**分类号**  **学号 M201973167**

**学校代码10487 密级**

****

**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** | **：** |  | **XXXX年XX月XX日** |  |  |

**A Dissertation Submitted in Partial Fulfillment of the Requirements**

**for the Master Degree in Engineering**

**Research on Self-Knowledge Distillation with Self-Attention**

**Candidate : GAO Ye**

**Major : Computer Software and Theory**

**Supervisor : Prof. He Kun**

**Huazhong University of Science and Technology**

**Wuhan 430074, P. R. China**

**May, 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。

保 密□，在 年解密后适用本授权书。

本论文属于

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学到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，基于特征的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的信息视为教师模型，把神经网络的浅层的信息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，忽略了各个浅层块特征图对最深层块特征图的不同影响。

通过分析BYOT模型的自蒸馏原理，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，以区分达到将各浅层块对最深层块的影响的目的。然后，将PD-BYOT模型与三种数据增强模型（cutout模型、SLA模型、mixup模型）相结合，作为进一步提升模型的性能。

在BYOT模型和PD-BYOT的基础上，提出一种全新的自知识蒸馏模型——基于自注意力机制的自知识蒸馏模型（Self-Knowledge Distillation with Self-Attention, SKDSA）。SKDSA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，为不同浅层块计算出不同的注意力权重，从而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。除了添加自注意力模块以外，SKDSA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SKDSA模型中的自注意力模块等价于集成学习中的袋装法，这证实了SKDSA模型具有更强的稳定性和抗过拟合能力。最后，将SKDSA模型与三种数据增强模型（cutout模型、SLA模型、mixup模型）相结合，作为进一步提升模型性能的备选方案。

实验结果表明，PD-BYOT模型相比于BYOT模型在性能上有一定的提升，结合数据增强之后，又有进一步的提升。（适当扩充）进一步地，SKDSA模型在多个图像数据集上取得了相比于现有其他的自知识蒸馏模型更强的性能。消融实验则验证了自注意力模块和自注意力模块中的知识蒸馏模块对提升SKDSA模型性能的必要性，以及结合SKDSA模型和数据增强模型可以进一步提升模型的性能。

**关键词：** 自知识蒸馏；自注意力机制；数据增强；集成模型；暗知识

# Abstract

Knowledge distillation is the process of transferring knowledge learned from one model to another. The model providing knowledge is called the teacher, and the model receiving knowledge is called the student. Unlike traditional knowledge distillation, the feature-based 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. Be your own teacher: Improve the performance of convolutional neural networks via self distillation, namely BYOT) treat all shallow layers equally, neglecting their different impact on the deepest layer.

In order to improve the performance of self-distillation, we first propose the decay-BYOT model, which utilizes the geometric progression of attenuation coefficient to differentiate each shallow layer’s impact on the deepest layer. Since decay-BYOT’s performance improvement is not obvious compared to the original model, we further put forward a novel framework, which we refer to as Self-Knowledge Distillation with Self-Attention (SKDSA). It adds attention links between each shallow layer and the deepest layer of BYOT, computes the attention score of each shallow layer, so as to quantify the contribution of each shallow layer’s feature map to the deepest layer’s feature map. Apart from the self-attention module, the SKDSA model modifies two BYOT’s loss functions to mine the network’s dark knowledge more efficiently. We then provide theoretical proof that the proposed approach is essentially an ensemble modeling strategy (namely bagging), which means SKDSA has the advantage of robustness and preventing overfitting. Moreover, we combine the SKDSA model with three data augmentation models (cutout, SLA, mixup) to further improve model performance.

The experimental results show that the SKDSA model outperforms all the self-distillation models on various image datasets. Moreover, the ablation study proves the necessity of self-attention links and temperature scaling in the self-attention module. The ablation study also shows that combining the SKDSA model and data augmentation techniques can further improve performance.

**Keywords:** Attention, Self-attention, Knowledge Distillation, Self-knowledge Distillation

# 目 录

[基于自注意力机制的自知识蒸馏研究 I](#_Toc100852462)

[A Dissertation Submitted in Partial Fulfillment of the Requirements II](#_Toc100852463)

[for the Master Degree in Engineering II](#_Toc100852464)

[摘 要 I](#_Toc100852465)

[Abstract III](#_Toc100852466)

[目 录 IV](#_Toc100852467)

[1 绪论 1](#_Toc100852468)

[1.1 研究背景与意义 1](#_Toc100852469)

[1.2 国内外研究现状 2](#_Toc100852470)

[1.3 论文主要内容 9](#_Toc100852471)

[2 BYOT模型问题分析 12](#_Toc100852472)

[2.1 BYOT模型介绍 12](#_Toc100852473)

[2.2 PD-BYOT模型 14](#_Toc100852474)

[2.3 实验结果与分析 16](#_Toc100852475)

[2.4 本章小结 19](#_Toc100852476)

[3 基于自注意力机制的自知识蒸馏 20](#_Toc100852477)

[3.1 自注意力模型的结构 20](#_Toc100852478)

[3.2 基于自注意力机制的自知识蒸馏模型 21](#_Toc100852479)

[3.3 SKDSA模型和袋装法的等价性证明 30](#_Toc100852480)

[3.4 SKDSA模型和数据增强的结合 31](#_Toc100852481)

[3.5 本章小结 33](#_Toc100852482)

[4 实验验证与分析 34](#_Toc100852483)

[4.1 实验设置 34](#_Toc100852484)

[4.2 实验结果与分析 38](#_Toc100852485)

[4.3 消融实验与分析 42](#_Toc100852486)

[4.4 本章小结 46](#_Toc100852487)

[5 总结与展望 48](#_Toc100852488)

[5.1 主要工作总结 48](#_Toc100852489)

[5.2 主要创新点 48](#_Toc100852490)

[5.3 未来工作展望 49](#_Toc100852491)

[致 谢 50](#_Toc100852492)

[参考文献 51](#_Toc100852493)

[附录1 攻读硕士学位期间取得的研究成果 56](#_Toc100852494)

[附录2 攻读学位期间参加的科研项目 57](#_Toc100852495)

[附录3 中英文缩写对照表 58](#_Toc100852496)

# 绪论

## 研究背景与意义

近年来深度神经网络技术获得了迅猛的发展和广阔的应用，在计算机视觉[1]、自然语言处理[2]、语音识别[3]等领域都取得了非凡的成果。为了获得越来越优异的性能，深度神经网络的规模与日俱增，如表1.1所示，GPT-3模型的参数量甚至达到了1750亿。大型神经网络的训练需要昂贵的计算和时间成本，比如GPT-3训练一次的时间约为355个GPU年，成本约为460万美元。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)。

表1.1 大型神经网络的参数量（百万）

|  |  |
| --- | --- |
| 大型神经网络模型 | 参数量 |
| ELMO | 94 |
| BERT | 340 |
| GPT-2 | 1542 |
| Megatron | 8 000 |
| T5 | 11 000 |
| Turing NLG | 17 000 |
| GPT-3 | 175 000 |

虽然直接训练小型神经网络的性能较差，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，就可以大幅提升小型神经网络的性能。受此启发，Hinton等人[4]提出知识蒸馏技术：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。然而，传统的知识蒸馏技术存在一系列问题。首先，预训练大型的教师模型需要更多的训练时间。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，人们提出一系列自知识蒸馏[6][7][8][9]技术，即不利用外部的教师模型，利用学生模型自身的知识实现知识蒸馏。

## 国内外研究现状

### 知识蒸馏模型

知识蒸馏的主要思想是让学生模型直接模仿教师模型的最终预测。由于知识蒸馏是一种简单有效的模型压缩方法，目前知识蒸馏已经广泛用于不同的任务和场景。Hinton等人[4]提出通过学习软化标签的概率分布，将知识从教师模型迁移到学生模型。软化的标签通过引入温度缩放来提升低概率类别的概率。这个模型在图像分类任务上取得了令人惊喜的效果。

Hinton等人[4]提出的知识蒸馏模型原理如图1.1所示，它在较大的教师模型的帮助下训练较小的学生模型。知识由教师模型迁移到学生模型，通过最小化教师模型产生的logits和学生模型产生的logits的差异来实现。不同于其他神经网络压缩方式，知识蒸馏可以无视教师模型和学生模型结构上的差异。



