**分类号**  **学号**

**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（Knowledge distillation, KD）是将从一个模型学到的知识迁移到另一个模型的过程，其中提供知识的结构被称为教师模型，学习知识的结构被称为学生模型。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的信息视为教师模型，把神经网络的一些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（比如Be your own teacher: Improve the performance of convolutional neural networks via self distillation，BYOT）将作为学生模型的各个浅层一视同仁，忽略了各个层提取的特征对结果的不同作用。

为了进一步提升自蒸馏模型在图像分类任务中的性能，本文提出一种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模型是对BYOT模型的进一步改进：它通过给BYOT模型中的每一个中间层和最深层之间添加自注意力连接，为作为学生模型的不同中间层计算出不同的注意力权重，从而更有效地从网络中提取出暗知识。SKDSA模型中的自注意力连接还可以从理论上证明了等价于集成学习中的袋装法，这意味着SKDSA模型具有更强的稳定性。

为了验证SKDSA模型的有效性，在多个图像数据集（包括CIFAR-100, Tiny ImageNet, Caltech-UCSD Birds 200, Stanford 40, Stanford Dogs, MIT Indoor Scene Recognition）上选取了目前流行的多个卷积模型（残差网络, 宽残差网络, 残差密集网络）和自知识蒸馏模型进行实验测试。实验结果表明，SKDSA模型相比于现有其他的自知识蒸馏模型均能达到更高的分类准确性，同时消耗更短的训练时间。此外，消融实验进一步证明了自注意力组件对SKDSA模型的必要性，以及SKDSA模型和数据增强模型的相容性。

**关键词：** 注意力机制；自注意力机制；知识蒸馏; 自知识蒸馏

# Abstract

Knowledge distillation (KD) is the process of transfering knowledge learned from one model to another, in which the model providing knowledge is called “teacher” and the model receiving knowledge is called “student”. Unlike traditional knowledge distillation, the self-knowledge distillation technique utilizes the student network's layers to distill itself without an extra teacher model. Current self-knowledge distillation techniques (i.e. Be your own teacher: Improve the performance of convolutional neural networks via self distillation, namly BYOT) treat all block layers (act as student) equally, neglecting their different impact to the final layer(act as teacher).

In order to improve the performance of self-distillation, we propose a novel Self-Knowledge Distillation with Self-Attention (SKDSA) framework. SKDSA is a modification of BYOT. It adds attention links between each intermediate layer and final layer of BYOT, and utilizes the self-attention mechanism to adaptively obtain the coefficients of distillation for each student layer of the network, so as to mine the network’s dark knowledge more efficiently. We also provide theoretical proof that the proposed approach is essentially an ensemble modeling strategy, which means SKDSA has the advantage of robustness.

We conduct experiments on various datasets (CIFAR-100, Tiny ImageNet, Caltech-UCSD Birds 200, Stanford 40, Stanford Dogs, MIT Indoor Scene Recognition) using multilple frameworks (ResNet, Wide ResNet, DenseNet). The experimental results showsthat SKDSA outperforms all current self-distillation techniques and achieve the state-of-the-art accuracy. Moreover, the ablation study proves the effectiveness of self-attention links.

**Key words:** Attention, Self-attention, Knowledge Distillation, Self-knowledge Distillation

# 目 录

[标题：宋体，英文Times New Roman，一号，加粗，不超30字 I](#_Toc99616222)

[A Dissertation Submitted in Partial Fulfillment of the Requirements II](#_Toc99616223)

[for the Master Degree in Engineering/Science （工学/理学硕士）/the Professional Master Degree (专业学位) II](#_Toc99616224)

[摘 要 I](#_Toc99616225)

[Abstract III](#_Toc99616226)

[目 录 IV](#_Toc99616227)

[1 绪论（黑体，居中，三号） 1](#_Toc99616228)

[1.1 研究背景与意义（黑体，四号） 2](#_Toc99616229)

[1.2 XXX国内外研究现状（请拟定具体的题目） 2](#_Toc99616230)

[1.3 存在的问题 2](#_Toc99616231)

[1.4 本文主要内容 2](#_Toc99616232)

[2 系统与控制理论类论文\* 4](#_Toc99616233)

[2.1 引言（引言标题可选） 4](#_Toc99616234)

[2.2 预备知识（可选，标题可自选） 4](#_Toc99616235)

[2.3 问题的描述（请拟定具体的题目） 5](#_Toc99616236)

[2.4 控制器设计与闭环系统分析（请根据所设计的控制器特点自行拟定具体的题目） 6](#_Toc99616237)

[2.5 数值仿真（请拟定具体的题目） 6](#_Toc99616238)

[2.6 本章小结 6](#_Toc99616239)

[3 理论/算法类研究类论文 7](#_Toc99616240)

[3.1 引言（引言标题可选） 7](#_Toc99616241)

[3.2 \*\*理论/算法 7](#_Toc99616242)

[3.3 \*\*仿真或算法实现 7](#_Toc99616243)

[3.4 理论/算法准确性的评估 7](#_Toc99616244)

[3.5 分析与讨论 7](#_Toc99616245)

[3.6 本章小结 7](#_Toc99616246)

[4 学位论文写作细则 8](#_Toc99616247)

[4.1 关于图 8](#_Toc99616248)

[4.2 关于表格 9](#_Toc99616249)

[4.3 名词、术语 10](#_Toc99616250)

[4.4 符号、单位的使用 11](#_Toc99616251)

[4.5 数字的使用 11](#_Toc99616252)

[4.6 其它应该注意的问题 12](#_Toc99616253)

[4.7 本章小结 13](#_Toc99616254)

[5 总结与展望 14](#_Toc99616255)

[5.1 本文主要内容及结论 14](#_Toc99616256)

[5.2 本文的主要创新点 14](#_Toc99616257)

[5.3 未来工作展望 14](#_Toc99616258)

[致 谢 15](#_Toc99616259)

[参考文献 16](#_Toc99616260)

[附录1 攻读硕士学位期间取得的研究成果 19](#_Toc99616261)

[附录2 攻读学位期间参加的科研项目 21](#_Toc99616262)

[附录3 中英文缩写对照表 22](#_Toc99616263)

[附录4 其它附录 23](#_Toc99616264)

# 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

|  |  |
| --- | --- |
| 大型神经网络模型 | 参数量（单位：百万） |
| 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（记为教师模型），然后把其中的知识压缩到一个小型的神经网络（记为学生模型）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技术，即不利用外部的教师模型，利用学生模型自身的知识来做知识蒸馏。

自知识蒸馏技术[6][7][8][9]使用神经网络中间层的信息来做知识蒸馏，以便免去使用额外教师模型的开销。然而，当前已有的自知识蒸馏模型也存在可以改进的空间。其中，Zhang等人[6]提出的BYOT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，BYOT模型将各个作为学生的神经网络浅层的信息一视同仁，可能会造成一些信息的丢失。

为了解决这个问题，提出一种新的基于自注意力机制的自知识蒸馏框架（Self-Knowledge Distillation with Self-Attention, SKDSA）。这个框架可以计算出各个学生层的重要性差异，并根据重要性的差异赋予它们不同的注意力分数。为了测试SKDSA模型的泛化能力，在第四章使用当前流行的两种卷积神经网络在一些视觉数据集上做了实验。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。

## 国内外研究现状

### 知识蒸馏模型

传统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，其示意图如图1.1所示。由于传统知识蒸馏对于模型压缩简单而有效，目前已经广泛用于不同的任务和应用。Hinton等人[4]提出通过学习软化标签的类分布，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。Ba等人[10]发现浅层前馈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。



图1.1 传统的知识蒸馏

### 自知识蒸馏模型

虽然传统的知识蒸馏模型取得了优异的效果，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，人们提出一些自知识蒸馏的模型。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。

