**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** | **：** |  | **2022年XX月XX日** |  |  |

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

**for the Master Degree in Engineering （工学硕士）**

**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是对BYOT方法的进一步改进：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具有更强的稳定性。

为了验证SKDSA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模型（残差网络, 宽残差网络, 残差密集网络）和自知识蒸馏方法进行实验测试。实验结果表明，SKDSA相比于现有其他的自知识蒸馏方法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的必要性，以及SKDSA和数据增强方法的相容性。

**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

[基于自注意力机制的自知识蒸馏研究 I](#_Toc99347333)

[A Dissertation Submitted in Partial Fulfillment of the Requirements II](#_Toc99347334)

[for the Master Degree in Engineering （工学硕士） II](#_Toc99347335)

[摘 要 I](#_Toc99347336)

[Abstract II](#_Toc99347337)

[目 录 III](#_Toc99347338)

[1 绪论 1](#_Toc99347339)

[1.1 研究背景与意义 1](#_Toc99347340)

[1.2 国内外研究现状分析 2](#_Toc99347341)

[1.3 论文主要内容 8](#_Toc99347342)

[2 基础理论与相关知识 11](#_Toc99347343)

[2.1 传统知识蒸馏方法 11](#_Toc99347344)

[2.2 代表性的自知识蒸馏的方法 12](#_Toc99347345)

[2.3 代表性的注意力方法 15](#_Toc99347346)

[2.4 本章小结 18](#_Toc99347347)

[3 基于自注意力机制的自知识蒸馏 19](#_Toc99347348)

[3.1 BYOT方法 19](#_Toc99347349)

[3.2 基于自注意力机制的自知识蒸馏 21](#_Toc99347350)

[3.3 SKDSA和袋装法的等价性证明 30](#_Toc99347351)

[3.4 本章小结 33](#_Toc99347352)

[4 实验结果与分析 34](#_Toc99347353)

[4.1 实验环境 34](#_Toc99347354)

[4.2 实验数据集 34](#_Toc99347355)

[4.3 超参数设置 36](#_Toc99347356)

[4.4 评估标准 36](#_Toc99347357)

[4.5 对比模型 37](#_Toc99347358)

[4.6 网络架构 37](#_Toc99347359)

[4.7 实验结果与分析 38](#_Toc99347360)

[4.8 消融实验与分析 41](#_Toc99347361)

[4.9 本章小结 44](#_Toc99347362)

[5 总结与展望 46](#_Toc99347363)

[5.1 主要工作总结 46](#_Toc99347364)

[5.2 主要创新点 46](#_Toc99347365)

[5.3 未来工作展望 47](#_Toc99347366)

[参考文献 49](#_Toc99347367)

[附录1 攻读硕士学位期间取得的研究成果 58](#_Toc99347368)

[附录2 攻读学位期间参加的科研项目 59](#_Toc99347369)

[附录3 中英文缩写对照表 60](#_Toc99347370)

# 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 | 8000 |
| T5 | 11000 |
| Turing NLG | 17000 |
| GPT-3 | 175000 |

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一些问题。其中，Zhang等人[6]把神经网络的深层视为教师模型，把神经网络的浅层视为学生模型，用神经网络深层的信息来训练神经网络浅层。然而，这种方法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比其他自蒸馏和自注意方法获得了更优异的分类准确性。

## 国内外研究现状

### 知识蒸馏

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。这个方法在图片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

虽然传统的知识蒸馏方法取得了优异的效果，但是它们还是有值得改进的空间。首先，传统的知识蒸馏方法效率低下，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方法。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[6]就是一种这样的方法，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方法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。这样的方法带来如下优点：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提炼，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方法，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(图片旋转和颜色变化)，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

### 自注意力机制

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方法。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片的某些部分，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片里，大熊猫只占图片的一部分)。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][32]、分割[33]、显着性检测[34]等等。Fu等人[24]提出了循环注意卷积神经网络（Recurrent attention convolutional neural network, RA-CNN），它以强化学习的方法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]提出了自注意力生成对抗网络（Self-Attention Generative Adversarial Network, SAGAN），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。

关于注意力机制在网络结构中的应用[35][36]，我们将在第二章介绍。

### 数据增强

有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。

为了构建有用的深度学习模型，验证误差必须随着训练误差而不断减小。数据增强是实现这一目标的一种非常强大的方法。增强的数据将代表一组更全面的可能数据点，从而最小化训练集和验证集之间的距离，以及任何未来的测试集。

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2.12 图像数据增强分类

SKDSA和数据增强结合的实验将在第四章的消融实验部分（4.3.2节）介绍。

## 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1.4 组织结构图

第一章 绪论：第一章首先介绍知识蒸馏的提出背景和意义。接下来介绍了与研究内容相关的（包括知识蒸馏、自注意力机制、数据增强）的国内外研究现状。最后说明了文章的组织结构。