图1.1 知识蒸馏原理

在许多情况下，教师模型logits的归一化指数的输出在正确类型上概率很高，在非正确类型上概率趋近于零。在这种情况下，除了数据集中已有的正确标签外，模型无法提供更多的信息。为了解决这个问题，Hinton[4]提出知识蒸馏的概念。假定代表由教师模型输出的向量，是蒸馏温度参数，则一张图像的属于第个类别概率为：

|  |  |  |
| --- | --- | --- |
|  |  | （1.1） |

当时，可以得到正规的归一化指数公式（也就是没有蒸馏）。随着增加，归一化指数函数产生的概率分布越来越分散，提供越来越多的关于教师认为哪个类别和预测的类别更为相似的信息。教师模型提供的信息被称为暗知识，它影响着蒸馏模型的信息流。当计算蒸馏损失函数时，使用和教师模型相同的温度来计算学生模型的logits。Hinton[4]还发现训练学生模型时把正确标签和教师的软化标签放在一起更有效。因此，学生模型的损失函数可以定义为学生模型预测出的标签和真实标签的差距。总损失函数由学生损失函数和蒸馏损失函数相加而成：

|  |  |  |
| --- | --- | --- |
|  |  | （1.2） |

其中函数代表交叉熵损失函数，代表学生模型预测的概率分布，代表真实标签的独热向量，代表知识蒸馏的系数，代表知识蒸馏的温度，代表蒸馏温度时教师模型的概率分布，代表蒸馏温度时学生模型的概率分布。

Ba等人[10]发现浅层神经网络可以学习以前由深度网络学习的复杂函数，并实现以前只有深度模型才能实现的精度。在某些情况下，浅层神经网络可以使用与原始深度模型相似的参数总数来学习这些深度函数。Kim等人[11]提出类距离损失，帮助教师网络形成密集聚类的向量空间，使学生网络更容易从中学习。

### 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。为了解决这些问题，研究者提出了自知识蒸馏的模型。自知识蒸馏的目标是让学生模型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网络比浅层网络包含更高阶、更抽象的信息，深层部分可以用来向浅层部分蒸馏知识。换言之，基于特征的自知识蒸馏通过在不同神经网络层之间添加辅助连接来实现利用中间层信息的目的。BYOT模型[6]根据深度将神经网络分成若干个浅层块，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，以便把信息从神经网络深层传递到神经网络浅层。Ji等人[8]提出了通过自知识蒸馏进行特征精炼模型（Feature Refinement via Self-Knowledge Distillation, FRSKD）。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。而在Hou等人[12]的工作中，相比于直接蒸馏特征，他们使用基于注意力的模型来迫使神经网络浅层去模仿神经网络深层的注意力图。Luan等人‍[13]将每个层分支打造成一个分类器，把最深层的分类器的特征图和logits蒸馏到浅层分类器。



图1.2 基于特征的自知识蒸馏

另一种类型的自知识蒸馏模型则是基于训练样本的数据增强，其原理如图1.3所示。Yun等人[7]提出按类别预测的正则化方法，并在此基础上提出了基于按类别预测正则化方法的自知识蒸馏模型（Regularizing Class-wise Predictions via 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由教师模型和学生模型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”的自知识蒸馏模型（Data-Distortion Guided Self-Distillation for Deep Neural Networks, DDSGD），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。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logits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

### 自注意力模型

基于神经网络的注意力机制最早由Bahdanau等人[16]提出，现已变成深度学习中的一种流行的架构。注意力在人工智能的诸多方向已经有了广泛的应用，比如自然语言处理，语音识别和计算机视觉。注意力机制类比于人体的生物学机制。比如，人体视觉系统倾向于选择性地关注图像的某些部分，同时忽视其他不相关的信息。类似的，在语言、语音或图像等任务中，输入的某些部分比其他部分更重要。比如，在机器翻译或摘要提取的任务中，只有输入里面的某些特定单词对预测下一个单词有帮助。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]、摘要归纳[23]等。Vaswani[17]等人提出Transformer架构具有划时代的意义：它彻底摒弃了递归和卷积，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RNN模型结合起来。Shen等人[19]提出了“定向自注意力网络” （Directional selfattention network, DiSAN）来学习句子嵌入，它仅基于所提出的注意力而没有任何 RNN/CNN 结构。DiSAN由具有时间顺序编码的定向自注意力组成，然后将序列压缩为向量表示的多维注意力。Xing等人[20]提出了“分层循环注意网络”（Hierarchical recurrent attention network, HRAN）来在一个统一的框架中对这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话语级别的注意力来提取话语内部和话语之间的重点。

注意力机制同样可以应用在计算机视觉任务中，比如动作识别、图像分类[24][25][26][27]、图像生成[28]、目标检测[29]、人物识别[30][31]、分割[32]、显着性检测[33]等等。Fu等人[24]提出了循环注意卷积神经网络，它用强化学习在多个尺度上递归地学习的判别区域注意力和基于区域的特征表示。Zhang等人[28]提出基于自注意力的生成对抗网络，它能够完成基于注意力的、远距离依赖的图像生成任务。Kong等人[29]提出特征重配架构，以高度非线性但有效的方式将低级表示与高级语义特征相结合。他们的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‍[30]提出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，以便优化不受控制图像中人员的识别。

对于注意力模块结构的改进也有了一系列的探讨[34][35]。Huang等人[34]提出的密集隐式注意网络(Dense-and-Implicit Attention Network, DIANet)[34]。已有的注意力模型将注意力模块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模块，以减少深度神经网络的参数，加强网络层之间信息的集成。Zhao等人[35]探讨了自注意结构两种类型的变体。第一种变体是成对自注意力。成对注意力与卷积不同有两点不同：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，成对注意力不会将固定权重附加到特定位置，并且对排列和基数保持不变。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任何影响。第二种变体是互补注意力。互补注意力是卷积的一种推广。它没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。

### 卷积网络模型

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也越来越多。然而，随着神经网络层数的不断增加，传统卷积网络越来越难以训练。为此，研究者对卷积网络提出一系列改进，包括残差神经网络[36]，宽残差神经网络[37]，残差密集网络[38]等等。

（1）残差网络模型

研究表明，层数越多的神经网络具有越强的表征能力。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。因此，如何训练越来越深的卷积网络成为了一个研究热点。

He等人[36]提出了深度残差网络（Deep residual networks, ResNet）来解决上述问题。残差模型不是期望每个网络层直接适配所需的底层映射，而是明确地让这些层去适配残差映射。记所需的底层映射为，残差模型让堆叠的非线性层拟合另一个映射。原始映射因此表示为。假设优化残差映射比优化原始映射容易。在极端情况下，如果恒等映射是最优的，则将残差设为零要比通过一系列非线性层拟合恒等映射更容易。

如图1.4所示，可以通过具有“快捷连接”的前馈神经网络来实现。快捷连接是指跳过一层或多层的连接。在残差网络中，快捷连接只是简单地执行恒等映射，它们的输出被添加到堆叠层的输出中。恒等快捷连接既不增加额外的参数，也不增加计算复杂性。



图1.4 残差块的结构

（2）宽残差网络模型

深度残差神经网络被证明能够扩展到数千层并且仍然具有改进性能的空间。然而，每提高百分之一的准确率都会增加几乎一倍的层数，因此训练非常深的残差神经网络存在减少特征重用的问题，这使得这些神经网络的训练速度非常慢。为了解决这些问题，宽残差神经网络（Wide Residual Networks, WRN）[37]在残差神经网络块的架构的基础上，使用了一种减少深度和增加残差神经网络宽度的新架构。实验证实即使是一个简单的 16 层的宽残差神经网络在准确性和效率上也优于所有以前的残差神经网络在多个数据集上的结果。

宽残差网络的结构如图1.5所示。神经网络的宽度由因子决定。原始残差网络的宽度默认为1。方括号中的内容代表一组又一组的卷积结构，乘数代表一组中块的个数。为了表示方便最终的分类层被省略。



图1.5 宽残差网络的结构[37]

（3）残差密集网络

研究表明，如果卷积网络在靠近输入的层和靠近输出的层之间包含较短的连接，则它们可以更有效地训练。Huang等人[38]根据这一观察提出残差密集卷积网络（Densely Connected Convolutional Networks, DenseNet），它以前馈方式将每一层连接到其他每一层。对于每一层，所有前面层的特征图都被用作输入，它自己的特征图被用作所有后续层的输入。残差密集网络缓解了梯度消失问题，加强了特征传播，加强特征重用，并大大减少了参数的数量。

残差密集网络为了确保网络中各层之间的最大信息流，将所有层直接相互连接。残差密集网络为了保持前馈特性，每一层都从所有前面的层获得额外的输入，并将其自己的特征图传递给所有后续层，这种结构如图1.6所示。和残差网络不同的是，残差密集网络在将特征传递到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，某一层的输入由所有先前卷积块的特征图组成。



图1.6 残差密集网络示意图

## 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对最深层块特征的不同影响。

为了提升BYOT模型的性能，首先提出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用等比数列的形式加以区分。随后进一步提出一种新的基于自注意力机制的自知识蒸馏模型（Self-Knowledge Distillation with Self-Attention, SKDSA）。SKDSA模型可以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分数。实验结果显示SKDSA模型比其他自蒸馏模型和自注意模型获得了更优异的分类准确性。