一种解决方案是做基于特征的自知识蒸馏。由于模型神经网络的深层比浅层包含更高阶、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额外的连接来实现利用中间层信息的目的，其思路如图1.2所示。BYOT模型[6]就是一种这样的模型，它通过加入一系列辅助的神经网络层来进行自知识蒸馏，以便把信息从神经网络深层传递到神经网络浅层。而Hou等人[12]的工作中，相比于直接蒸馏特征，他们使用的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，他们用最深层的分类器的信息(特征图和logits)蒸馏到浅层分类器。



图1.2 深层教导浅层的自知识蒸馏

基于特征的自知识蒸馏把关注点放在如何更有效地在自知识蒸馏过程中“教导监督”学生模型，用于训练学生的数据却没有被充分挖掘和利用。因此，另外一些人想到通过训练样本的数据增强来做知识蒸馏。这种策略具有如下优点：首先，优化一个没有其他模型或分支的单个学生神经网络更加迅速和有效；第二，通过数据到数据的自蒸馏，学生模型可以学到更多固有的泛化表示；第三，在计算和存储成本相对较低的前提下，学生模型的性能能够显著提高。Xu等人[14]设计了一种通过数据失真实现自知识蒸馏的策略，即使用一种一致规则化损失函数来迫使输入和它的变种能够产生相似的预测结果。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

注意力机制最早由Bahdanau等人[16]提出，现已变成神经网络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。对注意力机制的直观理解可以用人类的生物学机制来解释。比如说，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。比如说，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。类似地，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)。注意力机制能够让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，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(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。

### 数据增强模型

许多研究希望通过将深度卷积网络应用于计算机视觉任务来改进当前的基准。提高这些模型的泛化能力是最困难的挑战之一。泛化性是指模型在以前见过的数据（训练数据）与以前从未见过的数据（测试数据）上评估时的性能差异。泛化性差的模型过度拟合了训练数据。发现过拟合的一种方法是在训练期间绘制每个时期的训练和验证准确度。

数据增强是减少过度拟合的技术之一，它的特点是从问题的根源，即训练数据集开始处理过度拟合。这是基于可以通过增强从原始数据集中提取更多信息的假设。这种数据增强增强通过数据扭曲或过采样人为地增加了训练数据集的大小，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。过采样增强创建合成实例并将它们添加到训练集中。例如，GAN样本可以与随机裁剪堆叠在一起，以进一步扩大数据集。图像数据增强的大致分类如图2.12所示。



图1.4 图像数据增强分类

Lee等人[9]提出了通过输入转换进行自监督标签增强模型（Self-supervised Label Augmentation via Input Transformations, SLA）。SLA模型学习一个关于原始标签和自监督标签的联合分布的任务，而不是将其视为的两个独立任务。例如，当在 CIFAR-10（10 个标签）数据集上使用旋转自监督（4 个标签）进行训练时，SLA模型学习了所有可能组合（即 40 个标签）的联合概率分布。自监督标签增强 (SLA)在不假设原始标签和自监督标签之间的关系的情况下，不会具有任何转换不变性。此外，由于作者为每个转换分配不同的自监督标签，因此可以在测试时通过聚合所有转换来进行预测。这可以使用单个模型提供（隐式）集成效果。

## 论文主要内容

论文共分为5章内容，章节内容之间的关系图如图1.4所示。

图1.5 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 基础理论与相关知识：本章主要介绍了本篇论文中涉及到的基础理论和相关知识（包括典型卷积网络模型[36][37][38]、典型知识蒸馏模型、典型自注意力模型），为第三章正式提出新模型和第四章对新模型进行实验评估做铺垫。

第三章 基于自注意力机制的自知识蒸馏：本章在BYOT模型的基础上提出新模型SKDSA。详细介绍了SKDSA模型的原理框图，损失函数，算法流程。并且证明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挖剪法[40]结合起来。

第四章 实验结果和分析：本章对SKDSA模型在多个数据集上（包括CIFAR-100[41], Tiny ImageNet[42], Caltech-UCSD Birds 200[43], Stanford 40[44], Stanford Dogs[45], MIT Indoor Scene Recognition[46]）进行图像分类实验，证明模型的有效性和稳定性。接着做了消融研究，证明自注意力机制对于模型的不可或缺。

第五章 总结与展望：本章总结全文的主要工作，说明主要的创新点，并展望未来可能的改进方向。

# 基础理论与相关知识

由于SKDSA模型是知识蒸馏模型和注意力模型的结合，这一章中我们将分别介绍典型的知识蒸馏模型（2.2节）和注意力模型（2.3节），以便在第三章引出SKDSA模型。本章节提到的自蒸馏模型和注意力模型也会作为第四章实验评估的对比模型。

## 典型的卷积网络模型

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。

### 残差神经网络

越深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（如图2.1所示）。因此，如何如何训练越来越深的卷积网络成为了一个研究热点。



图2.1 神经网络层数的增加导致性能的退化[36]

He等人[36]通过引入深度残差网络来解决性能退化问题。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2.2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2.2 残差块的结构[36]

### 宽残差神经网络

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[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2.3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2.3 宽残差网络的结构[37]

### 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（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2.4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。

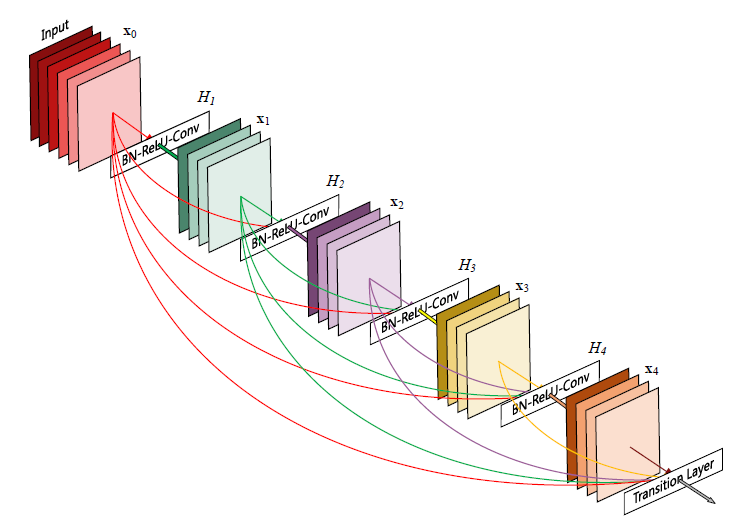


图2.4 残差密集网络示意图[38]

## 典型的自知识蒸馏的模型

在介绍典型自知识蒸馏模型之前，首先回顾传统的知识蒸馏模型。



图2.5 知识蒸馏示意图

传统知识蒸馏模型的示意图如图2.1所示，它是一项在大的教师模型的帮助下训练小的学生模型的技术。和其他神经网络压缩方式不同的是，知识蒸馏可以无视教师模型和学生模型结构上的不同。在Hinton等人[4]提出的模型中，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。

然而在许多情况下，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。假定代表由模型产生的logit向量，一张图像的属于第个类别概率为：

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

其中是蒸馏温度参数，代表教师模型中的蒸馏程度。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软，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信息。是暗信息在影响整体蒸馏的信息流。当计算蒸馏损失函数的时候，在教师模型中使用的相同的温度也用来计算学生模型中的logits。对于有正确标签的图像，Hinton[4]指出训练学生模型时把正确标签和教师的软化标签放在一起训练更有效。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：

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

其中函数代表的是由软标签和计算出的交叉熵损失函数，代表知识蒸馏的系数。

接下来我们介绍几种典型的自知识蒸馏模型的原理，它们也是我们在第四章对SKDSA模型进行实验评估时的对比模型。

### 数据失真引导的自蒸馏模型

Xu等人提出了数据失真引导的自蒸馏模型[14]（Data-Distortion Guided Self-Distillation for Deep Neural Networks, DDSGD）。它无需其他模型的帮助，即可根据数据表示的一致分布有效地优化单个网络。通常，常见的数据失真技术（图像的随机镜像/裁剪、旋转、多尺度等等）仅从现有训练样本的邻域扩大额外的虚拟样本，这使得网络可以看到更多失真的训练样本以提高模型泛化能力。一般认为，泛化能力要求模型能够在相同训练数据的不同版本中保持相似或不变的类概率预测（后验分布）和特征表示（特征分布）。因此，DDSGD通过数据失真引导自蒸馏机制，以最小化知识迁移的后验概率和特征分布的差异。具体说来，DDGSD利用相对熵来约束来衡量不同扭曲版本的后验概率分布之间的匹配，其方式类似于不同模型之间的知识蒸馏[4]，并采用经验最大平均差异作为衡量这些失真版本之间特征分布一致性的非参数指标。

