分类号 TP391 学号 M201973167
学校代码 10487 密级 公开

华中科技大学

硕士学位论文

（学术型 ☑ 专业型 □）

基于自注意力机制的自知识蒸馏研究

学位申请人：高也

学 科 专 业：计算机软件与理论

指 导 教 师：何琨 教授

答 辩 日 期：2022 年 5 月 21 日

A Dissertation Submitted in Partial Fulfillment of the Requirements for the Master Degree in Engineering

Research on Self-Knowledge Distillation with Self-Attention Mechanism

Candidate : GAO Ye

Major : Computer Software and Theory

Supervisor : Prof. HE Kun

Huazhong University of Science and Technology

Wuhan 430074, P. R. China

May 21, 2022

独创性声明

本人声明所呈交的学位论文是我个人在导师指导下进行的研究工作及取得的研究成果。尽我所知，除文中已经标明引用的内容外，本论文不包含任何其他个人或集体已经发表或撰写过的研究成果。对本文的研究做出贡献的个人和集体，均已在文中以明确方式标明。本人完全意识到本声明的法律结果由本人承担。

学位论文作者签名：

日期：　　年　　月　　日

学位论文版权使用授权书

本学位论文作者完全了解学校有关保留、使用学位论文的规定，即：学校有权保留并向国家有关部门或机构送交论文的复印件和电子版，允许论文被查阅和借阅。本人授权华中科技大学可以将本学位论文的全部或部分内容编入有关数据库进行检索，可以采用影印、缩印或扫描等复制手段保存和汇编本学位论文。

本论文属于　保　密□，在_____年解密后适用本授权书。

　　不保密 ☑。

（请在以上方框内打“ √ ”）

学位论文作者签名：　　指导教师签名：

日期：　　年　　月　　日　　日期：　　年　　月　　日

摘 要

知识蒸馏是将一个神经网络的隐含知识迁移到另一个神经网络的过程，其中提供知识的神经网络称为教师模型，学习知识的神经网络称为学生模型。相比于传统的知识蒸馏模型，自知识蒸馏模型不需要外部的教师模型，而是把神经网络最深层视为教师模型，把神经网络的浅层视为学生模型来达到知识蒸馏的效果。

现有的自知识蒸馏模型（如 BYOT 模型）将作为学生模型的各个浅层块一视同仁，忽略了各个浅层块对最深层块的不同影响。通过分析 BYOT 模型的原理和不足，首先提出了一种基于逐块衰减的改进 BYOT 模型（Per-block Decay based BYOT, PD-BYOT），通过在 BYOT 模型的各浅层块添加构成等比数列的衰减系数，以达到区分各浅层块对最深层块影响的目的。

然后，在 BYOT 模型和所提出的改进 PD-BYOT 模型的基础上，提出了一种新的自知识蒸馏模型，即基于自注意力机制的自知识蒸馏（Self-Knowledge Distillation with Self-Attention Mechanism, SKDSAM）模型。SKDSAM 模型通过在 BYOT 模型的每一个浅层块和最深层块之间添加自注意力连接，计算出相应的注意力权重，从而准确地量化各浅层块对最深层块的不同影响。SKDSAM 还修正了 BYOT 模型的两种损失函数，以便更有效地提取知识蒸馏模型中的暗知识（Dark Knowledge）。随后，从理论上证明了 SKDSAM 模型中的自注意力机制等价于集成学习中的装袋法，证实了 SKDSAM 模型具有更强的稳定性和抗过拟合能力。最后，将 SKDSAM 模型与三种数据增强技术（Cutout、SLA 及 Mixup）相结合，以进一步提升模型的性能。

实验结果表明，PD-BYOT 模型相比于 BYOT 模型提升分类准确率 0.98%。进一步地，SKDSAM 模型在多个图像数据集上取得了相比于现有的其他的自知识蒸馏模型更高的分类准确率。通过消融实验，说明了自注意力机制和自注意力机制中的知识蒸馏模块对提升 SKDSAM 模型性能的作用，以及结合数据增强技术能够进一步提升 SKDSAM 模型的性能。

关键词： 自知识蒸馏；自注意力；数据增强；装袋法

Abstract

Knowledge distillation is the process of transferring implicit knowledge from one neural network to another. The network providing knowledge is called the teacher, and the network receiving knowledge is called the student. Unlike traditional knowledge distillation, the self-knowledge distillation model distills knowledge from the deepest layer (acting as the teacher) to shallow layers (acting as the student) without an outside teacher model.

Current self-knowledge distillation techniques (i.e., BYOT) treat all shallow layers (acting as students) equally, neglecting their different impacts on the deepest layer. Through analyzing the mechanism and shortcomings of BYOT, we first propose the Per-block Decay based BYOT model (PD-BYOT), which utilizes the geometric progression of attenuation coefficient to differentiate each shallow layer's impact on the deepest layer.

On the basis of the model BYOT and the proposed improved model PD-BYOT, we further put forward a novel framework, which we refer to as Self-Knowledge Distillation with Self-Attention Mechanism (SKDSAM). It adds attention links between each shallow layer and the deepest layer of BYOT and computes the attention weight of each shallow layer so as to quantify the contribution of each shallow layer to the deepest layer. The SKDSAM model also modifies two BYOT's loss functions to mine the network's dark knowledge more efficiently. We then provide theoretical proof that the self-attention mechanism in SKDSAM is essentially an ensemble modeling strategy (namely Bagging), which means SKDSAM has the advantage of robustness and preventing overfitting. Moreover, we combine the SKDSAM model with three data augmentation techniques (Cutout, SLA and Mixup) to further improve model performance.

The experimental results show that the PD-BYOT model improves the classification accuracy by 0.98% over the BYOT model. Furthermore, the SKDSAM model outperforms all the self-distillation models on various image datasets. The ablation study proves the importance of the self-attention mechanism and temperature scaling in the self-attention mechanism. The ablation study also shows that combining the SKDSAM model and data augmentation techniques can further improve performance.

Keywords: Self-Knowledge Distillation, Self-Attention, Data Augmentation, Bagging

目 录

1 绪论

1.1 研究背景与意义

人工智能技术对人类的生产生活产生日益广泛而深远的影响。智能导航系统能够为司机规划出耗时最短的路线，避免车主进入繁忙拥堵的车道；人脸识别技术大幅简化了认证流程，使用户免于记忆复杂冗长的密码；智能教育能够为每名学生量身定制专属课程，因材施教提升教学效率；推荐系统准确揣摩用户的喜好，让消费者更容易买到心仪的货品。越来越多的有识之士认为，人工智能是第四次工业革命的先声，是科技竞争的制高点，在未来的经济、军事、教育、医学和生物等诸多领域发挥日益重要的作用。

作为人工智能皇冠上一颗耀眼的明珠，深度神经网络技术近年来获得了迅猛的发展和广阔的应用，在计算机视觉[1][2]、自然语言处理[3]、语音识别[4]等领域都取得了非凡的成果。为了追求越来越优异的性能，深度神经网络的规模与日俱增。然而，大型神经网络的训练需要昂贵的计算资源和时间成本，这使大型神经网络难以应用在计算资源紧缺的场合（比如移动设备或嵌入式设备）。因此，如何对大型神经网络进行压缩成为了研究者们关注的热点。

知识蒸馏是一种压缩大型神经网络的技术。虽然直接训练的小型神经网络性能较低，但是如果能够预先训练一个大型神经网络，再把得到的“知识”迁移到小型神经网络，便可显著提升小型神经网络的性能。受此启发，Hinton 等人[5]提出了知识蒸馏技术：首先训练一个大型的神经网络（称为教师模型），然后把其中的知识压缩到一个小型的神经网络（称为学生模型）上。知识蒸馏模型不仅在学术界常用的数据集上取得了令人惊喜的效果，在商用系统上也已经有了成功的实践。

然而，传统的知识蒸馏技术存在一系列不足之处。首先，预训练大型的教师模型依然需要大量的训练时间。其二，大的教师模型和小的学生模型存在容量上的差异，而且学生模型高度依赖于教师模型。因此，研究人员提出了一系列自知识蒸馏技术，即不利用外部的教师模型，仅利用神经网络自身的知识实现知识蒸馏。自知识蒸馏技

术进一步缩小了神经网络的规模，极大地降低了训练神经网络的时间成本。改进和完善自知识蒸馏模型，可以更有效地挖掘神经网络中隐含的知识，对于将神经网络应用在更广泛的场合具有重要的意义。

1.2 国内外研究现状

1.2.1 知识蒸馏模型

Hinton 等人提出的知识蒸馏模型原理如图 1.1 所示。训练样本分别进入教师模型和学生模型，教师模型的输出经过带有温度系数T的归一化指数层得到软标签，学生模型的输出分别经过带有温度系数T的归一化指数层和不带温度系数的归一化指数层得到软预测和硬预测。蒸馏损失为图 1.1 中软标签和软预测的差异，学生损失为图 1.1 中硬预测和硬标签（数据集中的真实标签）的差异。知识蒸馏的目标是最小化蒸馏损失和学生损失之和，从而使学生模型既获取了数据集中的知识，又得到了教师模型中隐含的知识。

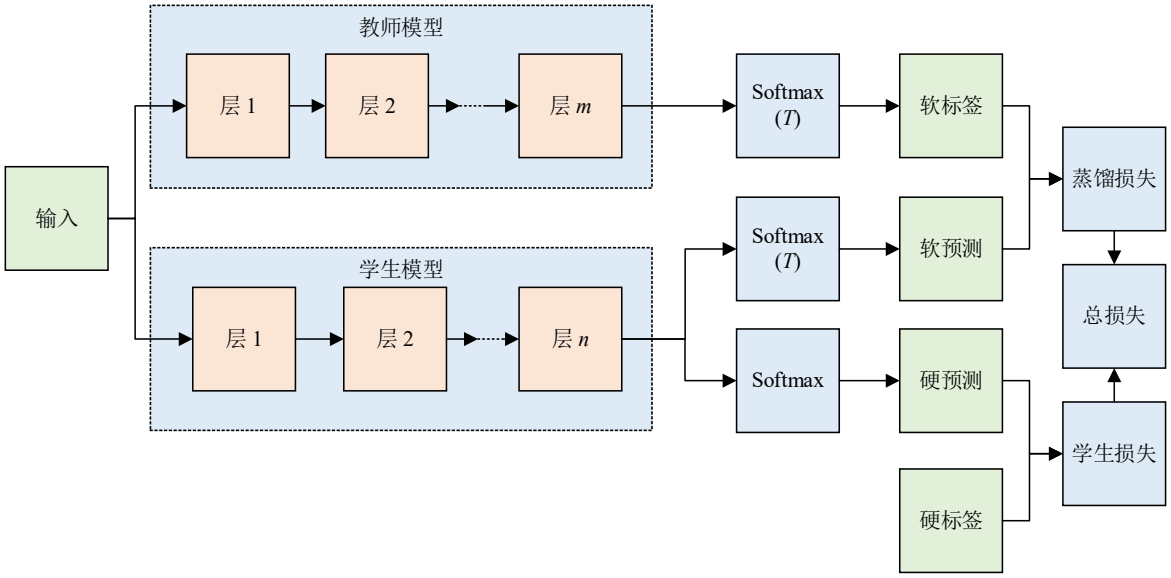

图 1.1 知识蒸馏原理示意

在不使用知识蒸馏的情况下（即$T = 1$），教师模型的输出在正确标签上的概率很高，在错误标签上的概率趋近于零，这使教师模型难以提供近似标签（比如猫和虎、狗和狼）的信息。为了使教师模型能够有效地向学生模型传递近似标签的知识，Hinton 提出了知识蒸馏的策略。假定$z = [z_1, \ldots, z_M]$代表由教师模型输出的向量，T是蒸馏温

度参数，则一张图像的属于第i个类别概率如式（1.1） 所示。

$$p(T)=\frac{exp(z_i/T)}{\sum_j exp(z_j/T)} \tag{1.1}$$

当$T=1$时，得到原始的归一化指数公式（也就是没有蒸馏）。随着T增加，归一化指数函数产生的概率分布更加分散，使教师模型提供更多相似类别的信息。教师模型提供的信息被称为暗知识（Dark Knowledge）。知识蒸馏模型的总损失函数由学生损失函数和蒸馏损失函数相加而成，如式（1.2）所示。

$$\begin{aligned} L &= L_{CE}+\lambda L_{KD} \\ &= CE(P_s,y)+\lambda T^2\cdot KL(P_s(T),P_t(T)) \end{aligned} \tag{1.2}$$

其中函数$CE(\cdot)$代表交叉熵损失函数，$KL(\cdot)$代表相对熵损失函数，P_s代表学生模型预测的概率分布（硬预测），y代表真实标签的独热向量（硬标签），T代表知识蒸馏的温度，$P_t(T)$代表蒸馏温度T时教师模型的概率分布（软标签），$P_s(T)$代表蒸馏温度T时学生模型的概率分布（软预测），λ代表平衡两个损失函数的超参数。

其他研究者进一步分析和改进了 Hinton 所提出的模型。Ba 等人[6]发现浅层神经网络也能够学习深度神经网络表示的复杂函数，并达到以前只有深度神经网络才能达到的精度。Kim 等人[7]提出了类距离损失函数，通过辅助教师模型形成密集聚类的向量空间，从而更有效地提升学生模型的性能。

1.2.2 自知识蒸馏模型

虽然传统的知识蒸馏模型取得了优异的成果，但是它们还是有值得改进的空间。首先，传统的知识蒸馏模型效率低下，因为学生模型很少需要用到教师模型的全部知识。第二，高容量的教师模型的训练过程需要大量的计算和存储资源。为了进一步提升模型的效率，研究者提出了自知识蒸馏模型。自知识蒸馏模型的特点是让神经网络蒸馏自己内部的知识，而不需要借助外部的教师模型。

第一种类型的自知识蒸馏模型是基于特征的自知识蒸馏模型，其原理如图 1.2 所示。由于深度神经网络的深层部分比浅层部分包含更高阶、更抽象的信息，因此可以利用神经网络的深层部分向浅层部分蒸馏暗知识。换言之，基于特征的自知识蒸馏模型通过在不同神经网络层之间添加辅助网络来实现利用深层信息的目的。基于辅助

分类器的自知识蒸馏（Be Your Own Teacher, BYOT）模型[8]根据深度将神经网络分成若干个浅层块和一个最深层块，通过给浅层块增加一系列辅助的分类器来进行自知识蒸馏，以便把暗知识从神经网络深层传递到神经网络浅层。Ji 等人[9]提出了基于自知识蒸馏的特征精炼（Feature Refinement via Self-Knowledge Distillation, FRSKD）模型。FRSKD 模型引入了一个辅助的自教师模块，先将分类器输出的特征输入自教师模块，再将自教师模块提炼出的特征图返回分类器网络。Hou 等人[10]提出了自注意力蒸馏（Self Attention Distillation, SAD）模型，SAD 模型引导神经网络浅层模仿神经网络深层的注意力图，使神经网络浅层学到深层精炼后的信息，从而强化了神经网络的表征能力。Luan 等人[11]将每个块的分支打造成一个分类器，挖掘同一个神经网络中不同分类器的知识从而提升每一个分类器的准确率。

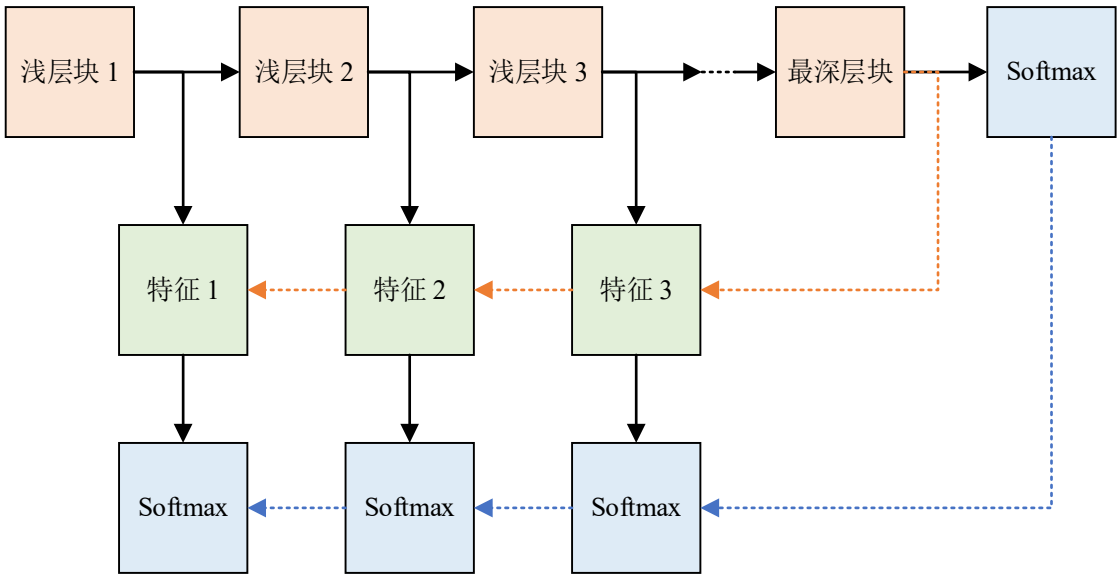

图 1.2　基于特征的自知识蒸馏示意

另一种类型的自知识蒸馏模型是基于训练样本的数据增强，其原理如图 1.3 所示。Lee 等人[12]提出的自监督标签增强（Self-supervised Label Augmentation, SLA）技术使分类器学习一个关于原始标签和自监督标签的联合概率分布，在推理阶段聚合得出预测结果；SLA 在推理阶段和自知识蒸馏相结合（Self-supervised Label Augmentation based Self-Distillation, SLA-SD）以便加速推理过程。Yun 等人[13]提出了按类别预测的正则化方法，并在此基础上提出了类级别的自知识蒸馏（Class-wise Self-Knowledge Distillation, CS-KD）。按类别预测的正则化方法迫使标签相同的不同

样本所输出的概率分布尽可能接近，具有防止过度自信预测的优点。CS-KD 模型将传统知识蒸馏模型中的损失函数由教师模型和学生模型对同一张图像的预测差异替换为对同一类图像的预测差异，从而显著提升了模型性能。Xu 等人[14]提出了基于数据扭曲的自知识蒸馏（Data-Distortion Guided Self-Distillation, DDGSD）模型，通过尽可能缩小输入数据和它的“扭曲版本”后验概率分布之间的差异达到自知识蒸馏的效果。Lee 等人[15]利用了两种类型的数据增强方式（图像旋转和颜色变化），又用集成学习把学生模型的输出聚合到一起，再由学生模型进行自知识蒸馏。