第二章 基础理论与相关知识：第二章主要介绍了本篇论文中涉及到的接触理论和相关知识（包括知识蒸馏、注意力机制），为第三章的方法提出和第四章的实验评估铺垫。

第三章 基于自注意力机制的自知识蒸馏：在BYOT的基础上提出SKDSA方法。详细介绍了SKDSA方法的原理框图，损失函数，卷积神经网络结构（包括残差网络‍[37]、宽残差网络[38]、残差密集网络[39]），伪代码。并且证明SKDSA方法和袋装法‍[40]的等价性。

第四章 实验结果和分析：在多个数据集上（包括CIFAR-100[41], Tiny ImageNet[42], Caltech-UCSD Birds 200[44], Stanford 40[45], Stanford Dogs[46], MIT Indoor Scene Recognition[47]）进行图像分类实验，发现SKDSA的性能比其他已有的自知识蒸馏方法结果都优异的性能。之后还做了消融研究。

第五章 总结与展望：SKDSA将自注意力机制和自知识蒸馏结合起来，获得了比其他已有的自知识蒸馏方法结果更优异的图像分类结果。随后证明了SKDSA和袋装法的等价性。随后进行了消融研究。

# 基础理论与相关知识

由于SKDSA方法是知识蒸馏方法和注意力方法的结合，这一章中我们将分别介绍有代表性的知识蒸馏方法（2.1节和2.2节）和注意力方法（2.3节），以便在第三章引出SKDSA方法。本章节提到的自蒸馏方法和注意力方法也会作为第四章实验评估的对比模型。

## 传统知识蒸馏方法



图2.1 知识蒸馏示意图

传统知识蒸馏的示意图如图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softmax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]提出标签软化的概念。假定代表由模型产生的logit向量，一张图片的属于第个类别概率为：

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

其中是蒸馏温度参数，代表教师模型中标签软化的程度。当时，可以得到正规的softmax公式(也就是没有蒸馏)。随着增加，softmax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。

## 代表性的自知识蒸馏的方法

这一节我们介绍几种代表性的自知识蒸馏方法的原理，它们也是我们在第四章对SKDSA进行实验评估时的对比模型。

### DDGSD

Xu等人[14]提出的DDGSD。

Xu等人提出了一种自蒸馏机制，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（例如，图片的随机镜像/裁剪、旋转、多尺度）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。在另一个观点中，有效的泛化要求模型能够在相同训练数据的不同扭曲版本中保持相似或不变的类概率预测（后验分布）和全局特征表示（特征分布）。因此，为了进一步挖掘单个网络的潜在容量，他们设计了一种新的数据失真引导自提炼机制，以最小化数据到数据知识转移的后验概率和特征分布的差异。具体说来，他们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(MMD)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。在训练基线分类任务时，他们仍然使用硬标签保留传统的交叉熵损失项（即 softmax 损失）。上述三个损失项可以集成在一起作为优化目标，以端到端的方式从头开始实现单网络的自蒸馏。这种新机制可以驱动单个网络自动学习更多固有表示进行泛化，有效抑制学习到的特征朝向单方向依赖。

### FRSKD

SKDSA对比试验的基准之一是Ji等人[8]提出的FRSKD。

过去的知识蒸馏方法依赖于辅助网络，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，基于数据增强的方法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。

为了解决现有自知识蒸馏的局限性，Ji等人[8]提出了一种新颖的自知识蒸馏方法，即通过自知识蒸馏（FRSKD）进行特征细化，它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 与现有知识蒸馏方法之间的差异。Ji等人[8]提出的方法 FRSKD 可以利用软标签和特征图蒸馏来进行自知识蒸馏。



图2.3 FRSKD和其他知识蒸馏方法的区别[8]

### SLA-SD

SKDSA对比试验的基准之一是Lee等人[9]提出的SLA-SD。

Lee等人[9]发现，多任务学习方法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方法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2.4 聚合和自蒸馏[9]

### CS-KD

SKDSA对比试验的基准之一是Yun等人[7]提出的FRSKD。

正则化深度神经网络的预测分布可能是有效的，因为它包含模型最简洁的知识。Yun等人[7]利用暗知识的概念（即DNN 做出的错误预测的知识）为深度模型开发一种新的输出正则化器。它的重要性通过知识蒸馏得到了证明。

虽然大多数工作使用知识蒸馏将教师网络学到的暗知识转移到学生网络，作者在训练单个网络时对暗知识本身进行正则化，即自知识蒸馏。具体来说，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 (CS-KD)，它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，如图2.5所示。可以预期，如果样本属于同一类，作者所提出的正则化方法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2.5 CS-KD示意图

## 代表性的注意力方法

SKDSA是添加了自注意机制的BYOT方法，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。

那么有什么方法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。

### DIANet

注意力机制也可以应用在网络结构中，比如Huang等人[35]提出的DIANet，这也是第四章对比实验的基准之一。

注意力是一种选择性关注一小部分信息而忽略其他可感知信息的认知过程，目前已被用于有效地缓解神经网络从句子中学习大量信息上下文的问题、图像和视频。特别是在计算机视觉中，深度神经网络结合了模仿注意力机制的特殊算子，可以有效地处理图像中的信息区域。这些算子被模块化并作为注意力模块插入网络。