### 通过自知识蒸馏的特征细化模型

过去的知识蒸馏模型依赖于辅助网络，其复杂度与分类器网络相同或更低；因此，无论是通过作为卷积层输出的特征还是分类器网络的软标签，都很难生成精炼的知识。此外，基于数据增强的模型容易丢失样本之间的局部信息，例如不同扭曲的样本或旋转的样本。因此，利用特征蒸馏面临一定的困难。

为此，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）。它通过引入了一个辅助的自学网络，可以将精炼的知识转移到分类器网络。图2.3显示了FRSKD与现有知识蒸馏模型之间的差异。FRSKD模型可以利用软标签和特征图蒸馏来进行自知识蒸馏。



图2.6 FRSKD和其他知识蒸馏模型的区别[8]

### 自监督标签增强的知识蒸馏模型

Lee等人[9]发现，多任务学习强制原始任务的主要分类器在自监督任务的转换方面保持不变。例如，当使用旋转作为自监督时（将每个图像旋转 0、90、180、270 度同时保留其原始标签），可以强制初级分类器学习对图像旋转不变的表示方法。强制这种不变性可能会导致任务的复杂性增加，因为转换可能会在很大程度上改变样本的特征和/或识别对象的有意义信息。因此，这可能会损害整体表示学习，并降低主要全监督模型的分类准确性（参见第 3.2 节中的表 1）。

为了解决这一问题，Lee等人[9]提出了一个简单而有效的想法，即学习一个关于原始标签和自监督标签的联合分布的统一任务，而不是像先前文献一样将其视为的两个独立任务。例如，当在 CIFAR-10（10 个标签）上使用旋转自监督（4 个标签）进行训练时，作者学习了所有可能组合（即 40 个标签）的联合概率分布。

作者将这种标签增强模型称为自监督标签增强 (SLA)，在不假设原始标签和自监督标签之间的关系的情况下，不会具有任何转换不变性。此外，由于作者为每个转换分配不同的自监督标签，因此可以在测试时通过聚合所有转换来进行预测，如图2.4所示。这可以使用单个模型提供（隐式）集成效果。



图2.7 聚合和自蒸馏[9]

### 通过自知识蒸馏对分类预测正则化模型

Yun等人[7]提出了一种新的正则化技术，即按类别自知识蒸馏 (Class-wise Predictions via Self-knowledge Distillation, CS-KD)。如图2.5所示，CS-KD模型匹配或提取相同标签的不同样本之间的深度神经网络的预测分布。可以预期，如果样本属于同一类，作者所提出的正则化策略会迫使深度神经网络产生类似的错误预测，而传统的交叉熵损失没有考虑预测分布的这种一致性。此外，它可以同时实现两个理想的目标：防止过度自信的预测和减少类内变化。



图2.8 CS-KD示意图

## 典型的注意力网络模型

SKDSA模型是添加了自注意机制的自知识蒸馏模型，所以本小结介绍注意力机制。

### 注意力机制的提出

在心理学中，注意力是选择性地专注于一件或几件事情而忽略其他事情的认知过程。比如，假设一个人正在查看一张小学生的合影照。照片里会有一群孩子列成几排，而教师会坐在中间的某个地方。如果想要知道照片中有几个人，他仅仅需要数一数照片中有几个脑袋就可以了，而不需要在意照片中的其他事物。如果想要知道照片中的教师，他就会去寻找照片中符合成年人特征的信息，而忽略了其他信息。这就是大脑的“注意力”机制。

深度学习中最早使用的注意力机制是对自然语言处理中基于编码器-解码器的神经机器翻译系统的改进。后来，注意力机制及其变种逐步被用于深度学习的其他领域，包括计算机视觉、语音处理等等。

在注意力机制被引入自然语言处理之前，机器翻译主要使用基于RNN的编码器-解码器机制。也就是说，编码器和解码器都是由一个个RNN单元组成的。它的主要过程是：编码器神经网络读取源语句并将其编码为固定长度的向量，然后解码器从编码向量输出翻译。整个编码器-解码器系统经过联合训练，以最大化给定源语句的正确翻译概率。

由于神经网络需要能够将源语句的所有必要信息压缩成一个固定长度的向量，这可能使神经网络难以处理长句子，尤其是那些比训练语料库中的句子更长的句子。这就是基于RNN机制面临的长距离依赖难题。产生长距离依赖难题的原因是梯度消失和爆炸问题。随着输入句子长度的增加，基于RNN的编码器-解码器网络的性能会迅速下降。基于RNN的机器翻译面临的另一个问题是，在翻译句子时，没有办法凸显不同单词重要性的差别。

那么有什么策略可以在创建上下文向量的同时保持输入句子中的所有相关信息完整无缺？Bahdanau 等人[16]提出了一个简单优雅的模型，此模型不仅可以在上下文向量中考虑所有输入词，还应该对每个输入词给予相对重要性。这个模型就是注意力机制。该模型在翻译中每生成一个单词时，它都会搜索源句中最相关信息集中的一组位置；然后，该模型根据与这些源位置和所有先前生成的目标词相关联的上下文向量来预测目标词。

### 自注意力网络模型

注意力机制在网络结构中的另一个应用是Zhao等人[35]提出的自注意力网络（self-attention network, SAN），这也是第四章对比试验的基准之一。

Zhao等人[35]在这篇文章中探索了自我注意算子的变体，并评估了它们作为图像识别模型基本构建块的有效性。他们探索了两种类型的自注意力。第一类是成对自注意力，它概括了自然语言处理中使用的标准点积注意力。成对注意力之所以引人注目，是因为与卷积不同，它本质上是一个集合算子，而不是一个序列算子。不同于卷积，它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。这使得自注意力算子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。Zhao等人提出了一些成对注意力的变体，它们比点积注意力具有更大的表达能力，同时保留了这些不变性。值得一提的是，他们的权重计算不会折叠通道维度，并允许特征聚合适应每个通道。第二类算子是互补注意力。互补注意力算子像卷积一样，能够唯一地识别其足迹内的特定位置。它没有成对注意力的排列和基数不变性，但严格来说比卷积更强大。

## 需要的损失函数

### 交叉熵

交叉熵是一种常用于机器学习中的损失函数。它是对给定随机变量或事件集的两个概率分布之间差异的度量。

在信息论中，低概率事件有更多的信息量，高概率事件有更少的信息量。假设事件的概率为，则事件包含的信息量如下所示：

|  |  |  |
| --- | --- | --- |
|  |  | （2-13） |

熵是表示某个随机事件所需的比特数。不均匀分布具有较低的熵，而均匀的概率分布具有较高的熵。

具有个离散状态中的一组的随机变量和它们的概率的熵值为：

|  |  |  |
| --- | --- | --- |
|  |  | （2-14） |

交叉熵建立在信息论中的熵的概念之上。它计算一个概率分布与另一个概率分布相比，表示平均事件所需的比特数。

两个概率分布和之间的交叉熵，可以记为，其中是交叉熵函数，代表是目标分布，代表目标分布的估计。由下式计算：

|  |  |  |
| --- | --- | --- |
|  |  | （2-15） |

此计算适用于离散概率分布。交叉熵值将是一个以比特为单位的正数，如果两个概率分布和相同，则结果将等于分布的熵。

交叉熵被广泛用作分类模型的损失函数。分类问题是涉及一个或多个输入变量和类别标签预测的问题。分类模型训练集中的每个样本都有一个已知的类标签，概率为1，所有其他的类标签的概率为0。模型可以估计样本属于每个类别标签的概率。 然后可以使用交叉熵来计算两个概率分布之间的差异。

