 强化学习

强化学习(Reinforcement Learning)的目的是使智能体根据环境状态、行动和奖励，学习出最佳策略. 强化学习的环境在已设置完毕的情况下，使用知识蒸馏传递的知识能产生性能更好的学生智能体. 基于强化学习的知识蒸馏主要有两种目的，第一种是用来加强深度学习网络模型的策略，其主要思想是通过知识蒸馏结合一个或多个教师的策略. 教师和学生模型都需要在环境的约束下，最大程度地获得学生模型的回报. 从而，学生模型能从多个教师智能体中获得更高性能的策略[103]，不同策略的学生模型也能在相互蒸馏中继续强化策略[104].

第二种目的是用来获取更轻量级的网络模型. 强化学习和知识蒸馏的结合能将强化学习网络模型中的策略知识迁移到轻量级的单个网络中[105]，或利用教师的策略知识来逐步减少学生模型中的冗余[106].

6.4 图卷积

图卷积(Graph Convolutional Network，GCN)是以一组节点及节点间的关系为特征的卷积网络，由于其具有强大的建模能力而受到广泛的应用. 图卷积的网络模型具有拓扑结构，知识蒸馏可以将图卷积教师模型的拓扑结构知识传递到学生模型中[107]. 同时，图卷积也能促进知识迁移. 知识蒸馏主要利用了图卷积强大的建模能力，将图卷积捕捉到教师模型某些领域的知识迁移到学生模型中，如空间的几何形状[108]，目标在空间和时间上的相互作用[109]和多教师间的互补性知识[110]. 另外，Zhang 等人[111]通过图卷积强调数据的可靠性，使学生模型专注于对可靠知识的学习.

6.5 其它压缩技术

网络参数在正常的情况下为 32 位的浮点数，也称为全精度网络. 网络量化是使用更低的精度去代替全精度的网络参数，例如使用 2 位或 8 位的整数来替代浮点数. 但是，网络量化压缩之后的准确率通常会稍微下降. 基于量化压缩的知识蒸馏方法是通过知识蒸馏来处理因量化压缩而导致的网络准确率下降问题. 通常，全精度的教师模型通过知识蒸馏来指导低精度的学生模型训练[112].

除了量化，其它较为常用的压缩技术如剪枝. 剪枝是把网络中“不重要”的连接删除，然后微调稀疏网络以恢复准确性，其要点是减少剪枝带来的性能误差. 在现有的工作中，Bai 等人[113]通过教师和学生的交叉蒸馏来减少逐层剪枝带来的误差. Li 等人[114]则在每个要剪枝的网络块后添加一个 1×1 卷积层，并通过知识蒸馏使学生的参数与教师对齐.

在其它的方法中，学生模型也可以是卷积核低秩分解后的网络[115]. 全精度教师和低精度学生模型存在的“代沟”问题，除了上述方法之外，还可以通过它们间的预先共同训练来解决.

6.6 自动编码器

自动编码器(Autoencoder, AE)是一种无监督的神经网络模型. 由于它在压缩比和迁移学习的性能较好，已被广泛应用在降维和生成模型的训练. 自动编码器通过特征的重构，从数据样本中自动学习隐含的特征，它的这一特性可协助知识蒸馏提高学生网络的性能. 例如，Li 等人[116]首先将自动编码器用于自动学习教师和学生网络精确的特征表示，然后通过知识蒸馏迁移教师的特征来提高学生网络的泛化能力. 在知识蒸馏的训练过程中，可以使用自动编码器实时地判断学生网络是否已经学习到丰富的特征知识. 除了学习特征，自动编码器也能自动地调节教师和学生网络结构上的差异，从而使知识更容易地迁移[117].

同样，知识蒸馏作为一种辅助性的技术，能帮助自动编码器学习到更具有鲁棒性的特征表示. 例如，Shen 等人[44]通过自编码器压缩多个教师的中间特征到单个网络中，然后通过知识蒸馏将所获得的融合特征作为学生网络参数的指导. Pan[118]等人使用知识蒸馏动态地微调自动编码器生成的软目标来减少预测特征中噪声. Kim 等人[119]将教师网络的输出信息进行编码，学生网络通过模仿教师的信息进行解码来获得更多的抽象特征.

6.7 集成学习

集成学习(Ensemble Learning)的核心思想是“三个臭皮匠，顶一个诸葛亮”. 它使用多个网络对同一个任务进行处理，其性能通常会比单个网络好. 集成学习和知识蒸馏相融合的一个重要应用方向是：使一个简单模型的性能能够和多个集成的网络相媲美，即本文在 5.2 节介绍的“多教师学习”. 此外，知识蒸馏也能用于增强具有多个学生网络组成的集成模型. 它主要是使多个学生网络通过同伴教学的方式，从头开始进行相互指导学习，最后融合成推理性能更强的集成网络. 一个网络中的每个分支结构都可以充当一个学生模型，从而使得原始的网络模型能够通过相互蒸馏，演变为性能更好的新网络[82]. 一个集成的网络也可以通过知识蒸馏压缩为多个不同架构的学生模型，每个学生模型学习独特的知识，训练结束后的各个分支学生网络能重新组合成性能更强的集成网络[120].

在不考虑模型复杂度的情况下，每个学生都可以是独立的网络结构. 集成模型融合了各个学生的多样性信息，并能通过知识蒸馏继续强化其性能. 例如，Guo 等人[121]使知识在各个不同能力的学生模型之间进行转移，并有效地集成多个学生输出的软目标作为各个学生网络的监督信息. Zaras 等人[122]通过知识蒸馏，使每一个参与训练的学生网络都在某个类别数据上具有高度专业化的知识，然后使用集成学习，汇集各个学生网络的多样性知识来加强任务处理的性能. Garcia 等人[123]将不同模态数据的网络在合作环境中从零开始学习，并相互加强，最后融合各个专家网络的预测来处理多模态数据.

6.8 联邦学习

联邦学习(Federated Learning)[124]是一种保障数据隐私的模型训练方法，能够打破传统企业等机构的数据边界，在医疗、金融等行业中具有广泛的应用前景. 联邦学习支持数据在本地训练，训练后局部模型的参数通过加密机制被发送到服务器. 服务器聚合所有参与者的局部模型，然后将更新好的全局模型通过加密机制再发送给每个参与者进行下一次训练，直到全局损失函数收敛或达到所需的训练精度为止. 然而，各个参与者的局部模型与服务器的全局模型之间需要频繁交换训练信息将导致通信开销剧增. 如果参与者的带宽有限或通信成本高昂，会导致联邦学习效率低下，甚至无法进行.

知识蒸馏可用于减少分布式的联邦学习训练时所占用的带宽，它在联邦学习的各个阶段，通过减少部分参数或者样本的传输的方式实现成本压缩. 一些工作通过只传输模型的预测信息而不是整个模型参数，各个参与者利用知识蒸馏，学习服务器聚合全局模型的预测信息来提高性能. 知识蒸馏和联邦学习通过只传输各个局部模型和全局模型的预测值，可以减少网络带宽、允许模型异构和保护数据的隐私. 例如，Sui 等人[125]从多个参与者的预测信息中学习一个紧凑的全局模型. Sattler 等人[126]提出量化、无损编码和双蒸馏的方法来进一步地节约传输软目标所占用的带宽. 另外，各个参与者的数据可以通过数据蒸馏的方式，压缩为少量的伪样本，传输并聚合这些伪样本来训练全局模型也能减少带宽的使用[127]. 不同于只传输网络的预测，另有工作关注于利用知识蒸馏高效地融合各个参与者的异构局部模型知识. 例如，Shen 等人[128]使用相互蒸馏，在联邦学习本地更新的过程中训练不同架构的模型. Lin 等人[129]利用未标记的数据或伪样本聚合所有异构参与者的模型知识.

6.9 小 节

知识蒸馏与上述的主流技术融合的优缺点如表 4 所示. 另外，还有其它非主流的技术方法能提高知

识蒸馏效率和性能. 例如，迁移学习和知识蒸馏的结合可以快速初始化一个学生网络，使蒸馏学习的过程并非是从头开始[31]. 注意力机制和知识蒸馏的结合能让学生获取到教师网络最重要的知识[130].

表 4 知识蒸馏与其它主流技术融合的优缺点

知识蒸馏的方法	优点	缺点
GAN	使学生网络能够更好学习到教师模型的特征分布.	训练的成本高，特别是当师生能力差异很大的时候，难以训练.
NAS	提高网络搜索的速度，质量和模型的可扩展性.	训练的成本高.
强化学习	提高模型的可扩展性.	需要设计高效的奖励函数.
图卷积	z 能利用图丰富的特征表示能力.	图表示仅限于特定类型的数据.
其它压缩技术	可以获得更加轻量的高效网络模型.	学生网络的性能可能还是会下降.
自动编码器	自动学习隐含的特征表示.	训练的成本高.
集成学习	“三个臭皮匠，顶一个诸葛亮”.	多个网络可能会互相干扰.
联邦学习	减少通信的成本和支持模型异构.	各参与者输出的预测信息会泄漏隐私.

7 知识蒸馏的应用进展

知识蒸馏的最初目的是压缩深度学习网络模型，这在资源受限的终端设备上具有广泛的应用. 但随着研究的新进展，知识蒸馏不仅可以用于压缩模型，还可以通过神经网络的互学习、自学习等优化策略和无标签、跨模态等数据资源对模型的性能增强也具有显著的提升效果. 目前知识蒸馏的主要应用领域有计算机视觉、自然语言处理、语音识别、推荐系统、信息安全、多模态数据和金融证券. 知识蒸馏在计算机视觉、自然语言处理、语音识别和推荐系统上的应用根据其目的的不同，可以分为模型压缩和模型增强. 模型压缩是为了获得简单而高效的网络模型，以方便部署于资源受限的设备. 而模型增强通常是利用其它资源(如无标签或跨模态的数据)来获取复杂的高性能网络.

7.1 计算机视觉

7.1.1 计算机视觉的模型压缩

计算机视觉领域有目标分类，目标检测和目标分割这三个主要的任务. 根据任务的研究热度，本小节将详细分析深度估计、密集预测(Dense Prediction)和低分辨率网络.

深度估计是估计图像中每个像素相对于拍摄源的距离，它在 3D 重构、自动驾驶和姿势估计等问题上具有重要的作用. 单目深度估计是仅通过一张或者一个视角下的 RGB 图像来评估深度，由于成本低和设备便利等优点而受到广泛的关注. 虽然单个 RGB 图像可能会有无数个真实的场景，但是人类却能根据自身丰富的经验来推断出单个图片的场景深度. 根据这一基本事实，可以将丰富经验的教师模型的知识迁移到单个 RGB 图像的场景深度估计中. 基于知识蒸馏的单目深度估计的核心思想是通过利用强大教师模型的知识重构出单个 RGB 图像的深度[48].