论文共分为5章内容，章节内容之间的关系如图1.7如所示。



图1.7 组织结构图

第一章 绪论：本章首先介绍知识蒸馏的提出背景和意义。接下来介绍了与文章内容相关的（包括知识蒸馏、自注意力机制、数据增强）的国内外研究现状。最后以图表和文字说明文章的组织结构。

第二章 BYOT模型问题分析：介绍了BYOT模型的网络结构和损失函数，分析了BYOT模型存在的问题，给BYOT模型增加衰减系数作为改进方案，并实验测试了PD-BYOT模型的性能。

第三章 基于自注意力机制的自知识蒸馏：在BYOT模型和PD-BYOT模型的基础上提出新模型SKDSA。详细介绍了新模型的网络结构，损失函数，训练流程，说明新模型相比BYOT模型所做的改进。随后证明SKDSA模型和袋装法‍[39]的等价性。最后将SKDSA方法和数据增强中的cutout模型[40]、SLA方法[9]、mixup方法[41]结合起来。

第四章 实验验证和分析：本章在多个数据集上（CIFAR-100[42], Tiny ImageNet[43], Caltech-UCSD Birds 200[44], Stanford 40[45], Stanford Dogs[46], MIT Indoor Scene Recognition[47]）测试了SKDSA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，FRSKD模型）和自注意力模型（DIANet模型，SAN模型）对比了的实验性能。接着在消融实验中，测试了去掉自注意力模块、去掉自注意力模块中知识蒸馏的对于模型性能的影响，分别测试了结合三种数据增强方法（cutout模型、SLA模型、mixup模型）后的SKDSA模型性能，测试了超参数的其它取值对实验结果的影响。

第五章 总结与展望：本章总结全文的主要工作，说明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。

# BYOT模型问题分析

## BYOT模型介绍

### BYOT模型的网络结构

BYOT模型[6]是一种基于特征的自知识蒸馏模型，它将模型的神经网络深层信息蒸馏到模型的神经网络浅层。BYOT模型的网络结构如图2.1所示：残差神经网络根据其深度分为几个浅层块（ResBlock 1, ResBlock 2, ResBlock 3）和一个最深层块（ResBlock 4），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

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的归一化指数输出视为学生模型，将最深层的归一化指数输出视为教师模型。在推理阶段，将每个浅层块之下添加的瓶颈层、全连接层和指数归一化层全部移除。

### BYOT模型的损失函数

记个样本组成的数据集为，其中。记所有类组成的集合为，其中。记BYOT模型中的各个分类器为，其中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。记为通过第个浅层块下面的全连接层后输出的概率分布，则它再通过加入温度系数的归一化指数层后，所输出向量属于第个类别的概率为：

|  |  |  |
| --- | --- | --- |
|  |  | （2.1） |

其中代表分类器中的蒸馏温度，初始被设为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）归一化指数输出的交叉熵，最深层分类器归一化指数输出概率分布和每个浅层分类器归一化指数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损失函数。

BYOT模型的损失函数1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）归一化指数输出概率分布的交叉熵。BYOT模型通过损失函数1把隐含在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模型的损失函数1由式（2.2）表示，其中代表某样本经过第个浅层分类器的加入温度系数的归一化指数层输出的概率分布，代表该样本在数据集中对应真实标签的独热向量，代表交叉熵损失函数：

|  |  |  |
| --- | --- | --- |
|  |  | （2.2） |

BYOT模型的损失函数2是最深层分类器归一化指数输出概率分布和每个浅层分类器归一化指数输出概率分布的相对熵。BYOT模型通过损失函数2将最深层分类器的影响传递到每个浅层分类器，迫使添加的浅层分类器模仿最深层分类器的预测结果。

BYOT模型的损失函数2由式（2.3）表示，其中代表第个浅层分类器经过加入温度系数的归一化指数层输出的概率分布，代表最深层分类器经过加入温度系数的归一化指数层输出的概率分布，代表相对熵损失函数：

|  |  |  |
| --- | --- | --- |
|  |  | （2.3） |

BYOT模型的损失函数3是最深层分类器特征图和每个浅层分类器特征图的损失函数。BYOT模型通过损失函数3将最深层分类器特征图中的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。

BYOT模型的损失函数3由式（2.4）表示，其中表示第个浅层分类器的特征图，表示最深层分类器的特征图，代表损失函数：

|  |  |  |
| --- | --- | --- |
|  |  | （2.4） |

综合上述三种损失函数（式（2.2）（2.3）（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，其中和是平衡式（2.2）（2.3）（2.4）的超参数，代表BYOT模型的总损失函数：

|  |  |  |
| --- | --- | --- |
|  |  | （2.5） |

## PD-BYOT模型

在卷积神经网络处理图像的任务中，卷积神经网络的中间层可以视为特征提取器，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网络层的信息一视同仁，可能会造成一些信息的丢失。为了改进这一点，区分各层对总损失函数的不同的贡献度很有必要。

为了使每个浅层块的损失函数对总损失函数的贡献度都有所不同，最简单直接的方法是给每个浅层块的损失函数1、损失函数2、损失函数3分别添加一个额外的超参数，，，这三个超参数都随浅层块数以等比的形式乘到原有的损失函数上。下文将此模型称为基于逐块衰减的BYOT模型（Per-block Decay based BYOT, PD-BYOT），三个额外的超参数，，称为衰减系数。PD-BYOT模型的网络结构如图2.2所示。



图2.2 PD-BYOT的网络结构

使用衰减系数修正后的第个块的损失函数1为：

|  |  |  |
| --- | --- | --- |
|  |  | （2.6） |

使用衰减系数修正后的第个块的损失函数2为：

|  |  |  |
| --- | --- | --- |
|  |  | （2.7） |

使用衰减系数修正后的第个块的损失函数3为：

|  |  |  |
| --- | --- | --- |
|  |  | （2.8） |

综合式（2.6）（2.7）（2.8），得到修正后的总损失函数为：

|  |  |  |
| --- | --- | --- |
|  |  | （2.9） |

## 实验结果与分析

### 实验环境

实验环境为在AutoDL公司外租的服务器。系统为Ubuntu 18.04版本，CPU为Intel(R) Xeon(R) Silver 4110 CPU @ 2.10GHz，内存15GB，GPU为Nvidia 2080ti，显存11GB。实验使用AutoDL自带的系统环境，运行模型需要的主要包配置如表4.1所示：

表2.1 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 |

### 实验数据集

数据集选择CIFAR-100，其中训练集有50000个样本，验证集有10000个样本。

### 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设为0.9，权重衰减（weight decay）设为，动量（momentum）设为0.9。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。一批训练样本的数量（batch size）设为128。

### 网络架构

使用残差网络（ResNet）实现PD-BYOT模型。设残差网络第一个卷积层的卷积核大小为，步长为1，填充为1。

### 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，样本总数为，预测概率最大的标签是正确标签的样本数量为，则分类准确率的计算公式为

|  |  |  |
| --- | --- | --- |
|  |  | （2.10） |

实验的优化的目标是尽可能提高验证集中的分类准确率。

### 对比模型

对比模型为原版的BYOT模型，它等价于将衰减系数，，均设为1的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，使用多组损失系数，，的在数据集CIFAR-100上进行实验，并且和原版模型进行对比。实验结果记录最后轮次的分类准确率。实验结果如表2.2所示，最优的结果用粗体标注。

从表2.2可以看出，在通用图像数据集上，PD-BYOT模型可以略微提升原BYOT模型的性能。具体来说，在取值0.6，取值1.1，取值1（或取值1，取值0.5，取值1）时，可以取得最优的实验结果，比原BYOT模型提升准确率0.98%。

这个结果说明PD-BYOT模型有两点不足。第一，PD-BYOT模型相比原BYOT模型性能提升并不明显。第二，深度神经网络作为一个复杂的动力系统，很难证明式（2.6）（2.7）（2.8）中的等比数列能够准确刻画不同浅层块对最深层块的真实影响。第三，衰减系数的变化对模型性能的影响没有规律可循，这使寻找最优的衰减系数非常困难。为了解决这些问题，在第三章提出一种性能更优，同时更有效区分不同浅层块贡献度的方法。