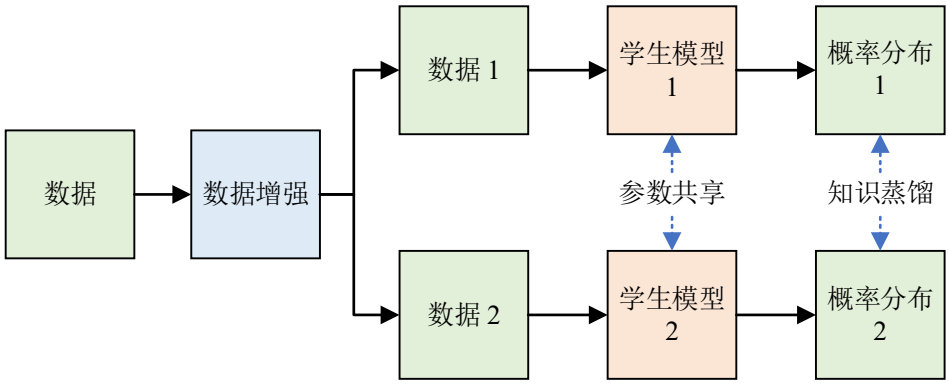

图 1.3　基于数据增强的自知识蒸馏示意

1.2.3　注意力机制

注意力机制最早由 Bahdanau 等人[16]提出，现已成为深度学习中的一种流行的架构。注意力机制已经广泛应用在人工智能的诸多方向，比如自然语言处理、语音识别和计算机视觉。深度学习中的注意力机制可类比于人脑的生物学机制。比如，人类视觉系统倾向于选择性地关注图像的重点部分，同时忽视其他不相关的信息。类似的，在计算机处理语言、语音和图像等任务中，输入信息的某些部分比其他部分更为关键。比如，在机器翻译或摘要提取的任务中，输入文本只有一些关键词汇对预测下文有所帮助；在图像处理任务中，输入图像只有的某些部分和图像标签有所关联（比如一张大熊猫的图像里，大熊猫只占全图像的一部分）。注意力机制能够使模型格外关注输入信息的重点部分，从而使模型更加有效。

在自然语言处理领域，注意力机制在很多任务中都扮演着至关重要的角色，比如机器翻译[17]、语言建模[18]、自然语言推断[19]、智能问答[20]、情感分析[21]、语义分析[22]

和摘要归纳[23]等。Vaswani[17]等人提出的 Transformer 架构具有划时代的意义：它的编码器彻底摒弃了循环神经网络和卷积神经网络，完全基于注意力机制进行建模。Dehghani 等人[18]提出了 Universal Transformer（UT）作为 Transformer 模型的推广，它将前馈序列模型（如 Transformer）的并行性和全局视野与循环神经网络结合起来。Shen 等人[19]提出了定向自注意力网络（Directional Self-Attention Network, DiSAN）以便学习句子嵌入，DiSAN 由具有时间顺序编码的定向自注意力组成，其后继是由序列压缩而成的向量所表示的多维注意力。Xing 等人[20]提出了分层循环注意力网络（Hierarchical Recurrent Attention Network, HRAN），在一个统一的框架中对上下文和关联回答进行建模。HRAN 分别使用单词级别的注意力和句子级别的注意力来提取句子内部和句子之间的重点。

注意力机制同样能够应用在计算机视觉任务中，比如动作识别[24]、图像分类[25][26]、图像生成[27]、目标检测[28]、人物识别[29][30]、分割[31]和显着性检测[32]等。张宇等人[24]提出了一种基于注意力机制的动作识别方法：该方法在数据预处理阶段使用数据增强技术以降低模型过拟合的风险，在特征提取阶段采用结合注意力机制的残差网络来增强模型的特征提取能力。Zhang 等人[27]提出了基于自注意力机制的生成对抗网络，它能够完成需要注意力导向和远距离依赖的图像生成任务。Kong 等人[28]提出了一种新颖的特征配置架构，以高度非线性但有效的方式将低级表征与高级语义特征相结合。此架构由全局注意力和局部重构组成，能够在全局和局部的不同空间位置和尺度上收集面向任务的特征。Li 等人[29]提出了 Harmonious Attention CNN（HA-CNN）模型，通过联合学习软像素注意力和硬区域注意力以及同时优化特征表示，以便改善对失控图像中人员的识别。

对于注意力机制的网络结构也有了一系列的探讨[33][34]。Huang 等人[33]提出了密集隐式注意力网络（Dense-and-Implicit Attention Network, DIANet）[33]。已有的注意力模型将注意力机制插入深度神经网络的每一层，而 DIANet 模型则在不同的网络层中共享一个注意力机制，从而减少了神经网络的参数，加强了网络层之间信息的集成。Zhao 等人[34]探讨了自注意力网络结构两种类型的变体。第一种变体是成对自注意力。成对注意力与卷积有两点不同：第一，成对注意力本质上是一个集合算子，而

不是一个序列算子；第二，成对注意力不会给特定位置附加固定的权重，并且对排列和基数具有不变性。这使得自注意力的“足迹”可以增加甚至变得不规则，而不会对参数数量产生影响。第二种变体是互补注意力。互补注意力是卷积的一种推广，它不具有成对注意力的排列和基数不变性，但具有比原始残差网络更强大的性能。

1.2.4 卷积网络模型

由于后续的章节需要代码实现所提出的模型，本小节对实现模型所需的卷积网络框架进行简单的论述。

卷积神经网络已成为计算机图像识别的主要方法。随着软硬件水平的提升，卷积神经网络的层数越来越多，表征能力越来越强。然而，一方面受困于梯度消失和爆炸问题，训练深层神经网络面临严重的困难；另一方面，随着网络层数的增加，神经网络的性能开始饱和甚至退化。为此，研究者对卷积网络提出了一系列改进，包括深度残差网络[35]、宽残差网络[36]、残差密集网络[37]等等。

（1）深度残差网络

何恺明等人[35]提出了深度残差网络（Deep Residual Networks, ResNet）来解决深度卷积神经网络的性能退化问题。ResNet 不是令每个网络层直接适配所需的底层映射，而是令这些网络层去适配残差映射。记所需的底层映射为$H(x)$，ResNet 让堆叠的网络层拟合另一个映射$F(x):=H(x)-x$，原始映射因此表示为$F(x)+x$。这种做法基于以下假设：优化残差映射比优化原始映射容易。在最极端的情况下，假设恒等映射是最优解，那么将残差直接设为零要比通过一系列非线性神经网络层拟合恒等映射更容易。

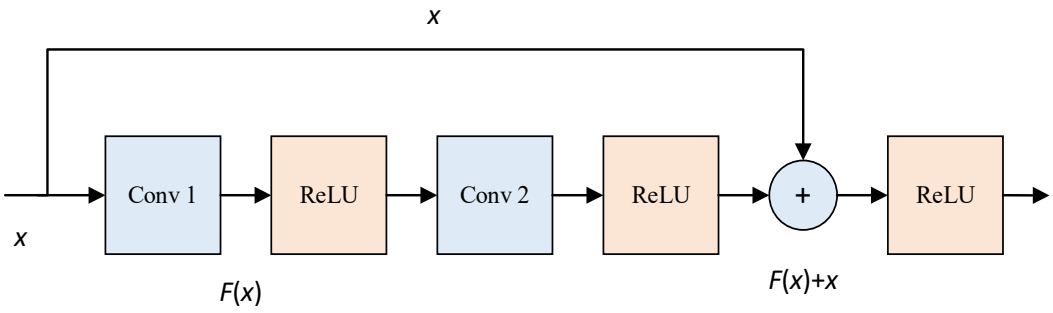

图 1.4　残差块的结构示意

一个残差块的结构如图 1.4 所示，$F(x)+x$能够通过具有“捷径连接”的前馈神经网络实现，其中捷径连接是指跳过一层或多层神经网络的连接。在 ResNet 中，捷径连接只是简单地实现恒等映射，它的输出被添加到所跨越的网络层的最终输出中。恒等捷径连接既不给卷积网络增加额外的参数，也不增加卷积网络的计算复杂性。实验结果证实，ResNet 比同等规模的传统卷积网络更容易优化；随着神经网络层数的增加，ResNet 性能的提升比传统卷积网络更为显著。

（2）宽残差网络

ResNet 被证明能够扩展到数千层并且仍然具有提升性能的空间。然而，每提高百分之一的性能需要增加几乎一倍的神经网络层数，这使训练 ResNet 非常缓慢。尽管许多研究者认为增加神经网络的深度比扩展神经网络的宽度更有效，Zagoruyko 等人发现[36]如果使用合适的架构，扩展残差块的宽度相比增加神经网络的深度更能有效提升神经网络的性能。他们据此提出了宽残差网络（Wide Residual Networks, WRN），其所需的层数仅仅是相同性能 ResNet 的 2%，训练速度比相同性能 ResNet 快 2 倍。这证实了 ResNet 中起主要作用的是残差块，而更多的神经网络层数仅起到辅助作用。

WRN 的结构如表 1.1 所示，按前馈顺序依次是原始的卷积层（卷积层 1）、三个残差网络层（卷积层 2、卷积层 3、卷积层 4）、平均池化层。其中 WRN 的宽度由因子k决定，方括号中的内容代表一组又一组的卷积结构，乘数N代表一组中残差块的个数。

表 1.1　WRN 的结构示意

组名称	组输出维数	组内部结构
卷积层 1	32×32	$[3\times3, 16]$
卷积层 2	32×32	$\begin{bmatrix}3\times3, & 16\times k\\ 3\times3, & 16\times k\end{bmatrix}\times N$
卷积层 3	16×16	$\begin{bmatrix}3\times3, & 32\times k\\ 3\times3, & 32\times k\end{bmatrix}\times N$
卷积层 4	8×8	$\begin{bmatrix}3\times3, & 64\times k\\ 3\times3, & 64\times k\end{bmatrix}\times N$
平均池化层	1×1	$[8\times8]$

（3）残差密集网络

研究表明，如果将卷积网络靠近输入的层和靠近输出的层连接起来，则卷积网络

的性能会显著提升。黄高等人[37]根据这一观察提出了残差密集卷积网络（Densely Connected Convolutional Networks, DenseNet），它将神经网络中的每一层都连接到前馈方向的所有网络层。DenseNet 缓解了梯度消失问题，促进了特征传播，加强了特征重用，并大大减少了神经网络所需参数的数量。

DenseNet 的结构如图 1.5 所示，为了最大化神经网络中各层之间的信息流，DenseNet 将所有神经网络层直接相互连接。DenseNet 中的每一神经网络层都从所有前面的神经网络层获得额外的输入，并将其自己的特征图传递给所有后续的神经网络层。不同于 ResNet，DenseNet 在将特征传递到神经网络层之前不是通过求和来组合特征，而是通过连接来组合特征。因此，某一神经网络层的输入由它所有先前神经网络层的特征图组成。DenseNet 具有以下优点：由于 DenseNet 不需要冗余的特征图，所以 DenseNet 相比传统的卷积网络需要更少的参数；由于 DenseNet 提升了神经网络层间信息流的传递，使每一层都能够直接接触到损失函数的导数和原始输入信号，所以 DenseNet 的训练更为容易；由于密集连接具有正则化的功能，所以 DenseNet 训练小模型时具有更强的抗过拟合能力。

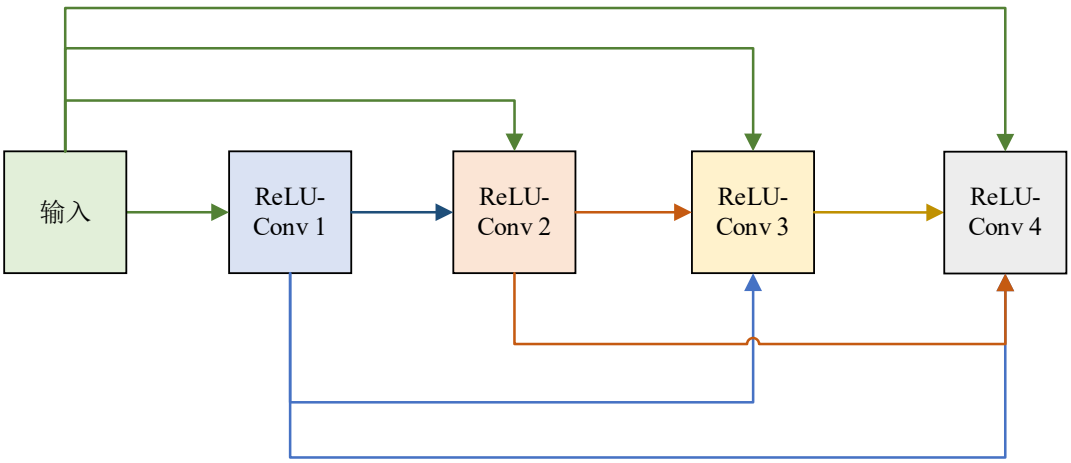

图 1.5　DenseNet 的结构示意

1.3　论文主要内容

基于特征的自知识蒸馏模型把神经网络的深层块视为教师模型，把神经网络的

浅层块视为学生模型，用神经网络深层块的信息来辅助训练神经网络的浅层块。然而，现有的基于特征的自知识蒸馏模型（如 BYOT 模型）将作为学生模型的各个浅层块的特征一视同仁，忽略了各浅层块特征图对最深层块特征图的不同影响。

为了提升 BYOT 模型的性能，本文首先提出了基于逐块衰减辅助分类器的自知识蒸馏（Per-block Decay based Be Your Own Teacher, PD-BYOT）模型，通过给 BYOT 模型的各浅层块添加衰减系数将各个浅层块对最深层块的影响以等比数列的形式加以区分。随后，进一步提出了一种新的基于自注意力机制的自知识蒸馏（Self-Knowledge Distillation with Self-Attention Mechanism, SKDSAM）模型。SKDSAM 模型能够计算出各个浅层块特征图重要性的差异，并根据重要性的差异赋予它们不同的注意力权重。实验结果显示 SKDSAM 模型比其他自蒸馏模型和自注意模型具有更优异的分类准确率。

论文共分为五章内容，章节内容之间的关系如图 1.6 所示。

第一章首先介绍知识蒸馏模型的背景和意义，接下来论述了与本文内容相关的（包括知识蒸馏、注意力机制和卷积神经网络）的国内外研究现状，最后说明文章的组织结构。

第二章论述了 BYOT 模型的网络结构和损失函数，分析了 BYOT 模型存在的不足，然后通过增加衰减系数提出了改进的 PD-BYOT 模型。

第三章在 BYOT 模型和 PD-BYOT 模型的基础上，提出了新模型 SKDSAM；详细说明了 SKDSAM 模型的网络结构、损失函数和训练流程，说明 SKDSAM 模型相比 BYOT 模型和 PD-BYOT 模型所做的改进；随后证明 SKDSAM 模型和装袋法[38]的等价性；最后将 SKDSAM 模型分别和数据增强技术 Cutout[39]、SLA[12]及 Mixup[40]结合起来。

第四章首先比较了 PD-BYOT 模型和原 BYOT 模型的实验性能；然后在多个数据集上（CIFAR-100[41]、Tiny ImageNet[42]、Caltech-UCSD Birds 200[43]、Stanford 40 Actions[44]、Stanford Dogs[45]和 MIT Indoor Scene Recognition[46]）测试了 SKDSAM 模型的分类准确率，并和几种当前流行的自知识蒸馏模型（DDGSD 模型、BYOT 模型、CS-KD 模型、SLA-SD 模型和 FRSKD 模型）和自注意力模型（DIANet 模型和 SAN

模型）对比了实验性能；最后在消融实验中，分别测试了移除自注意力机制和移除自注意力机制中知识蒸馏模块对于模型性能的影响，测试了结合数据增强技术后的SKDSAM 模型性能，测试了超参数的其它取值对实验结果的影响。

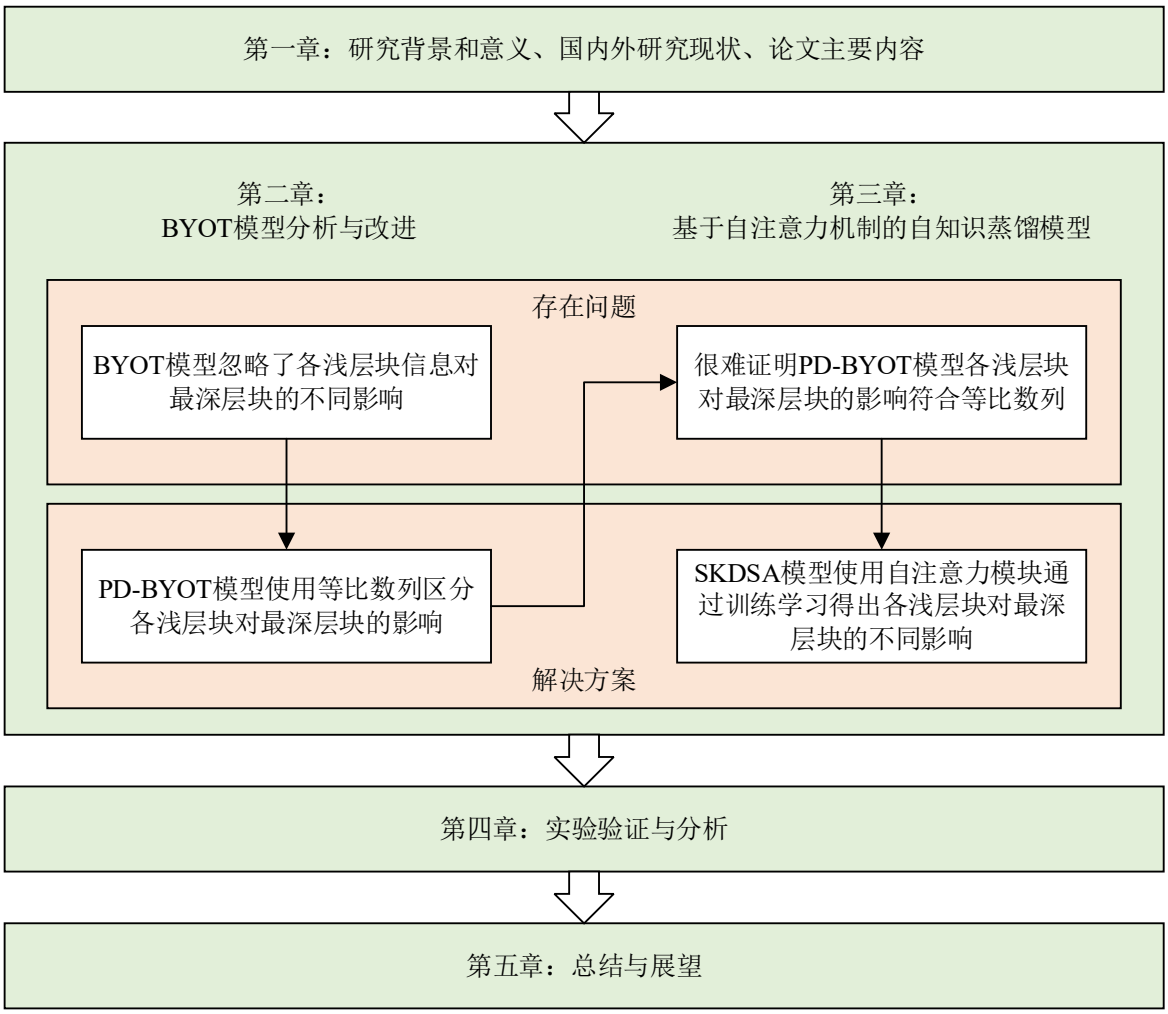

图 1.6　组织结构图

第五章总结了全文的主要工作，列举了所提出模型的主要创新点，并展望所提出模型在未来可能的改进方向。

2 BYOT 模型分析与改进

本章首先论述了 BYOT 模型的网络结构，说明了 BYOT 模型损失函数的计算方法；随后分析了 BYOT 模型存在的不足并提出初步的改进方案，即基于逐块衰减辅助分类器的自知识蒸馏（Per-block Decay based Be Your Own Teacher, PD-BYOT）模型。