密集预测是预测出图像中每个像素点标签的任务统称，如语义分割和目标检测. 基于知识蒸馏的密集预测是将每个像素充当软标签并对其单独进行知识蒸馏，但是由于需要空间上下文的结构语义，密集预测通常被认为是结构化预测的问题[22,23]. 样本内部的特征关系可以作为结构特征知识，如 Hou 等人[22]将教师中学习到的场景结构迁移给学生执行道路标记分割任务，其传递的是一个样本中不同区域之间的结构关系. 多种蒸馏的知识形式能充当结构特征知识，例如师生网络间的中间特征，关系特征和输出特征的结构化知识能应用于目标分割[23].

高分辨率的大尺寸图片包含着目标更详细的信息，因而可以使深度学习网络获得更好的性能. 但是高分辨率图片在深度学习网络中的计算量和内存量是巨大的，导致在资源受限的边缘计算设备上运行困难. 同时，在高分辨率图片上训练的网络模型无法用于预测低分辨率的图片[131]. 因此，使用低分辨率的小尺寸图片作为输入进行训练出来的网络模型在多种场合下都有使用的价值. 借助于知识蒸馏的方法，可以将高分辨率的复杂教师模型所学习到的知识迁移到低分辨率的高效学生模型中. 例如，Fu 等人[132]将教师模型所学习的空间和时间知识迁移到低分辨率的轻量级时空网络中来执行视频注意预测任务.

7.1.2 计算机视觉的模型增强

知识蒸馏在计算机视觉上的应用除了可以获取到高效的网络模型外，还可以通过加入其它知识来提高某个复杂网络的性能. 例如，Wang 等人[133]提出从

动作识别中提取知识用于早期动作预测，这是通过完整视频的知识来提升部分视频的学习. Yu 等人[134]通过教师模型提供的内部(数据注释)和外部(互联网上的公共文本)语言知识来指导学生模型的视觉关系识别. 另外有一些工作关注于利用任务的相关性来实现模型的增强. 比如，Dou 等人[135]利用共享的跨模态信息来同时提高计算机断层扫描和磁共振成像的图像分割的性能. Orbes-Arteainst 等人[136]使用知识蒸馏去减少不同医学数据间的域差距来提高医学图形分割的性能.

7.2　自然语言处理

7.2.1　自然语言处理的模型压缩

高效的自然语言处理(Natural Language Processing, NLP)网络模型主要包括高效的预训练语言模型和压缩的 NLP 任务. 预训练语言模型是通过大规模文本语料库训练而来. 由于预训练的语言模型包含丰富的句法和语义等信息，只需要使用少量的数据集对特定任务进行微调就能获得高性能的 NLP 模型. 因此在 NLP 领域，预训练语言模型具有重要的地位并取得了一些显著的进展，如 BERT[137]. 但是这些预训练语言模型是计算密集的而难以应用于资源受限的环境，因此需要对其进行压缩以满足具体任务的要求. BERT 从提出开始就在 11 项 NLP 基准任务上取得了最高的性能，从而使得 BERT 成为热门的研究方向. 表 5 中归纳了基于知识蒸馏的方法来加速 BERT 模型的主流工作，PD[140]的对比模型为 $\text{BERT}_{\text{LARGE}}$，其余对比模型为 $\text{BERT}_{\text{BASE}}$. 表中的“-”表示作者在原文中没有显示的，他们的一些工作要么只显示模型大小，要么只显示参数量. 然而参数量和模型大小并不直接等价[101]，即模型大小并不是只存储参数量.

高效的 NLP 网络既可以通过高效预训练语言模型的迁移学习获得，也可以通过压缩高性能的复杂 NLP 教师模型来获得. 这里压缩 NLP 任务只介绍关于知识蒸馏方法的，即通过模仿高性能复杂 NLP 教师模型的知识来获得高效的 NLP 学生模型. Yang 等人[147]结合多个教师的知识来执行问答匹配问题. Arora 等人[148]提出模仿教师的特性和匹配输出来减少问答任务的计算和内存需求. 另外，BERT 还可以充当教师模型来指导轻量级语言模型的训练，如 mBERT[149]直接用于获取轻量级的多语言序列标记模型[150]和多语言命名实体识别模型[151].

7.2.2　自然语言处理的模型增强

与计算机视觉上的应用类似，可以使用知识蒸馏的方法来提高某个具体 NLP 任务的性能. 表 6 给出了这部分工作所使用的主流方法和类别.

7.3　语音识别

7.3.1　语音识别的模型压缩

语音识别主要是依靠 RNN(Recurrent Neural

表 5　基于知识蒸馏的高效 BERT

模型	描述	GLUE 得分	速度	模型大小	参数量
MINILM[130]	MINILM 深入模仿 BERT 模型的最后一个 Transformer 层的自我注意模块来加速.	–1.1	2. 0×	–	66M
MiniBERT[138]	MiniBERT 使用无标签的数据从头开始蒸馏一个预训练的多语言模型.	–2.4	27×	0.167×	–
FastBERT[139]	FastBERT 自适应调整每个样本的计算量，使较容易的样本尽早结束来减少推理的时间.	–	1–12×	–	–
PD[140]	通过利用预训练和特定任务蒸馏之间的相互作用来缩小 BERT.	–3.3	1.25×	–	110.1M
DistilBERT[141]	DistilBERT 在没有下一个句子的预测目标时，使用动态掩码方法在大批次中利用梯度累积进行蒸馏.	–2.5	1.6×	0.4×	66M
Zhao 等人[142]	通过共享变量投影将教师的知识逐层迁移到词汇量较小，嵌入和隐藏状态维度较低的学生模型.	–13	76.9×	0.016×	1.7M
MobileBERT [143]	为了利用更细粒度的知识，MobileBERT 假设师生网络的层数相同，并引入瓶颈模块以保持其隐藏层的大小相同.	+0.2	4.0×	0.233×	25.3M
Patient-KD [144]	Patient-KD 采用增量提取过程，同时学习了 BERT 的中间特征和预测输出.	–2.24	1.94×	–	67M
TinyBERT[145]	为了利用更细粒度的知识，TinyBERT 采用统一函数来确定师生网络层之间的映射，同时使用参数矩阵对学生模型网络层的隐藏状态进行线性变换.	–3	9.4×	0.133×	66M
BERT-of-Theseus[146]	BERT-of-Theseus 逐步将 BERT 中的原始模块替换为参数更少的替代模块.	–1.4	1.94×	–	66M

Network)的序列化建模，尤其是 LSTM(Long Short Term Memory)[156]网络. 然而 LSTM 是计算量密集型的，导致在低资源的环境下很难使用. 因此，一个重要的研究方向是使用知识蒸馏压缩 LSTM 网络. 表 7 给出了这部分工作所使用的主流方法和类别.

RNN 的计算是顺序的，即 RNN 的相关算法只能从从左到右或从右到左的单向运算，这限制了模型的并行能力和难以缓解特别长期的依赖问题[161]. Transformer[161]提出使用注意力机制在长时序列建模上的优势来解决 RNN 顺序计算的上述两个问题. 由于预测的精度高等优点，Transformer 已成为最热门的研究方向之一，如 BERT[137]. 然而，Transformer 的参数量大会增加内存的消耗并减慢了推理的过程. 为方便 Transformer 的部署，Aguilar 等人[162]把教师的多个 Transformer 层的特征知识压缩到学生的单个 Transformer 层中. Gaido 等人[163]通过引入自适应层来压缩 Transformer 结构.

除了将大规模语音识别网络的知识迁移到小型语音识别网络中之外，基于跨模态的知识蒸馏能将其它领域的知识用于协助小型语音识别网络的学习. 在利用语音教师模型知识的方式上，结合其它的技术方法能提高学生模型的性能，如共同利用知识蒸馏和量化来压缩声学事件检测模型[164].

7.3.2 语音识别的模型增强

同样，可以使用知识蒸馏的方法来提高某个具体语音识别任务的性能. 表 8 给出了这部分工作所使用的主流方法和类别.

7.4 推荐系统

7.4.1 推荐系统的模型压缩

推荐系统是解决因信息量庞大而无法准确和高效地获取到用户感兴趣物品(Item)的最有效工具之一[181]. 性能越好的推荐系统往往需要更多的参数量，并且其参数数量通常是规模庞大的[182-185]. 这一缺陷导致了高性能的推荐系统通常无法实时地在线部署于终端设备上. 知识蒸馏是解决这个问题的一种高效方法，其将离线训练的功能强大的教师模型中的知识迁移到在线推荐的简单学生模型中. 推荐系统通常需要推荐用户最感兴趣的一些物品，一个简单的方法是将复杂教师模型输出的软目标知识迁移到简单学生模型中[182,183]. 其中，Tang 等人[182]仅传递教师模型软标签中排名最高的几个物品，而 Lee 等人[183]为前几个重要物品分配高权重并同时传递了排名低物品的相关信息. 除了可以利用教师模型的各种知识形式，还能结合知识蒸馏的方法和融合的技术去获得高效的推荐系统模型. 例如，Zhang 等人[184]将基于路径的可区别推荐模型的结构化知识通过相互蒸馏迁移到基于嵌入的简单推荐系统模型中. Wang 等人[185]将图卷积推荐系统中关于排名信息的知识迁移到二值化的推荐网络中来提高在线推荐的效率.

表 6 基于模型增强的 NLP 例子

模型增强的类型	方法的描述	具体相关工作的描述
直接迁移知识	直接地将复杂教师模型的知识迁移到学生模型中.	Hu 等人[152]从多个教师模型中获取知识来提高阅读理解任务的性能. Chen 等人[153]利用 BERT 的双向特性指导 Seq2Seq 模型的语言生成.
间接迁移知识	通过提高数据集的质量或数量来提高模型增强的效果.	Wu 等人[154]利用标签丰富的语言来帮助其他标签贫乏的语言进行训练. Feng 等人[155]通过相互蒸馏学习来避免对话匹配任务中大规模人工标注和数据中的噪音.