表2.2 衰减系数，，在不同取值的分类准确性（%）

| d\_ce | d\_kd | d\_feature | 准确率（%） |
| --- | --- | --- | --- |
| 1 | 1 | 1 | 76.01 |
| 1.1 | 0.5 | 1 | 76.46 |
| 0.9 | 1 | 1 | 76.19 |
| 0.9 | 0.5 | 1 | 75.61 |
| 0.8 | 1 | 1 | 76.26 |
| 0.7 | 1 | 1 | 76.58 |
| 0.7 | 1.1 | 1 | 75.38 |
| 0.7 | 1.2 | 1 | 76.45 |
| 0.6 | 0.5 | 1 | 76.53 |
| 0.6 | 0.6 | 1 | 74.92 |
| 0.6 | 0.7 | 1 | 73.68 |
| 0.6 | 0.8 | 1 | 75.77 |
| 0.6 | 1 | 1 | 76.64 |
| **0.6** | **1.1** | **1** | **76.99** |
| 0.6 | 1.2 | 1 | 75.63 |
| 0.5 | 1 | 1 | 75.91 |
| 1 | 0.4 | 1 | 74.86 |
| **1** | **0.5** | **1** | **76.99** |
| 1 | 0.6 | 1 | 72.24 |
| 1 | 0.9 | 1 | 75.26 |
| 1 | 1.1 | 1 | 76.05 |
| 1 | 1.2 | 1 | 76.55 |
| 1 | 1.3 | 1 | 75.63 |
| 1 | 1.5 | 1 | 75.7 |
| 1 | 1 | 0.5 | 75.34 |
| 1 | 1 | 0.9 | 75.43 |
| 1 | 1 | 1.1 | 76.02 |
| 1 | 1 | 1.2 | 76.12 |
| 1 | 1 | 1.3 | 76.11 |
| 1 | 1 | 1.5 | 75.78 |

## 本章小结

本章2.1节介绍了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。

2.2节指出BYOT模型存在的问题是忽略了各个浅层块信息对最深层块的不同影响，并提出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用等比数列的形式加以区分。

2.3节通过实验证实了PD-BYOT模型仅能略微改进原BYOT模型的性能，为第三章提出进一步的改进方案——结合自注意力机制的BYOT模型做铺垫。

# 基于自注意力机制的自知识蒸馏模型

在第二章中，原始的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PD-BYOT模型则把各个浅层块对最深层块的影响用等比数列的形式加以区分，很难证明这与各个浅层块对最深层块的真实影响相符合。那么，怎样才能准确区分出各个浅层块对最深层块的不同贡献度呢？本章的解决方案是给BYOT模型增加了自注意力模块，让自注意力模块通过训练“学习”得出各个浅层块对最深层块的不同贡献度，从而使不同深度的网络层能够更有效地聚合。

## 自注意力模型的结构

常规的自注意力的计算流程如图3.1所示，这种架构最早由Vaswani[17]等人在著名的论文“Attention Is All You Need”提出。自注意模型的输入为，，。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，爱奇艺所有的视频为，按相关性检索出的结果为。自注意力模块的计算流程可以视为信息检索的流程。

对于某个，首先计算出它和的相似度（一般为两者点积）：

|  |  |  |
| --- | --- | --- |
|  |  | （3.1） |

再由相似度计算得出相应的注意力权重：

|  |  |  |
| --- | --- | --- |
|  |  | （3.2） |

最后将与相应的注意力权重相乘再相加，即可得到注意力模块的输出：

|  |  |  |
| --- | --- | --- |
|  |  | （3.3） |



图3.1 自注意力模型的计算流程

## 基于自注意力机制的自知识蒸馏模型

### SKDSA模型的网络结构

为了准确量化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[34][35]的启发，本节将自注意力模型和自知识蒸馏模型相结合，提出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信息的不同重要性。

SKDSA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模块，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。自注意力模块的结构将在3.2.2详述。



图3.2 SKDSA模型的网络结构

每个浅层块和相应的瓶颈层、全连接层、加入温度系数（默认设为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（默认设为4）的归一化指数层组成最深层分类器。在训练阶段，SKDSA模型将所有浅层分类器的归一化指数输出视为学生模型，而将最深层的归一化指数输出视为教师模型。在推理阶段，将每个浅层块之下添加的瓶颈层、全连接层和指数归一化层全部移除，将最深层块和每个浅层块之间的自注意力模块也全部移除。

对比将SKDSA模型的网络结构（图3.2）和BYOT模型的框架（图2.1），SKDSA模型比BYOT模型增加了自注意力模块（上方红色部分）。SKDSA模型通过自注意力模块把最深层块的特征图和浅层块的特征图联系起来，以便提升模型的稳定性和泛化能力。

### SKDSA模型中的自注意力模块

类似于图3.1中自注意力模型的计算流程，SKDSA中自注意力模块的计算流程如图3.3所示。记个样本组成的数据集为，其中。记所有类组成的集合为，其中。记SKDSA模型中的各个分类器为，其中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。记为通过第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个类别的概率为：

|  |  |  |
| --- | --- | --- |
|  |  | （3.4） |

其中代表分类器中的蒸馏温度，默认设为4。将式（3.4）记为函数，即蒸馏函数。

如图3.3所示，自注意力模块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。记第个浅层分类器的特征图为，最深层分类器的特征图为。在浅层分类器的特征图和最深层分类器的特征图后分别添加一个非线性投影层，分别得到结果和。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，二是为了使和的维数相同（记为），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，类似于式（3.4）中对知识蒸馏的定义，将投影层输出向量的蒸馏函数定义为：

|  |  |  |
| --- | --- | --- |
|  |  | （3.5） |

其中是对自注意力模块的投影层输出向量的蒸馏温度。值得一提的是，投影层输出的蒸馏温度和分类器输出的蒸馏温度是彼此独立的。在第四章的4.3.3部分将专门探讨对模型性能的影响。



图3.3 SKDSA自注意力模块的计算流程

根据3.1节的介绍，求解自注意力模型需要、、。对所有投影层输出的结果统一使用蒸馏函数（式（3.5）），得到SKDSA模型中第个浅层分类器的输出和最深层分类器的输出，它们分别对应自注意模块的和：

|  |  |  |
| --- | --- | --- |
|  |  | （3.6） |

其中。自注意模块需要的将根据下文的式（3.11）求取。

利用和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，即可得到第个浅层分类器和最深层分类注意力连接的对应的权重系数：

|  |  |  |
| --- | --- | --- |
|  |  | （3.7） |

其中。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系数之和满足。



图3.4 SKDSA模型中自注意力模块的直观理解

对SKDSA模型的直观理解如图3.4所示，它在第个浅层块（学生模型）和最深层块（教师模型）上增加一个自注意力模块，用来提取自注意力模型中的和；然后通过和计算出第个浅层块对应的注意力权重系数。计算得出的注意力权重系数即代表第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。

### SKDSA模型的损失函数

表3.1 BYOT模型和SKDSA模型损失函数的区别

|  | BYOT模型 | SKDSA模型 |
| --- | --- | --- |
| 损失函数1 | 计算所有分类器 | 只计算最深层分类器 |
| 损失函数2 | 直接计算相对熵 | 相对熵乘以温度的平方 |
| 损失函数3 | 损失函数 | 归一化损失函数 |
| 总损失函数 | 与超参数相乘后直接相加 | 与注意力权重系数相乘 |

类似于BYOT模型，SKDSA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归一化指数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损失函数。为了获得更优异的实验性能，SKDSA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。

SKDSA模型的损失函数1是真实标签的独热向量与最深层分类器归一化指数输出概率分布的交叉熵。SKDSA模型通过损失函数1把隐含在数据集中的知识从真实标签引入到模型的最深层分类器。

SKDSA模型的损失函数1由式（3.8）表示，其中代表某样本经过最深层分类器的加入温度系数的归一化指数层输出的概率分布，代表该样本在数据集中对应的真实标签形成的独热向量，代表交叉熵损失函数：

|  |  |  |
| --- | --- | --- |
|  |  | （3.8） |

交叉熵函数具备半正定性。当它的值为0时，意味着概率分布和相同；当它的值越大，意味着概率分布和的差异越大。由于概率分布和都是归一化指数层的输出，的海森矩阵正定的，所以是一个凸函数，这意味着它的局部最优等于全局最优。

SKDSA模型损失函数1（式（3.8））和BYOT模型损失函数1（式（2.2））的区别是：BYOT模型损失函数1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SKDSA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。实验证实这样可以取得更高的分类准确率。

SKDSA模型的损失函数2是最深层分离器归一化指数输出概率分布和每个浅层分类器归一化指数输出概率分布的相对熵。SKDSA模型通过损失函数2将最深层分类器的影响传递到每个浅层分类器，迫使添加的浅层分类器模仿最深层分类器的预测结果。