### 相对熵

在很多情况下需要比较两个概率分布。具体来说，我们可能有一个随机变量和该变量的两个不同概率分布，例如真实分布和对真实分布的估计。在这种情况下，量化分布之间的差异可能很有必要——这通常被称为计算两个概率分布之间的距离的问题。

容易想到的一种方法是计算两个概率分布之间的散度。散度是不对称的，计算分布和的散度将给出与和散度不同的值。

散度是信息论中许多不同计算的重要基础，普遍用于机器学习中的计算。 例如，互信息（信息增益）和交叉熵。散度也直接用作复杂建模问题的工具，例如在优化生成对抗网络 (GAN) 模型时逼近目标的概率分布。

相对熵就是信息论中常用的一种散度。它量化了一个概率分布与另一个概率分布的差异程度。两个分布和之间的相对熵通常表示为。它的计算方式为：

|  |  |  |
| --- | --- | --- |
|  |  | （2-16） |

相对熵的直观理解是，当中某个事件的概率很大，而中相同事件的概率很小时，则存在很大的散度；当来自的概率很小而来自的概率很大时，也存在很大的散度，但没有第一种情况那么大。它可用于测量离散和连续概率分布之间的差异，在后一种情况下，计算事件的积分而不是离散事件的概率之和。

## 本章小结

本章介绍了SKDSA模型所涉及的基础理论和相关知识（包括典型的自知识蒸馏模型和自注意力模型），以方便在后续章节进一步阐述。另外，由于DIANet和SAN揭示了注意力可以用在网络结构中，这也为我们将自蒸馏模型和自注意力模型结合在一起提供了灵感。

# 基于自注意力机制的自知识蒸馏

本章将首先自知识蒸馏中的BYOT模型（3.1节），然后阐述在BYOT模型基础上改进得到的基于自注意力机制的自知识蒸馏模型（3.2节），最后证明SKDSA模型和袋装法的等价性（3.3节）。

## BYOT模型

由于SKDSA模型是对BYOT模型的改进，本节对自知识蒸馏的BYOT模型进行介绍。BYOT模型也是第四章实验评估中的对比基准之一。

### BYOT模型的原理框图

基于特征的自知识蒸馏的是把模型的神经网络深层的信息蒸馏到模型的神经网络浅层，Zhang等人提出的BYOT模型[6]就是对上述思路的一种实现。他们根据以下思维方式构建自知识蒸馏框架：（1）目标卷积神经网络根据其深度分为几个浅层。例如，将残差网络（具有50层）根据残差块分为4个部分。（2）在每个浅层块之下设置一个瓶颈层，再结合一个仅用于训练并且可以在推理中移除的全连接层，，这组成多个分类器（包括浅层分类器和最深层分类器）。添加瓶颈层的主要考虑是减轻每个浅层块之间的影响，并添加来自学生层信息的引导。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，而将最深层的归一化指数输出视为教师模型。BYOT模型的框架图如图3.1所示。



图3.1 BYOT模型的框架图[6]

### BYOT模型的损失函数

记个样本组成的集合为，其中包含个类。记相应的类组成的集合为，其中。模型中的各个浅层分类器被记为，其中代表卷积网络层中浅层块的个数。在每一个分类器后都设置一个归一化指数层，其输出为：

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

其中代表通过全连接层后的输出。代表第个浅层分类器后属于第个类别的概率。是知识蒸馏温度，初始被设为1。值越大，概率分布越“分散”。

BYOT模型在训练过程中引入了三种损失函数：真实标签与每个分类器（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L2损失函数。

BYOT模型的损失函数1：真实标签与每个分类器归一化指数输出的交叉熵。

BYOT模型的损失函数1通过来自训练数据集的真实标签和每个分类器（包括最深层分类器和所有浅层分类器）的归一化指数输出的交叉熵计算。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（3.2）表示，其中代表某样本经过第个浅层分类器的归一化指数层的输出，代表该样本在数据集中对应的真实标签，代表交叉熵损失函数：

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

BYOT模型的损失函数2：最深层分类器归一化指数输出和每个浅层分类器归一化指数输出的相对熵。

BYOT模型的损失函数2通过最深层分类器的归一化指数输出和每个浅层分类器的归一化指数的相对熵计算。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（3.3）表示，其中代表第个浅层分类器经过归一化指数层的输出，代表最深层分类器经过归一化指数层的输出，代表相对熵损失函数：

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

BYOT模型的损失函数3：最深层分类器特征图和每个浅层分类器特征图的L2损失函数。

BYOT模型的损失函数3通过最深层分类器的特征图和每个浅层分类器的特征图之间的L2损失计算。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（3.4）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L2损失：

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

所有新添加的层（图3.2中浅层块下方的部分）仅在训练期间应用。它们不对推理过程产生影响。

综合上述三种损失函数（式（3.2）（3.3）（3.4）），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（3.2）（3.3）（3.4）的超参数，代表BYOT模型的总损失函数：

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

## 基于自注意力机制的自知识蒸馏

### SKDSA模型的原理框图

正如3.1节所提到的，一个自知识蒸馏模型的浅层的特征图对于提升模型的稳定性非常重要。然而，式子（3.5）中的函数只是把由浅层和最深层计算出的损失函数直接相加，而没有考虑各个浅层蒸馏信息的不同重要性，这可能会导致一定的信息损失。为了更有效地挖掘模型中的暗知识，同时考虑到注意力机制在网络结构中已有成功的应用[34][35]，很有必要将自注意力机制和自知识蒸馏机制相结合，提出一种基于自注意力机制的自知识蒸馏模型来考虑自知识蒸馏模型中各个浅层的不同重要性。



图3.2 SKDSA模型的框架图

SKDSA模型的框架图如图3.2所示，（1）神经网络根据深度被分成几个浅层块。（2）在每一个浅层块下方，添加上额外的瓶颈层和全连接层，这组成多个分类器（包括浅层分类器和最深层分类器）。（3）然后，加上注意力连接，用来衡量每个浅层和最深层特征的相关性，注意力模块的作用是计算浅层和最深层的相似度。（4）其中每个分类器都是单独训练，训练采用三种类型的监督方法。预测出的分数是作为衡量每个浅层对总损失函数贡献度的系数。

通过将SKDSA模型的框架图（图3.2）和BYOT模型的框架图（图3.1）对比，可以发现SKDSA模型比BYOT模型增加了自注意力连接。SKDSA模型通过自注意力机制把模型的表示和浅层的语义信息联系起来，自注意力机制得出的系数(或者说权重)来更有效地从模型浅层获取信息，以便提升模型的稳定性和泛化能力。

### 自注意力权重系数

计算自注意力权重系数的流程图如图3.3所示。类似于3.1.2节中的记号，记个样本组成的集合为，其中包含个类。记相应的类组成的集合为，其中。SKDSA模型中的各个分类器被记为，其中代表卷积网络层中分类器的个数。