表 7 知识蒸馏用于压缩 LSTM 语音识别网络的例子

压缩的类型	方法的描述	具体相关工作的描述
简单压缩	在不需要数据对齐的蒸馏任务中,直接执行知识的迁移.	Geras 等人[157]将 LSTM 网络中的知识迁移到卷积神经网络(Convolutional Neural Networks，CNN)的语音识别模型中. Lu 等人[158]将大规模 LSTM 教师模型的软标签来指导小型 LSTM 的语音识别网络的学习.
复杂压缩	在需要数据对齐的蒸馏任务中，需要采用序列级的知识蒸馏或者将数据对齐之后再执行帧级的知识蒸馏.	端到端的自动语音识别框架是不需要数据对齐的，如 CTC(Connectionist Temporal Classification)[159]模型. 然而在知识蒸馏过程中，师生网络的序列数据需要对齐. 对于无对齐的序列数据，一种方案是将数据对齐之后再执行帧级的 CTC 声学模型压缩[160]. 另一种方案是序列级的 CTC 模型压缩[24,25]，即学生模型的后验概率仍在序列水平上进行优化，其核心是获取最优教师模型的输出分布.

表 8　基于模型增强的语音识别例子

论文	描述	蒸馏的知识形式/方法
Markov 等人[165]	使用干净的语音来训练教师，学生模型使用教师相应干净的语音信息来从嘈杂语音中学习鲁棒性的语音识别.	跨模态蒸馏
Joy 等人[166]	利用标准化的语音教师的特权信息来指导学生模型非标准化语音学习训练.	跨模态蒸馏
Abraham 等人[167]	从复杂的拥有丰富数据的高资源模型中提取知识来改进数据量少的低资源声学模型.	跨模态蒸馏
Shen 等人[168]	将基于长话语的教师模型的特征表示知识转移到基于短话语的学生模型中.	跨模态蒸馏
Das 等人[169]	将知识从训练有素的多语种 DNN 迁移到使用转录训练的目标语言 DNN 中.	跨模态蒸馏
Subramanian 等人[170]	通过模仿多通道输入的软掩码来获得单通道输入的学生模型.	跨模态蒸馏
Heo 等人[171]	通过模仿教师模型的软标签来学习不同声学场景之间的相似性.	输出特征知识
Kurata 等人[172]	融合多个不同结构 CTC 模型的知识来提高语音识别的准确率.	多教师学习
Li 等人[173]	将丰富的音频数据用于学习视听语音识别.	跨模态蒸馏
Zhang 等人[174]	使用单个语音识别的网络产生的软目标来协助多个说话者的自动语音识别训练.	输出特征知识
Suzuki 等人[175]	将近距离麦克风的知识迁移到喉头麦克风中来提高喉头麦克风的语音识别性能.	跨模态蒸馏
Bai 等人[176]	使用大的文本语言模型产生的软标签来指导 seq2seq 模型的语音识别训练.	跨模态蒸馏
Afouras 等人[177]	使用自动语音识别的教师模型指导无标签唇读视频的学生模型训练.	跨模态蒸馏
Moriya 等人[178]	将基于注意的编解码模型的注意力知识来提高基于时序分类的 CTC 自动语音识别性能.	中间特征知识
Wang 等人[179]	使用正常语音的信息来指导异常语音的重建.	跨模态蒸馏
Futami 等人[180]	将 BERT 语言模型产生的软标签来指导语音模型学习.	跨模态蒸馏

7.4.2　推荐系统的模型增强

模型增强同样能应用于推荐系统中，即将知识蒸馏用于获取高性能的推荐系统而不管其模型复杂度和大小. 知识蒸馏能将教师模型的某些知识迁移到学生模型中并提高其推荐或预测的性能，如用户评论的外部知识[186]和用户停留时长的知识[187]. 另外，知识蒸馏能通过改善数据的质量来间接性地提高推荐系统的质量. 例如，Pan 等人[118]使用知识蒸馏消除离散输入数据之间存在的噪声来间接性地提高推荐系统的性能. Liu 等人[188]使用知识蒸馏通过构建均匀分布的统一数据来提高推荐系统的性能.

7.5　安全和隐私问题

深度学习虽然取得了巨大的成就，但在信息安全问题上存在着较大的风险. 威胁和攻击经常出现，它们不仅威胁到深度学习模型，还会威胁人类的隐私. 基于知识蒸馏的安全保护是将知识蒸馏用于保护网络模型的安全和隐私等方面. 在深度学习网络模型的安全方面，仅仅在小车的图像中增加少量的扰动，网络就能将小车识别为猫[189]. 因此在对安全敏感的应用场景中(如无人驾驶)，必须考虑对抗性样本可能会带来的安全隐患. 为了解决这个问题，Papernot 等人[189]在蒸馏中将提取有关训练点的其他知识(如软标签)反馈到训练方案中，以降低对抗性样本对网络模型的有效性. Goldblum 等人[190]通过最小化教师在普通图像上的输出和学生在对抗性图像上的输出之间的差异来将教师的对抗性知识转移到学生模型中. 另外，同互联网攻击方法日新月异一样，网络模型的攻击也会存在时效问题，从而生成对抗性样本的速度也是一种重要的因素. 针对这一问题，Gil 等人[191]从多个样本的优化中获得知识来提高对抗性样本的生成速度.

直接利用数据必然会侵犯到个体的隐私，尤其是使用医学的数据是非法的. 除了当前热门的联邦学习[124]之外，知识蒸馏也可以应用于保护用户数据的隐私. 这两种方法的共同点是私有的数据并不直接进行共享. 基于联邦学习的思想是获取敏感数据的网络参数，而基于知识蒸馏的主要思想是将敏感数据学习到的特征通过知识蒸馏迁移到学生模型中. 在具体的应用中，Wang 等人[192]通过公共数据集将教师模型私有数据集学习到的特征迁移到学生模型中. 而 Vongkulbhisal 等人[193]将多个独立训练的分类器通过知识蒸馏训练为统一的分类器.

7.6　多模态数据

人类往往是通过视觉、听觉、嗅觉、味觉和触觉这五种感官信息来感受世界，这五种感官信息对应着 5 种不同的模态信息. 机器学习和人类认识世界的方式一样，都是通过利用多模态的信息来提高处理任务的性能. 在一些应用领域，数据是由不同数据结构的多模态数据所组成的，如包含着文本、图片和语音数据的视频. 因此，这些涉及到多模态

数据的应用领域需要使用多模态学习(Multimodal Learning). 应用在多模态学习中的知识蒸馏称为多模态蒸馏(Multimodal Distillation). 跨模态蒸馏和多模态蒸馏都涉及到将知识蒸馏应用于多种模态数据的学习. 它们的主要思想是利用不同模态数据的信息为目标任务提供互补的线索，从而提高学生网络的性能. 然而,它们的应用方式是不同的. 跨模态蒸馏通常是将不同模态数据的特征隐含地嵌入在单个模态数据的学生网络中，提高使用单个模态数据作为输入的学生网络在预测时的性能. 多模态蒸馏则是使用知识蒸馏整合多种模态数据的信息，提高模型泛化的能力，学生网络在预测时可以使用不同的模态数据. 因此，在通常情况下，跨模态蒸馏是多模态蒸馏的一个特例.

多模态蒸馏是从不同模态数据域中提取能为同一主题提供互补信息的异构特征，并在知识蒸馏的学习中将它们关联起来，为目标任务提供更多的多样化知识，其难点是：在训练如何时高效地获取多模态数据的互补性特征. 由于多模态数据的来源不同，导致各位置的数据并不一定是配对的. 配对多模态数据间的同位置像素具有相关性，能够容易地将一种模态学习到的特征知识迁移到另一种模态数据的网络中. 例如，Cioppa 等人[194]提出利用同一个场景的不同视角下的多个模态信息来训练单个网络. Wu 等人[195]从视频中提取不同模态数据关系的集成特征，提高暴力事件检测的性能. Xu 等人[45]通过产生一组中间辅助任务，提取不同任务所需要的特征，为学习目标任务提供丰富的多模态信息.

未配对多模态数据间的同位置像素并没有相关性，一种简单的方法是集成各个模态数据的输出特征来提高目标任务的性能. 例如，Garcia 等人[123]使用相互蒸馏增强不同模态数据网络的性能，并集成各模态网络的互补性特征，增强视频动作识别. Vielzeuf 等人[196]通过知识蒸馏，将不同模态数据的网络特征迁移到单个学生网络中. Wang 等人[197]独立训练各模态数据的教师网络，为学生网络提供多种模态的补充信息.

只使用各个模态数据的输出特征可能会限制多模态数据融合的整体性能. 为了克服这个问题，一些工作提出使用更多蒸馏的知识形式或知识蒸馏的方法去提高未配对的多模态数据网络的泛化能力. 例如，Dou 等人[135]通过特征归一化和共享卷积内核提出了未配对的多模态图像分割的蒸馏方案. Chen 等人[198]设计了特定于类的距离矩阵编码特征知识来避免特征的空间对齐问题，从而获取到各模态数据网络隐含层中的特征知识. Mun 等人[199]动态分配训练样本，解决多模态数据不足的问题，间接地提高集成模型泛化和专业化的能力.

7.7 金融证券

深度学习可用于金融证券领域中的自动化建模，进行风险预测和欺诈检测等任务的处理. 具体地，它能够在模拟的交易环境中直接优化与任务相关的指标并对其准确地建模，如利润. 然而，金融数据通常会包含大量不稳定的噪声而限制了网络模型的泛化能力. 例如，交易政策的微小变化会导致模型的性能差别很大. 人类在金融决策之前往往能通过自身的经验去清除噪声. 但是在深度学习网络中，噪声已变为网络要学习的特征并且会淹没掉有用的特征. 为了减轻噪声对金融证券的影响，Fang 等人[200]引入了财务经验作为先验知识来减少非平稳的噪声. Tsantekidis 等人[201]引入了一个多样化的教师模型集合来对不同货币的交易进行训练，学生网络通过知识蒸馏，学习教师集合最有利政策中的共同观点. 在股票趋势预测中，市场回报和超额回报这两种收益相关的因素在有噪声的数据中会存在着干扰特征，它们会相互地干扰着对方预测的性能. 为此，Tang 等人[202]通过自蒸馏动态地控制不同样本知识的重要性来过滤掉对目标任务有干扰性的特征.

深度学习模型能够直接从数据信息中自动学习到决策的方法，人类却往往不能理解其决策的过程. 然而，在金融系统中，决策的可解释性至关重要. 知识蒸馏技术能够将学习到的暗知识从网络中转移到可微分的树模型中，以方便解释测试样本的决策过程[203]. 对于金融的异常活动检测，深度学习也会起到一定的作用. 在金融领域中，可能出现异常的非法活动，如欺诈和洗钱. 因此，检测金融服务业中的异常现象是一项重要的任务. 在信息安全领域中，犯罪者的手段会在短短几天甚至更短的时间内进行更换，导致监控和反制犯罪的方法会存在着时效性. 对于该问题，知识蒸馏能够将金融历史数据的知识迁移到新的网络模型上，从而提高检测异常现象的速度和性能[204]. 此外，知识蒸馏能够通过引入相关的外部知识提高金融证券领域某些任务的性能，如从外部知识库中学习背景知识来提高从新闻中生成财务报告的质量[205].