SKDSA模型的损失函数2由式（3.9）表示，其中代表第个浅层分类器经过加入温度系数的归一化指数层输出的概率分布，代表最深层分类器经过加入温度系数的归一化指数层输出的概率分布，代表相对熵损失函数，代表蒸馏温度（默认设为4）。

|  |  |  |
| --- | --- | --- |
|  |  | （3.9） |

相对熵函数具备半正定性。当它的值为0时，意味着概率分布和概率分布相同；当它的值越大，意味着概率分布和的差异越大。由于标签和都是归一化指数层输出的概率分布，的海森矩阵正定的，所以是一个凸函数，这意味着它的局部最优等于全局最优。另外，由于相对熵函数具有不对称性，学生模型和教师模型也具有不对称性，因此相对熵很适合描述学生模型和教师模型的差异。

SKDSA模型损失函数2（式（3.9））和BYOT模型损失函数2（式（2.3））的区别是：BYOT模型直接计算最深层分离器归一化指数输出概率分布和每个浅层分类器归一化指数输出概率分布的相对熵，而SKDSA模型在计算出最深层分离器归一化指数输出概率分布和每个浅层分类器归一化指数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。实验证实这样可以取得更高的分类准确率。

SKDSA模型的损失函数3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损失函数。SKDSA模型通过损失函数3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。

SKDSA模型的损失函数3由式（3.10）表示，其中表示第个浅层分类器的特征图，表示最深层分类器的特征图，代表归一化损失函数：

|  |  |  |
| --- | --- | --- |
|  |  | （3.10） |

损失函数具备半正定性。当它的值为0时，意味着特征图和特征图相同；当它的值越大，意味着特征图和的差异越大。由于所有的范数都满足三角不等式，即所有的范数都是凸函数，所以损失函数作为范数的一种也是一个凸函数。这意味着它的局部最优等于全局最优。

SKDSA模型损失函数3（式（3.10））和BYOT模型损失函数3（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损失函数，而SKDSA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损失函数。实验证实这样可以取得更高的分类准确率。

综合式（3.9）和（3.10），可以得到第个浅层分类器的损失函数为：

|  |  |  |
| --- | --- | --- |
|  |  | （3.11） |

其中代表平衡式（3.9）和（3.10）的超参数。式（3.11）对应于自注意模块的输入。

将第个浅层分类器的损失函数（式（3.11））预期对应的注意力权重系数（式（3.7））相乘后累加，即可得到自注意力模块的损失函数之和：

|  |  |  |
| --- | --- | --- |
|  |  | （3.12） |

其中即为自注意力模块（图3.3中）的最终输出。

综合式（3.10）和（3.14），即可得出SKDSA模型的总损失函数：

|  |  |  |
| --- | --- | --- |
|  |  | （3.13） |

其中代表式（2.2）中提到的知识蒸馏损失系数。超参数和的取值会在4.3.4部分测试。

由于交叉熵、相对熵、归一化范数都是凸函数，作为它们线性组合的总损失函数也是凸函数。这意味着它的局部最优等于全局最优，容易通过梯度下降法找到模型的最优解。

SKDSA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SKDSA模型则将损失函数2和损失函数3结合后的结果乘以对应的注意力权重，再和损失函数1与超参数的乘积相加。实验证实这样可以取得更高的分类准确率（详见4.3.1部分）。

### 训练流程

SKDSA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，输出为损失函数最小的模型参数。

| **算法3.1** SKDSA训练流程伪代码 |
| --- |
| **输入**： 训练集，随机初始化的模型参数  **输出**： 最优的模型参数 |
| 1: **while** 模型参数尚未收敛 **do**：  2: 从训练集中挑出一批数据记为。  3: 将放入SKDSA模型进行前向传播，获得第个浅层分类器的特征图和最深层分类器的特征图。  4: 根据式（3.6），利用和计算和。  5: 根据式（3.7），利用和计算注意力权重系数。  6: 根据式（3.8）（3.9）（3.10）分别计算损失函数1、损失函数2、损失函数3。  7: 根据式（3.11）计算第个浅层分类器的损失函数。  8: 根据式（3.12）和注意力权重计算所有浅层分类器的损失函数之和。  9: 根据式（3.13）中总损失函数。  10: 为了最小化，对使用随机梯度下降法，更新模型参数。  11: **end while**  12: **return** |

由于训练数据集的庞大，难以全部装入显存之中，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集装入显存进行计算。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（3.6）（3.7）（3.8）（3.9）（3.10）（3.11）（3.12）（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。

## SKDSA模型和袋装法的等价性证明

本节将证明SKDSA模型和袋装法[39]的等价性。

袋装法是一种是机器学习中的一种集成算法，旨在提高统计分类和回归类型的机器学习算法的稳定性和准确性，减少其方差以避免复杂模型的过拟合。袋装法的过程如图3.5所示：首先，创建多个拔靴（bootstrap）样本，以便每个新的样本拔靴将充当另一个独立的数据集；然后，为每个样本拟合一个弱学习器，最后聚合它们，这样就可以“平均”它们的输出，从而获得一个方差小于单个弱学习器的集成模型。“平均”弱学习器的输出不会改变预期结果，但可以减少其方差。



图3.5 袋装法原理

袋装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。对于有投票机制的袋装法，假设有个神经网络记为，它们的投票权重依次记为，最终预测结果可以表示成：

|  |  |  |
| --- | --- | --- |
|  |  | （3.14） |

因为SKDSA模型使用注意力机制将各个浅层块的信息汇总到最深层块，SKDSA模型的最终预测表示可以写成：

|  |  |  |
| --- | --- | --- |
|  |  | （3.15） |

其中代表第个浅层块的注意力权重，代表第个浅层分类器经过加入温度系数的归一化指数层的输出。

令式（3.14）中的，，再令，

|  |  |  |
| --- | --- | --- |
|  |  | （3.16） |

把式（3.15）和（3.16）代入（3.14），可以得到：

|  |  |  |
| --- | --- | --- |
|  |  | （3.17） |

由此证明了SKDSA模型的注意力机制可以被视为袋装法，而注意力权重系数又恰恰对应于集成学习中的投票权重。这意味着SKDSA模型具有袋装法的优点：更强的泛化能力和抗过拟合能力。

## SKDSA模型和数据增强的结合

数据增强是减少过度拟合的技术之一，它的特点是通过处理训练数据集来缓解过拟合问题。数据增强通过数据扭曲或过采样人为增加训练数据集的大小，方式包几何变换、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SKDSA模型的性能，在3.2节给模型添加了自注意力模块之后，本节将SKDSA模型与数据增强模型相结合，包括cutout[40]、SLA[9]、mixup[41]三种数据增强模型。第四章的实验将分别记录SKDSA模型使用原数据集的性能，和结合三种数据增强模型之后的性能。

### cutout模型

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用“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。这种策略可以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。

 

图3.6 原图像（左）和cutout后的图像（右）

### SLA模型

以图3.7中标签为猫的图像为例，过去的数据增强模型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（Self-supervised Label Augmentation）模型并不直接强迫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）

### mixup模型

mixup模型是一种简单有效的数据增强模型。记和是输入向量，对应标签的独热向量分别是和，则它们混合生成的样本和对应的向量是：

|  |  |  |
| --- | --- | --- |
|  |  | （3.18） |

其中和是从训练集中随机挑选的样本，取值范围是[0, 1]。mixup的假设是特征向量的线性插值可以导出对应目标的线性插值。

## 本章小结

本章首先分析了BYOT模型和PD-BYOT模型的不足，提出SKDSA模型作为改进方案，随后在3.1节介绍了典型自注意力模型的计算流程。

3.2节比较了SKDSA模型和BYOT模型在网络结构、损失函数上的不同，说明了SKDSA模型相比于BYOT模型所做的改进。

3.3节从理论上证明了SKDSA模型和袋装法的等价性，这意味着SKDSA模型具备避免复杂模型过拟合的优点。

3.4节将SKDSA模型与三种数据增强模型（cutout模型、SLA模型、mixup模型）相结合，作为进一步提升SKDSA模型性能的备选方案。

# 实验验证与分析

本章的安排如下：

在实验设置部分（4.1节），首先介绍了设置实验所需的软硬件环境，各种图像数据集，说明了模型超参数的设置、网络架构，介绍了和SKDSA模型做性能对比的其他模型。

在实验结果和分析部分（4.2节），首先对比了在SKDSA模型和其他自知识蒸馏模型的性能，然后对比了SKDSA模型和其他自注意力模型的性能。

在消融实验与分析部分（4.3节），对比了添加和撤除自注意力模块对模型性能的影响，对比了使用和撤除自注意力模块中知识蒸馏模块对SKDSA模型性能的影响，对比了结合与不结合数据增强对SKDSA模型性能的影响，探讨了使用不同超参数时的SKDSA模型性能。