2.1 BYOT 模型分析

2.1.1 BYOT 模型的网络结构

基于辅助分类器的自知识蒸馏（Be Your Own Teacher, BYOT）模型[8]是一种基于特征的自知识蒸馏模型，它将神经网络的深层信息蒸馏到神经网络浅层。BYOT 模型的网络结构如图 2.1 所示：神经网络根据其深度划分为几个浅层块（浅层块 1、浅层块 2 和浅层块 3）和一个最深层块，在每个浅层块之下设置一个瓶颈层，瓶颈层下再结合一个全连接层和加入温度系数的归一化指数层。

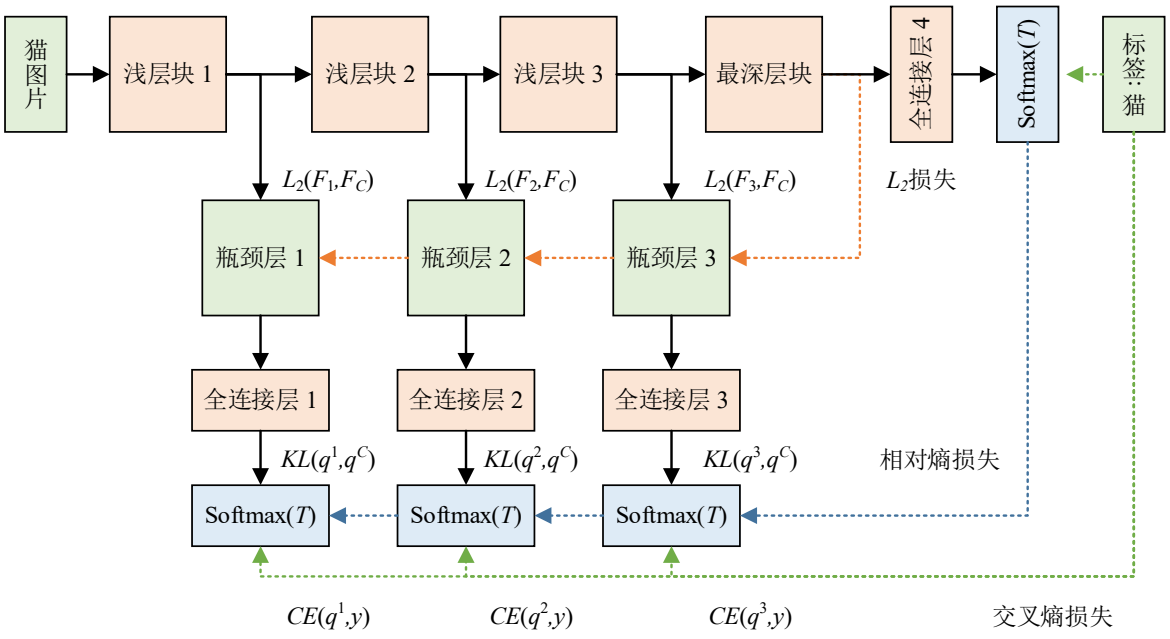

图 2.1 BYOT 模型的网络结构示意

每个浅层块和相应的瓶颈层、全连接层、加入温度系数的归一化指数层组成一个浅层分类器，最深层块和其后的全连接层、加入温度系数的归一化指数层组成最深层

分类器。添加瓶颈层的目的是减轻每个浅层块之间的影响，并增加来自最深层信息的引导。在训练阶段，BYOT 模型将所有浅层分类器视为学生模型，将最深层分类器视为教师模型，将最深层分类器的暗知识蒸馏到每一个浅层分类器。在推理阶段，将每个浅层块之下添加的瓶颈层、全连接层和归一化指数层全部移除。

2.1.2 BYOT 模型的损失函数

记N个样本组成的数据集为$X = \{x_i\}_{i=1}^{N}$，其中$x_i \in \{1,2,\ldots,N\}$。记数据集中所有类组成的集合为$Y = \{y_i\}_{i=1}^{M}$，其中$y_i \in \{1,2,\ldots,M\}$。记 BYOT 模型中的各个分类器为$\Theta = \{\theta_{i/C}\}_{i=1}^{C}$，其中$C$代表卷积网络层中分类器（包括最深层分类器和所有浅层分类器）的个数。记z^c为通过第$\theta_{c/C}$个浅层块下面的全连接层后的输出，则它再通过加入温度系数的归一化指数层后，所输出向量属于第i个类别的概率如式（2.1）所示。

$$q_i^c = \frac{exp(z_i^c/T)}{\sum_j^M exp(z_j^c/T)} \tag{2.1}$$

其中T代表分类器中的蒸馏温度，初始值为 1。

BYOT 模型在训练过程中引入了三种损失函数：真实标签的独热向量与每个分类器（包括最深层分类器和所有浅层分类器）输出的交叉熵，最深层分类器输出的概率分布和每个浅层分类器输出概率分布的相对熵，最深层分类器特征图和每个浅层分类器特征图的L_2损失函数。

BYOT 模型的第一种损失函数是真实标签的独热向量与每个分类器（包括最深层分类器和所有浅层分类器）输出概率分布的交叉熵。BYOT 模型通过交叉熵把数据集中的知识从真实标签引入到模型中的所有分类器。

BYOT 模型的交叉熵损失函数由式（2.2）表示，其中q^i代表某样本经过第i个浅层分类器输出的概率分布，y代表该样本在数据集对应真实标签的独热向量，$CE(\cdot)$代表交叉熵损失函数。

$$L_{BYOT-CE} = CE(q^i, y) \tag{2.2}$$

BYOT 模型的第二种损失函数是最深层分类器输出的概率分布和每个浅层分类器输出的概率分布的相对熵。BYOT 模型通过相对熵将最深层分类器中的暗知识传递到每个浅层分类器，引导浅层分类器模仿最深层分类器的预测结果。

BYOT 模型的相对熵由式（2.3）表示，其中q^i代表第i个浅层分类器输出的概率分布，q^C代表最深层分类器输出的概率分布，$KL(\cdot)$代表相对熵损失函数。

$$L_{BYOT-KL} = KL(q^i, q^C) \tag{2.3}$$

BYOT 模型的第三种损失函数是最深层分类器特征图和每个浅层分类器特征图的L_2损失函数。BYOT 模型通过L_2损失函数将最深层分类器特征图中的信息引入每个浅层分类器，从而引导所有的浅层分类器的特征图去模仿最深层分类器的特征图。

BYOT 模型的L_2损失函数由式（2.4）表示，其中F_i表示第i个浅层分类器的特征图，F_C表示最深层分类器的特征图，$||\cdot||_2^2$代表L_2损失函数。

$$L_{BYOT-feature} = ||F_i - F_C||_2^2 \tag{2.4}$$

综合上述三种损失函数（式（2.2）、（2.3）和（2.4）），再将所有分类器（包括最深层分类器和所有浅层分类器）的损失函数相加，即可得到 BYOT 模型的整体的损失函数，如式（2.5）所示，其中α和λ是平衡式（2.2）、（2.3）和（2.4）的超参数，L_{BYOT}代表 BYOT 模型的总损失函数。

$$L_{BYOT} = \sum_{i=1}^{C}((1-\alpha)\cdot CE(q^i, y) + \alpha \cdot KL(q^i, q^C) + \lambda \cdot ||F_i - F_C||_2^2) \tag{2.5}$$

2.2 BYOT 模型改进

在卷积神经网络的图像处理任务中，可将卷积神经网络的中间层视为特征提取器，其中网络浅层提取图像的边缘信息，网络深层提取图像中更高维、更抽象的信息。然而在计算损失函数时，BYOT 模型将各个作为学生模型的不同浅层块的信息一视同仁，可能会造成一些暗知识的损失。因此，区分各浅层块对最深层块的不同的贡献度很有必要。

2.2.1 PD-BYOT 模型的网络结构

为了区分各浅层块对最深层块的不同影响，提出基于逐块衰减辅助分类器的自知识蒸馏（Per-block Decay based Be Your Own Teacher, PD-BYOT）模型。PD-BYOT 模型的网络结构如图 2.2 所示：卷积网络根据深度被分成几个浅层块和一个最深层块，在每个浅层块下方添加一个瓶颈层，瓶颈层下再结合一个全连接层和加入温度系

数的归一化指数层。和原 BYOT 模型网络结构（图 2.1）不同的是，PD-BYOT 模型在每个浅层块和最深层块之间的三种损失函数上各添加一个成等比数列的衰减系数（分别记为d_{CE}、d_{KL}和d_{feture}）。

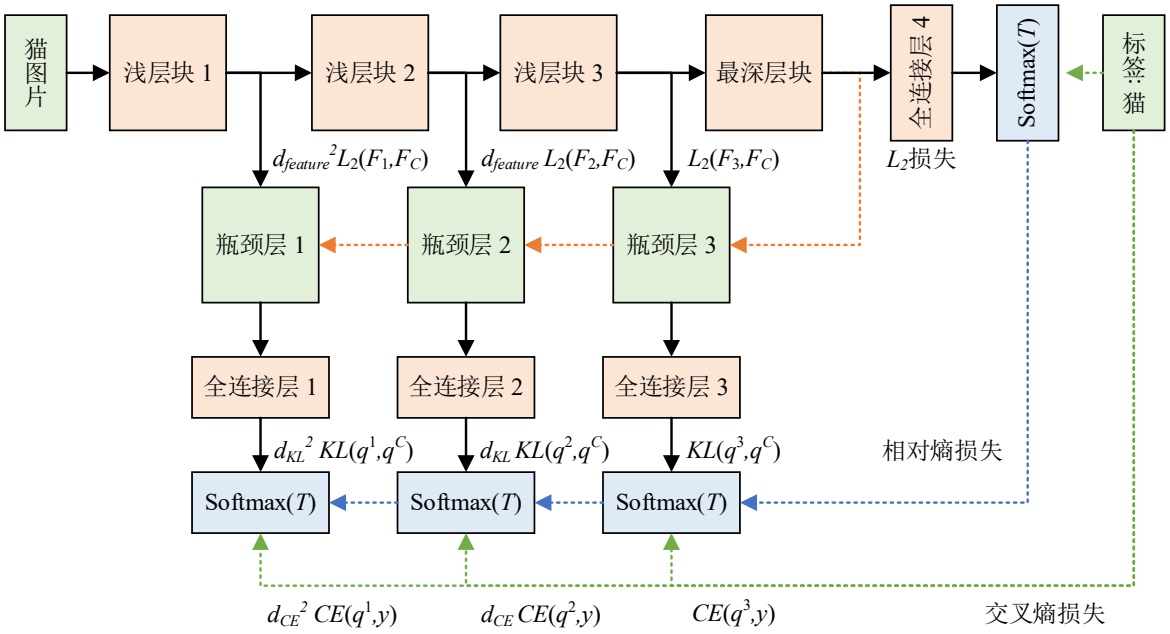

图 2.2　PD-BYOT 模型的网络结构示意

2.2.2　PD-BYOT 模型的损失函数

为了使每个浅层块的损失函数对总损失函数的影响都有所不同，PD-BYOT 模型给每个浅层块的交叉熵、相对熵和L_2损失函数分别添加一个衰减系数d_{CE}、d_{KL}和d_{feture}，它们都随浅层块的序数以等比数列的形式与原有的损失函数相乘。

使用衰减系数d_{CE}修正后的第i个块的交叉熵如式（2.6）所示。

$$L_{PDBYOT-CE} = d_{CE}^{C-i} \cdot CE(q^i, y) \tag{2.6}$$

使用衰减系数d_{KL}修正后的第i个块的相对熵如式（2.7）所示。

$$L_{PDBYOT-KL} = d_{KL}^{C-i} \cdot KL(q^i, q^C) \tag{2.7}$$

使用衰减系数d_{feture}修正后的第i个块的L_2损失函数如式（2.8）所示。

$$L_{PDBYOT-feature} = d_{feture}^{C-i} \cdot ||F_i - F_C||_2^2 \tag{2.8}$$

综合式（2.6）、（2.7）和（2.8），得到 PD-BYOT 模型的总损失函数如式（2.9）所示。

$$L_{PDBYOT} = (1-\alpha)\sum_{i=1}^{C} d_{CE}^{C-i} \cdot CE\left(q^i, y\right) + \alpha \sum_{i=1}^{C} d_{KL}^{C-i} \cdot KL\left(q^i, q^C\right) + \lambda \sum_{i=1}^{C} d_{feture}^{C-i} \cdot ||F_i - F_C||_2^2 \quad (2.9)$$

2.2.3 PD-BYOT 模型的训练步骤

PD-BYOT 模型的训练步骤如算法 2.1 所示。输入为训练集数据和随机初始化的模型参数，输出为使总损失函数最小化的模型参数。

算法 2.1 PD-BYOT 的训练步骤

输入：训练集D，随机初始化的模型参数θ

输出：使总损失函数最小化的模型参数θ

1: **while** 模型参数θ尚未收敛 **do**

2: 从训练集D中挑出一批数据记为B

3: 将B放入 PD-BYOT 模型前向传播

4: $L_{PDBYOT-CE} \leftarrow d_{CE}^{C-i} \cdot CE\left(q^i, y\right)$ /*式（2.6）*/

5: $L_{PDBYOT-KL} \leftarrow d_{KL}^{C-i} \cdot KL\left(q^i, q^C\right)$ /*式（2.7）*/

6: $L_{PDBYOT-feature} \leftarrow d_{feture}^{C-i} \cdot ||F_i - F_C||_2^2$ /*式（2.8）*/

7: $L_{PDBYOT} \leftarrow (1-\alpha) \cdot L_{PDBYOT-CE} + \alpha \cdot L_{PDBYOT-KL} + \lambda \cdot L_{PDBYOT-feature}$ /*式（2.9）*/

8: 为了最小化L_{PDBYOT}，对L_{PDBYOT}使用随机梯度下降法，更新模型参数θ

9: **end while**

10: **return** θ

由于庞大的训练数据集难以一次全部装入显存之中，所以事先将训练集分为多个批次，每轮循环将一个批次的数据装入显存进行计算。在一次循环里的前向传播阶段，通过输入的训练集数据和现有的模型参数，依次根据式（2.6）、（2.7）、（2.8）和（2.9）计算出模型的总损失函数。在一次循环里的反向传播阶段，对总损失函数使用随机梯度下降法，更新模型参数。直到总损失函数收敛到一个区间范围内，停止

循环。

2.3 本章小结

本章第 2.1 节论述了 BYOT 模型的网络结构和损失函数，它是一种典型的基于特征的自知识蒸馏方法。第 2.2 节分析了 BYOT 模型的不足，它忽略了各个浅层块信息对最深层块的不同影响，并提出了 PD-BYOT 模型：通过给 BYOT 模型的各浅层块添加衰减系数将各个浅层块对最深层块的影响以等比数列的形式加以区分，为第三章提出进一步的改进方案做准备。

3 基于自注意力机制的自知识蒸馏模型

在第二章中，基于辅助分类器的自知识蒸馏（Be Your Own Teacher, BYOT）模型将各个浅层块的信息一视同仁，忽略了各浅层块信息对最深层块的不同影响；而基于逐块衰减辅助分类器的自知识蒸馏（Per-block Decay based Be Your Own Teacher, PD-BYOT）模型则把各个浅层块对最深层块的影响以等比数列的形式加以区分，很难证明其与各个浅层块对最深层块的真实影响相符合。

因此，本章提出给 BYOT 模型增加自注意力机制，使自注意力机制通过训练“学习”得出各个浅层块对最深层块的不同贡献度，从而使不同深度网络层的信息能够更有效地聚合。称这种新的自知识蒸馏模型为基于自注意力机制的自知识蒸馏（Self-Knowledge Distillation with Self-Attention Mechanism, SKDSAM）模型。

3.1 典型自注意力机制的网络结构

典型自注意力机制的网络结构如图 3.1 所示。自注意力机制的输入为$Query$、$Key_i\ (1 \le i \le n)$、$Value_i\ (1 \le i \le n)$，其中n为Key_i和$Value_i$的个数。三者的概念由信息检索引申而来，比如，当在爱奇艺网站检索视频时，搜索框中输入的查询词汇为$Query$，爱奇艺所有的视频为Key，按相关性检索出的结果为$Value$。自注意力机制的计算流程同样可视为信息检索的流程。