在每一个分类器后都设置一个归一化指数层，其输出为：

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

其中代表通过全连接层后的输出。代表第个分类器后属于第个分类器的概率，代表蒸馏温度。我们将式（3.6）记为函数，即蒸馏函数。

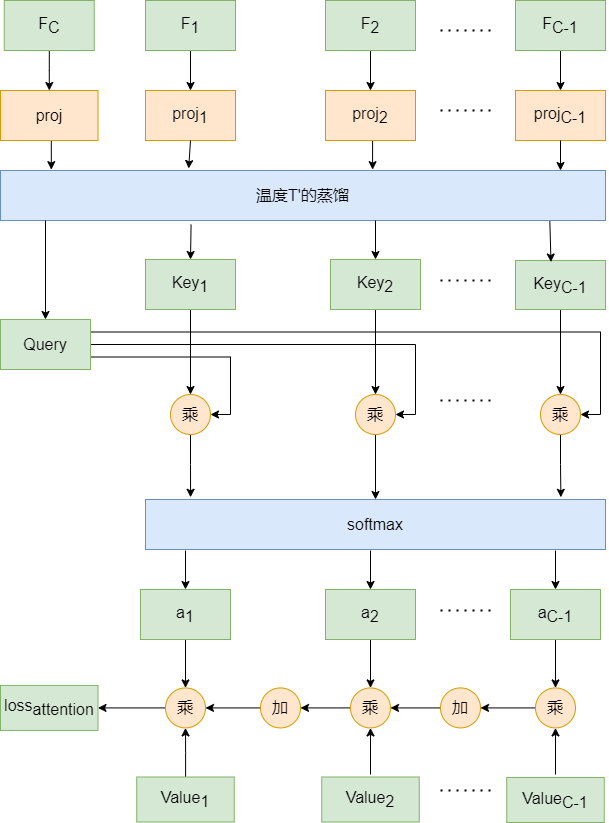


图3.3 自注意力权重计算流程

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。为了能够降维并提取出特征图中更高阶、更本质的特征，我们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，使两者的维数的相同（记为）。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，二是为了使查询和键的维数相同。

记特征图通过投影层的输出向量为，类似于式（3.6）中对知识蒸馏的定义，我们对投影层输出向量的蒸馏函数定义为：

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

其中是对投影层输出向量的蒸馏温度。值得一提的是，对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。在实验部分4.3.3中将探讨对模型性能的影响。

对所有投影层输出的结果统一使用蒸馏函数（式（3.7）），即可得到SKDSA模型中第个浅层分类器生成的键（记为）和最终层分类器生成的查询（记为）：

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

其中。

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：

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

其中。

根据归一化指数函数的定义，不难得出所有浅层分类器的注意力权重系数之和满足。

对SKDSA模型的直观理解如图3.4所示，通过在浅层块（学生模型）和最深层块（教师模型）上加一个注意力模块，用来提取查询和键；然后通过查询和键计算出每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分别代表每一个浅层块对于最深层块的蒸馏重要性。



图3.4 SKDSA模型中自注意力模块的直观理解

### SKDSA模型的损失函数

类似于BYOT模型，SKDSA模型也引入了三种损失函数：真实标签与最深层分类器归一化指数输出的交叉熵，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L1损失函数。为了获得更优异的实验性能，SKDSA模型对BYOT模型的三种损失函数进行了细微的调整，差别如表格3.1所示。

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 | L2损失函数 | 归一化L1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 |

SKDSA模型的损失函数1：真实标签与最深层分类器的交叉熵。

SKDSA模型的损失函数1通过来自训练数据集的真实标签和最深层分类器的归一化指数输出来计算。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（3.9）表示，其中代表某样本经过最深层分类器的归一化指数层输出，代表该样本在数据集中对应的真实标签，代表交叉熵损失函数：

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

SKDSA模型损失函数1和BYOT模型损失函数1的区别是：BYOT模型的损失函数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）的输出与真实标签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与真实标签的交叉熵损失。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：最深层分离器归一化指数输出和每个浅层分类器归一化指数输出的相对熵。

SKDSA模型的损失函数2通过最深层分类器的归一化指数输出和每个浅层分类器的归一化指数的相对熵计算。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模型的损失函数由式(3.10)表示，其中代表第个浅层分类器经过归一化指数层的输出，代表最深层分类器经过归一化指数层的输出，代表相对熵损失函数，代表蒸馏温度。

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

SKDSA模型损失函数2和BYOT模型损失函数2的区别是：BYOT模型直接计算最深层分离器归一化指数输出和每个浅层分类器归一化指数输出的相对熵，而SKDSA模型在计算出最深层分离器归一化指数输出和每个浅层分类器归一化指数输出的相对熵之后，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：最深层分类器特征图和每个浅层分类器特征图的归一化L1损失函数。

SKDSA模型的损失函数3通过最深层分类器的特征图和每个浅层分类器的特征图之间的归一化L1损失计算。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（3.11）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L1损失函数：

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

SKDSA模型损失函数3和BYOT模型损失函数3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L2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L1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（3.10）和（3.11），可以得到对于第个浅层分类器的损失函数为：

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

其中代表调整式（3.10）和（3.11）的超参数。式（3.12）对应于图3.1中的。

将第个浅层分类器的损失函数（式（3.12））预期对应的注意力权重系数（式（3.8））相乘后累加，即可得到自注意力模块的损失函数之和：

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

其中即为图3.3中的最终输出。

综合式（3.9）和式（3.13），即可得出SKDSA模型的总损失函数：

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

其中代表式（2.2）中提到的知识蒸馏损失系数。

SKDSA模型总损失函数和BYOT模型总损失的区别是：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）。

### 算法流程

SKDSA模型的训练步骤如表3.1所示。

表3.2 SKDSA模型的训练步骤

| 基于自注意力机制的自知识蒸馏 |
| --- |
| **输入**： 训练集; 随机初始化的模型参数。  **输出**： 经过充分的训练和更新的模型参数。  1： 循环：  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（3.7），利用和计算和。  5： 根据式（3.8），利用和计算注意力权重系数。  6： 根据式（3.9）（3.10）（3.11）分别计算损失函数1、损失函数2、损失函数3。  7： 根据式（3.12）计算第个浅层分类器的损失函数。  8： 根据式（3.13）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（3.14）中总损失函数。  10： 对梯度反向传播来更新参数。  11：循环直到模型参数收敛。 |

## 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。

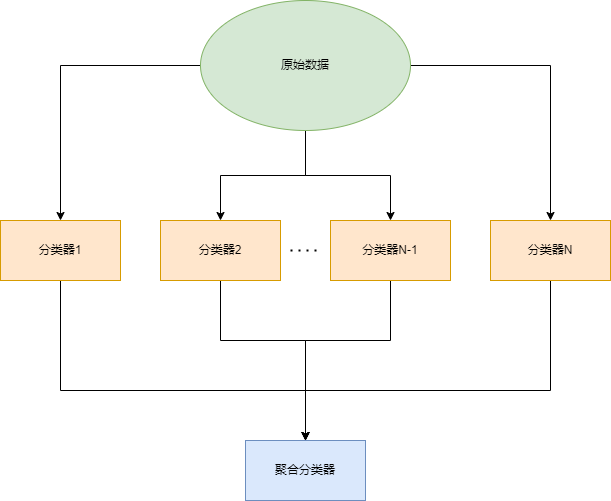


图3.5 袋装法示意图

袋装法是一种是机器学习中的一种集成算法，旨在提高用于统计分类和回归的机器学习算法的稳定性和准确性，减少方差并有助于避免复杂过度拟合。袋装法的过程如图3.8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（几乎）独立的数据集，该数据集来自真实分布；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，但会减少其方差。

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下式：

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

因为SKDSA模型使用注意力机制去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：

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

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归一化指数层的输出。

令式(3.5)中的，，再令，

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

把式（3.16）和（3.17）代入（3.15），可以得到：

|  |  |  |
| --- | --- | --- |
|  |  | （3.19） |

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模型的泛化能力和抗过拟合能力。

## SKDSA模型和数据增强的结合

为了进一步提升SKDSA模型的性能，在3.2节给模型添加了注意力模块之后，本节考虑将SKDSA模型与数据增强模型相结合。常见的图像数据增强方法有裁剪、变色、旋转、注入噪声、MixUp等等。本节仅讨论将挖剪法（cutout）[40]这种数据增强策略。

以一张标签为猫的图片为例（如图3.6所示）。挖剪图像策略将原图像中随机大小的一块正方形区域用“挖掉”（如图3.7所示），再将新图像添加到原有的数据集中，标签保持不变。如果原图像是用矩阵表示，那么挖剪法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）和挖剪后的图像（右）