## 实验设置

### 实验环境

实验环境为在AutoDL公司外租的服务器。系统为Ubuntu 18.04版本，CPU为Intel(R) Xeon(R) Silver 4110 CPU @ 2.10GHz，内存15GB，GPU为Nvidia 2080ti，显存11GB。实验使用AutoDL自带的系统环境，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 |

### 实验数据集

1. 在CIFAR-100[42]数据集上的实验

为了测试SKDSA模型对通用图像的分类性能，在数据集CIFAR-100上进行实验。CIFAR-10和CIFAR-100是由8000万个微小图像数据集的标记子集。它们由 Alex Krizhevsky、Vinod Nair和Geoffrey Hinton收集而成。

CIFAR-10数据集由10个类别的60000个32x32彩色图像组成，每个类别包含 6000个图像。有50000个训练图像和10000个测试图像。

数据集分为五个训练批次和一个测试批次，每个批次有10000张图像。测试批次包含来自每个类别的1000个随机选择的图像。训练批次包含随机顺序的剩余图像，但一些训练批次可能包含来自一个类的图像多于另一个。在它们之间，训练批次包含来自每个类别的5000张图像。这些类是完全互斥的。比如，汽车类和卡车类之间没有重叠。汽车类包括轿车、SUV 之类的东西，卡车类只包括大卡车。两者都不包括皮卡车。

CIFAR-100数据集类似于CIFAR-10，不同之处是它有100个类，每个类包含 600 张图像。每个类有 500 个训练图像和 100 个测试图像。CIFAR-100 中的 100 个类分为 20 个超类。每个图像都带有一个“精细”标签(它所属的类)和一个“粗略”标签(它所属的超类)。

在实验中，CIFAR-100被分为50000个数据的训练集和10000个数据的验证集。

1. 在Tiny ImageNet[43]数据集上的实验

为了测试SKDSA模型对通用图像的分类性能，在数据集Tiny ImageNet上进行实验。ImageNet[48]项目是一个大型视觉数据集，用于视觉对象识别软件研究。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规模太大，研究者提出Tiny ImageNet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。

1. 在Caltech-UCSD Birds 200数据集[44]上的实验

为了测试SKDSA模型对细粒度图像（鸟类）的分类性能，在数据集Caltech-UCSD Birds 200 (CUB-200)上进行实验。CUB-200是一个鸟类的图像数据集，包含200 种鸟类（主要分布在北美）的6033张照片。

1. 在Stanford 40 Actions[45]数据集上的实验

为了测试SKDSA模型对细粒度图像（人类动作）的分类性能，在Stanford 40 Actions[45]数据集上进行实验。Stanford 40 Actions数据集包含人类40种动作的图像。数据集总共有 9532 张图像，每类动作有 180-300张图像。

1. 在Stanford Dogs数据集[46]上的实验

为了测试SKDSA模型对细粒度图像（狗类）的分类性能，在Stanford 40 Actions[45]数据集上进行实验。Stanford Dogs数据集包含来自世界各地的120种狗的图像。该数据集是使用ImageNet中的图像和标注构建的，用于细粒度图像分类任务。数据集中包括120种类别的20580张图像。

1. 在MIT Indoor Scene Recognition数据集[47]上的实验

为了测试SKDSA模型对细粒度图像（室内场景）的分类性能，在MIT Indoor Scene Recognition数据集上进行实验。室内场景识别是高层次视觉中一个具有挑战性的问题。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，虽然一些室内场景（例如走廊）可以通过全局空间属性有效地表征，但其他的（例如书店）可以通过它们包含的对象来有效地表征。更一般地说，为了解决室内场景识别问题，需要一个可以利用局部和全局判别信息的模型。

该数据集包含67个室内类别，共15620张图像。图像的数量因类别而异，但每个类别至少有100张图像。

### 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设为0.9，权重衰减（weight decay）设为，动量（momentum）设为0.9。轮次总数为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。设一批训练样本的数量（batch size）为128。实验只使用了一种常规的数据增强策略，即随机剪裁（cutout）。知识蒸馏损失系数设为1.5，超参数设为100。

### 网络架构

为了测试SKDSA模型的性能，分别使用当前流行的卷积神经网络架构残差网络（ResNet）[36]，宽残差网络（WRN）[37]，残差密集网络（DenseNet）[38]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，步长为1，填充为1。

### 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，样本总数为，预测概率最大的标签是正确标签的样本数量为，则分类准确率的计算公式为：

|  |  |  |
| --- | --- | --- |
|  |  | （4.1） |

实验的优化的目标是尽可能提高验证集中的分类准确率。

### 对比模型

首先参加对比的模型为不添加自知识蒸馏模块，也不添加自注意力模块，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，实验将SKDSA模型的性能和五种典型的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：做你自己的教师(BYOT)[6]，数据失真引导的自知识蒸馏(DDGSD)[14]，通过教育自己来提纯自己(FRSKD)[8]，通过输入转换来做自监督标签增强的自知识蒸馏(SLA-SD)[9]，通过自知识蒸馏预测类别(CS-KD)[7]。

BYOT [6]模型通过辅助的分类器来利用浅层的输出，再通过真实标签和神经网络自身的信号(比如预测的logit和特征图)来训练这些辅助的分类器。

DDGSD[14]模型对于同一个样本产生不同的变形，在训练神经网络让同一个样本的不同变形产生一致性的预测。

FRSKD[8]模型利用一个辅助的自教育神经网络为分类神经网络传递精炼后的知识。为了进行自知识蒸馏它同时使用了软标签和特征图蒸馏。

SLA-SD[9]模型通过输入转换和把输出聚合成教师标签实现自监督标签增强。

CS-KD[7]模型使用隶属于同一类别的其他样本的logits来做正则化，训练目标是让同一类别的样本产生相似的预测结果。

两种注意力模型分别是：密集隐含的注意力神经网络(DIANet)[34]，自注意力神经网络(SAN)[35]。

DIANet[34]模型给不同的神经网络层添加一个注意力组件，目的是更有效地利用神经网络层间的信息。

SAN[35]模型利用成对的注意力组件去抽取更有用的信息来引导模型训练。

实验中的FRSKD[8]模型和SLA-SD[9]模型使用了其作者提供的官方代码，根据相应的文章实现了其对应的模型。

## 实验结果与分析

### 在通用图像数据集上的实验结果

为了测试SKDSA模型在通用图像数据集上的性能，分别使用基于残差网络框架（ResNet）、宽残差网络框架（WRN）、残差密集网络框架（DenseNet）的SKDSA模型在数据集CIFAR-100和Tiny-ImageNet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、CS-KD模型）、结合了数据增强模型（cutout模型）的SKDSA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4.2、表4.3、表4.4所示，最优的结果用粗体标注。

值得一提的是，此处SKDSA模型结合的数据增强模型是cutout模型，对结合另外两种数据增强模型的讨论详见4.3.2。

从表4.2、表4.3、表4.4可以看出，在通用图像数据集上，SKDSA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SKDSA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SKDSA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SKDSA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4.2 基于ResNet框架的各模型在通用图像数据集上的分类准确率（%）

| 基于ResNet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 |
| SKDSA | 80.51 | 60.42 |
| SKDSA+数据增强 | **81.77** | **62.61** |

表4.3 基于WRN框架的各模型在通用图像数据集上的分类准确率（%）

| 基于WRN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 |
| SKDSA | 74.80 | 53.21 |
| SKDSA+数据增强 | **76.03** | **53.42** |

从表4.2和表4.3还可以看出，在通用图像数据集上，SKDSA模型和数据增强模型（cutout）的结合具有比单一的SKDSA模型更卓越的性能。具体来说，基于ResNet的SKDSA模型在CIFAR-100和Tiny ImageNet数据集上相比基于ResNet的交叉熵模型分别提升准确率6.97%和8.01%，基于WRN的SKDSA模型在CIFAR-100和Tiny ImageNet数据集上相比基于WRN的交叉熵模型分别提升准确率5.61%和2.17%。

表4.4 基于DenseNet框架的各模型在通用图像数据集上的分类准确率（%）

| 基于DenseNet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 |
| SKDSA | **81.55** | **62.79** |

### 在细粒度图像数据集上的实验结果

为了测试SKDSA模型在细粒度图像数据集上的性能，分别使用基于残差网络框架（ResNet）、残差密集网络框架（DenseNet）在一些细粒度图像数据集（Caltech-UCSD Bird (CUB200)，Stanford 40，Stanford Dogs (Dogs)，MIT Indoor Scene Recognition (MIT67)）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、CS-KD模型）、结合了数据增强模型（cutout模型）的SKDSA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4.5和表4.6所示，最优的结果用粗体标注。