7.8 小 节

知识蒸馏的主要思想是将教师网络中的知识迁移到学生网络中来改善目标领域的性能，这一特性

有着广泛的运用. 无论是在哪个应用领域，知识蒸馏都需要首先明确要蒸馏教师网络的哪些知识，其次才是知识蒸馏的具体方法. 不同任务可以使用相同的知识蒸馏方法，然而它们的知识形式往往是不相同的. 例如，计算机视觉一般侧重于平面的空间知识，NLP 通常关注于上下文的语义知识，而语音识别更强调序列化的时间知识. 因此，在实际的应用中，需要根据特定任务的特点和联系来挖掘合适且足够的知识来提高学生网络的性能.

8 知识蒸馏的研究趋势展望

知识蒸馏是一个新兴的研究领域，它仍有许多值得深入探索和亟待解决的问题. 在这一节中，我们提出一些值得进一步深入探讨的研究点，也是我们今后需要解决完善的研究方向.

(1) 如何确定何种知识是最佳的. 知识蒸馏中的知识是一个抽象的概念，网络参数，网络的输出和网络的中间特征等都可以理解为知识. 但是何种知识是最佳的，或者哪些知识以互补的方式能成为最佳的通用知识表示？为了回答这个问题，我们需要了解每种知识以及不同种类组合知识的作用. 比如说，基于特征的知识通常用于模仿教师特征产生的过程，基于关系的知识常用于捕获不同样本之间或网络层之间特征的关系. 当教师和学生的模型容量(“代沟”)较小的时候，学生只模仿教师的软目标就可以获得有竞争力的性能. 而当师生的“代沟”较大时，需要将多种蒸馏的知识形式和方法结合来表示教师模型. 虽然能明白多种知识的组合方式通常能提高学生网络的性能，但是使用哪些知识形式，方法和技术的组合是最优的，还尚无定论.

(2) 如何确定何处的知识是最佳的. 一些工作随机选择中间网络的某层特征作为知识，比如 FitNets[27]将教师前几层的网络特征作为特征蒸馏的位置. 然而他们都没有提供一个理由，即为什么能够成为代表性知识. 这主要是由于教师和学生模型结构的不一致导致的，即教师模型通常比学生模型拥有更多的网络层. 因此，需要筛选教师模型中最具有代表性的特征. 然而教师模型中哪些特征层是最具有代表性的？这也是一个未解决的问题. 在基于关系的知识蒸馏中，也一样无法解释该选择哪些层的关系知识作为学生模仿的对象. 如 FSP 矩阵[31]随机选择教师模型的两个网络层作为关系蒸馏的位置. 关系知识蒸馏是容量无关的，即关系蒸馏仅仅需要获取的是网络层间或样本间的关系知识. 因此这不是师生间的“代沟”问题，而是归咎于知识其实是一个“黑盒”问题.

(3) 如何定义最佳的师生结构. 知识蒸馏传递的并不是参数，而是抽取到的知识. 因此知识蒸馏是网络架构无关的，即任何学生都可以向任何教师学习. 通常，容量更大的学生模型可以学习更多的知识，但复杂度过大会延长推理的时间. 容量更大的教师模型隐含着较多的知识和更强的能力，但是并非能力越强的教师就能产生更佳的学生模型[13]. 同时，每一个教师模型都有一个最强学生结构[100]. 因此，我们只能在给定的教师模型的前提下，找到最佳的学生模型. 然而在未指定教师模型的情况下，目前还无法确定最佳的学生模型.

(4) 如何衡量师生间特征的接近程度. 知识蒸馏是要将教师网络中的知识迁移到学生模型中，迁移效果的好坏最终可以通过学生网络性能来体现. 然而在网络训练的过程中，只能通过损失函数去判断教师和学生之间特征的接近程度. 因此需要提前设计好知识蒸馏的损失函数，如 KL 散度、均方误差(Mean Squared Error，MSE)和余弦相似性. 而损失函数的选取受算法和离群点等因素的影响，并且，不同损失函数的作用范围是不一样的. 例如，通过 KL 散度衡量的两个随机分布上的相似度是非对称的. 余弦相似性强调两个向量的特征在方向上的差异，却没有考虑向量大小. MSE 在高维特征中的作用不明显，且很容易被随机特征混淆[4]. 因此，衡量师生间特征接近程度的方法是多样化的，我们需要根据特定的问题和场景选取最合适的损失函数.

(5) 蒸馏的知识形式、方法和融合技术还需要深入探索. 原始知识蒸馏将类间的关系信息作为知识，但这在“代沟”较大的师生网络中效果不佳. 为了解决这一问题，后续的研究者寻找不同形式的“知识”来充实知识的内涵，如关系知识. 其知识的来源应该是多样化的，可以来自于单个或多个的样本和网络本身. 同样，知识蒸馏的方法和融合技术也能缓解甚至解决师生间的“代沟”问题，它们强调充分地利用知识来提高模型的表征能力. 新的知识形式、方法和融合技术的发现可能会伴随着新的应用场景，这将丰富知识蒸馏的理论框架和实践的应用.

(6) 模型压缩和模型增强的深度融合. 模型压缩是将强大的复杂教师模型中的“知识”迁移到简单的学生模型中以满足低资源设备的应用要求，而模型增强用于获取高性能的复杂网络. 模型压缩和模型增强的融合是将教师模型中的“特权信息”迁移或继续强化轻量级学生模型的性能. 例如，Liu 等

人[206]通过从文本翻译模型中迁移“特权信息”来改进轻量级的语音翻译模型. 在未来的工作中，甚至能将无标签或其它领域数据的“特权信息”来继续加强一个轻量级学生模型的性能.

(7) 知识蒸馏在数据样本增强上的应用. 深度学习是数据驱动型的，往往需要大规模的数据集才能避免过度拟合. 由于隐私和法律等原因，在一些领域上，通常无法获取大规模的原始数据集，如医疗数据. 知识蒸馏需要足够的数据，才能将教师网络中的知识迁移到学生网络中. 换句话说，数据是连接教师网络和学生网络的桥梁. 先前的研究已经证明了知识蒸馏在数据样本增强上的广阔应用前景，如通过知识蒸馏产生原始数据集的近似样本[207]、使用其它相关数据的知识来减轻对目标数据集的依赖[208]以及教师和学生间部分网络的共同训练来提高具有小样本学生网络的性能[114]. 未来的工作需要继续探索知识蒸馏在数据样本增强上的应用场景和高效的蒸馏方法来实现小样本学习(Few-Shot Learning)或零样本学习(zero-shot learning).

(8) 知识蒸馏在数据标签上的应用. 给数据上标签需要特定领域的专业知识、大量的时间和成本. 可以利用知识蒸馏减少标注训练数据的麻烦，解决数据标签的问题. 如果该领域存在着强大的教师网络，能通过知识蒸馏给无标签的数据增加注释. 具体地，教师网络对未标记数据进行预测，并使用它们的预测信息充当学生模型数据的自动标注[209]. 以无标签数据作为输入的教师网络会产生软标签，这恰好能为学生网络提供学习的指导信息. 即使该领域没有强大的教师网络，也可以通过跨模态知识蒸馏，将其它领域的知识充当无标签数据的监督信号[66]. 因此，知识蒸馏能够减少对数据标签的依赖，需要继续研究它在半监督或无监督学习上的应用.

9 结束语

从 2015 年至今，知识蒸馏在模型压缩和模型增强方面受到了学术界和工业界的广泛关注和深入研究，并且取得了许多瞩目的研究成果. 知识蒸馏既可以获得轻量级的网络模型以便应用于资源受限的设备上，也可以通过传递不同形式的知识来强化深度学习模型的性能. 本文全面调研、归纳和分析了知识蒸馏相关的国内外研究现状，包括其由来、作用机制、知识形式、关键方法、与其它技术的融合以及应用进展等，并讨论了今后的发展方向，希望能对相关领域的研究人员和工程技术人员提供有益的帮助.

参 考 文 献

[1] Elsken T, Metzen J H, Hutter F. Neural architecture search: A survey. Journal of Machine Learning Research, 2019, 20(55): 1-21

[2] Hinton G, Vinyals O, Dean J. Distilling the knowledge in a neural network. arXiv preprint arXiv:1503. 02531, 2015

[3] Gou J, Yu B, Maybank S J, Tao D. Knowledge distillation: A survey. arXiv preprint arXiv:2006. 05525, 2020

[4] Wang L, Yoon K J. Knowledge distillation and student-teacher learning for visual intelligence: A review and new outlooks. arXiv preprint arXiv:2004. 05937, 2020

[5] Tian Y, Krishnan D, Isola P. Contrastive representation distillation//Proceedings of the 8th International Conference on Learning Representations. Addis Ababa, Ethiopia, 2019: 1-19

[6] Liu Y, Cao J, Li B, Yuan C, Hu W, Li Y, Duan Y. Knowledge distillation via instance relationship graph//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, USA, 2019: 7096-7104

[7] Pan S J, Yang Q. A survey on transfer learning. IEEE Transactions on Knowledge and Data Engineering, 2009, 22(10): 1345-1359

[8] Bucilǎ C, Caruana R, Niculescu-Mizil A. Model compression// Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Philadelphia, USA, 2006: 535-541

[9] Liang P, Daumé III H, Klein D. Structure compilation: Trading structure for features//Proceedings of the 25th International Conference on Machine Learning. Helsinki, Finland, 2008: 592-599

[10] Ba J, Caruana R. Do deep nets really need to be deep?// Proceedings of the 27th International Conference on Neural Information Processing Systems. Montreal, Canada, 2014: 2654-2662

[11] Li J, Zhao R, Huang J T, Gong Y. Learning small-size DNN with output-distribution-based criteria//Proceedings of the 15th Annual Conference of the International Speech Communication Association. Singapore, 2014: 1910-1914

[12] Furlanello T, Lipton Z C, Tschannen M, Itti L, Anandkumar A. Born again neural networks//Proceedings of the 35th International Conference on Machine Learning. Stockholm, Sweden, 2018: 1607-1616

[13] Cho J H, Hariharan B. On the efficacy of knowledge distillation//Proceedings of the IEEE International Conference on Computer Vision. Seoul, Korea (South), 2019: 4794-4802