第四章的实验将分别记录SKDSA模型使用原数据集的性能，和结合挖剪图像数据增强策略之后的性能。

## 本章小结

本章在BYOT模型的基础上提出了SKDSA模型。本章比较了SKDSA模型和BYOT模型在原理框图、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。紧接着从理论上证明了SKDSA模型和袋装法的等价性。最后将SKDSA模型与数据增强策略相结合。

# 实验结果与分析

实验使用一个较小的、自知识蒸馏的神经网络在几个图像分类任务上测试SKDSA模型。为了更进一步测试SKDSA的性能还进行了消融研究。

## 实验准备

### 实验设置

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I 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

实验的目的在于验证本文提出的SKDSA模型能否在图像分类任务中取得更高的准确率。为了测试SKDSA模型的普适性，我们在多个流行的图像数据集上做了实验。

1. 在CIFAR-100[41]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，并和几种典型的自知识蒸馏模型和自注意力模型比较分类准确性。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-10和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[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Tiny ImageNet上进行实验。ImageNet[47]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[43]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[44]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[44]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[45]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[44]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[46]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。初始学习率(learning rate)设为0.1，到第100个epoch时，学习率降为最初的1/10; 到第150个epoch时，学习率降为最初的1/100; epoch总数为250。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）。蒸馏温度从{1，2，3，4}中选取，超参数从{0.5，1，1.5，2，3}中选取，超参数从{50，100，200，500}中选取。

### 评估标准

实验使用分类准确率作为衡量模型性能的指标。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

参加对比的基准线模型为不添加自知识蒸馏模块、也不添加自注意力模块的模型。

除基准线模型之外，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]。

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]。

* **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。
* **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。这里的实验使用了相同的训练集并且微调了基准线模型的超参数。

### 网络架构

为了测试SKDSA模型，在当前流行的卷积神经网络架构残差网络[36]，宽残差网络[37]，残差密集网络[38]上进行了实验。设置包含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。优化函数选择随机梯度下降法（Stochastic gradient descent, SGD）。

## 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、宽残差网络框架、残差密集网络框架的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、结合了数据增强模型的SKDSA模型进行对比。每个实验都重复3次，实验结果记录最后epoch的分类准确率的平均数。实验结果分别如表4.1、表4.2、表4.3所示，最优的结果用粗体标注。

表4.2 各模型在通用图像数据集上的分类准确率（残差网络框架）

| 残差网络框架 | CIFAR-100 | TinyImageNet |
| --- | --- | --- |
| 基准线 | 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 各模型在通用图像数据集上的分类准确率（宽残差网络框架）

| 宽残差网络框架 | CIFAR-100 | TinyImageNet |
| --- | --- | --- |
| 基准线 | 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%** |

表4.4 各模型在通用图像数据集上的分类准确率（残差密集网络框架）

| 残差密集网络框架 | CIFAR-100 | TinyImageNet |
| --- | --- | --- |
| 基准线 | 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%** |

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，SKDSA模型使用残差网络框架在CIFAR-100和TinyImageNet数据集上相比基准线模型分别提升准确率5.71%和5.82%，使用宽残差网络框架在CIFAR-100和TinyImageNet数据集上相比基准线模型分别提升准确率4.38%和1.96%，使用残差密集网络框架在CIFAR-100和TinyImageNet数据集上相比基准线模型分别提升准确率3.78%和2.01%。

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的结合（表4.2和4.3使用的都是“挖剪法”数据增强模型）具有比单一的SKDSA模型更卓越的性能。具体来说，使用残差网络框架在CIFAR-100和TinyImageNet数据集上相比基准线模型分别提升准确率6.97%和8.01%，使用宽残差网络框架在CIFAR-100和TinyImageNet数据集上相比基准线模型分别提升准确率5.61%和2.17%。

### 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、残差密集网络框架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、结合了数据增强模型的SKDSA模型进行对比。每个实验都重复3次，实验结果记录最后epoch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 各模型细粒度图像数据集上的分类准确率（残差网络框架）

| 残差网络框架 | CUB200 | MIT67 | Dogs | Stanford40 |
| --- | --- | --- | --- | --- |
| 基准线 | 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%** |

表4.6 各模型在细粒度图像数据集上的分类准确率（残差密集网络）

| 残差密集网络框架 | CUB200 | MIT67 | Dogs |
| --- | --- | --- | --- |
| 基准线 | 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，SKDSA模型使用残差网络框架在CUB200、MIT67、Dogs、Stanford40数据集上相比基准线模型分别提升准确率12.48%, 5.83%, 6.46%和13.72%，使用残差密集网络框架在CUB-200、MIT67、 Dogs数据集上相比基准线模型分别提升准确率13.40%, 5.99%和7.89%。

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（表4.5中使用的是“挖剪法”数据增强模型）具有比单一的SKDSA模型更卓越的性能。具体来说，使用残差网络框架在CUB200、MIT67、Dogs、Stanford40数据集上相比基准线模型分别提升准确率14.31%, 9.70%, 8.09% 和18.31%。

### 不同自注意力机制的实验结果

为了比较SKDSA模型和两种注意力模型（DIANet, SAN）的性能，分别使用三种残差网络框架（ResNet18, ResNet34, ResNet50）在CIFAR-100数据集上进行实验，并且和其他两种注意力模型进行对比。每个实验都重复3次，实验结果记录最后epoch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。

表4.7 各模型在通用图像数据集上的分类准确率（残差网络框架）

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

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，SKDSA模型使用ResNet18框架相比SAN和DIANet分别提升准确率4.61%和3.89%，使用ResNet34框架相比SAN和DIANet分别提升准确率4.41%和3.71%，使用ResNet50框架相比SAN和DIANet分别提升准确率3.93%和2.53%。

## 消融实验与分析

### 自注意力机制的重要性

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、宽残差网络框架的去掉注意力机制的SKDSA模型、SKDSA模型在CIFAR-100和TinyImageNet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epoch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 SKDSA模型添加和不添加自注意力模块的分类准确率（使用残差网络框架）

| 残差网络框架 | CIFAR-100 | TinyImageNet |
| --- | --- | --- |
| w/o SA | 77.80% | 57.20% |
| SKDSA | **80.51%** | **60.02%** |

表4.9 SKDSA模型添加和不添加自注意力模块的分类准确率（宽残差网络框架）

| 宽残差网络框架 | CIFAR-100 | TinyImageNet |
| --- | --- | --- |
| 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，一个模型在添加自注意力模块后具有比原模型更优异的性能。具体来说，使用残差网络框架时，SKDSA模型在CIFAR-100和TinyImaget数据集上比去掉自注意力模块的SKDSA模型分别提升准确率2.71%和2.82%；使用宽残差网络框架时，SKDSA模型在CIFAR-100和TinyImaget数据集上比去掉自注意力模块的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，对模型性能的发挥不可或缺。

### 自注意力模型与数据增强模型的相容性

为了证明自注意力模型和数据增强模型的相容性，分别使用基于残差网络框架、宽残差网络框架的MixUp模型、添加自注意力模块的MixUp模型、SLA模型、添加自注意力机制的SLA模型在CIFAR-100和TinyImageNet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epoch的分类准确率的平均数。实验结果如表4.10和表4.11所示，最优的实验结果用粗体字标注。

表4.10 数据增强模型添加和不添加自注意力模块的分类准确率（使用残差网络框架）

| 残差网络框架 | CIFAR-100 | TinyImageNet |
| --- | --- | --- |
| MixUp | 78.33% | 58.43% |
| MixUp+SA | **81.41%** | **61.44%** |
| SLA | 77.52% | 58.48% |
| SLA+SA | **82.81%** | **63.02%** |

表4.11 数据增强模型添加和不添加自注意力模块的分类准确率（宽残差网络框架）

| 宽残差网络框架 | CIFAR-100 | TinyImageNet |
| --- | --- | --- |
| MixUp | 72.21% | 52.82% |
| MixUp+SA | **76.73%** | **53.68%** |
| SLA | 73.00% | 50.77% |
| SLA+SA | **76.83%** | **53.89%** |