对于某个Key_i，首先计算出它和$Query$的相似度（一般为两者点积），如式（3.1）所示。

$$sim = Query \cdot Key_i \tag{3.1}$$

再对式（3.1）得出的相似度使用归一化指数函数，计算相应的注意力权重，如式（3.2）所示。

$$a_i = \text{softmax}(Query \cdot Key_i) = \frac{exp(Query \cdot Key_i)}{\sum_{j=1}^{n} exp(Query \cdot Key_j)} \tag{3.2}$$

最后将$Value_i$与相应的注意力权重相乘后再累加，即可得到自注意力机制的输出，如式（3.3）所示。

$$Output = \sum_{i=1}^{n} Value_i \cdot a_i = \sum_{i=1}^{n} Value_i \cdot \frac{exp(Query \cdot Key_i)}{\sum_{j=1}^{n} exp(Query \cdot Key_j)} \quad (3.3)$$

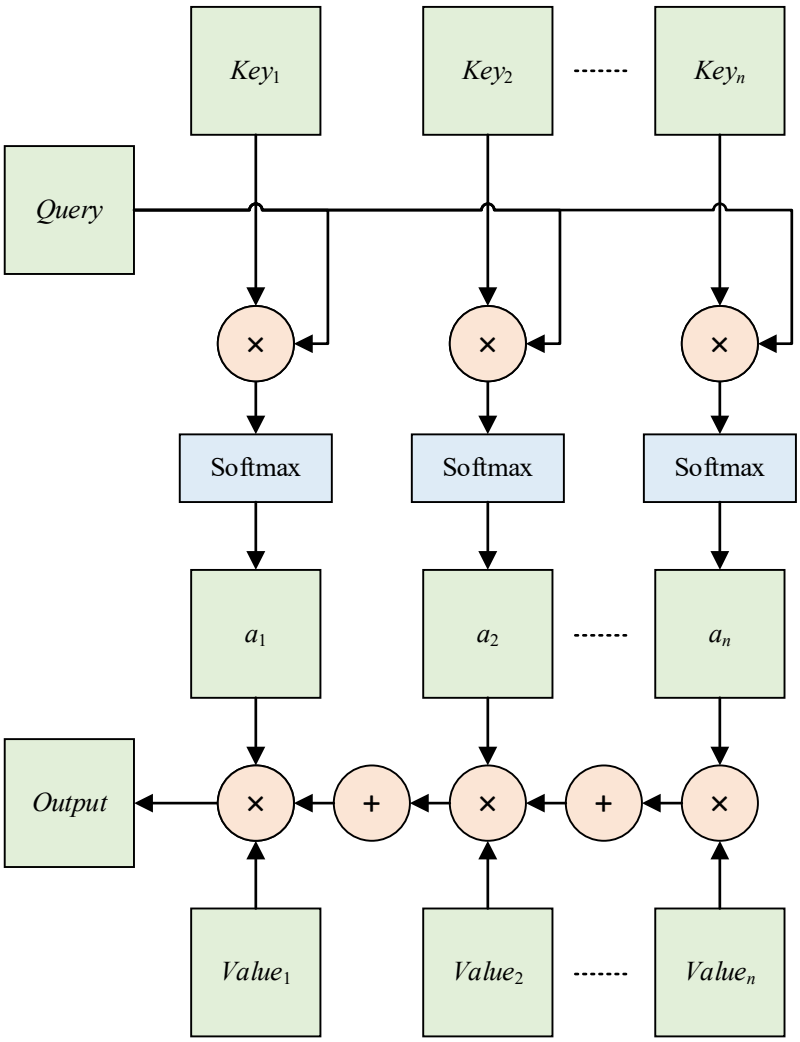

图 3.1　自注意力机制的网络结构示意

3.2　基于自注意力机制的 SKDSAM 模型

3.2.1　SKDSAM 模型的网络结构

为了准确量化 BYOT 模型中各个浅层块对最深层块的不同贡献度，受注意力机制在网络结构中应用[33][34]的启发，本节将自注意力机制和自知识蒸馏模型相结合，提出了一种基于自注意力机制的自知识蒸馏模型来有效区分自知识蒸馏模型中各个浅层块信息的不同重要性。

SKDSAM 模型的网络结构如图 3.2 所示：卷积网络根据深度被分成几个浅层块和一个最深层块，在每个浅层块下方添加一个瓶颈层，瓶颈层下再结合一个全连接层和加入温度系数的归一化指数层；在每个浅层分类器和最深层分类器之间加上自注意力连接，作用是计算每一个浅层块特征图和最深层块特征图的相似度。自注意力机制的网络结构将在第 3.2.2 小节详述。

每个浅层块和相应的瓶颈层、全连接层、加入温度系数（默认值为 4）的归一化指数层组成一个浅层分类器，最深层块和其后的全连接层、加入温度系数（默认值为 4）的归一化指数层组成最深层分类器。在训练阶段，SKDSAM 模型将所有浅层分类器视为学生模型，将最深层分类器视为教师模型，将最深层分类器的暗知识蒸馏到每一个浅层分类器。在推理阶段，将每个浅层块之下添加的瓶颈层、全连接层和归一化指数层全部移除，将最深层块和每个浅层块之间的自注意力连接也全部移除。

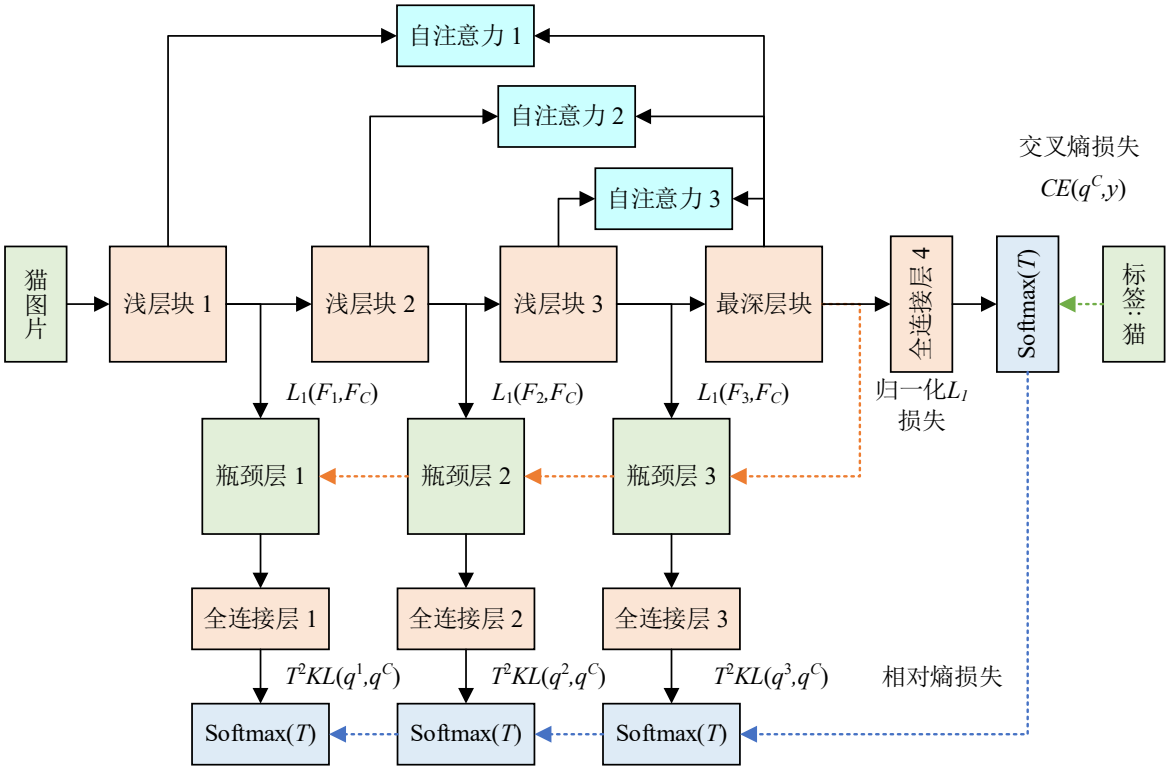

图 3.2　SKDSAM 模型的网络结构示意

对比将 SKDSAM 模型的网络结构（图 3.2）和 BYOT 模型的网络结构（图 2.1），SKDSAM 模型比 BYOT 模型增加了自注意力机制。SKDSAM 模型通过自注意力机制把最深层块的特征图和浅层块的特征图联系起来，以便提升自蒸馏模型的稳定性

和泛化能力。

3.2.2 SKDSAM 模型中的自注意力机制

类似于图 3.1 中自注意力机制的网络结构，SKDSAM 中自注意力机制的网络结构如图 3.3 所示。记N个样本组成的数据集为$X = \{x_i\}_{i=1}^{N}$，其中$x_i \in \{1,2,\ldots,N\}$。记所有类组成的集合为$Y = \{y_i\}_{i=1}^{M}$，其中$y_i \in \{1,2,\ldots,M\}$。记 SKDSAM 模型中的各个分类器为$\Theta = \{\theta_{i/C}\}_{i=1}^{C}$，其中$C$代表卷积网络层中分类器（包括最深层分类器和所有浅层分类器）的个数。记z^c为通过第$\theta_{c/C}$个浅层块下面的全连接层后的输出，则它再通过加入温度系数的归一化指数层后，则所输出向量属于第i个类别的概率如式（3.4）所示。

$$q_i^c = \frac{exp(z_i^c/T)}{\sum_j^M exp(z_j^c/T)} \tag{3.4}$$

其中T代表分类器中的蒸馏温度，默认值为 4。将式（3.4）记为$soft(\cdot)$函数，即蒸馏函数。

如图 3.3 所示，自注意力机制的输入为模型中所有分类器（包括最深层分类器和所有浅层分类器）的特征图。记第i个浅层分类器的特征图为F_i，最深层分类器的特征图为F_C。在浅层分类器的特征图F_i和最深层分类器的特征图F_C后分别添加一个非线性投影层，分别得到结果$proj_i(F_i)$和$proj_C(F_C)$。使用投影层的目的有两点，一是为了提取出特征图里更高阶、更本质的特征，二是为了使$Query$和Key_i的维数相同（记为D_C），方便后续计算它们的相似度。

记特征图通过投影层后的输出向量为$p = (p_1, p_2, \ldots, p_{D_C})$，类似于式（3.4）中对知识蒸馏的定义，投影层输出向量的蒸馏函数定义如式（3.5）所示。

$$soft(p) = \frac{exp(p_i/T')}{\sum_j^{D_c} exp(p_j/T')} \tag{3.5}$$

其中T'是对自注意力机制的投影层输出向量的蒸馏温度。对投影层输出的蒸馏温度T'和对辅助分类器输出的蒸馏温度T是彼此独立的。在第四章的第 4.4.2 小节将分别探讨T和T'对 SKDSAM 模型性能的影响。

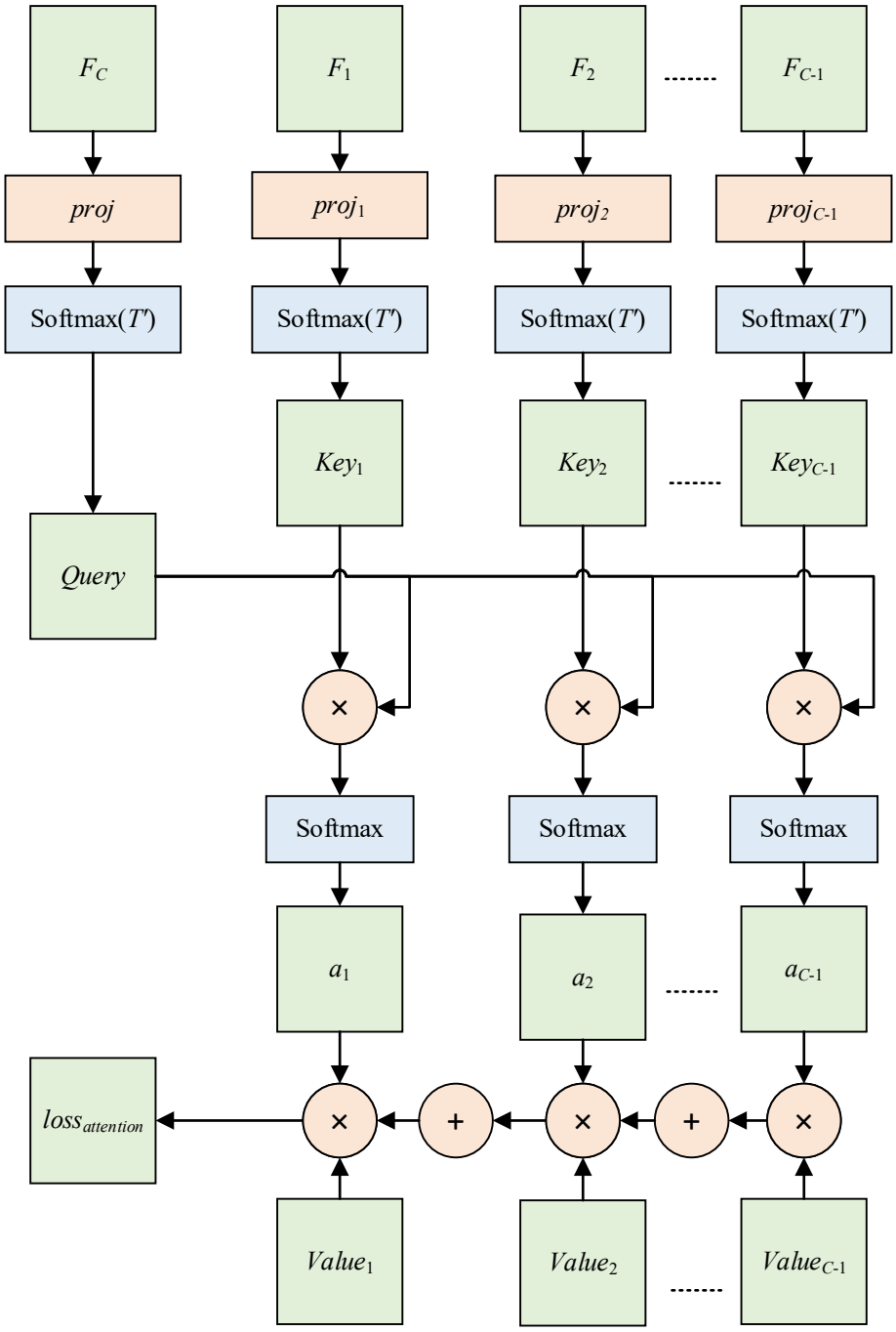

图 3.3　SKDSAM 自注意力机制的网络结构示意

根据 3.1 节的内容，求解自注意力机制需要$Query$、Key_i、$Value_i$。对所有投影层输出的结果统一使用蒸馏函数（式（3.5））, 得到 SKDSAM 模型中最深层分类器的输出和第i个浅层分类器的输出，它们分别对应自注意力机制的$Query$和Key_i，如式（3.6）所示。

$$Query = soft\big(proj_C(F_C)\big)$$
$$Key_i = soft(proj_i(F_i)) \qquad (3.6)$$

其中$1 \le i < C$。自注意力机制需要的$Value_i$将根据下文的式（3.11）求取。

利用$Query$和Key_i的乘积计算出它们的相似度，再对所有相似度应用归一化指数函数，即可得到第i个浅层分类器和最深层分类器注意力连接的对应的权重a_i，如式（3.7）所示。

$$a_i = \mathrm{softmax}(Query \cdot Key_i) = \frac{exp(\,Query \cdot Key_i)}{\sum_{j=1}^{C-1} exp(\,Query \cdot Key_j)} \qquad (3.7)$$

其中$1 \le i < C$。根据归一化指数函数的定义，易知所有浅层分类器的注意力权重之和满足$\sum_{i=1}^{C-1} a_i = 1$。

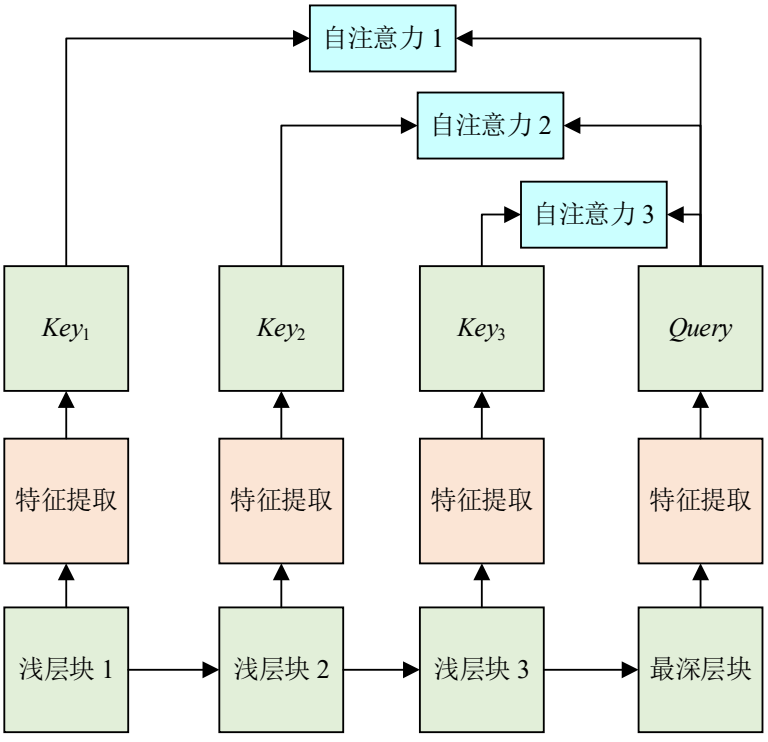

图 3.4　SKDSAM 模型中自注意力机制的直观理解

对 SKDSAM 模型的直观理解如图 3.4 所示，它在第i $(1 \le i < C)$个浅层块（学生模型）和最深层块（教师模型）之间增加一个自注意力连接，经特征提取后得到自注意力机制的输入$Query$和Key_i；然后通过$Query$和Key_i计算出第i个浅层块对应的注意力权重a_i。计算得出的注意力权重a_i即代表第i个浅层块对于最深层块的重要性