[14] Urban G, Geras K J, Kahou S E, Aslan Ö, Wang S, Caruana R, Mohamed A, Philipose M, Richardson M, Caruana M. Do deep convolutional nets really need to be deep and convolutional?// Proceedings of the 5th International Conference on Learning Representations. Toulon, France, 2017:1-13

[15] Tang Z, Wang D, Zhang Z. Recurrent neural network training with dark knowledge transfer//Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Shanghai, China, 2016: 5900-5904

[16] Yuan L, Tay F E H, Li G, Wang T, Feng J. Revisiting knowledge distillation via label smoothing regularization//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA, 2020: 3903-3911

[17] Vapnik V, Vashist A. A new learning paradigm: learning using privileged information. Neural Networks, 2009, 22(5-6): 544-557

[18] Phuong M, Lampert C. Towards understanding knowledge distillation//Proceedings of the 36th International Conference on

Machine Learning. Long Beach, USA, 2019: 5142-5151
[19] Cheng X, Rao Z, Chen Y, Zhang Q. Explaining knowledge distillation by quantifying the knowledge//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA, 2020: 12925-12935
[20] Chen G, Choi W, Yu X, Han T, Chandraker M. Learning efficient object detection models with knowledge distillation//Proceedings of the 30th International Conference on Neural Information Processing Systems. Long Beach, USA, 2017: 742-751
[21] Wang T, Yuan L, Zhang X, Feng J. Distilling object detectors with fine-grained feature imitation//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Long Beach, USA, 2019: 4933-4942
[22] Hou Y, Ma Z, Liu C, Hui T-W, Loy C C. Inter-Region affinity distillation for road marking segmentation//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA, 2020: 12486-12495
[23] Liu Y, Chen K, Liu C, Qin Z, Luo Z, Wang J. Structured knowledge distillation for semantic segmentation//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Long Beach, USA, 2019: 2604-2613
[24] Takashima R, Sheng L, Kawai H. Investigation of sequence-level knowledge distillation methods for CTC acoustic models//Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Brighton, UK, 2019: 6156-6160
[25] Huang M, You Y, Chen Z, Qian Y, Yu K. Knowledge distillation for sequence model//Proceedings of the 19th Annual Conference of the International Speech Communication Association. Hyderabad, India, 2018: 3703-3707
[26] Gotmare A, Keskar N S, Xiong C, Socher R. A closer look at deep learning heuristics: learning rate restarts, warmup and distillation//Proceedings of the 7th International Conference on Learning Representations. New Orleans,USA, 2019:1-16
[27] Romero A, Ballas N, Kahou S E, Chassang A, Gatta C, Bengio Y. Fitnets: hints for thin deep nets//Proceedings of the 3rd International Conference on Learning Representations. San Diego, USA, 2015:1-13
[28] Zagoruyko S, Komodakis N. Paying more attention to attention: Improving the performance of convolutional neural networks via attention transfer//Proceedings of the 5th International Conference on Learning Representations. Toulon, France, 2017:1-13
[29] Li X, Xiong H, Wang H, Rao Y, Liu L, Huan J. Delta: deep learning transfer using feature map with attention for convolutional networks//Proceedings of the 7th International Conference on Learning Representations. New Orleans, USA, 2019:1-13
[30] Passalis N, Tefas A. Learning deep representations with probabilistic knowledge transfer//Proceedings of the European Conference on Computer Vision (ECCV). Munich, Germany, 2018: 268-284
[31] Yim J, Joo D, Bae J, Kim J. A gift from knowledge distillation: fast optimization, network minimization and transfer learning// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, USA, 2017: 4133-4141
[32] Park W, Kim D, Lu Y, Cho M. Relational knowledge distillation// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Long Beach, USA, 2019: 3967-3976
[33] Srinivas S, Fleuret F. Knowledge transfer with Jacobian Matching//Proceedings of the 35th International Conference on Machine Learning. Stockholm, Sweden, 2018: 4723-4731
[34] Lee S H, Kim D H, Song B C. Self-supervised knowledge distillation using singular value decomposition//Proceedings of the 15th European Conference on Computer Vision (ECCV). Munich, Germany, 2018: 339-354
[35] Chen Y, Wang N, Zhang Z. Darkrank: accelerating deep metric learning via cross sample similarities transfer//Proceedings of the AAAI Conference on Artificial Intelligence. New Orleans, USA, 2018: 2852-2859
[36] Peng B, Jin X, Liu J, Zhou S, Wu Y, Liu J, Zhang Z, Liu Y. Correlation congruence for knowledge distillation//Proceedings of the IEEE International Conference on Computer Vision. Seoul, Korea, 2019: 5006-5015
[37] Lee S, Song B C. Graph-based knowledge distillation by multi-head attention network//Proceedings of the 30th British Machine Vision Conference. Cardiff, UK, 2019: 141
[38] Bajestani M F, Yang Y. Tkd: Temporal knowledge distillation for active perception//Proceedings of the IEEE Winter Conference on Applications of Computer Vision. Snowmass Village, USA, 2020: 953-962
[39] Deng J, Pan Y, Yao T, Zhou W, Li H, Mei T. Relation distillation networks for video object detection//Proceedings of the IEEE International Conference on Computer Vision. Seoul, Korea, 2019: 7022-7031
[40] Wang H, Li Y, Wang Y, Hu H, Yang M-H. Collaborative distillation for ultra-Resolution universal style transfer// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA, 2020: 1857-1866
[41] Liu Y, Shu C, Wang J, Shen C. Structured knowledge distillation for dense prediction. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2020, (42): 1-15
[42] Xu X, Zou Q, Lin X, Huang Y, Tian Y. Integral knowledge distillation for multi-Person pose estimation. IEEE Signal Processing Letters, 2020, (27): 436-440
[43] Goodfellow I J, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S, Courville A, Bengio Y. Generative adversarial nets// Proceedings of the Advances in Neural Information Processing Systems. Montreal, Canada, 2014: 2672-2680
[44] Shen C, Wang X, Song J, Sun L, Song M. Amalgamating knowledge towards comprehensive classification//Proceedings of the AAAI Conference on Artificial Intelligence. Honolulu, USA, 2019, 33: 3068-3075
[45] Xu D, Ouyang W, Wang X, Sebe N. Pad-net: multi-tasks guided prediction-and-distillation network for simultaneous depth estimation and scene parsing//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA, 2018: 675-684
[46] Shen C, Xue M, Wang X, Song J, Sun L, Song M. Customizing student networks from heterogeneous teachers via adaptive knowledge amalgamation//Proceedings of the IEEE International Conference on Computer Vision. Seoul, Korea, 2019: 3504-3513
[47] Ye J, Wang X, Ji Y, Ou K, Song M. Amalgamating filtered knowledge: learning task-customized student from multi-task teachers//Proceedings of the 28th International Joint Conference on Artificial Intelligence. Macao, China, 2019: 4128-4134
[48] Ye J, Ji Y, Wang X, Ou K, Tao D, Song M. Student becoming the master: Knowledge amalgamation for joint scene parsing, depth estimation, and more//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Long Beach, USA,

2019: 2829-2838

[49] You S, Xu C, Xu C, Tao D. Learning from multiple teacher networks//Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Halifax, Canada, 2017: 1285-1294

[50] Fukuda T, Suzuki M, Kurata G, Thomas S, Cui J, Ramabhadran B. Efficient knowledge distillation from an ensemble of teachers//Proceedings of the 18th Annual Conference of the International Speech Communication Association. Stockholm, Sweden, 2017: 3697-3701

[51] Wu A, Zheng W S, Guo X, Lai J H. Distilled person re-identification: towards a more scalable system//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Long Beach, USA, 2019: 1187-1196

[52] Xiang L, Ding G. Learning from multiple experts: self-paced knowledge distillation for long-tailed classification. arXiv preprint arXiv:2001. 01536, 2020

[53] Jiang L, Wen Z, Liang Z, Wang Y, Melo G, Li Z, Ma L, Zhang J, Li X, Qi L. Long short-term sample distillation// Proceedings of the AAAI Conference on Artificial Intelligence. New York, USA, 2020: 4345-4352

[54] Jin X, Lan C, Zeng W, Chen Z. Uncertainty-aware multi-shot knowledge distillation for image-based object re-identification// Proceedings of the AAAI Conference on Artificial Intelligence. New York, USA, 2020: 11165-11172

[55] Hou Y, Ma Z, Liu C, Loy C C. Learning to steer by mimicking features from heterogeneous auxiliary networks//Proceedings of the AAAI Conference on Artificial Intelligence. Honolulu, USA, 2019, 33: 8433-8440

[56] Wang Y, Xu C, Xu C, Tao D. Adversarial learning of portable student networks//Proceedings of the AAAI Conference on Artificial Intelligence. New Orleans, USA, 2018: 4260-4267

[57] Mirzadeh S I, Farajtabar M, Li A, Levine N, Matsukawa A, Ghasemzadeh H. Improved knowledge distillation via teacher assistant//Proceedings of the AAAI Conference on Artificial Intelligence. New Orleans, USA, 2020, 34(04): 5191-5198

[58] Passalis N, Tzelepi M, Tefas A. Heterogeneous knowledge distillation using information flow modeling//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA, 2020: 2339-2348

[59] Gao M, Wang Y, Wan L. Residual error based knowledge distillation. Neurocomputing, 2021(433): 154-161

[60] Albanie S, Nagrani A, Vedaldi A, Zisserman A. Emotion recognition in speech using cross-modal transfer in the wild//Proceedings of the 26th ACM international Conference on Multimedia. Seoul, Korea, 2018: 292-301

[61] Su J C, Maji S. Adapting models to signal degradation using distillation//Proceedings of the British Machine Vision Conference. London, UK, 2017: 1-14

[62] Aytar Y, Vondrick C, Torralba A. Soundnet: learning sound representations from unlabeled video//Proceedings of the Advances in Neural Information Processing Systems. Barcelona, Spain, 2016: 892-900

[63] Girdhar R, Tran D, Torresani L, Ramanan D. Distinit: learning video representations without a single labeled video// Proceedings of the IEEE International Conference on Computer Vision. Seoul, Korea, 2019: 852-861

[64] Liu Y, Sheng L, Shao J, Yan J, Xiang S, Pan C. Multi-label image classification via knowledge distillation from weakly-supervised detection//Proceedings of the 26th ACM international Conference on Multimedia. Seoul, Korea, 2018: 700-708

[65] Gupta S, Hoffman J, Malik J. Cross modal distillation for supervision transfer//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, USA, 2016: 2827-2836