从表4.10和表4.11可以看出，在通用图像数据集上，数据增强在添加自注意力模块后具有比原模型更优异的性能。具体来说，使用残差网络框架时，添加自注意力模块的MixUp模型在CIFAR-100和TinyImaget数据集上比原MixUp模型分别提升准确率3.08%和3.01%，添加自注意力模块的SLA模型在CIFAR-100和TinyImaget数据集上比原SLA模型分别提升准确率5.29%和4.54%；使用宽残差网络框架时，添加自注意力模块的MixUp模型在CIFAR-100和TinyImaget数据集上比原MixUp模型分别提升准确率4.52%和0.86%，添加自注意力模块的SLA模型在CIFAR-100和TinyImaget数据集上比原SLA模型分别提升准确率3.83%和3.12%。

以上结果说明，自注意力模块可以广泛地提升其他基于数据增强的自知识蒸馏模型的性能。这证实了自注意力模型和数据增强模型的相容性。

### 自注意力模块中的蒸馏温度分析

为了研究自注意力蒸馏温度对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。每个实验都重复三次，实验结果记录最后epoch的分类准确率的平均数。实验结果如图4.1所示，其中横坐标代表自注意力蒸馏温度，绿柱代表分类器蒸馏温度，黄柱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。



图4.1 不同蒸馏温度下的结果以及是否使用温度缩放的结果

从图4.1中可以看出，。具体来说，在不使用分类器蒸馏时，SKDSA模型在自注意力蒸馏温度为2, 3和4时比不使用自注意力蒸馏分别提升准确率1.85%, 1.52%和2.18%；在使用分类器蒸馏时，SKDSA模型在自注意力蒸馏温度为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对提升模型性能非常必要。

### 敏感性分析

这部分实验测试了两个超参数和在不同取值下对分类准确率的影响。测试的取值范围是{0.5，1，1.5，2，3}，取值范围是{50，100，200，500}。实验使用残差网络模型在数据集CIFAR-100和TinyImageNet上做了实验。图4.2说明了不同超参数下SKDSA模型的分类准确性。除了超参数以外，所有的设置都和文章4.3部分中的设置相同。可以发现SKDSA模型在超参数的变化下表现稳定，表现最优的超参数是，。

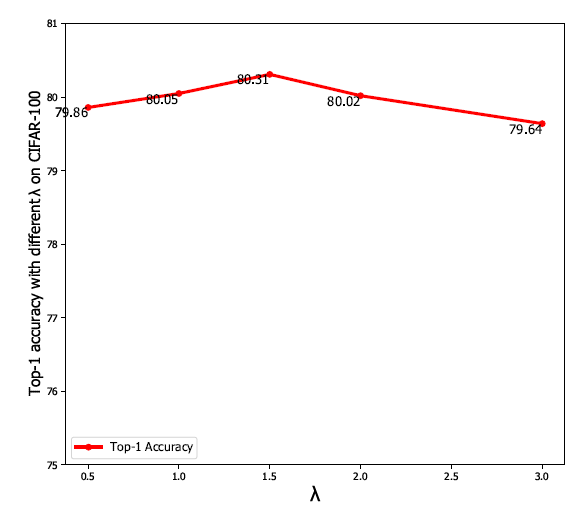


图4.2 使用不同值时的分类准确率

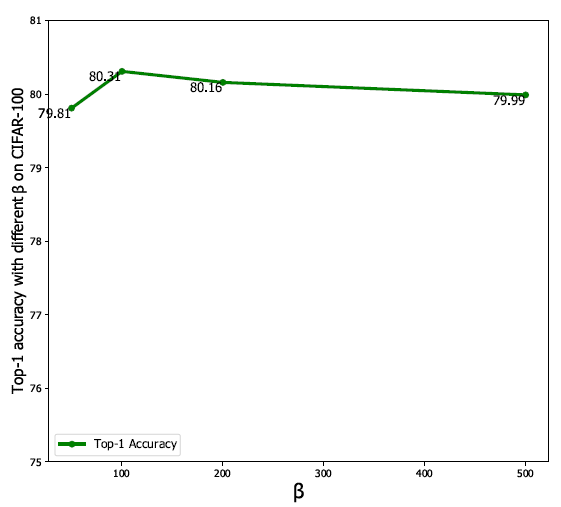


图4.3 使用不同时的分类准确率

## 本章小结

本章使用SKDSA模型在多个通用图像数据集和细粒度图像数据集(包括CIFAR-100[41]，Caltech-UCSD Bird(CUB200)[43]，Stanford 40[44]，Stanford Dogs (Dogs)[45]，and MIT Indoor Scene Recognition (MIT67))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（包括DDGSD模型，BYOT模型，CS-KD模型，SLA-SD模型，FRSKD模型）作为对比。可以看出SKDSA模型在多个数据集上的性能相比现有的自知识蒸馏模型结果都有令人振奋的提升，证实了SKDSA模型的有效性。

另外，这一章还对SKDSA模型中的自注意力模块、自注意力机制与数据增强机制的相容性、蒸馏温度等分别进行了消融实验分析。消融分析证明了自注意机制对提升自知识蒸馏性能很重要，SKDSA模型和数据增强策略相容，温度缩放可以增强模型的泛化能力，以及找到了相对较优的超参数。

# 总结与展望

## 主要工作总结

在BYOT模型的基础上，提出一种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（SKDSA），神经网络最深层扮演教师模型和神经网络浅层扮演的学生模型通过自注意力机制联系起来，通过这种方式把神经网络深层的知识迁移到浅层。SKDSA模型还可以从理论上被证明等价于集成学习。

SKDSA模型进一步被证明了可以视为一种集成学习机制并为此提供了理论证明，这意味着具有更优的泛化性能。虽然这个证明看起来显而易见，但是这个证明是必要的，因为它展示SKDSA模型与袋装法的相似之处，其中弱学习器和投票过程可以提供更稳定和可靠的模型。

SKDSA模型在多个数据集上的图像分类任务上进行了实验并和当前主流的自蒸馏效果进行了对比。SKDSA模型不仅比其他自知识蒸馏性能更优，而且训练时间也更短。消融实验进一步证明了SKDSA模型的注意力机制对提升性能很有必要。

## 主要创新点

由于过去的基于特征的自知识蒸馏模型忽视了不同层特征信息之间重要性的差别，这可能会导致大量暗知识的损失。SKDSA模型给位于神经网络深层的教师和神经网络浅层的学生添加自注意力连接，以便区分不同位置的浅层对最深层特征表示的不同影响。多项实验结果表明SKDSA模型的图像分类准确性比现有的其他自知识蒸馏模型更高。

SKDSA模型看似简单直接，但有充分的根据。尽管过去人们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模型是第一个将它们结合起来的。SKDSA模型使用注意力机制作为一种手段来考虑每个中间层对最终特征表示的不同的重要性，从而挖掘端到端深度学习中的暗知识。

## 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。

1. 自动调整新引入的超参数。

为了平衡交叉熵损失函数、相对熵损失函数和提示损失函数，引入了两个超参数和，如式（3.6）所示。实验发现这两个超参数对模型性能有影响。由于计算资源有限而且论文截止时间临近，SKDSA模型还没有在更广大的范围搜索合适的超参数。在不久的将来，可以使用类似学习率衰减或受动量启发的算法来探索两个超参数的自动调整。

第二，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。

第三，尝试更多的数据增强方式。

因为时间所限，实验只采用了图像裁剪这一种数据增强的方式。未来可能尝试采用更多的数据增强（比如几何变换、颜色转换、随机擦除、对抗训练和风格迁移等）策略，进一步提高实验结果。

第四，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[48], VGG[49]等。

第五，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。另外，FRSKD模型和SLA模型的结合在一些数据集上已经取得了更优异的效果。未来可以考虑将BYOT模型和SLA模型结合在一起。