的量化，从而解决了本章一开始提出的问题。

3.2.3 SKDSAM 模型的损失函数

类似于 BYOT 模型，SKDSAM 模型也引入了三种损失函数：真实标签的独热向量与最深层分类器输出概率分布的交叉熵、最深层分类器和每个浅层分类器的相对熵、最深层分类器特征图和每个浅层分类器特征图的归一化L_1损失函数。为了追求更优异的实验性能，SKDSAM 模型修改 BYOT 模型的三种损失函数，两者的差别如表 3.1 所示。

表 3.1 BYOT 模型和 SKDSAM 模型损失函数的区别

损失函数	BYOT 模型	SKDSAM 模型
交叉熵	计算所有分类器	只计算最深层分类器
相对熵	直接计算相对熵	相对熵乘以温度的平方
L_n损失函数	L_2损失函数	归一化L_1损失函数
总损失函数	与超参数相乘后直接相加	与注意力权重相乘

SKDSAM 模型的第一种损失函数是真实标签的独热向量与最深层分类器输出概率分布的交叉熵。SKDSAM 模型通过交叉熵把数据集中的知识从真实标签引入到模型的最深层分类器。

SKDSAM 模型的交叉熵由式（3.8）表示，其中q^C代表某样本经过最深层分类器输出的概率分布，y代表该样本在数据集中对应的真实标签形成的独热向量，$CE(\cdot)$代表交叉熵损失函数。

$$L_{SKDSA-CE} = CE(q^C, y) \tag{3.8}$$

SKDSAM 模型交叉熵（式（3.8））和 BYOT 模型交叉熵（式（2.2））的区别是：BYOT 模型的交叉熵计算的是所有分类器（包括最深层分类器和所有浅层分类器）输出概率分布与真实标签独热向量的交叉熵损失再统计它们的和，而 SKDSAM 模型仅仅计算最深层分类器输出概率分布与真实标签独热向量的交叉熵损失。实验证实这样能够取得更高的分类准确率。

SKDSAM 模型的第二种损失函数是最深层分离器输出概率分布和每个浅层分类

器输出概率分布的相对熵。SKDSAM 模型通过相对熵将最深层分类器中的暗知识传递到每个浅层分类器，引导浅层分类器模仿最深层分类器的预测结果。

SKDSAM 模型的相对熵由式（3.9）表示，其中q^i代表第i个浅层分类器输出的概率分布，q^C代表最深层分类器输出的概率分布，$KL(\cdot)$代表相对熵损失函数，T代表分类器蒸馏温度（默认值为 4）。

$$L_{SKDSA-KL} = T^2 KL(q^i, q^C) \tag{3.9}$$

SKDSAM 模型相对熵（式（3.9））和 BYOT 模型相对熵（式（2.3））的区别是：BYOT 模型直接计算最深层分离器输出概率分布和每个浅层分类器输出概率分布的相对熵，而 SKDSAM 模型在计算出最深层分离器输出概率分布和每个浅层分类器输出概率分布的相对熵之后，还要将结果和蒸馏温度的平方相乘。实验证实这样能够取得更高的分类准确率。

SKDSAM 模型的第三种损失函数是最深层分类器特征图和每个浅层分类器特征图的归一化L_1损失函数。SKDSAM 模型通过归一化L_1损失函数将最深层分类器特征图中隐含的知识引入每个浅层分类器，从而引导所有的浅层分类器的特征图去模仿最深层分类器的特征图。

SKDSAM 模型的归一化L_1损失函数由式（3.10）表示，其中F_i表示第i个浅层分类器的特征图，F_C表示最深层分类器的特征图，$||\cdot||_1$代表L_1损失函数。

$$L_{SKDSA-feature} = || \frac{F_i}{||F_i||_2} - \frac{F_C}{||F_C||_2} ||_1 \tag{3.10}$$

SKDSAM 模型归一化L_1损失函数（式（3.10））和 BYOT 模型L_2损失函数（式（2.4））的区别是：BYOT 模型计算的是最深层分类器特征图和每个浅层分类器特征图的L_2损失函数，而 SKDSAM 模型计算的则是最深层分类器特征图和每个浅层分类器特征图的归一化L_1损失函数。实验证实这样能够取得更高的分类准确率。

综合式（3.9）和（3.10），得到第i个浅层分类器的损失函数如式（3.11）所示。

$$L_{i-classifier} = T^2 KL(q^i, q^C) + \beta || \frac{F_i}{||F_i||_2} - \frac{F_C}{||F_C||_2} ||_1 \tag{3.11}$$

其中β代表平衡式（3.9）和（3.10）的超参数。式（3.11）对应于自注意力机制的输入$Value_i$。

将所有浅层分类器的损失函数（式（3.11））与对应的注意力权重（式（3.7））相乘后累加，即可得到自注意力机制的损失函数之和，如式（3.12）所示。

$$L_{attention} = \sum_{i}^{C-1} a_i (T^2 KL(q^i, q^C) + \beta || \frac{F_i}{||F_i||_2} - \frac{F_C}{||F_C||_2} ||_1) \tag{3.12}$$

其中$L_{attention}$即为自注意力机制（图 3.3 中）的最终输出。

综合式（3.8）和（3.12），即可得出 SKDSAM 模型的总损失函数，如式（3.13）所示。

$$\begin{aligned} L_{SKDSA} = CE(q^C, y) + \lambda \sum_{i=1}^{C-1} a_i (T^2 KL(q^i, q^C) \\ + \beta \cdot || \frac{F_i}{||F_i||_2} - \frac{F_C}{||F_C||_2} ||_1) \end{aligned} \tag{3.13}$$

其中λ代表平衡式（3.8）和（3.12）的超参数。超参数λ和β的取值会在第 4.4.4 小节调试。

SKDSAM 模型总损失函数和 BYOT 模型总损失函数的区别是：BYOT 模型将三种损失函数乘以相应的超参数后直接相加，而 SKDSAM 模型则将相对熵和归一化L_1损失函数结合后的结果乘以对应的注意力权重，再和交叉熵与超参数的乘积相加。实验证实这样能够取得更高的分类准确率（详见第 4.4.1 小节）。

3.2.4 SKDSAM 模型的训练步骤

SKDSAM 模型的训练步骤如算法 3.1 所示。输入为训练集数据和随机初始化的模型参数，输出为使总损失函数最小化的模型参数。

由于庞大的训练数据集难以一次全部装入显存之中，所以事先将训练集分为多个批次，每轮循环将一个批次的数据装入显存进行计算。在一次循环里的前向传播阶段，使用输入的训练集数据和现有的模型参数，依次根据式（3.6）、（3.7）、（3.8）、（3.9）、（3.10）、（3.11）、（3.12）和（3.13）计算出模型的总损失函数。在一次循环里的反向传播阶段，对总损失函数使用随机梯度下降法，更新模型参数。直到总损失函数收敛到一个区间范围内，停止循环。

算法 3.1 SKDSAM 的训练步骤

输入：训练集D，随机初始化的模型参数θ

输出：使总损失函数最小化的模型参数θ

while 模型参数θ尚未收敛 **do**

  $B \leftarrow minibatch(D)$ /*从训练集D中挑出一批次数据记为B */

  将B放入 SKDSAM 模型前向传播，得到第i浅层分类器特征图F_i和最深层分类器特征图F_C

  $Query \leftarrow soft\left(proj_C(F_C)\right)$ /*式（3.6）*/

  $Key_i \leftarrow soft(proj_i(F_i))$ /*式（3.6）*/

  $a_i \leftarrow \text{softmax}(Query \cdot Key_i)$ /*式（3.7）*/

  $L_{i-classifier} \leftarrow T^2 KL(q^i, q^C) + \beta || \frac{F_i}{||F_i||_2} - \frac{F_C}{||F_C||_2} ||_1$ /*式（3.11）*/

  $L_{attention} \leftarrow \sum_i^{C-1} a^i \cdot loss_{i-classifier}$ /*式（3.12）*/

  $L_{SKDSA} \leftarrow CE(q^C, y) + \lambda \cdot L_{attention}$ /*式（3.13）*/

  为了最小化L_{SKDSA}，对L_{SKDSA}使用随机梯度下降法，更新模型参数θ

end while

return θ

3.3 SKDSAM 模型和装袋法的等价性证明

本节将证明 SKDSAM 模型和装袋法（Bagging）[38]的等价性。

装袋法是机器学习中的一种集成算法，旨在提高机器学习算法的稳定性和准确率，减少模型方差以避免复杂模型的过拟合。装袋法的过程如图 3.5 所示：首先，创建多个拔靴（Bootstrap）样本，每个拔靴样本充当一个独立的数据集；然后，为每个拔靴样本拟合一个弱学习器，最后通过聚合平均分类器的预测结果，从而获得一个方差小于单个弱学习器的集成模型。聚合弱学习器的结果不会改变其期望，但能够减少其方差。

装袋法的核心思想是通过分开训练多个弱学习者，再通过一个权重投票机制得到一个更加稳定和强大的模型。对于有投票机制的装袋法，假设有C个神经网络记为$f_1, \ldots, f_i, \ldots, f_C$，它们的投票权重依次记为$v_1, \ldots, v_i, \ldots, v_M$，则最终预测结果如式（3.14）所示。

$$f = \sum_{i=1}^{C} v_i f_i \tag{3.14}$$

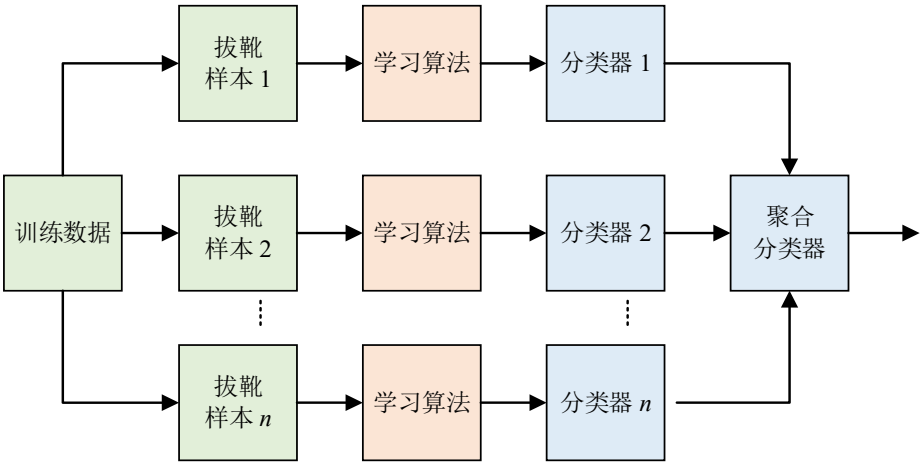

图 3.5　装袋法原理示意

因为 SKDSAM 模型使用自注意力机制将各个浅层块的信息汇总到最深层块，所以 SKDSAM 模型的最终预测如式（3.15）所示。

$$q^C = \sum_{i=1}^{C-1} a_i q^i \tag{3.15}$$

其中a_i代表第i个浅层块的注意力权重，q^i代表第i个浅层分类器的输出。

令式（3.14）中的$v_i = \frac{1}{2} a_i (i < C)$，$a_C = \frac{1}{2}$，再令$f_i$如式（3.16）所示。

$$f_i = \begin{cases} q^i, & if\ i \in [1, C-1] \\ q^C, & if\ i = C \end{cases} \tag{3.16}$$

把式（3.15）和（3.16）代入（3.14），得到结果如式（3.17）所示。

$$\begin{aligned} f &= \sum_{i=1}^{C} v_i f_i \\ &= \frac{1}{2} q^C + \frac{1}{2} \sum_{i=1}^{C-1} a_i\, q^i \\ &= \frac{1}{2} q^C + \frac{1}{2} q^C \\ &= q^C \end{aligned} \tag{3.17}$$

由此证明了 SKDSAM 模型的自注意力机制能够被视为装袋法，而 SKDSAM 中的注意力权重又恰恰对应于装袋法中的投票权重。这意味着 SKDSAM 模型和装袋法一样具有更强的泛化能力和抗过拟合能力。

3.4 结合数据增强技术的 SKDSAM 模型

数据增强是一种在不收集新数据的前提下增加原始数据多样性的技术，其方式包含对于原图像进行几何变换、颜色转换、随机擦除、对抗训练和神经风格迁移等等。为了进一步提升 SKDSAM 模型的性能，在 3.2 节给模型添加了自注意力机制之后，本节将 SKDSAM 模型与数据增强技术相结合，包括 Cutout[39]、自监督标签增强（Self-supervised Label Augmentation, SLA）[12]和 Mixup[40]三种数据增强技术。

3.4.1 Cutout

Cutout 数据增强技术由 DeVries 等人[40]提出。以图 3.6（左）中标签为猫的图像为例，Cutout 策略将原图像中随机大小的一块正方形区域“挖掉”（如图 3.6（右）所示），再将新图像添加到原有的数据集中，原标签保持不变。如果原图像是用矩阵表示，那么 Cutout 就是将矩阵中的一个子方块矩阵内部的全部元素替换为 0。这种策略能够有效提升视觉模型对遮挡物体的识别能力，以及提升视觉模型根据目标细节识别标签的能力。

在第 4.3.1 小节和第 4.3.2 小节，SKDSAM 模型使用 Cutout 数据增强后的数据集，以进一步提升模型性能。

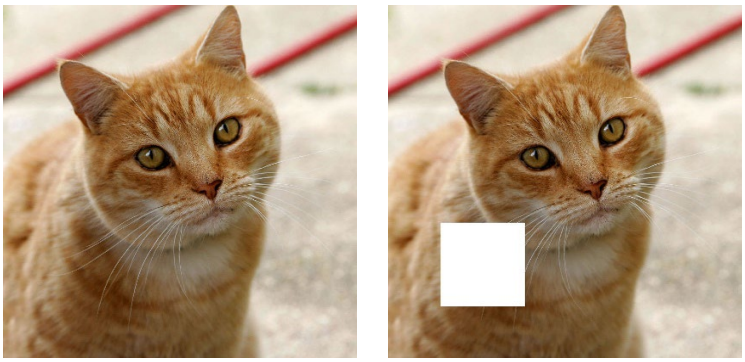

图 3.6 原图像（左）和 Cutout 后的图像（右）

3.4.2 SLA

SLA 数据增强技术由 Lee 等人[12]提出。以图 3.7 中标签为猫的图像为例，已有的数据增强技术是迫使分类器对原图像（图 3.7（左一））和旋转后的图像（图 3.7（左二、右二、右一））输出标签趋于一致。这种做法会极大地增加模型的复杂度，因为这会极大地改变样本的特征。SLA 技术并不直接迫使分类器学习原始图像和它变种的转换不变性，而是让分类器学习一个关于原始标签和自监督标签的联合概率分布，在推理阶段聚合得出预测结果。

图 3.7 原图像（左一）旋转 90°（左二）旋转 180°（右二）旋转 270°（右一）

在第 4.4.3 小节，SKDSAM 模型使用 SLA 数据增强后的数据集，以进一步提升模型性能。

3.4.3 Mixup

Mixup 数据增强技术由 Zhang 等人[40]提出。记x_i和x_j是输入向量，对应标签的独热向量分别是y_i和y_j，则它们混合生成的样本和对应的向量是如式（3.18）所示。

$$\begin{aligned}\tilde{x} &= \lambda x_i + (1-\lambda)x_j \\ \tilde{y} &= \lambda y_i + (1-\lambda)y_j\end{aligned} \tag{3.18}$$

其中(x_i, y_i)和(x_j, y_j)是从训练集中随机挑选的样本，λ服从 Beta 分布，取值范围是[0, 1]。Mixup 的假设是特征向量的线性插值能够导出对应标签的线性插值。

在第 4.4.3 小节，SKDSAM 模型使用 Mixup 数据增强后的数据集，以进一步提升模型性能。

3.5 本章小结

本章首先分析了 BYOT 模型和 PD-BYOT 模型的不足，提出了基于自注意力机制的自知识蒸馏（Self-Knowledge Distillation with Self-Attention Mechanism, SKDSAM）

模型作为改进方案。随后，在第 3.1 节论述了典型自注意力机制的网络结构。第 3.2 节比较了 SKDSAM 模型和 BYOT 模型在网络结构、损失函数上的不同，说明了 SKDSAM 模型相比于 BYOT 模型和所提出的 PD-BYOT 模型作出的改进。第 3.3 节从理论上证明了 SKDSAM 模型和装袋法的等价性，这意味着 SKDSAM 模型具备避免复杂模型过拟合的优点，具有更强的稳定性和泛化能力。第 3.4 节将 SKDSAM 模型与三种数据增强技术（Cutout、SLA 和 Mixup）相结合，作为进一步提升 SKDSAM 模型性能的备选方案。

4 实验结果与分析

本章首先说明了实验设置的具体细节，然后比较了 PD-BYOT 模型和原 BYOT 模型的性能，接下来比较了 SKDSAM 模型和其他自知识蒸馏模型及其他自注意力模型的性能，最后探讨了自注意力机制对 SKDSAM 模型性能所起的作用。

4.1 实验设置

4.1.1 实验环境

实验在本实验室的服务器上进行。服务器系统为 Ubuntu 18.04 版本，CPU 为 Intel(R) Xeon(R) Silver 4110 CPU @ 2.10GHz，内存 16GB，GPU 为 Nvidia 2080ti，显存 11GB。运行模型需要的主要包配置如表 4.1 所示。

表 4.1 运行模型需要的包配置

包名称	版本	说明
python	3.6	程序编写语言
numpy	1.9	矩阵运算库
pytorch	1.7	机器学习框架
torchvision	0.12	图像转换工具

4.1.2 实验数据集

（1）CIFAR-100 数据集[41]

CIFAR-100 数据集是一种常用的具有标注的微小图像数据集，它由 Alex Krizhevsky、Vinod Nair 和 Geoffrey Hinton 收集而成。

CIFAR-100 数据集由 100 个类别、共计 60 000 张 32×32 彩色图像组成；每个类别包含 600 张图像，每个类别的图像又分为 500 张训练图像和 100 张测试图像。这些类别之间是完全互斥的。比如，汽车类和卡车类之间没有重叠：汽车类包括轿车、运动型多用途车之类的东西，卡车类只包括大卡车，两者都不包括皮卡车。

本章使用 CIFAR-100 数据集测试模型对通用图像的分类性能。

（2）Tiny ImageNet 数据集[42]

ImageNet 数据集[47]是一个大型视觉数据集，常用于计算机视觉和深度学习研究。ImageNet 包含 20 000 多个类别，每个典型类别由数百张图像组成。ImageNet 现已成为图像识别的标准基准之一。