表4.5 基于ResNet的各模型细粒度图像数据集上的分类准确率（%）

| 基于ResNet模型 | CUB200 | MIT67 | Dogs | Stanford40 |
| --- | --- | --- | --- | --- |
| 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 |
| SKDSA | 65.20 | 61.90 | 71.21 | 58.02 |
| SKDSA+数据增强 | **67.03** | **65.77** | **72.84** | **62.61** |

从表4.5和表4.6中可以看出，在细粒度图像数据集上，SKDSA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SKDSA模型在CUB200、MIT67、Dogs、Stanford40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SKDSA模型在CUB-200、MIT67、 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4.6 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（%）

| 基于DenseNet模型 | CUB200 | MIT67 | 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 |
| SKDSA | **71.10** | **64.20** | **74.50** |

从表4.5还可以看出，在细粒度图像数据集上，SKDSA模型和数据增强模型的结合（cutout）具有比单一的SKDSA模型更卓越的性能。具体来说，基于ResNet框架的SKDSA模型在CUB200、MIT67、Dogs、Stanford40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%。

### 不同自注意力模型的实验结果

为了比较SKDSA模型和两种注意力模型（DIANet, SAN）的性能，分别使用基于三种残差网络框架（ResNet18, ResNet34, ResNet50）的SKDSA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，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。实验结果分别如表4.7所示，最优的结果用粗体标注。

从表4.7可以看出，在CIFAR-100数据集上，SKDSA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SKDSA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SKDSA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SKDSA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%。

表4.7 基于ResNet框架的各模型在通用图像数据集上的分类准确率（%）

| 基于ResNet模型 | ResNet18 | ResNet34 | ResNet50 |
| --- | --- | --- | --- |
| SAN | 75.90% | 76.40% | 77.20% |
| DIANet | 76.62% | 77.10% | 78.60% |
| SKDSA | **80.51%** | **80.81%** | **81.13%** |

## 消融实验与分析

### 自注意力模块的重要性

为了证明自注意力模块对于SKDSA模型的重要性，分别使用基于残差网络框架（ResNet）、宽残差网络框架（WRN）的去除自注意力模块的SKDSA模型、SKDSA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。实验结果如表4.8和表4.9所示，最优的实验结果用粗体字标注。

表4.8 基于ResNet框架的SKDSA模型和去除自注意力模块后的分类准确率（%）对比

|  |  |  |
| --- | --- | --- |
| 基于ResNet模型 | CIFAR-100 | Tiny ImageNet |
| w/o SA | 77.80 | 57.20 |
| SKDSA | **80.51** | **60.02** |

表4.9 基于WRN框架的SKDSA模型和去除自注意力模块后的分类准确率（%）对比

|  |  |  |
| --- | --- | --- |
| 基于WRN模型 | CIFAR-100 | Tiny ImageNet |
| w/o SA | 71.71 | 51.92 |
| SKDSA | **74.80** | **53.21** |

从表4.8和表4.9可以看出，在通用图像数据集上，SKDSA模型具有比去除自注意力模块的SKDSA模型更优异的性能。具体来说，基于ResNet框架的SKDSA模型在CIFAR-100和Tiny ImageNet数据集上相比去除注意力模块的基于ResNet框架的SKDSA模型分别提升准确率2.71%和2.82%，基于WRN框架的SKDSA模型在CIFAR-100和Tiny ImageNet数据集上相比去除注意力模块的基于WRN框架的SKDSA模型分别提升准确率3.09%和1.29%。

以上结果说明，自注意力模块在SKDSA模型中起到了重要作用，对模型性能的提升不可或缺。

### 自注意力模块中的知识蒸馏模块的重要性

为了证明自注意力模块中的知识蒸馏模块对SKDSA模型性能的重要性，使用基于残差网络框架的SKDSA模型在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。实验结果如表4.10所示，其中第一列代表自注意力模块的蒸馏温度，第二列代表分类器蒸馏温度，第三列代表分类器蒸馏温度。需要说明的，温度设为1等价于不使用知识蒸馏模型。

表4.10 SKDSA模型在不同知识蒸馏温度下的分类准确率（%）

|  |  |  |
| --- | --- | --- |
| 自注意力模块的蒸馏温度 | 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 |

从表4.10中可以看出，自注意力模块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4.10第二列），SKDSA模型在自注意力模块蒸馏温度为2, 3和4时比不使用自注意力模块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4.10第三列），SKDSA模型在自注意力模块的蒸馏温度为2, 3和4时比不使用自注意力蒸馏分别提升准确率1.73%, 1.76%和2.38%。

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

### SKDSA模型与数据增强模型的相容性

4.2.1小结和4.2.2小结的实验证实了结合cutout数据增强方法可以提升进一步提升SKDSA模型的性能。本小节将SKDSA模型和另外两种数据增强模型（mixup模型和SLA模型）结合并测试了其性能。

为了证明SKDSA模型和数据增强模型的相容性，分别使用基于ResNet框架的SKDSA模型结合mixup模型、基于ResNet框架的SKDSA模型、基于ResNet框架的交叉熵模型结合mixup模型、基于WRN框架的SKDSA模型结合mixup模型、基于WRN框架的SKDSA模型、基于WRN框架的交叉熵模型结合mixup模型、基于ResNet框架的SKDSA模型结合SLA模型、基于ResNet框架的交叉熵模型结合SLA模型、基于WRN框架的SKDSA模型结合SLA模型、基于WRN框架的交叉熵模型结合SLA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。实验结果如表4.11、表4.12、表4.13、表4.14所示，最优的实验结果用粗体字标注。

表4.11 结合mixup模型对基于ResNet模型准确率（%）的影响

|  |  |  |
| --- | --- | --- |
| 基于ResNet模型 | CIFAR-100 | Tiny ImageNet |
| SKDSA | 80.51 | 60.42 |
| mixup+交叉熵 | 78.33 | 58.43 |
| mixup+SKDSA | **81.41** | **61.44** |

表4.12 结合mixup模型对基于WRN模型准确率（%）的影响

|  |  |  |
| --- | --- | --- |
| 基于WRN模型 | CIFAR-100 | Tiny ImageNet |
| SKDSA | 74.80 | 53.21 |
| mixup+交叉熵 | 72.21 | 52.82 |
| mixup+SKDSA | **76.73** | **53.68** |

表4.13 结合SLA模型对基于ResNet模型准确率（%）的影响

|  |  |  |
| --- | --- | --- |
| 基于ResNet模型 | CIFAR-100 | Tiny ImageNet |
| SKDSA | 80.51 | 60.42 |
| SLA+交叉熵 | 77.52 | 58.48 |
| SLA+SKDSA | **82.81** | **63.02** |

表4.14 结合SLA模型对基于WRN模型准确率（%）的影响

|  |  |  |
| --- | --- | --- |
| 基于WRN模型 | CIFAR-100 | Tiny ImageNet |
| SKDSA | 74.80 | 53.21 |
| SLA+交叉熵 | 73.00 | 50.77 |
| SLA+SKDSA | **76.83** | **53.89** |

从表4.11、表4.12、表4.13、表4.14可以看出，数据增强模型结合SKDSA模型具有比原始SKDSA模型更优异的性能，也具有比结合交叉熵模型更优异的性能。具体来说，基于ResNet框架的SKDSA模型结合mixup模型在CIFAR-100和Tiny ImageNet数据集上相比基于ResNet框架的SKDSA模型分别提升准确率0.90%和1.02%，相比基于ResNet框架的交叉熵模型结合mixup模型分别提升准确率3.08%和3.01%（表4.11）；基于WRN框架的SKDSA模型结合mixup模型在CIFAR-100和Tiny ImageNet数据集上相比基于WRN框架的SKDSA模型分别提升准确率1.93%和0.47%，相比基于WRN框架的交叉熵模型结合mixup模型分别提升准确率4.52%和0.86%（表4.12）；基于ResNet框架的SKDSA模型结合SLA模型在CIFAR-100和Tiny ImageNet数据集上相比基于ResNet框架的SKDSA模型分别提升准确率2.3%和2.6%，相比基于ResNet框架的交叉熵模型结合SLA模型分别提升准确率5.29%和4.54%（表4.13）；基于WRN框架的SKDSA模型结合SLA模型在CIFAR-100和Tiny ImageNet数据集上相比基于WRN框架的SKDSA模型分别提升准确率2.03%和0.68%，相比基于WRN框架的交叉熵模型结合SLA模型分别提升准确率3.83%和3.12%（表4.14）。