[66] Wang C, Kong C, Lucey S. Distill knowledge from nrsfm for weakly supervised 3d pose learning//Proceedings of the IEEE International Conference on Computer Vision. Seoul, Korea, 2019: 743-752

[67] Luo Z, Hsieh J T, Jiang L, Niebles J, Li F. Graph distillation for action detection with privileged modalities//Proceedings of the European Conference on Computer Vision (ECCV). Munich, Germany, 2018: 166-183

[68] Liu Q, Xie L, Wang H, Yuille A. Semantic-aware knowledge preservation for zero-shot sketch-based image retrieval// Proceedings of the IEEE International Conference on Computer Vision. Seoul, Korea, 2019: 3662-3671

[69] Hoffman J, Gupta S, Darrell T. Learning with side information through modality hallucination//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, USA, 2016: 826-834

[70] Aditya S, Saha R, Yang Y, Baral C. Spatial knowledge distillation to aid visual reasoning//Proceedings of the 2019 IEEE Winter Conference on Applications of Computer Vision (WACV). Waikoloa Village, USA, 2019: 227-235

[71] Ye H J, Lu S, Zhan D C. Distilling cross-task knowledge via relationship matching//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA, 2020: 12396-12405

[72] Yuan M, Peng Y. Ckd: Cross-task knowledge distillation for text-to-image synthesis. IEEE Transactions on Multimedia. 2020, 22(8): 1955-1968

[73] Li K, Yu L, Wang S, Heng P. Towards cross-modality medical image segmentation with online mutual knowledge distillation// Proceedings of the AAAI Conference on Artificial Intelligence. New York, USA, 2020: 775-783

[74] Piao Y, Rong Z, Zhang M, Ren W, Lu H. A2dele: adaptive and attentive depth distiller for efficient RGB-D salient object detection//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Seattle, USA, 2020: 9060-9069

[75] Garcia N C, Morerio P, Murino V. Modality distillation with multiple stream networks for action recognition//Proceedings of the European Conference on Computer Vision (ECCV). Munich, Germany, 2018: 103-118

[76] Tavakolian M, Tavakoli H R, Hadid A. Awsd: adaptive weighted spatiotemporal distillation for video representation//Proceedings of the IEEE International Conference on Computer Vision. Seoul, Korea, 2019: 8020-8029

[77] Gu X, Ma B, Chang H, Shan S, Chen X. Temporal knowledge propagation for image-to-video person re-identification//Proceedings of the IEEE International Conference on Computer Vision. Seoul, Korea, 2019: 9647-9656

[78] Zhao L, Peng X, Chen Y, Kapadia M, Metaxas D N. Knowledge as priors: cross-modal knowledge generalization for datasets without superior knowledge//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA, 2020: 6528-6537

[79] Hu H, Xie L, Hong R, Tian Q. Creating something from nothing:

unsupervised knowledge distillation for cross-modal hashing// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA, 2020: 3123-3132

[80] Li D, Yu X, Xu C, Petersson L, Li H. Transferring cross-domain knowledge for video sign language recognition//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA, 2020: 6205-6214

[81] Zhang Y, Xiang T, Hospedales T M, Lu H. Deep mutual learning//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA, 2018: 4320-4328

[82] Cui J, Kingsbury B, Ramabhadran B, Saon G, Sercu T, Audhkhasi K, Sethy A, Nußbaum-Thom M, Rosenberg A. Knowledge distillation across ensembles of multilingual models for low-resource languages//Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). New Orleans, USA, 2017: 4825-4829

[83] Chen D, Mei J P, Wang C, Feng Y, Chen C. Online knowledge distillation with diverse peers//Proceedings of the AAAI Conference on Artificial Intelligence. New York, USA, 2020: 3430- 3437

[84] Anil R, Pereyra G, Passos A, Ormándi R, Dahl G E, Hinton G E. Large scale distributed neural network training through online distillation//Proceedings of the 6th International Conference on Learning Representations. Vancouver, Canada, 2018:1-12

[85] Li Z, Hoiem D. Learning without forgetting. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017, 40(12): 2935-2947

[86] Hou S, Pan X, Change Loy C, Wang Z, Lin D. Lifelong learning via progressive distillation and retrospection//Proceedings of the European Conference on Computer Vision (ECCV). Munich, Germany, 2018: 437-452

[87] Zhai M, Chen L, Tung F, He J, Nawhal M, Mori G. Lifelong gan: Continual learning for conditional image generation//Proceedings of the IEEE International Conference on Computer Vision. Seoul, Korea, 2019: 2759-2768

[88] Mobahi H, Farajtabar M, Bartlett P L. Self-distillation amplifies regularization in hilbert space//Proceedings of the Advances in Neural Information Processing Systems. 2020:1-33

[89] Zhang Z, Sabuncu M. Self-distillation as instance-specific label smoothing//Proceedings of the Advances in Neural Information Processing Systems. 2020:1-16

[90] Yun S, Park J, Lee K, Shin J. Regularizing class-wise predictions via self-knowledge distillation//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA, 2020: 13876-13885

[91] Xu T B, Liu C L. Data-distortion guided self-distillation for deep neural networks//Proceedings of the AAAI Conference on Artificial Intelligence. Honolulu, USA, 2019, 33: 5565-5572

[92] Zhang L, Song J, Gao A, Chen J, Bao C, Ma K. Be your own teacher: improve the performance of convolutional neural networks via self distillation//Proceedings of the IEEE International Conference on Computer Vision. Seoul, Korea, 2019: 3713-3722

[93] Nie X, Li Y, Luo L, Zhang N, Feng J. Dynamic kernel distillation for efficient pose estimation in videos//Proceedings of the IEEE International Conference on Computer Vision. Seoul, Korea, 2019: 6942-6950

[94] Haidar M A, Rezagholizadeh M. Textkd-gan: text generation using knowledge distillation and generative adversarial networks//Proceedings of the Canadian Conference on Artificial Intelligence. Kingston, Canada, 2019: 107-118

[95] Wang X, Zhang R, Sun Y, Qi J. Kdgan: knowledge distillation with generative adversarial networks//Proceedings of the Advances in Neural Information Processing Systems. Montreal, Canada, 2018: 775-786

[96] Belagiannis V, Farshad A, Galasso F. Adversarial network compression//Proceedings of the European Conference on Computer Vision (ECCV). Munich, Germany, 2018: 431-449

[97] Heo B, Lee M, Yun S, Choi J Y. Knowledge distillation with adversarial samples supporting decision boundary//Proceedings of the AAAI Conference on Artificial Intelligence. Honolulu, USA, 2019, 33: 3771-3778

[98] Bashivan P, Tensen M, DiCarlo J J. Teacher guided architecture search//Proceedings of the IEEE International Conference on Computer Vision. Seoul, Korea, 2019: 5320-5329

[99] Li C, Peng J, Yuan L, Wang G, Liang X, Lin L, Chang X. Block-wisely supervised neural architecture search with knowledge distillation//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA, 2020: 1989-1998

[100] Liu Y, Jia X, Tan M, Vemulapalli R, Zhu Y, Green B, Wang X. Search to distill: pearls are everywhere but not the eyes// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA, 2020: 7539-7548

[101] Yang S, Gong Z, Ye K, Wei Y, Huang Z. H, Huang Z. EdgeRNN: A compact speech recognition network with spatio-temporal features for edge computing. IEEE Access, 2020(8): 81468-81478

[102] Kang M, Mun J, Han B. Towards oracle knowledge distillation with neural architecture search//Proceedings of the AAAI Conference on Artificial Intelligence. New York, USA, 2020: 4404-4411

[103] Berseth G, Xie C, Cernek P, Panne M. Progressive reinforcement learning with distillation for multi-skilled motion control//Proceedings of the 6th International Conference on Learning Representations. Vancouver, Canada, 2018，1-15

[104] Lai K H, Zha D, Li Y, Hu X. Dual policy distillation//Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence. Yokohama, Japan, 2020: 3146-3152

[105] Rusu A A, Colmenarejo S G, Gülçehre C, Desjardins G, Kirkpatrick J, Pascanu R, Mnih V, Kavukcuoglu K, Hadsell R. Policy distillation//Proceedings of the 4th International Conference on Learning Representations. San Juan, Puerto Rico, 2016:1-13

[106] Ashok A, Rhinehart N, Beainy F, Kitani K M. N2N learning: network to network compression via policy gradient reinforcement learning//Proceedings of the 6th International Conference on Learning Representations. Vancouver, Canada, 2018:1-20

[107] Yang Y, Qiu J, Song M, Tao D, Wang X. Distilling knowledge from graph convolutional networks//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA, 2020: 7074-7083

[108] Lassance C, Bontonou M, Hacene G B, Gripon V, Tang J, Ortega A. Deep geometric knowledge distillation with graphs// Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing. Barcelona, Spain, 2020: 8484-8488

[109] Pan B, Cai H, Huang D A, Lee K H, Gaidon A, Adeli E, Niebles J C. Spatio-temporal graph for video captioning with knowledge distillation//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA, 2020: 10870-10879

[110] Zhang C, Peng Y. Better and faster: knowledge transfer from multiple self-supervised learning tasks via graph distillation for

video classification//Proceedings of the 27th International Joint Conference on Artificial Intelligence. Stockholm, Sweden, 2018: 1135-1141

[111] Zhang W, Miao X, Shao Y, Jiang J, Chen L, Ruas O, Cui B. Reliable data distillation on graph convolutional network// Proceedings of the 2020 ACM SIGMOD International Conference on Management of Data. Portland, USA, 2020: 1399-1414

[112] Polino A, Pascanu R, Alistarh D. Model compression via distillation and quantization//Proceedings of the 6th International Conference on Learning Representations. Vancouver, Canada, 2018: 1-21

[113] Bai H, Wu J, King I, Lyu M R. Few shot network compression via cross distillation//Proceedings of the AAAI Conference on Artificial Intelligence. New York, USA, 2020: 3203-3210

[114] Li T, Li J, Liu Z, Zhang C. Few sample knowledge distillation for efficient network compression//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA, 2020: 14639-14647

[115] Wang Z, Deng Z, Wang S. Accelerating convolutional neural networks with dominant convolutional kernel and knowledge pre-regression//Proceedings of the European Conference on Computer Vision. Amsterdam, The Netherlands, 2016: 533-548

[116] Li L, Sun K, Zhu J. A novel multi-knowledge distillation approach. IEICE Transactions on Information and Systems, 2021, 104(1): 216-219

[117] He T, Shen C, Tian Z, Gong G, Sun C, Yan Y. Knowledge adaptation for efficient semantic segmentation//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, USA, 2019: 578-587