由于 ImageNet 数据集规模太大，研究者提出了 Tiny ImageNet 数据集作为 ImageNet 数据集的替代。它相比于 ImageNet 数据集规模较小而且图像类别较少。Tiny ImageNet 数据集包含 200 个图像类，训练集包含 100 000 张图像，验证集包含 10 000 张图像，测试集包含 10 000 张图像。所有图像的大小均为 64×64。

本章使用 Tiny ImageNet 数据集测试模型对通用图像的分类性能。

（3）Caltech-UCSD Birds 200 数据集[43]

Caltech-UCSD Birds 200（CUB-200）数据集是一个鸟类的图像数据集，包含 200 种鸟类（主要分布在北美）共计 6033 张图像。

本章使用 CUB-200 数据集测试模型对细粒度图像的分类性能。

（4）Stanford 40 Actions 数据集[44]

Stanford 40 Actions（Stanford-40）数据集包含人类 40 种动作的图像。Stanford-40 数据集共计 9532 张图像，每类动作包含 180-300 张图像。

本章使用 Stanford-40 数据集测试模型对细粒度图像的分类性能。

（5）Stanford Dogs 数据集[45]

Stanford Dogs（Dogs）数据集包含来自世界各地的 120 种狗的图像。该数据集由 ImageNet 中的图像和标注构建，用于细粒度图像分类任务。Dogs 数据集共计 20580 张图像。

本章使用 Dogs 数据集测试模型对细粒度图像的分类性能。

（6）MIT Indoor Scene Recognition 数据集[46]

室内场景识别是视觉识别中一个极具挑战性的任务。大多数适用于室外场景的场景识别模型在室内领域表现不佳。主要困难在于，虽然某些室内场景（例如走廊）可通过全局空间属性有效地表征，但其他一些场景（例如书店）必须通过它们包含的对象才能有效地表征。所以，为了更有效地识别室内场景，要求模型能够同时利用局

部和全局的判别信息。

MIT Indoor Scene Recognition（MIT-67）数据集包含 67 个室内场景类别，共计 15620 张图像。每类图像的数量因类别而异，但每个类别至少包含 100 张图像。

本章使用 MIT-67 数据集测试模型对细粒度图像的分类性能。

4.1.3 超参数设置

所有的神经网络都是从零开始训练。优化函数中随机梯度下降的动量值（momentum）设为 0.9，权重衰减（weight decay）设为 5×10^{-4}。轮次总数设为 250。初始学习率（learning rate）设为 0.1，到第 100 个轮次时，学习率降为最初的 1/10; 到第 150 个轮次时，学习率降为最初的 1/100。设一批次训练样本的数量（batch size）为 128。SKDSAM 模型中的超参数λ设为 1.5，超参数β设为 100。

4.1.4 网络结构设置

（1）PD-BYOT 模型的网络结构设置

使用残差网络（Deep Residual Networks, ResNet）框架实现 PD-BYOT 模型。设 ResNet 第一个卷积层的卷积核大小为3×3，步长为 1，填充为 1。

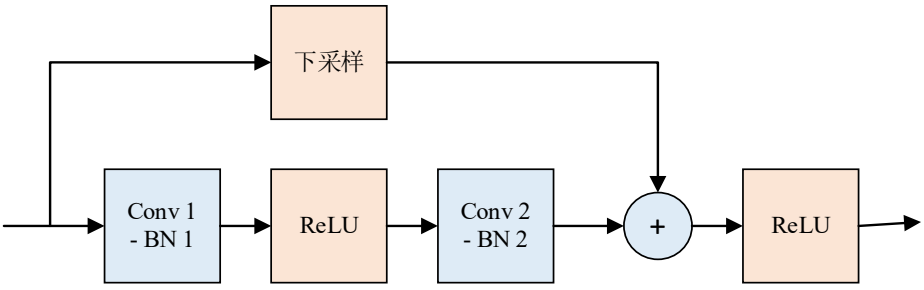

图 4.1　一个浅层块的结构示意

设置 PD-BYOT 模型中的每个浅层块的结构如图 4.1 所示，按前馈方向依次是卷积层 1、批量标准化（Batch normalization, BN）层 1、线性整流（Rectified linear unit, ReLU）层、卷积层 2、批量标准化层 2、与最初输入信号的下采样相加及线性整流层。

设置 PD-BYOT 模型中的每个瓶颈层的结构如图 4.2 所示，按前馈方向依次是卷积层 1、批量标准化层 1、线性整流层、卷积层 2、批量标准化层 2、线性整流层、卷

积层 3、批量标准化层 3 和线性整流层。

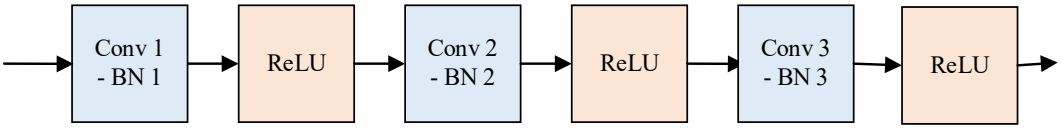

图 4.2　瓶颈层的结构示意

（2）SKDSAM 模型的网络结构设置

为了测试 SKDSAM 模型的性能，分别使用当前流行的卷积神经网络框架深度残差网络（Deep Residual Networks, ResNet）[35]，宽残差网络（Wide Residual Networks, WRN）[36]，残差密集网络（Densely Connected Convolutional Networks, DenseNet）[37]实现其功能。具体设置为 64 个滤波器，第一个卷积层的卷积核大小为3×3，步长为 1，填充为 1。

设置 SKDSAM 模型有三个浅层块，每一个浅层块的结构如图 4.3 所示，按前馈方向依次是卷积层 1、批量标准化（Batch Normalization, BN）层 1、线性整流（Rectified Linear Unit, ReLU）层、卷积层 2、批量标准化层 2、与最初输入信号的下采样相加及线性整流层。

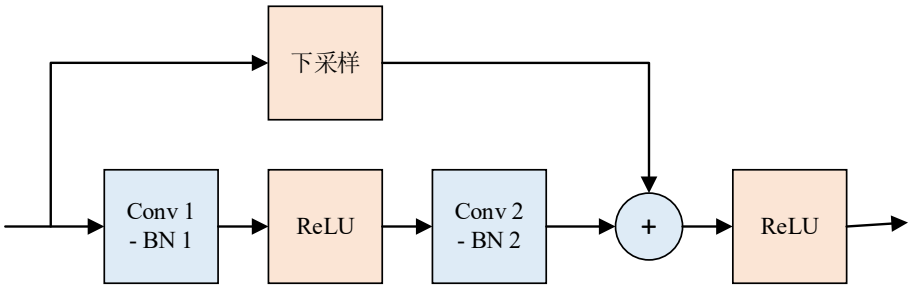

图 4.3　一个浅层块的结构示意

设置一个自适应分离卷积（Adaptive Separable Convolution, SepConv）[48]单元的结构如图 4.4 所示，按前馈方向依次是卷积层 1、卷积层 2、批量标准化层 1、线性整流层、卷积层 3、卷积层 4、批量标准化层 2 和线性整流层。

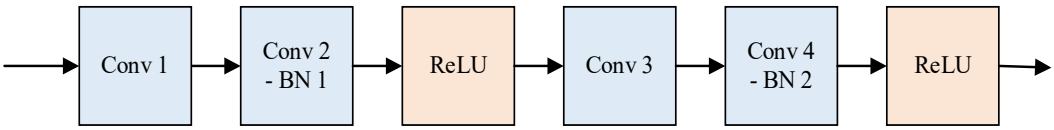

图 4.4　一个自适应分离卷积单元的结构示意

设置第一个浅层块所连接的瓶颈层的结构如图 4.5 所示，按前馈方向依次是自适

应分离卷积层 1、自适应分离卷积层 2、自适应分离卷积层 3 和平均池化层。

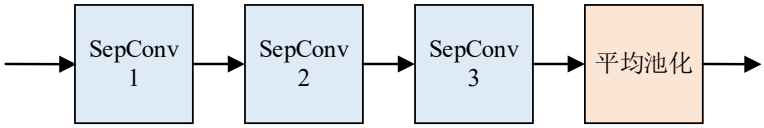

图 4.5　第一个瓶颈层的结构示意

设置第二个浅层块所连接的瓶颈层的结构如图 4.6 所示，按前馈方向依次是自适应分离卷积层 1、自适应分离卷积层 2 和平均池化层。

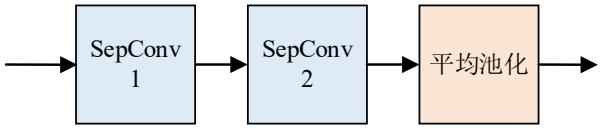

图 4.6　第二个瓶颈层的结构示意

设置第三个浅层块所连接的瓶颈层的结构如图 4.7 所示，按前馈方向依次是自适应分离卷积层 1 和平均池化层。

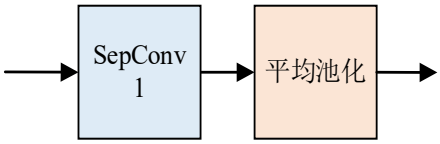

图 4.7　第三个瓶颈层的结构示意

设置每个浅层块所连接的自注意力机制的结构如图 4.8 所示，按前馈方向依次是自适应分离卷积层、批量标准化层、线性整流层、上采样层和非线性激活层。

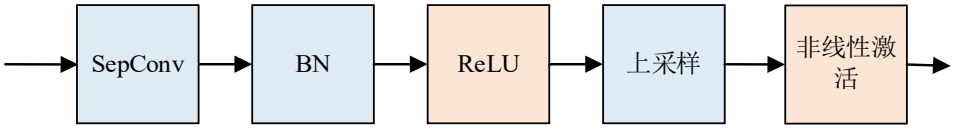

图 4.8　自注意力连接的结构示意

4.1.5　评估标准

实验使用图像的分类准确率作为衡量模型性能的指标。记分类准确率为Acc，样本总数为N，预测概率最大的标签是正确标签的样本数量为T，则分类准确率的计算公式如式（4.1）所示。

$$Acc = \frac{T}{N} \tag{4.1}$$

实验的优化目标是尽可能提高验证集中的分类准确率。

4.1.6 对比模型

首先进行对比的模型为不添加自知识蒸馏模块，也不添加自注意力机制，仅仅计算最深层分类器交叉熵损失函数的模型（记为“交叉熵模型”）。除交叉熵模型之外，实验还将 SKDSAM 模型的性能和五种典型的自知识蒸馏模型和两种典型的注意力模型相比较。

五种自知识蒸馏模型分别是：基于辅助分类器的自知识蒸馏（Be Your Own Teacher, BYOT）[8]、数据失真引导的自知识蒸馏（Data-Distortion Guided Self-Distillation, DDGSD）[14]、基于自知识蒸馏的特征精炼（Feature Refinement via Self-Knowledge Distillation, FRSKD）[9]、基于自标签增强的自知识蒸馏（Self-supervised Label Augmentation based Self-Distillation, SLA-SD）[12]、类级别的自知识蒸馏（Class-wise Self-Knowledge Distillation, CS-KD）[13]。

（1）BYOT 模型首先通过辅助分类器得到浅层块的特征，再通过数据集中的真实标签和最深层块的暗知识来训练这些辅助分类器。

（2）DDGSD 模型先使同一个样本产生不同的变种，再引导神经网络对同一个样本的不同变种输出一致的预测。

（3）FRSKD 模型引入了一个辅助的自教师模块，先将分类器输出的特征输入自教师模块，再将自教师模块提炼出的特征图返回分类器网络。

（4）SLA-SD 首先学习一个关于原始标签和自监督标签的联合概率分布，再在推理阶段结合自知识蒸馏聚合得出预测结果。

（5）CS-KD 模型引导标签相同的不同样本所输出的概率分布尽可能接近，将知识蒸馏模型的损失函数定义为由教师模型和学生模型对同一类图像的预测差异。

两种注意力模型分别是：密集隐含的注意力神经网络（Dense-and-Implicit Attention Network, DIANet）[33]和自注意力神经网络（Self-Attention Network, SAN）[34]。

（1）DIANet 模型给不同的神经网络层之间添加注意力连接，目的是更有效地利用神经网络层间的信息。

（2）SAN 模型利用成对的自注意力机制去抽取更重要的信息来引导模型训练。

4.2 PD-BYOT 模型实验结果与分析

为了测试 PD-BYOT 模型在通用图像数据集上的性能，使用多组衰减系数d_{CE}、d_{KL}和d_{feture}在 CIFAR-100 数据集上进行实验，并且和原 BYOT 模型（即d_{CE}、d_{KL}和d_{feture}均设为 1）进行对比。实验结果记录最后轮次的分类准确率。实验结果如表 4.2、表 4.3、表 4.4 和表 4.5 所示，最优的结果用粗体标注。

表 4.2 衰减系数在不同取值的分类准确率（%）

d_{CE}	d_{KL}	d_{feture}	准确率
1.0	1.0	1.0	76.01
0.9	1.0	1.0	76.19
0.8	1.0	1.0	76.26
0.7	**1.0**	**1.0**	**76.58**
0.5	1.0	1.0	75.91

表 4.3 衰减系数在不同取值的分类准确率（%）

d_{CE}	d_{KL}	d_{feture}	准确率
1.0	0.4	1.0	74.86
1.0	**0.5**	**1.0**	**76.99**
1.0	0.6	1.0	72.24
1.0	0.9	1.0	75.26
1.0	1.1	1.0	76.05
1.0	1.2	1.0	76.55
1.0	1.3	1.0	75.63
1.0	1.5	1.0	75.70

表 4.2、表 4.3、表 4.4 和表 4.5 的实验结果表明， PD-BYOT 模型能够在一定程度上提升原 BYOT 模型的性能。其中最优结果在d_{CE}取值 0.6、d_{KL}取值 1.1 和d_{feture}取值 1.0（或d_{CE}取值 1.0、d_{KL}取值 0.5 和d_{feture}取值 1.0）时取得，相比原 BYOT 模型提升准确率 0.98%。

表 4.4　衰减系数在不同取值的分类准确率（%）

d_{CE}	d_{KL}	d_{feture}	准确率
0.6	0.6	1.0	74.92
0.6	0.7	1.0	73.68
0.6	0.8	1.0	75.77
0.6	1.0	1.0	76.64
0.6	**1.1**	**1.0**	**76.99**
0.6	1.2	1.0	75.63

表 4.5　衰减系数在不同取值的分类准确率（%）

d_{CE}	d_{KL}	d_{feture}	准确率
1.0	1.0	0.5	75.34
1.0	1.0	0.9	75.43
1.0	1.0	1.1	76.02
1.0	**1.0**	**1.2**	**76.12**
1.0	1.0	1.3	76.11
1.0	1.0	1.5	75.78

4.3　SKDSAM 模型实验结果与分析

4.3.1　在通用图像数据集上的实验结果

为了测试 SKDSAM 模型在通用图像数据集上的性能，分别使用基于 ResNet、WRN 和 DenseNet 框架的 SKDSAM 模型在 CIFAR-100 和 Tiny-ImageNet 数据集上进行实验，并且和其他自知识蒸馏模型（BYOT 模型、DDGSD 模型、FRSKD 模型、SLA-SD 模型和 CS-KD 模型）及结合了数据增强技术（Cutout）的 SKDSAM 模型进行对比。每个实验都重复 3 次，实验结果记录最后轮次的分类准确率的平均数。

实验结果分别如表 4.6、表 4.7 和表 4.8 所示，其中第一列代表基于某种框架实现的 SKDSAM 模型和作为对比的各种模型，第二列代表各种模型在 CIFAR-100 数据集上的分类准确率，第三列代表各种模型在 Tiny ImageNet 数据集上的分类准确率，

最优的结果用粗体标注。

表 4.6　基于 ResNet 框架的各模型在通用图像数据集上的分类准确率（%）

模型	CIFAR-100	Tiny ImageNet
交叉熵	74.80	54.60
DDGSD	76.68	57.76
BYOT	76.87	56.76
CS-KD	78.01	57.72
SLA-SD	77.88	58.67
FRSKD	78.61	59.61
SKDSAM	80.51	60.42
SKDSAM+Cutout	**81.77**	**62.61**

表 4.7　基于 WRN 框架的各模型在通用图像数据集上的分类准确率（%）

模型	CIFAR-100	Tiny ImageNet
交叉熵	70.42	51.25
DDGSD	71.98	52.30
BYOT	70.28	51.43
CS-KD	72.64	52.23
SLA-SD	73.00	51.64
FRSKD	73.27	53.08
SKDSAM	74.80	53.21
SKDSAM+Cutout	**76.03**	**53.42**

表 4.6、表 4.7 和表 4.8 的实验结果表明，在通用图像数据集上，SKDSAM 模型具有比其他自知识蒸馏模型更优异的性能。具体来说，基于 ResNet 框架的 SKDSAM 模型在 CIFAR-100 和 Tiny ImageNet 数据集上相比基于 ResNet 框架的交叉熵模型分别提升准确率 5.71%和 5.82%，基于 WRN 框架的 SKDSAM 模型在 CIFAR-100 和 Tiny ImageNet 数据集上相比基于 WRN 框架的交叉熵模型分别提升准确率 4.38%和 1.96%，基于 DenseNet 框架的 SKDSAM 模型在 CIFAR-100 和 Tiny ImageNet 数据集上相比基于 DenseNet 框架的交叉熵模型分别提升准确率 3.78%和 2.01%。

表 4.8　基于 DenseNet 框架的各模型在通用图像数据集上的分类准确率（%）

模型	CIFAR-100	Tiny ImageNet
交叉熵	77.77	60.78
DDGSD	78.20	61.58
BYOT	78.07	61.12
CS-KD	79.39	62.04
SLA-SD	79.76	61.76
FRSKD	80.55	61.12
SKDSAM	**81.55**	**62.79**

表 4.6 和表 4.7 的实验结果表明，在通用图像数据集上，结合数据增强技术的 SKDSAM 模型具有比单一的 SKDSAM 模型更卓越的性能。具体来说，基于 ResNet 框架的 SKDSAM 模型在 CIFAR-100 和 Tiny ImageNet 数据集上相比基于 ResNet 框架的交叉熵模型分别提升准确率 6.97%和 8.01%，基于 WRN 框架的 SKDSAM 模型在 CIFAR-100 和 Tiny ImageNet 数据集上相比基于 WRN 框架的交叉熵模型分别提升准确率 5.61%和 2.17%。