以上结果说明，SKDSA模型可以提升基于数据增强模型的性能，数据增强模型也可以提升SKDSA模型的性能。这证实了SKDSA模型和数据增强模型的相容性。

### 敏感性分析

SKDSA模型的总损失函数（式（3.13））包含两个重要的超参数和，在前面的实验中取值1.5，取值100，本小节测试两者其他取值对模型性能的影响。

为了测试了超参数在不同取值下对SKDSA模型性能的影响，使用基于残差网络框架的SKDSA模型在CIFAR-100数据集上进行实验。测试的取值范围是{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。实验结果如表4.15所示，最优的实验结果用粗体字标注，可以看出当时，SKDSA模型的分类准确率最高。

表4.15 SKDSA模型使用不同值的分类准确率（%）

|  | 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 |

表4.16 SKDSA模型使用不同值的分类准确率（%）

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

为了测试了超参数在不同取值下对SKDSA模型性能的影响，使用基于残差网络框架的SKDSA模型在CIFAR-100数据集上进行实验。测试的取值范围是{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4.16所示，可以看出当时分类准确率最高。

## 本章小结

本章使用SKDSA模型在多个通用图像数据集（CIFAR-100[42]，Tiny ImageNet[43]）和细粒度图像数据集(Caltech-UCSD Bird(CUB200)[44]，Stanford 40[45]，Stanford Dogs (Dogs)[46]，MIT Indoor Scene Recognition (MIT67)[47])上进行图像分类实验，并和当前几种主流的自知识蒸馏模型（DDGSD模型，BYOT模型，CS-KD模型，SLA-SD模型，FRSKD模型）和自注意力模型（DIANet模型，SAN模型）作为对比。

从实验结果（4.2节）可以看出，SKDSA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SKDSA模型的有效性，证实了使用自注意力模块区分各个浅层块对最深层块的不同贡献度的正确性。

消融实验与分析（4.3节）还证明，自注意模块可以有效提升自知识蒸馏模型性能，自注意力模块中的知识蒸馏模块对SKDSA模型起到了重要作用，SKDSA模型可以和多种数据增强模型相容，以及找到了较优的超参数，。

# 总结与展望

## 主要工作总结

主要工作可以总结如下：

第一，指出BYOT模型存在的问题是忽略了各个浅层块信息对最深层块的不同影响，提出PD-BYOT模型作为初步改进方案：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用等比数列的形式加以区分。实验证实添加衰减系数仅能略微改进BYOT模型的性能。

第二，在BYOT模型和PD-BYOT模型的基础上，为了准确量化各浅层块特征图对最深层块特征图的不同贡献度，提出基于自注意力机制的自知识蒸馏（SKDSA）模型。详细说明了SKDSA模型的网络结构、损失函数，列举了新模型相比原BYOT模型所做的改进。随后用数学理论证明SKDSA模型可以视为集成学习中的袋装法，这意味着SKDSA具有优异的抗过拟合性能。最后将模型与三种数据增强模型（cutout模型、SLA模型、mixup模型）相结合，作为进一步提升模型性能的备选方案。

第三，在多个图像数据集（CIFAR-100[42]，Tiny ImageNet[43]，Caltech-UCSD Bird[44]，Stanford 40[45]，Stanford Dogs[46]，MIT Indoor Scene Recognition[47]）上测试了SKDSA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（DDGSD模型，BYOT模型，CS-KD模型，SLA-SD模型，FRSKD模型）和自注意力模型（DIANet模型，SAN模型）更优异的实验结果。消融实验证实了SKDSA模型的自注意力模块和自注意力模块中的标签软化模块对提升性能很有必要，以及SKDSA模型结合数据增强模型可以进一步提升模型的性能，找到了使SKDSA模型性能较优的超参数取值。

## 主要创新点

主要创新点可以归纳如下：

第一，提出PD-BYOT模型，通过给BYOT模型中不同深度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用等比数列的形式加以区分。

第二，尽管研究者已经对自知识蒸馏模型和自注意力模型都有了充分的研究，但SKDSA模型是第一个将两者以一种简单有效的方式结合起来。SKDSA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添加自注意力连接，准确量化了神经网络不同深度的浅层特征对最深层特征表示的不同影响，从而更有效地利用知识蒸馏中的暗知识。

## 未来工作展望

SKDSA模型虽然在实验上取得了不错的效果，但仍然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。

第一，尝试把学生层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和作为教师模型的最深层常常大小相近。之后或许可以考虑把学生模型映射到新的较小的空间中，或许可以取得更优异的性能。

第二，尝试其他深度网络架构。

因为时间所限，SKDSA模型的卷积网络架构只尝试了残差网络、宽残差网络、密集残差网络三种架构。未来可以尝试更多的网络模型，比如ResNeXt[49], VGG[50]等。

第三，在更多数据集上进行实验。

因为实验室计算资源和资金有限，SKDSA模型只在几个较小的数据集上进行了实验。未来可以在更大的数据集上进行实验，比如ImageNet。

第四，在更多类型的任务上尝试SKDSA模型。

虽然SKDSA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它的可行性，未来可以思考将SKDSA模型应用到语言模型上的策略。后续的工作也可以考虑将SKDSA模型应用于半监督学习和弱监督学习任务。

# 致 谢

# 参考文献

[1] 徐卫军. 深度学习在计算机视觉领域中的应用. 电子技术, 2021

[2] 车万翔，郭江，张伟男，刘铭. 基于深度学习的自然语言处理. 中文信息学报, 2021

[3] 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

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

[5] Pan, Sinno Jialin, Qiang Yang. A survey on transfer learning. IEEE Transactions on knowledge and data engineering, 2009, 22(10):1345-1359

[6] Zhang, Linfeng, 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

[7] Yun, Sukmin, 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

[8] Ji, Mingi, 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

[9] Lee, Hankook, 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

[10] Ba, Jimmy, 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

[11] Kim, Seung Wook, 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

[12] Hou, Saihui, Xinyu Pan, Chen Change Loy, Zilei Wang, Dahua Lin. Learning a unified classifier incrementally via rebalancing. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (ICCV 2019), Seoul, Korea, October 27-November 2, 2019: 831–839

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

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

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

[19] Shen, Tao, Tianyi Zhou, Guodong Long, Jing Jiang, Shirui Pan, Chengqi Zhang. Disan: Directional selfattention 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] Xing, Chen, 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] Wang, Yequan, 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] She, Heng, 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] Chopra, Sumit, Michael Auli, Alexander M. 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] Fu, Jianlong, Heliang Zheng, Tao Mei. Look closer to see better: Recurrent attention convolutional neural network for fine-grained image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR 2017), Honolulu, Hawaii, USA, July 21-26, 2017: 4438-4446

[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] 乔思波，庞善臣，王敏，翟雪，于世行，丁桐. 基于残差混合注意力机制的脑部 CT 图像分类卷积神经网络模型. 电子学报, 2021, 49(5), 984

[28] Zhang, Han, 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

[29] Kong, Tao, 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

[30] Li, Wei, 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

[31] Liu, Hao, 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

[32] Fu, Jun, 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

[33] Kuen, Jason, 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

[34] Huang, Zhongzhan, 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

[35] Zhao, Hengshuang, 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

[36] He, Kaiming, 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

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

[38] Huang, Gao, Zhuang Liu, Laurens Van Der Maaten, Kilian Q. Weinberger. Weinberger and L. Maaten. 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

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

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

[41] Zhang, Hongyi, Moustapha Cisse, Yann N. Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. arXiv preprint arXiv:1710.09412, 2017

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

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

[44] Wah, Catherine, Steve Branson, Peter Welinder, Pietro Perona, Serge Belongie. The Caltech-UCSD Birds-200-2011 Dataset. Technical report, 2011

[45] Yao, Bangpeng, Xiaoye Jiang, Aditya Khosla, Andy Lai Lin, Leonidas Guibas, Li Fei-Fei. 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

[46] Khosla, Aditya, 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)

[47] Quattoni, Ariadna, 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

[48] Deng, Jia, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, Li Fei-Fei. 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

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

# 附录1 攻读硕士学位期间取得的研究成果

**投稿论文**

[1] XXX, **XXX**, XXX\*. Self-Knowledge Distillation with Self-Attention. International Joint Conference on Artificial Intelligence, 2022, under review. (CCF A；华中科技大学为第一署名单位)

# 附录2 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：项目成员

# 附录3 中英文缩写对照表

BYOT Be your own teacher

CS-KD Class-wise self-knowledge distillation

DDGSD Data-distortion guided self-distillation for deep neural networks

DenseNet Densely connected convolutional networks

DIANet Dense-and-implicit attention network

FRSKD Feature refinement via self-knowledge distillation

GAN Generative adversarial networks

LSTM Long term short memory

ResNet Deep residual networks

RNN Recurrent neural network

SAN Self-attention network

SKDSA Self-knowledge distillation based on self-attention

SLA Self-supervised label augmentation

SLA-SD Self-supervised label augmentation and self-distillation

WRN Wide residual networks