[118] Pan Y, He F, Yu H. A novel enhanced collaborative autoencoder with knowledge distillation for top-*N* recommender systems. Neurocomputing, 2019(332): 137-148

[119] Kim J, Park S U, Kwak N. Paraphrasing complex network: Network compression via factor transfer//Proceedings of the 32nd International Conference on Neural Information Processing Systems. Montreal, Canada, 2018: 2765-2774

[120] Walawalkar D, Shen Z, Savvides M. Online ensemble model compression using knowledge distillation//Proceedings of the European Conference on Computer Vision. Glasgow, UK, 2020: 18-35

[121] Guo Q, Wang X, Wu Y, Yu Z, Liang D, Hu X, Luo P. Online knowledge distillation via collaborative learning//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA, 2020: 11020-11029

[122] Zaras A, Passalis N, Tefas A. Improving Knowledge Distillation using Unified Ensembles of Specialized Teachers. Pattern Recognition Letters, 2021(146): 215-221

[123] Garcia N C, Bargal S A, Ablavsky V, Morerio P, Murino V, Sclaroff S. Distillation multiple choice learning for multimodal action recognition//Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. 2021: 2755-2764

[124] McMahan B, Moore E, Ramage D, Hampson S, Arcas B A. Communication-efficient learning of deep networks from decentralized data//Proceedings of the 20th International Conference on Artificial Intelligence and Statistics. Fort Lauderdale, USA, 2017: 1273-1282

[125] Sui D, Chen Y, Zhao J, Jia Y, Xie Y, Sun W. FedED: Federated learning via ensemble distillation for medical relation extraction//Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). 2020: 2118-2128

[126] Sattler F, Marban A, Rischke R, Samek W. Communication-efficient federated distillation. arXiv preprint arXiv: 2012. 00632, 2020

[127] Zhou Y, Pu G, Ma X, Li X, Wu D. Distilled One-Shot Federated Learning. arXiv preprint arXiv:2009. 07999, 2020

[128] Shen T, Zhang J, Jia X, Zhang F, Huang G, Zhou P, Wu F, Wu C. Federated mutual learning. arXiv preprint arXiv:2006. 16765, 2020

[129] Lin T, Kong L, Stich S U, Jaggi M. Ensemble distillation for robust model fusion in federated learning//Proceedings of the Advances in Neural Information Processing Systems. 2020:1-26

[130] Wang W, Wei F, Dong L, Bao H, Yang N, Zhou M. Minilm: Deep self-attention distillation for task-agnostic compression of pre-trained transformers//Proceedings of the Advances in Neural Information Processing Systems. 2020:1-15

[131] Zhu M, Han K, Zhang C, Lin J, Wang Y. Low-resolution visual recognition via deep feature distillation//Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing. Brighton, UK, 2019: 3762-3766

[132] Fu K, Shi P, Song Y, Ge S, Lu X, Li J. Ultrafast video attention prediction with coupled knowledge distillation//Proceedings of the AAAI Conference on Artificial Intelligence. New York, USA, 2020: 10802-10809

[133] Wang X, Hu J F, Lai J H, Zhang J, Zheng W S. Progressive teacher-student learning for early action prediction//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Long Beach, USA, 2019: 3556-3565

[134] Yu R, Li A, Morariu V I, Davis L S. Visual relationship detection with internal and external linguistic knowledge distillation// Proceedings of the IEEE International Conference on Computer Vision. Venice, Italy, 2017: 1974-1982

[135] Dou Q, Liu Q, Heng P A, Glocker B. Unpaired multi-modal segmentation via knowledge distillation. IEEE Transactions on Medical Imaging, 2020, 39(7): 2415-2425

[136] Orbes-Arteainst M, Cardoso J, Sørensen L, Igel C, Ourselin S, Modat M, Nielsen M, Pai A. Knowledge distillation for semi-supervised domain adaptation//Proceedings of the OR 2. 0 Context-Aware Operating Theaters and Machine Learning in Clinical Neuroimaging. Shenzhen, China, 2019: 68-76

[137] Devlin J, Chang M W, Lee K, Toutanova K. Bert: Pre-training of deep bidirectional transformers for language understanding// Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Minneapolis, USA, 2019: 4171-4186

[138] Tsai H, Riesa J, Johnson M, Arivazhagan N, Li X, Archer A. Small and practical BERT models for sequence labeling// Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). Hong Kong, China, 2019: 3623-3627

[139] Liu W, Zhou P, Wang Z, Zhao Z, Deng H, Ju Q. FastBERT: A self-distilling BERT with adaptive inference time//Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. 2020:6035-6044

[140] Turc I, Chang M W, Lee K, Toutanova K. Well-read students learn better: on the importance of pre-training compact models. arXiv

preprint arXiv:1908. 08962, 2019
[141] Sanh V, Debut L, Chaumond J, Wolf T. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. arXiv preprint arXiv:1910. 01108, 2019
[142] Zhao S, Gupta R, Song Y, Zhou D. Extreme language model compression with optimal subwords and shared projections. arXiv preprint arXiv:1909. 11687, 2019
[143] Sun Z, Yu H, Song X, Liu R, Yang Y, Zhou D. MobileBERT: Task-sgnostic compression of BERT by progressive knowledge transfer. Stroudsburg, USA: ACL. 2019
[144] Sun S, Cheng Y, Gan Z, Liu J. Patient knowledge distillation for BERT model compression//Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). Hong Kong, China, 2019: 4314-4323
[145] Jiao X, Yin Y, Shang L, Jiang X, Chen X, Li L, Wang F, Liu Q. TinyBERT: Distilling BERT for natural language understanding// Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Findings. 2020: 4163-4174
[146] Xu C, Zhou W, Ge T, Wei F, Zhou M. BERT-of-theseus: Compressing BERT by progressive module replacing//Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). 2020: 7859-7869
[147] Yang Z, Shou L, Gong M, Lin W, Jiang D. Model compression with multi-task knowledge distillation for web-scale question answering system. arXiv preprint arXiv:1904. 09636, 2019
[148] Arora S, Khapra M M, Ramaswamy H G. On knowledge distillation from complex networks for response prediction//Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Minneapolis, USA, 2019: 3813-3822
[149] Pires T, Schlinger E, Garrette D. How multilingual is multilingual BERT?//Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. Florence, Italy, 2019: 4996-5001
[150] Wang X, Jiang Y, Bach N, Wang T, Huang F, Tu K. Structure-level knowledge distillation for multilingual sequence labeling//Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. 2020: 3317-3330
[151] Mukherjee S, Awadallah A H. XtremeDistil: Multi-stage distillation for massive multilingual models//Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. 2020: 2221-2234
[152] Hu M, Peng Y, Wei F, Huang Z, Li D, Yang N, Zhou M. Attention-guided answer distillation for machine reading comprehension// Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. Brussels. Belgium, 2018: 2077-2086
[153] Chen Y C, Gan Z, Cheng Y, Liu J Z, Liu J J. Distilling knowledge learned in BERT for text generation//Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. 2020: 7893-7905
[154] Wu Q, Lin Z, Karlsson B F, Lou J, Huang B. Single-/multi-source cross-lingual NER via teacher-student learning on unlabeled data in target language//Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. 2020: 6505-6514
[155] Feng J, Tao C, Wu W, Feng Y, Zhao D, Yan R. Learning a matching model with co-teaching for multi-turn response selection in retrieval-based dialogue systems//Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. Florence, Italy, 2019: 3805-3815
[156] Hochreiter S, Schmidhuber J. Long short-term memory. Neural Computation, 1997, 9(8): 1735-178
[157] Geras K J, Mohamed A, Caruana R, Urban G, Wang S, Aslan O, Philipose M, Richardson M, Sutton C. Blending lstms into CNNs. arXiv preprint arXiv:1511. 06433, 2015
[158] Lu L, Guo M, Renals S. Knowledge distillation for small- footprint highway networks//Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing. New Orleans, USA, 2017: 4820-4824
[159] Graves A, Fern á ndez S, Gomez F, Schmidhuber J. Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks//Proceedings of the 23rd International Conference on Machine Learning. Pittsburgh, USA, 2006: 369-376
[160] Ding H, Chen K, Huo Q. Compression of CTC-trained acoustic models by dynamic frame-wise distillation or segment-wise n-best hypotheses imitation//Proceedings of the 20th Annual Conference of the International Speech Communication Association. Graz, Austria, 2019: 3218-3222
[161] Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez A, Kaiser L, Polosukhin I. Attention is all you need//Proceedings of the 31st International Conference on Neural Information Processing Systems. Long Beach, USA, 2017: 6000-6010
[162] Aguilar G, Ling Y, Zhang Y, Yao B, Fan X, Guo C. Knowledge distillation from internal representations//Proceedings of the AAAI Conference on Artificial Intelligence. New York, USA, 2020: 7350-7357
[163] Gaido M, Di Gangi M A, Negri M, Turchi M. End-to-end speech-translation with knowledge distillation: FBK@ IWSLT2020// Proceedings of the 17th International Conference on Spoken Language Translation. 2020: 80-88
[164] Shi B, Sun M, Kao C C, Rozgic V, Matsoukas, Wang C. Compression of acoustic event detection models with quantized distillation//Proceedings of the 20th Annual Conference of the International Speech Communication Association. Graz, Austria, 2019: 3639-3643
[165] Markov K, Matsui T. Robust speech recognition using generalized distillation framework//Proceedings of the 17th Annual Conference of the International Speech Communication Association. San Francisco, USA, 2016: 2364-2368
[166] Joy N M, Kothinti S R, Umesh S, Abraham B. Generalized distillation framework for speaker normalization//Proceedings of the 18th Annual Conference of the International Speech Communication Association. Stockholm, Sweden, 2017: 739-743
[167] Abraham B, Seeram T, Umesh S. Transfer learning and distillation techniques to improve the acoustic modeling of low resource languages//Proceedings of the 18th Annual Conference of the International Speech Communication Association. Stockholm, Sweden, 2017: 2158-2162
[168] Shen P, Lu X, Li S, Kawai H. Feature representation of short utterances based on knowledge distillation for spoken language identification//Proceedings of the 19th Annual Conference of the International Speech Communication Association. Hyderabad, India, 2018: 1813-1817
[169] Das A, Hasegawa-Johnson M. Improving DNNs trained with non-native transcriptions using knowledge distillation and target interpolation//Proceedings of the 19th Annual Conference of the

International Speech Communication Association. Hyderabad, India, 2018: 2434-2438

[170] Subramanian A S, Chen S J, Watanabe S. Student-teacher learning for BLSTM mask-based speech enhancement//Proceedings of the 19th Annual Conference of the International Speech Communication Association. Hyderabad, India, 2018: 3249-3253