4.3.2　在细粒度图像数据集上的实验结果

为了测试 SKDSAM 模型在细粒度图像数据集上的性能，分别使用基于 ResNet 和 DenseNet 框架的 SKDSAM 模型在一些细粒度图像数据集（CUB-200、Stanford-40、Dogs 和 MIT-67）上进行实验，并且和其他自知识蒸馏模型（BYOT 模型、DDGSD 模型、FRSKD 模型、SLA-SD 模型和 CS-KD 模型）及结合了数据增强技术（Cutout）的 SKDSAM 模型进行对比。每个实验都重复 3 次，实验结果记录最后轮次的分类准确率的平均数。

实验结果分别如表 4.9 和表 4.10 所示，其中第一列代表基于某种框架实现的 SKDSAM 模型和作为对比的各种模型，第二列代表各种模型在 CUB-200 数据集上的分类准确率，第三列代表各种模型在 MIT-67 数据集上的分类准确率，第四列代表各种模型在 Dogs 数据集上的分类准确率，第五列（仅限于表 4.9）代表各种模型在 Stanford-40 数据集上的分类准确率，最优的结果用粗体标注。

表 4.9　基于 ResNet 框架的各模型细粒度图像数据集上的分类准确率（%）

模型	CUB-200	MIT-67	Dogs	Stanford-40
交叉熵	52.72	56.07	64.75	44.30
DDGSD	58.79	59.68	69.20	45.84
BYOT	58.86	58.41	68.92	48.51
CS-KD	64.86	57.42	69.02	47.34
SLA-SD	56.47	61.59	67.64	54.67
FRSKD	65.39	61.61	70.77	56.00
SKDSAM	65.20	61.90	71.21	58.02
SKDSAM+Cutout	**67.03**	**65.77**	**72.84**	**62.61**

表 4.9 和表 4.10 的实验结果表明，在细粒度图像数据集上，SKDSAM 模型具有比其他自知识蒸馏模型更优异的性能。具体来说，基于 ResNet 框架的 SKDSAM 模型在 CUB-200、MIT-67、Dogs 和 Stanford-40 数据集上相比基于 ResNet 框架的交叉熵模型分别提升准确率 12.48%、5.83%、6.46%和 13.72%，基于 DensetNet 框架的 SKDSAM 模型在 CUB-200、MIT-67 和 Dogs 数据集上相比基于 DensetNet 框架的交叉熵模型分别提升准确率 13.40%、5.99%和 7.89%。

表 4.10　基于 DenseNet 框架的各模型在细粒度图像数据集上的分类准确率（%）

模型	CUB-200	MIT-67	Dogs
交叉熵	57.70	58.21	66.61
DDGSD	65.35	59.10	70.48
BYOT	66.80	58.80	71.14
CS-KD	69.17	59.98	72.19
SLA-SD	68.88	61.11	73.19
FRSKD	69.60	63.35	74.00
SKDSAM	**71.10**	**64.20**	**74.50**

表 4.9 还表明在细粒度图像数据集上，结合数据增强技术的 SKDSAM 模型具有比原始 SKDSAM 模型更卓越的性能。具体来说，基于 ResNet 框架的 SKDSAM 模型在 CUB-200、MIT-67、Dogs 和 Stanford-40 数据集上相比基于 ResNet 框架的交叉熵

模型分别提升准确率 14.31%、9.70%、8.09%和 18.31%。

4.3.3 不同注意力模型的实验结果

为了比较 SKDSAM 模型和两种注意力模型（DIANet 和 SAN）的性能，分别使用基于三种残差网络框架（ResNet18、ResNet34 和 ResNet50）的 SKDSAM 模型在 CIFAR-100 数据集上进行实验，并且和其他两种注意力模型（DIANet 模型和 SAN 模型）进行对比。每个实验都重复 3 次，实验结果记录最后轮次的分类准确率的平均数。

实验结果如表 4.11 所示，其中第一列代表基于 ResNet 框架实现的 SKDSAM 模型和作为对比的各种模型，第二列代表各种模型基于 ResNet18 框架的分类准确率，第三列代表各种模型基于 ResNet34 框架的分类准确率，第四列代表各种模型基于 ResNet50 框架的分类准确率，最优的结果用粗体标注。

表 4.11 基于 ResNet 框架的各模型在通用图像数据集上的分类准确率（%）

模型	ResNet18	ResNet34	ResNet50
SAN	75.90	76.40	77.20
DIANet	76.62	77.10	78.60
SKDSAM	**80.51**	**80.81**	**81.13**

表 4.11 的实验结果表明，在 CIFAR-100 数据集上，SKDSAM 模型具有比其他两种注意力模型更优异的性能。具体来说，基于 ResNet18 框架的 SKDSAM 模型相比基于 ResNet18 框架的 SAN 和 DIANet 分别提升准确率 4.61%和 3.89%，基于 ResNet34 框架的 SKDSAM 模型相比基于 ResNet34 框架的 SAN 和 DIANet 分别提升准确率 4.41%和 3.71%，基于 ResNet50 框架的 SKDSAM 模型相比基于 ResNet50 框架的 SAN 和 DIANet 分别提升准确率 3.93%和 2.53%。

4.4 SKDSAM 模型消融实验与分析

4.4.1 自注意力机制的重要性

为了证明自注意力机制对于 SKDSAM 模型的重要性，分别使用基于 ResNet 和

WRN 框架的移除自注意力机制的 SKDSAM 模型和 SKDSAM 模型在 CIFAR-100 和 Tiny ImageNet 数据集上进行实验。每个实验都重复三次，实验结果记录最后轮次的分类准确率的平均数。

表 4.12　基于 ResNet 框架的 SKDSAM 模型和其移除自注意力机制后的分类准确率（%）

模型	CIFAR-100	Tiny ImageNet
w/o SA	77.80	57.20
SKDSAM	**80.51**	**60.02**

实验结果如表 4.12 和表 4.13 所示，其中第一列代表基于 ResNet 框架实现的 SKDSAM 模型和移除自注意力机制的 SKDSAM 模型，第二列代表两种模型在 CIFAR-100 数据集上的分类准确率，第三列代表两种模型在 Tiny ImageNet 数据集上的分类准确率，最优的实验结果用粗体字标注。

表 4.13　基于 WRN 框架的 SKDSAM 模型和其移除自注意力机制后的分类准确率（%）

模型	CIFAR-100	Tiny ImageNet
w/o SA	71.71	51.92
SKDSAM	**74.80**	**53.21**

表 4.12 和表 4.13 的实验结果表明，在通用图像数据集上，SKDSAM 模型具有比移除自注意力机制的 SKDSAM 模型更优异的性能。具体来说，基于 ResNet 框架的 SKDSAM 模型在 CIFAR-100 和 Tiny ImageNet 数据集上相比移除注意力机制的基于 ResNet 框架的 SKDSAM 模型分别提升准确率 2.71%和 2.82%，基于 WRN 框架的 SKDSAM 模型在 CIFAR-100 和 Tiny ImageNet 数据集上相比移除注意力机制的基于 WRN 框架的 SKDSAM 模型分别提升准确率 3.09%和 1.29%。

以上结果说明，自注意力机制在 SKDSAM 模型中起到了重要作用，对模型性能的发挥不可或缺。

4.4.2　自注意力机制中的知识蒸馏模块的重要性

为了证明自注意力机制中的知识蒸馏模块对 SKDSAM 模型的重要性，同时测试不同蒸馏温度对 SKDSAM 模型性能的影响，使用基于 ResNet 框架的 SKDSAM 模型

于不同的知识蒸馏温度下在 CIFAR-100 数据集上进行实验。每个实验都重复三次，实验结果记录最后轮次的分类准确率的平均数。

表 4.14　SKDSAM 模型在不同知识蒸馏温度下的分类准确率（%）

自注意力机制的蒸馏温度	分类器蒸馏温度设为 1	分类器蒸馏温度设为 4
1	78.13	78.13
2	79.98	79.86
3	79.65	79.89
4	80.31	80.51

实验结果如表 4.14 所示，其中第一列代表自注意力机制中的蒸馏温度$T' \in \{1,2,3,4\}$，第二列代表分类器蒸馏温度$T = 1$时的分类准确率，第三列代表分类器蒸馏温度$T = 4$时的分类准确率。其中温度设为 1 等价于不使用知识蒸馏模型。

表 4.14 的实验结果表明，自注意力机制在添加知识蒸馏模块后具有比原模型更优异的性能。具体来说，在不使用分类器蒸馏时（表 4.14 的第二列），SKDSAM 模型在自注意力机制中的蒸馏温度T'为 2、3 和 4 时比不使用自注意力机制中的知识蒸馏分别提升准确率 1.85%、1.52%和 2.18%；在使用分类器蒸馏时（表 4.14 的第三列），SKDSAM 模型在自注意力机制的蒸馏温度T'为 2、3 和 4 时比不使用自注意力机制中的知识蒸馏分别提升准确率 1.73%、1.76%和 2.38%。

以上结果说明，自注意力机制中的蒸馏模块在 SKDSAM 模型中起到了重要作用，对模型性能的发挥不可或缺。随着自注意力机制中的蒸馏温度逐步升高，模型分类性能也在逐步提升。

4.4.3　SKDSAM 模型与数据增强技术的相容性

第 4.3.1 小节和第 4.3.2 小节的实验证实了结合 Cutout 数据增强技术能够提升进一步提升 SKDSAM 模型的性能。本小节将 SKDSAM 模型分别和另外两种数据增强技术（Mixup 和自监督标签增强（Self-supervised Label Augmentation, SLA））结合并测试了其性能。

为了证明 SKDSAM 模型和数据增强技术的相容性，分别使用基于 ResNet 框架的结合 Mixup 技术的 SKDSAM 模型、基于 ResNet 框架的 SKDSAM 模型、基于

ResNet 框架的结合 Mixup 技术的交叉熵模型、基于 WRN 框架的结合 Mixup 技术的 SKDSAM 模型、基于 WRN 框架的 SKDSAM 模型、基于 WRN 框架的结合 Mixup 技术的交叉熵模型、基于 ResNet 框架的结合 SLA 技术的 SKDSAM 模型、基于 ResNet 框架的结合 SLA 技术的交叉熵模型、基于 WRN 框架的结合 SLA 技术的 SKDSAM 模型及基于 WRN 框架的结合 SLA 技术的交叉熵模型在 CIFAR-100 和 Tiny ImageNet 数据集上进行实验。每个实验都重复三次，实验结果记录最后轮次的分类准确率的平均数。

表 4.15　结合与不结合 Mixup 技术的各模型（基于 ResNet 框架）分类准确率（%）

模型	CIFAR-100	Tiny ImageNet
SKDSAM	80.51	60.42
Mixup+交叉熵	78.33	58.43
Mixup+SKDSAM	**81.41**	**61.44**

表 4.16　结合与不结合 Mixup 技术的各模型（基于 WRN 框架）分类准确率（%）

模型	CIFAR-100	Tiny ImageNet
SKDSAM	74.80	53.21
Mixup+交叉熵	72.21	52.82
Mixup+SKDSAM	**76.73**	**53.68**

表 4.17　结合与不结合 SLA 技术的各模型（基于 ResNet 框架）分类准确率（%）

模型	CIFAR-100	Tiny ImageNet
SKDSAM	80.51	60.42
SLA+交叉熵	77.52	58.48
SLA+SKDSAM	**82.81**	**63.02**

表 4.18　结合与不结合 SLA 技术的各模型（基于 WRN 框架）分类准确率（%）

模型	CIFAR-100	Tiny ImageNet
SKDSAM	74.80	53.21
SLA+交叉熵	73.00	50.77
SLA+SKDSAM	**76.83**	**53.89**

实验结果如表 4.15、表 4.16、表 4.17 和表 4.18 所示，其中第一列代表基于某种框架实现的 SKDSAM 模型、结合数据增强技术的交叉熵模型及结合数据增强技术的 SKDSAM 模型，第二列代表三种模型在 CIFAR-100 数据集上的分类准确率，第三列代表三种模型在 Tiny ImageNet 数据集上的分类准确率，最优的实验结果用粗体字标注。

表 4.15、表 4.16、表 4.17 和表 4.18 的实验结果表明，结合数据增强技术的 SKDSAM 模型具有比原始 SKDSAM 模型更优异的性能，也具有比结合数据增强技术的交叉熵模型更优异的性能。具体来说，基于 ResNet 框架的结合 Mixup 技术的 SKDSAM 模型在 CIFAR-100 和 Tiny ImageNet 数据集上相比基于 ResNet 框架的 SKDSAM 模型分别提升准确率 0.90%和 1.02%，相比基于 ResNet 框架的结合 Mixup 技术的交叉熵模型分别提升准确率 3.08%和 3.01%（表 4.15）；基于 WRN 框架的结合 Mixup 技术的 SKDSAM 模型在 CIFAR-100 和 Tiny ImageNet 数据集上相比基于 WRN 框架的 SKDSAM 模型分别提升准确率 1.93%和 0.47%，相比基于 WRN 框架的结合 Mixup 技术的交叉熵模型分别提升准确率 4.52%和 0.86%（表 4.16）；基于 ResNet 框架的结合 SLA 技术的 SKDSAM 模型在 CIFAR-100 和 Tiny ImageNet 数据集上相比基于 ResNet 框架的 SKDSAM 模型分别提升准确率 2.3%和 2.6%，相比基于 ResNet 框架的结合 SLA 技术的交叉熵模型分别提升准确率 5.29%和 4.54%（表 4.17）；基于 WRN 框架的结合 SLA 技术的 SKDSAM 模型在 CIFAR-100 和 Tiny ImageNet 数据集上相比基于 WRN 框架的 SKDSAM 模型分别提升准确率 2.03%和 0.68%，相比基于 WRN 框架的结合 SLA 技术的交叉熵模型分别提升准确率 3.83%和 3.12%（表 4.18）。

以上结果说明，SKDSAM 模型能够进一步提升基于数据增强技术的性能，数据增强技术也能够进一步提升 SKDSAM 模型的性能。这证实了 SKDSAM 模型和数据增强技术的相容性。

4.4.4 敏感性分析

SKDSAM 模型的总损失函数（式（3.13））包含两个重要的超参数λ和β，在前面的实验中λ取值 1.5，β取值 100，本小节测试λ和β取其他数值对模型分类准确率的影响。

为了测试了超参数λ在不同取值下对 SKDSAM 模型性能的影响，使用基于 ResNet 框架的 SKDSAM 模型在 CIFAR-100 数据集上进行实验。进行测试的λ取值范围是$\{0.5, 1, 1.5, 2, 3\}$，其他超参数保持不变。每个实验都重复三次，实验结果记录最后轮次的分类准确率的平均数。

表 4.19　SKDSAM 模型取不同λ值时的分类准确率（%）

λ	分类准确率
0.5	79.86
1.0	80.05
1.5	**80.51**
2.0	80.02
3.0	79.64

实验结果如表 4.19 所示，其中第一列代表超参数λ的不同取值，第二列代表 SKDSAM 模型在特定λ取值时的分类准确率，最优的实验结果用粗体字标注。实验结果表明当$\lambda = 1.5$时，SKDSAM 模型的分类准确率最高。

为了测试了超参数β在不同取值下对 SKDSAM 模型性能的影响，使用基于 ResNet 框架的 SKDSAM 模型在 CIFAR-100 数据集上进行实验。进行测试的β取值范围是$\{50, 100, 200, 500\}$。每个实验都重复三次，实验结果记录最后轮次的分类准确率的平均数。

实验结果如表 4.20 所示，其中第一列代表超参数β的不同取值，第二列代表 SKDSAM 模型在特定β取值时的分类准确率，实验结果表明当$\beta = 100$时分类准确率最高，达到 80.51%。

表 4.20　SKDSAM 模型取不同β值时的分类准确率（%）

β	分类准确率
50	79.81
100	**80.51**
200	80.16
500	79.99

4.5 本章小结

为了测试 PD-BYOT 模型和 SKDSAM 模型的性能，在第 4.1 节说明了实验的具体设置。PD-BYOT 模型的实验结果（第 4.2 节）表明，PD-BYOT 模型相比原 BYOT 模型性能有一定的提升。SKDSAM 模型的实验结果（第 4.3 节）表明，SKDSAM 模型的性能相比现有的自知识蒸馏模型和自注意力模型都有令人振奋的提升，这证实了 SKDSAM 模型的有效性，证实了使用自注意力机制区分各个浅层块对最深层块的不同贡献度的正确性。SKDSAM 模型的消融实验与分析（第 4.4 节）还说明，自注意力机制能够有效提升自知识蒸馏模型性能，自注意力机制中的知识蒸馏模块对 SKDSAM 模型起到了重要作用，SKDSAM 模型能够和多种数据增强技术相容，以及找到了较优的超参数$\lambda = 1.5$，$\beta = 100$。

5 总结与展望

5.1 主要工作总结

深度神经网络技术在各行各业的应用日益广泛，但是大型神经网络的训练需要昂贵的计算资源和时间成本。为了压缩大型神经网络，研究者提出了知识蒸馏技术，将大型神经网络（教师模型）隐含的信息迁移到小型神经网络（学生模型），从而显著提升小型神经网络的性能。

自知识蒸馏模型是对传统知识蒸馏模型的改进，它不需要外部的大型神经网络，而是利用小型神经网络自身的信息实现知识蒸馏。这不仅使小型神经网络摆脱了对外部大型教师模型的依赖，也使知识蒸馏所需的时间明显缩短。本文旨在进一步提升自知识蒸馏模型的性能，主要做了以下工作：

（1）分析了基于辅助分类器的自知识蒸馏（Be Your Own Teacher, BYOT）模型，它的不足是忽略了各个浅层块信息对最深层块的不同影响。然后，提出了基于逐块衰减辅助分类器的自知识蒸馏（Per-block Decay based Be Your Own Teacher, PD-BYOT）模型作为初步改进方案，即通过给 BYOT 模型的各浅层块添加衰减系数将各个浅层块对最深层块的影响以等比数列的形式加以区分。实验表明添加衰减系数能略微改进 BYOT 模型的性能。