第六，尝试把SKDSA模型应用到半监督任务和弱监督任务中。

SKDSA模型可能适用于半监督学习和弱监督学习任务。未来有机会的话可以测试这一点。

第七，在语言模型上尝试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。

# 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] G. Hinton, O. Vinyals, J. Dean. Distilling the knowledge in a neural network. arXiv preprint arXiv:1503.02531, 2015

[5] S. J. Pan, Q. Yang. A survey on transfer learning. IEEE Transactions on knowledge and data engineering, 2009, 22(10):1345–1359

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

[7] S. Yun, J. Park, K. Lee, J. Shin. Regularizing class-wise predictions via self-knowledge distillation. In: Conference on Computer Vision and Pattern Recognition, 2020: 13876-13885

[8] M. Ji, S. Shin, S. Hwang, G. Park, I. Moon. Refine myself by teaching myself: Feature refinement via self-knowledge distillation. In: Conference on Computer Vision and Pattern Recognition, 2021: 10664-10673

[9] H. Lee, S. J. Hwang, J. Shin. Self-supervised label augmentation via input transformations. In: International Conference on Machine Learning, 2020: 5714-5724

[10] Ba J, Caruana R. Do deep nets really need to be deep? In: Conference on Neural Information Processing Systems, 2014: 27

[11] Kim, S. W., Kim, H. E.. Transferring knowledge to smaller network with class-distance loss. In: International Conference on Learning Representations workshop, 2017

[12] S. Hou, X. Pan, C. C. Loy, Z. Wang, D. Lin. Learning a unified classifier incrementally via rebalancing. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019: 831–839

[13] Luan, H. Zhao, Z. Yang, and Y. Dai. Msd: Multiself-distillation learning via multi-classifiers within deep neural networks. arXiv preprint arXiv:1911.09418, 2019

[14] T. Xu, C. Liu. Data-distortion guided self-distillation for deep neural networks. In: Association for the Advancement of Artificial Intelligence, 2019, 33(1): 5565-5572

[15] H. Lee, S. J. Hwang, J. Shin. Rethinking data augmentation: Self-supervision and self-distillation. In: International Conference on Machine Learning, 2020

[16] D. Bahdanau, K. Cho, Y. Bengio. Neural machine translation by jointly learning to align and translate. arXiv preprint arXiv:1409.0473, 2014

[17] Vaswani A., et al. Attention is all you need. Advances in neural information processing systems, 2017: 30

[18] M. Dehghani, S. Gouws, O. Vinyals, J. Uszkoreit, Ł. Kaiser. Universal transformers. arXiv preprint arXiv:1807.03819, 2018

[19] T. Shen, T. Zhou, G. Long, J. Jiang, S. Pan, C. Zhang. Disan: Directional selfattention network for rnn/cnn-free language understanding. In: Proceedings of the AAAI Conference on Artificial Intelligence, 2018, 32(1).

[20] C. Xing, Y. Wu, W. Wu, Y. Huang, M. Zhou. Hierarchical recurrent attention network for response generation. In: Proceedings of the AAAI Conference on Artificial Intelligence, 2018, 32(1).

[21] Y. Wang, M. Huang, X. Zhu, L. Zhao. Attention-based lstm for aspect-level sentiment classification. In: Proceedings of the 2016 conference on empirical methods in natural language processing, 2016: 606–615

[22] H. She, B. Wu, B. Wang, R. Chi. Distant supervision for relation extraction with hierarchical attention and entity descriptions. In: International Joint Conference on Neural Networks, 2018: 1–8

[23] S. Chopra, M. Auli, A. 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, 2016: 93-98

[24] J. Fu, H. Zheng, T. 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, 2017: 4438-4446

[25] K. Han, J. Guo, C. Zhang, M. Zhu. Attribute-aware attention model for fine-grained representation learning. In: Proceedings of the 26th ACM international conference on Multimedia, 2018: 2040-2048

[26] P. Sermanet, A. Frome, E. Real. Attention for fine-grained categorization. arXiv preprint arXiv:1412.7054, 2014

[27] T. Hackel, M. Usvyatsov, S. Galliani, J. D. Wegner, K. Schindler. Inference, learning and attention mechanisms that exploit and preserve sparsity in convolutional networks. International Journal of Computer Vision. 2020, 128(4):1047-59

[28] H. Zhang, I. Goodfellow, D. Metaxas, A. Odena. Self-attention generative adversarial networks. In: International conference on machine learning Proceedings of Machine Learning Research, 2019: 7354–7363

[29] T. Kong, F. Sun, C. Tan, H. Liu, W. Huang. Deep feature pyramid reconfiguration for object detection. In: Proceedings of the European conference on computer vision, 2018: 169–185

[30] W. Li, X. Zhu, S. Gong. Harmonious attention network for person re-identification. In: The IEEE Conference on Computer Vision and Pattern Recognition, 2018: 2285-2294

[31] H. Liu, J. Feng, M. Qi, J. Jiang, S. Yan. End-to-end comparative attention networks for person re-identification. IEEE Transactions on Image Processing, 2017, 26(7):3492–3506

[32] J. Fu, J. Liu, H. Tian, Z. Fang, H. Lu. Dual attention network for scene segmentation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2019: 3146-3154

[33] J. Kuen, Z. Wang, G. Wang. Recurrent attentional networks for saliency detection. In: Proceedings of the IEEE Conference on computer Vision and Pattern Recognition, 2016: 3668–3677

[34] Z. Huang, S. Liang, M. Liang, H. Yang. Dianet: Dense-and-implicit attention network. In: Association for the Advancement of Artificial Intelligence, 2020, 34(4): 4206-4214

[35] H. Zhao, J. Jia, V. Koltun. Exploring self-attention for image recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020: 10076-10085

[36] K. He, X. Zhang, S. Ren, J. Sun. Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition, 2016: 770-778

[37] S. Zagoruyko, N. Komodakis. Wide residual networks. arXiv preprint arXiv:1605.07146, 2016

[38] G. Huang, Z. Liu, K. Q. Weinberger and L. Maaten. Densely Connected Convolutional Networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition, 2017: 4700-4708

[39] Breiman L. 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] A. Krizhevsky, G. Hinton. Learning multiple layers of features from tiny images. 2009

[42] Wu J., Zhang Q., Xu G.. Tiny imagenet challenge. Technical report, 2017

[43] C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie. The Caltech-UCSD Birds-200-2011 Dataset. Technical report, 2011

[44] B. Yao, X. Jiang, A. Khosla, A. L. Lin, L. Guibas, F. F. Li. Human action recognition by learning bases of action attributes and parts. In: International conference on computer vision, 2011: 1331–1338

[45] A. Khosla, N. Jayadevaprakash, B. Yao, F. F. 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, 2011, 2(1)

[46] A. Quattoni, and A.Torralba. Recognizing Indoor Scenes. IEEE Conference on Computer Vision and Pattern Recognition, 2009: 413-420

[47] Deng, J. et al., 2009. Imagenet: A large-scale hierarchical image database. In: IEEE conference on computer vision and pattern recognition, 2009: 248–255

[48] S. Xie, R. Girshick, P. Dollar, Z. Tu and K. He. Aggregated Residual Transformations for Deep Neural Networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition, 2017: 1492-1500

[49] Simonyan K, Zisserman A. 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. In International Joint Conference on Artificial Intelligence, 2022, under review. (CCF A类会议；署名单位：华中科技大学)

# 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：知识蒸馏的研究成员

# 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

DIANet Dense-and-Implicit Attention Network

DNN Deep neural network

FRSKD Feature Refinement via Self-Knowledge Distillation

GAN Generative adversarial networks

LSTM Long term short memory

MMD Maximum Mean Discrepancy

RNN Recurrent neural network

SAN Self-attention network

SGD Stochastic gradient descent

SKDSA Self-knowledge distillation based on self-attention

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