[171] Heo H S, Jung J, Shim H, Yu H. Acoustic scene classification using teacher-student learning with soft-labels//Proceedings of the 20th Annual Conference of the International Speech Communication Association. Graz, Austria, 2019: 614-618

[172] Kurata G, Audhkhasi K. Guiding CTC posterior spike timings for improved posterior fusion and knowledge distillation//Proceedings of the 20th Annual Conference of the International Speech Communication Association. Graz, Austria, 2019: 1616-1620

[173] Li W, Wang S, Lei M, Siniscalchi S M. Lee C-H. Improving audio-visual speech recognition performance with cross-modal student-teacher training//Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Brighton, UK, 2019: 6560-6564

[174] Zhang W, Chang X, Qian Y. Knowledge distillation for End-to-end monaural multi-talker ASR system//Proceedings of the 20th Annual Conference of the International Speech Communication Association. Graz, Austria, 2019: 2633-2637

[175] Suzuki T, Ogata J, Tsunakawa T, Nishida M, Nishimura M. Knowledge distillation for throat microphone speech recognition// Proceedings of the 20th Annual Conference of the International Speech Communication Association. Graz, Austria, 2019: 461-465

[176] Bai Y, Yi J, Tao J, Tian Z, Wen Z. Learn spelling from teachers: transferring knowledge from language models to sequence-to-sequence speech recognition//Proceedings of the 20th Annual Conference of the International Speech Communication Association. Graz, Austria, 2019: 3795-3799

[177] Afouras T, Chung J S, Zisserman A. ASR is all you need: cross-modal distillation for lip reading//Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Barcelona, Spain, 2020: 2143-2147

[178] Moriya T, Sato H, Tanaka T, Ashihara T, Masumura R, Shinohara Y. Distilling attention weights for CTC-based ASR systems// Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Barcelona, Spain, 2020: 6894-6898

[179] Wang D, Yu J, Wu X, Liu S, Sun L, Liu X, Meng H. End-to-end voice conversion via cross-modal knowledge distillation for dysarthric speech reconstruction//Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Barcelona, Spain, 2020: 7744-7748

[180] Futami H, Inaguma H, Ueno S, Mimura M, Sakai S, Kawahara T. Distilling the knowledge of BERT for sequence-to-sequence ASR//Proceedings of the 21th Annual Conference of the International Speech Communication Association. Shanghai, China, 2020: 3635-3639

[181] Huang Z, Zhang J, Tian C, Sun S, Xiang Y. Survey on learning-to-rank based recommendation algorithms. Journal of Software, 2016, 27(3): 691-713 (in Chinese)
(黄震华, 张佳雯, 田春岐, 孙圣力, 向阳. 基于排序学习的推荐算法研究综述. 软件学报, 2016, 27(3): 691-713)

[182] Tang J, Wang K. Ranking distillation: Learning compact ranking models with high performance for recommender system//Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. London, UK, 2018: 2289-2298

[183] Lee J, Choi M, Lee J, Shim H. Collaborative distillation for top-n recommendation//Proceedings of the 2019 IEEE International Conference on Data Mining (ICDM). Beijing, China, 2019: 369-378

[184] Zhang Y, Xu X, Zhou H, Zhang Y. Distilling structured knowledge into embeddings for explainable and accurate recommendation//Proceedings of the 13th International Conference on Web Search and Data Mining. Houston, USA, 2020: 735-743

[185] Wang H, Lian D, Ge Y. Binarized collaborative filtering with distilling graph convolutional networks. arXiv preprint arXiv: 1906. 01829, 2019

[186] Chen X, Zhang Y, Xu H, Qin Z, Zha H. Adversarial distillation for efficient recommendation with external knowledge. ACM Transactions on Information Systems (TOIS), 2018, 37(1): 1-28

[187] Xu C, Li Q, Ge J, Gao J, Yang X, Pei C, Sun F, Wu J, Sun H, Ou W. Privileged features distillation at taobao recommendations// Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. 2020: 2590-2598

[188] Liu D, Cheng P, Dong Z, He X, Pan W, Ming Z. A general knowledge distillation framework for counterfactual recommendation via uniform data//Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval. 2020: 831-840

[189] Papernot N, McDaniel P, Wu X, Jha S, Swami A. Distillation as a defense to adversarial perturbations against deep neural networks//Proceedings of the 2016 IEEE Symposium on Security and Privacy (SP). San Jose, USA, 2016: 582-597

[190] Goldblum M, Fowl L, Feizi S, Goldstein T. Adversarially robust distillation//Proceedings of the AAAI Conference on Artificial Intelligence. New York, USA, 2020: 3996-4003

[191] Gil Y, Chai Y, Gorodissky O, Berant J. White-to-black: efficient distillation of black-box adversarial attacks// Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Minneapolis, USA, 2019: 1373-1379

[192] Wang J, Bao W, Sun L, Zhu X, Cao B, Yu P S. Private model compression via knowledge distillation//Proceedings of the AAAI Conference on Artificial Intelligence. Honolulu, USA, 2019, 33: 1190-1197

[193] Vongkulbhisal J, Vinayavekhin P, Visentini-Scarzanella M. Unifying heterogeneous classifiers with distillation//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Long Beach, USA, 2019: 3175-3184

[194] Cioppa A, Deliege A, Huda N U, Gade R, Droogenbroeck M V, Moeslund T B. Multimodal and multiview distillation for real-time player detection on a football field//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. Seattle, USA, 2020: 880-881

[195] Wu P, Liu J, Shi Y, Sun Y, Shao F, Wu Z, Yang Z. Not only look, but also listen: Learning multimodal violence detection under weak supervision//Proceedings of the European Conference on Computer Vision. Glasgow, UK, 2020: 322-339

[196] Vielzeuf V, Lechervy A, Pateux S, Jurie F. Towards a general model of knowledge for facial analysis by multi-source transfer learning. arXiv preprint arXiv:1911. 03222, 2019

[197] Wang Q, Zhan L, Thompson P, Zhou Z. Multimodal learning

with incomplete modalities by knowledge distillation//Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. 2020: 1828-1838

[198] Chen J, Li W, Li H, Zhang J. Deep class-specific affinity-guided convolutional network for multimodal unpaired image segmentation//Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention. Lima, Peru, 2020: 187-196

[199] Mun J, Lee K, Shin J, Han B. Learning to specialize with knowledge distillation for visual question answering//Proceedings of the Advances in Neural Information Processing Systems. Montr é al, Canada, 2018: 8092-8102

[200] Fang J, Lin J. Prior knowledge distillation based on financial time series. arXiv preprint arXiv:2006. 09247, 2020

[201] Tsantekidis A, Passalis N, Tefas A. Diversity-driven knowledge distillation for financial trading using deep reinforcement learning. Neural Networks. 2021(140):193-202

[202] Tang H, Wu L, Liu W, Bian J. ADD: Augmented disentanglement distillation framework for improving stock trend forecasting. arXiv preprint arXiv:2012. 06289, 2020

[203] Li J, Li Y, Xiang X, Xia S, Dong S, Cai Y. TNT: An interpretable tree-network-tree learning framework using knowledge distillation. Entropy, 2020, 22(11): 1203

[204] Shen H, Kursun E. Label augmentation via time-based knowledge distillation for financial anomaly detection. arXiv preprint arXiv:2101. 01689, 2021

[205] Ren Y, Wang Z, Wang Y, Zhang X. Generating long financial report using conditional variational autoencoders with knowledge distillation. arXiv preprint arXiv:2010. 12188, 2020

[206] Liu Y, Xiong H, Zhang J, He Z, Wu H, Wang H, Zong C. End-to-end speech translation with knowledge distillation//Proceedings of the 20th Annual Conference of the International Speech Communication Association. Graz, Austria, 2019: 1128-1132

[207] Nayak G K, Mopuri K R, Shaj V, Radhakrishnan V B, Chakraborty A. Zero-shot knowledge distillation in deep networks//Proceedings of the International Conference on Machine Learning. Long Beach, USA, 2019: 4743-4751

[208] Li Y, Yang J, Song Y, Cao L, Luo J, Li L. Learning from noisy labels with distillation//Proceedings of the IEEE International Conference on Computer Vision. Venice, Italy, 2017: 1910-1918

[209] Chen T, Kornblith S, Swersky K, Norouzi M, Hinton G E. Big self-supervised models are strong semi-supervised learners//Proceedings of the Advances in Neural Information Processing Systems. 2020: 1

HUANG Zhen-Hua, Ph.D., professor, Ph.D. supervisor. His research interests mainly include deep learning, recommendation system, and data mining.

YANG Shun-Zhi, Ph.D. candidate. His research focuses on efficient network model, knowledge distillation, and image recognition.

LIN Wei, M.S. candidate. His research focuses on recommendation system and knowledge distillation.

NI Juan, master, lecturer. Her research focuses on educational big data and knowledge graph.

SUN Sheng-Li, Ph.D., associate professor. His research interests mainly include deep learning, data mining, and database.

CHEN Yun-Wen, Ph.D., senior engineer. His research interests mainly include deep learning, data mining, and NLP.

TANG Yong, Ph.D., professor, Ph.D. supervisor. His research interests mainly include educational big data, data mining, and database.

Background

Because of its robustness to object diversity, deep learning has received widespread attention and rapid development in recent years. But the performing of deep learning networks often requires resources, which are limited in resource-constrained devices such as the Internet of Things and mobile internet. Thus researchers have begun to consider efficient deep learning network models. Knowledge distillation is a new method to obtain efficient small-scale networks. Its main idea is to transfer the "knowledge" from complex teacher networks with a strong learning ability to simple student networks. From 2015 to the present, Knowledge distillation has been widely studied and applied in model compression and enhancement. The purpose of model compression is to reduce network parameters and calculations without significantly reducing network accuracy. Along with the in-depth study, the researchers found that they can continue to improve the performance of a deep learning model by transferring data knowledge in different fields through knowledge distillation.

For researchers and engineers in related fields to better understand the theory and research status of knowledge distillation, this paper conducts a comprehensive investigation of knowledge distillation from the aspects of basic knowledge, theoretical methods, and applications. Specifically, the basic knowledge details the source and mechanism of knowledge distillation. The theoretical method summarizes the form and method of knowledge, and the integration with other technologies. Furthermore, this paper provides applications, challenges and insights for knowledge distillation.

This work is partially supported by the National Natural Science Foundation of China (No. 61772366, No. U1811263, and No. 61972328), which aims to research on knowledge distillation based recommendation system, and the Natural Science Foundation of Shanghai (No. 17ZR1445900), which aims to research on performance enhanced knowledge distillation.