（2）在 BYOT 模型和 PD-BYOT 模型的基础上，提出了基于自注意力机制的自知识蒸馏（Self-Knowledge Distillation with Self-Attention Mechanism, SKDSAM）模型，以便更准确地量化各浅层块特征图对最深层块特征图的不同贡献度。详细说明了 SKDSAM 模型的网络结构、损失函数，列举了 SKDSAM 模型相比原 BYOT 模型所做的改进。随后从理论上证明 SKDSAM 模型可以被视为集成学习中的装袋法，这意味着 SKDSAM 具有优异的抗过拟合性能。最后，将模型与三种数据增强技术（Cutout、自监督标签增强（Self-supervised label augmentation, SLA）及 Mixup）相结合，作为进一步提升模型性能的备选方案。

（3）在多个图像数据集（CIFAR-100 数据集、Tiny ImageNet 数据集、Caltech-UCSD Bird 数据集、Stanford 40 Actions 数据集、Stanford Dogs 数据集、MIT Indoor

Scene Recognition 数据集）上测试了 SKDSAM 模型的分类准确率，取得了比几种当前流行的自知识蒸馏模型和注意力模型更优异的实验结果。消融实验说明了 SKDSAM 模型的自注意力机制和自注意力机制中的知识蒸馏模块对提升性能的重要作用，以及 SKDSAM 模型结合数据增强技术能够进一步提升模型的性能，找到了使 SKDSAM 模型性能较优的超参数取值。

5.2 主要创新点

基于辅助分类器的自知识蒸馏（Be Your Own Teacher, BYOT）模型将神经网络中各个浅层块的信息一视同仁，可能会造成一些暗知识的损失。为了区分各浅层块信息对最深层块的不同影响，提出了两种解决方案：

（1）提出了 PD-BYOT 模型，通过给 BYOT 模型中不同深度的浅层块添加衰减系数将各个浅层块对最深层块的影响以等比数列的形式加以区分。

（2）提出了 SKDSAM 模型，将自知识蒸馏模型和自注意力机制以一种简单有效的方式结合起来。SKDSAM 模型给作为教师模型的最深层块和作为学生模型的各浅层块之间添加自注意力连接，准确量化了神经网络不同深度的浅层块对最深层块的不同影响，从而更有效地利用知识蒸馏中的暗知识。

5.3 未来工作展望

SKDSAM 模型虽然在实验中取得了显著的成效，但还有很多地方值得改进和扩展。未来可能的改进大致有以下方案。

（1）尝试把学生模型映射到较小的空间：在传统的知识蒸馏模型中，教师模型比学生模型大很多；但是在自知识蒸馏中，作为学生模型的浅层块和作为教师模型的最深层块常常大小相近。未来可考虑把学生模型映射到新的较小的空间中，以便进一步提升模型性能。

（2）尝试其他深度网络框架：因为时间所限，实现 SKDSAM 模型的卷积网络框架只尝试了深度残差网络（Deep Residual Networks, ResNet）、宽残差网络（Wide Residual Networks, WRN）和残差密集卷积网络（Densely Connected Convolutional

Networks, DenseNet）三种。未来将尝试用更多的网络框架实现 SKDSAM 模型，比如聚合残差变换网络（Aggregated Residual Transformations for Deep Neural Networks, ResNeXt）[49]，视觉几何组（Visual Geometry Group, VGG）[50]等。

（3）在更多数据集上进行实验：因为实验室计算资源有限，SKDSAM 模型只在几个较小的数据集上进行了实验。未来将在更大的数据集上进行实验，比如 ImageNet 数据集。

（4）在更多类型的任务上尝试 SKDSAM 模型：虽然 SKDSAM 模型在图像分类任务上取得了成功，但是还没有在语言模型上测试其可行性，未来将思考把 SKDSAM 模型应用到语言模型上。后续的工作也将考虑如何把 SKDSAM 模型应用于半监督学习和弱监督学习任务。

致 谢

时光飞逝，岁月如梭，转眼之间，我的硕士生涯即将结束。在这三年的时间里，我得到了许多人的指导和帮助，他们渊博的学识和丰富的经验使我获益良多。我在此向他们致以最衷心的感谢。

感谢我的导师何琨教授的耐心指导。我作为一名跨专业考来计算机专业的学生，本身人工智能的基础较为薄弱，对于自身研究方向的选择也一度迷茫。何老师带我尝试了很多方向的小组，包括自然语言处理、知识图谱、社交网络，最终来到了知识蒸馏的小组，并且遇到了意气相投的张硕玺学长一起做课题。在硕士论文初稿完成后，何老师多次给出中肯的修改建议，又让另一位博士生学长陈劲松帮忙纠错了两次，发现了很多我自己难以发现的问题。

感谢何老师的博士生张硕玺学长和陈劲松学长。张硕玺学长数学系出身，理论功底深厚实战能力强。本篇文章的选题建议是张学长提出来的，在代码实现和调试超参数的过程中也提出了很多有价值的建议。兴趣爱好上我和张师兄也有很多共同点，包括英超、欧冠、世界杯、篮球、古代史、游戏、金庸小说、娱乐圈八卦等。这些共同的兴趣爱好在科研之余给了我们很多乐趣，深化了我们的友谊，这反过来也让我们科研的合作更加顺畅。陈劲松学长虽然和我不是一个方向，但还是在硕士论文的写作格式、论文图表的规范性和严谨的书面表达上给了我很多有价值的建议，对于我硕士论文的顺利完成帮助很大。

感谢父母对我一直以来物质和精神方面的支持。像我这么大年纪的人，一般的家长早就逼着赶紧结婚生孩子了。但是家父家母非常开明，支持我勇敢追逐梦想，做自己想做的事。我也由衷地希望他们身体健康，幸福吉祥。

三年的研究生生涯紧张而充实，回首往事感慨良多。这三年的学习生涯令我从一个人工智能的门外汉到对这个学科有了略微的了解，也令我对人工智能的热爱进一步加深。虽然我的博士阶段研究方向将是组合优化和电路设计，我还是希望能够将所学到的人工智能知识应用到未来的研究中，为人工智能的发展、为了人类心智的荣耀做出应有的贡献。

参考文献

[1] Junyi Chai, Hao Zeng, Anming Li, Eric Ngai. Deep learning in computer vision: a critical review of emerging techniques and application scenarios. Machine Learning with Applications, 2021, 6: 100134

[2] 田萱，王亮，丁琪. 基于深度学习的图像语义分割方法综述. 软件学报，2019, 30(2): 440-468

[3] 鄂海红，张文静，肖思琪，程瑞，胡莺夕，周筱松等. 深度学习实体关系抽取研究综述. 软件学报，2019, 30(6): 1793-1818

[4] 张丹. 深度学习神经网络在语音识别中的应用探讨. 电子世界，2021, 6: 67-68

[5] Geoffrey Hinton, Oriol Vinyals, Jeff Dean. Distilling the knowledge in a neural network. arXiv preprint arXiv:1503.02531, 2015

[6] Jimmy Ba, Rich Caruana. Do deep nets really need to be deep? In: Conference on Neural Information Processing Systems (NIPS 2014), Montreal, QC, Canada, December 8-13, 2014: 27

[7] Seung Wook Kim, Hyo-Eun Kim. Transferring knowledge to smaller network with class-distance loss. In: International Conference on Learning Representations workshop (ICLR 2017), Toulon, France, April 24-26, 2017

[8] Linfeng Zhang, Jiebo Song, Anni Gao, Jingwei Chen, Chenglong Bao, Kaisheng Ma. Be your own teacher: improve the performance of convolutional neural networks via self distillation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV 2019), Seoul, Korea, October 27-November 2, 2019: 3713-3722

[9] Mingi Ji, Seungjae Shin, Seunghyun Hwang, Gibeom Park, Il-Chul Moon. Refine myself by teaching myself: feature refinement via self-knowledge distillation. In: Conference on Computer Vision and Pattern Recognition (CVPR 2021), Nashville, TN, USA, June 19-25, 2021: 10664-10673

[10] Yuenan Hou, Zheng Ma, Chunxiao Liu, Chen Change Loy. Learning lightweight lane detection cnns by self attention distillation. In: Proceedings of the IEEE/CVF international conference on computer vision (ICCV 2019), Seoul, Korea, October 27-

November 2, 2019: 1013-1021

[11] Yunteng Luan, Hanyu Zhao, Zhi Yang, Yafei Dai. Msd: multiself-distillation learning via multi-classifiers within deep neural networks. arXiv preprint arXiv:1911.09418, 2019

[12] Hankook Lee, Sung Ju Hwang, Jinwoo Shin. Self-supervised label augmentation via input transformations. In: International Conference on Machine Learning (ICML 2020), Sydney, NSW, Australia, July 12-18, 2020: 5714-5724

[13] Sukmin Yun, Jongjin Park, Kimin Lee, Jinwoo Shin. Regularizing class-wise predictions via self-knowledge distillation. In: Conference on Computer Vision and Pattern Recognition (CVPR 2020), Seattle, WA, USA, June 14-19, 2020: 13876-13885

[14] Tingbing Xu, Chenglin Liu. Data-distortion guided self-distillation for deep neural networks. In: Association for the Advancement of Artificial Intelligence (AAAI 2019), Honolulu, Hawaii, USA, January 27-February 1, 2019, 33(1): 5565-5572

[15] Hankook Lee, Sung Ju Hwang, Jinwoo Shin. Rethinking data augmentation: self-supervision and self-distillation. In: International Conference on Machine Learning (ICML 2020), Sydney, NSW, Australia, July 12-18, 2020

[16] Dzmitry Bahdanau, Kyunghyun Cho, Yoshua Bengio. Neural machine translation by jointly learning to align and translate. arXiv preprint arXiv:1409.0473, 2014

[17] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, et al. Attention is all you need. In: Advances in neural information processing systems (NIPS 2017), Long Beach, CA, USA, December 4-9, 2017: 30

[18] Mostafa Dehghani, Stephan Gouws, Oriol Vinyals, Jakob Uszkoreit, Łukasz Kaiser. Universal transformers. arXiv preprint arXiv:1807.03819, 2018

[19] Tao Shen, Tianyi Zhou, Guodong Long, Jing Jiang, Shirui Pan, Chengqi Zhang. Disan: Directional self-attention network for RNN/CNN-free language understanding. In: Proceedings of the AAAI Conference on Artificial Intelligence (AAAI 2018), New Orleans, LA, USA, February 2-7, 2018, 32(1)

[20] Chen Xing, Yu Wu, Wei Wu, Yalou Huang, Ming Zhou. Hierarchical recurrent attention network for response generation. In: Proceedings of the AAAI Conference on Artificial Intelligence (AAAI 2018) , New Orleans, LA, USA, February 2-7, 2018,

32(1)

[21] Yequan Wang, Minlie Huang, Xiaoyan Zhu, Li Zhao. Attention-based LSTM for aspect-level sentiment classification. In: Proceedings of the 2016 conference on empirical methods in natural language processing (EMNLP 2016), Austin, TX, USA, November 1-4, 2016: 606–615

[22] Heng She, Bin Wu, Bai Wang, Renjun Chi. Distant supervision for relation extraction with hierarchical attention and entity descriptions. In: International Joint Conference on Neural Networks (IJCNN 2018), Rio de Janeiro, Brazil, July 8-13, 2018: 1-8

[23] Sumit Chopra, Michael Auli, Alexander Rush. Abstractive sentence summarization with attentive recurrent neural networks. In: Proceedings of the conference of the North American chapter of the association for computational linguistics: human language technologies (NAACL 2016), San Diego, CA, USA, June 12-17, 2016: 93-98

[24] 张宇，张雷. 融入注意力机制的深度学习动作识别. Telecommunication Engineering, 2021, 61(10): 1205-1212

[25] 徐从安，吕亚飞，张筱晗，刘瑜，崔晨浩，顾祥岐. 基于双重注意力机制的遥感图像场景分类特征表示方法. 电子与信息学报, 2021, 43(3): 683-691

[26] 张祥东，王腾军，朱劭俊，杨耘. 基于扩张卷积注意力神经网络的高光谱图像分类. 光学学报, 2021, 41(3): 0310001

[27] Han Zhang, Ian Goodfellow, Dimitris Metaxas, Augustus Odena. Self-attention generative adversarial networks. In: International conference on machine learning Proceedings of Machine Learning Research (PMLR 2019), Long Beach, CA, USA, June 9-15, 2019: 7354-7363

[28] Tao Kong, Fuchun Sun, Chuanqi Tan, Huaping Liu, Wenbing Huang. Deep feature pyramid reconfiguration for object detection. In: Proceedings of the European conference on computer vision (ECCV 2018), Munich, Germany, September 8-14, 2018: 169-185

[29] Wei Li, Xiatian Zhu, Shaogang Gong. Harmonious attention network for person re-identification. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2018), Salt Lake City, Utah, USA, June 18-22, 2018: 2285-2294

[30] Hao Liu, Jiashi Feng, Meibin Qi, Jianguo Jiang, Shuicheng Yan. End-to-end comparative attention networks for person re-identification. IEEE Transactions on Image Processing, 2017, 26(7): 3492–3506

[31] Jun Fu, Jing Liu, Haijie Tian, Yong Li, Yongjun Bao, Zhiwei Fang, Hanqing Lu. Dual attention network for scene segmentation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR 2019), Long Beach, CA, USA, Jun 16-20, 2019: 3146-3154

[32] Jason Kuen, Zhenhua Wang, Gang Wang. Recurrent attentional networks for saliency detection. In: Proceedings of the IEEE Conference on computer Vision and Pattern Recognition (CVPR 2016), Las Vegas, NV, USA, June 26-July 1, 2016: 3668–3677

[33] Zhongzhan Huang, Senwei Liang, Mingfu Liang, Haizhao Yang. Dianet: dense-and-implicit attention network. In: Association for the Advancement of Artificial Intelligence (AAAI 2020), Hilton New York Midtown, NY, USA, February 7-12, 2020, 34(4): 4206-4214

[34] Hengshuang Zhao, Jiaya Jia, Vladlen Koltun. Exploring self-attention for image recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR 2020), Seattle, WA, USA, June 14-19, 2020: 10076-10085

[35] Kaiming He, Xiangyu Zhang, Shaoqing Ren, Jian Sun. Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR 2016), Las Vegas, NV, USA, June 26-July 1, 2016: 770-778

[36] Sergey Zagoruyko, Nikos Komodakis. Wide residual networks. arXiv preprint arXiv:1605.07146, 2016

[37] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, Kilian Weinberger. Densely connected convolutional networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR 2017), Honolulu, Hawaii, USA, July 21-26, 2017: 4700-4708

[38] Leo Breiman. Bagging predictors. Machine Learning, 1996, 24(2): 123–140

[39] Terrance DeVries, Graham Taylor. Improved regularization of convolutional neural networks with cutout. arXiv preprint arXiv:1708.04552, 2017

[40] Hongyi Zhang, Moustapha Cisse, Yann Dauphin, David Lopez-Paz. Mixup: beyond empirical risk minimization. arXiv preprint arXiv:1710.09412, 2017

[41] Alex Krizhevsky, Geoffrey Hinton. Learning multiple layers of features from tiny images. 2009: 7

[42] Jiayu Wu, Qixiang Zhang, Guoxi Xu. Tiny imagenet challenge. Technical report, 2017

[43] Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, Serge Belongie. The Caltech-UCSD birds-200-2011 dataset. Technical report, 2011

[44] Bangpeng Yao, Xiaoye Jiang, Aditya Khosla, Andy Lai Lin, Leonidas Guibas, Fei-Fei Li. Human action recognition by learning bases of action attributes and parts. In: International conference on computer vision (ICCV 2011), Barcelona, Spain, Nov 6-13, 2011: 1331–1338

[45] Aditya Khosla, Nityananda Jayadevaprakash, Bangpeng Yao, Fei-Fei Li. Novel dataset for fine-grained image categorization. In: First Workshop on Fine-Grained Visual Categorization, Institute of Electrical and Electronics Engineers Conference on Computer Vision and Pattern Recognition (CVPR 2011), Colorado Springs, CO, USA, June 20-25, 2011, 2(1)

[46] Ariadna Quattoni, Antonio Torralba. Recognizing indoor scenes. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2009), Miami Beach, Florida, USA, June 20-26, 2009: 413-420

[47] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, Fei-Fei Li. Imagenet: a large-scale hierarchical image database. In: IEEE conference on computer vision and pattern recognition (CVPR 2009), Miami Beach, Florida, USA, June 20-26, 2009: 248-255

[48] Simon Niklaus, Long Mai, Feng Liu. Video frame interpolation via adaptive separable convolution. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV 2017), Venice, Italy, October 22-29, 2017: 261-270

[49] Saining Xie, Ross Girshick, Piotr Dollár, Zhuowen Tu, Kaiming He. Aggregated residual transformations for deep neural networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR 2017), Honolulu, Hawaii, USA, July 21-26, 2017: 1492-1500

[50] Karen Simonyan, Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556, 2014

附录 1　攻读学位期间参加的科研项目

1. 中央高校基本科研业务费学科交叉专项

项目名称：大数据智能与先进计算基础研究平台建设

项目编号：No. 2019kfyXKJC021

起止时间：2019 年 8 月至 2021 年 8 月

担任角色：参与人

附录 2　中英文缩写对照表

BYOT	Be Your Own Teacher（基于辅助分类器的自知识蒸馏）
CS-KD	Class-wise Self-knowledge Distillation（类级别的自知识蒸馏）
DDGSD	Data-Distortion Guided Self-Distillation（基于数据扭曲的自知识蒸馏）
DenseNet	Densely Connected Convolutional Networks（残差密集网络）
DIANet	Dense-and-Implicit Attention Network（密集隐式注意力网络）
FRSKD	Feature Refinement via Self-Knowledge Distillation（基于自知识蒸馏的特征精炼）
PD-BYOT	Per-block Decay based Be Your Own Teacher（基于逐块衰减辅助分类器的自知识蒸馏）
ResNet	Deep Residual Networks（深度残差网络）
SAN	Self-Attention Network（自注意力网络）
SKDSAM	Self-Knowledge Distillation with Self-Attention Mechanism（基于自注意力机制的自知识蒸馏）
SLA	Self-supervised Label Augmentation（自监督标签增强）
SLA-SD	Self-supervised Label Augmentation based Self-Distillation（基于自标签增强的自知识蒸馏）
WRN	Wide Residual Networks（宽残差网络）