过去的注意力方法是设计一个新的注意力模块并将它们单独插入网络。一般来说，注意力模块可以分为三个部分：提取、处理和重新校准。首先，插件模块提取网络的内部特征，可以压缩通道信息或空间信息。接下来，该模块处理提取并生成掩码，以通过全连接层、卷积层来衡量特征的重要性。最后，将掩码重新校准特征。以前的注意力模型将重点放在设计有效的方法来处理提取的特征，他们的共同特点是将注意力模块被单独插入整个深度神经网络的每一层。

Huang等人[35]提出了一个新颖而简单的框架，该框架在不同的网络层中共享一个注意力模块，以鼓励分层信息的集成，这个参数共享模块被称为密集和隐式注意 (DIA) 单元。DIA 单元的结构和计算流程如图2.6所示。在 DIA 单元中也有三个部分：(1)提取、(2)处理、(3)重新校准。其中第二部分是 DIA 单元中对网络注意力进行建模的主要模块，是DIANet方法的关键创新，其中注意力模块的参数是共享的。

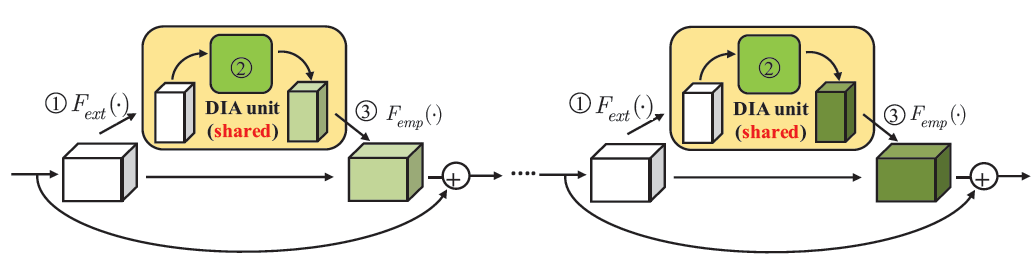


图2.6 残差网络中的DIA单元[35]

此方法的优点如下：(1) 如图2.6所示，DIA 单元与网络主干平行放置，并与同一阶段的所有层共享（具有相同空间大小的连续层的集合)以改善不同深度层的交互。 (2) 由于 DIA 单元是共享的，随着网络深度的增加，来自 DIA 单元的参数增量的数量大致保持不变。

### SAN

注意力机制在网络结构中的另一个应用是Zhao等人[36]提出的SAN，这也是第四章对比试验的基准之一。

Zhao等人[36]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。

## 本章小结

本章介绍了SKDSA方法所涉及的方法和技术（包括有代表性的自知识蒸馏方法和自注意力方法），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方法和自注意力方法结合在一起提供了启发。

# 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方法（3.1节），然后阐述在BYOT基础上改进得到的基于自注意力机制的自知识蒸馏方法（3.2节），最后证明SKDSA和袋装法的等价性（3.3节）。

## BYOT方法

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[6]提出的BYOT方法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：首先，目标卷积神经网络根据其深度分为几个浅层。例如，将ResNet50 根据ResBlocks分为4个部分。第二，在每个浅层之后设置一个分类器，结合一个瓶颈层和一个仅用于训练并且可以在推理中移除的全连接层。添加瓶颈层的主要考虑是减轻每个浅分类器之间的影响，并添加来自学生层信息的引导。在训练阶段，BYOT将所有具有相应分类器的浅层视为学生模型，而将最深层视为教师模型。BYOT的框架图如图2.2所示。



图2.2 BYOT的框架图[6]

记为一系列标签实例，其中N是标签的数量; 记为最深层的特征，记为最深层的输出。假设该模型有个学生层，记为第个学生层的特征，记为第个学生层的输出(的范围是从1到)。

为了提高学生模型的性能，在训练过程中引入了三种损失函数：

损失函数1： 真实标签与每个分类器的交叉熵损失。它使用来自训练数据集的真实标签和每个分类器（包括所有浅层分类器和最深层分类器）的输出来计算。通过这种方法，BYOT把隐含在数据集中的知识从真实标签引入到最深层分类器。

损失函数1由式（2.3）表示。它由第个的分类器输出和真实标签计算得出：

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

损失函数2： 教师指导下的相对熵损失函数。相对熵使用学生和教师之间的softmax输出计算，并引入每个浅层分类器的softmax层。通过引入相对熵，自知识蒸馏框架将教师神经网络(最深的神经网络)的影响传递到每个浅层分类器。

损失函数2由式（2.4）表示。它是和的相对熵，目的是让添加的浅层分类器能够模拟最深层的预测结果，其中代表第个分类器经过softmax层的输出，代表最深层分类器经过softmax层的输出，代表蒸馏温度。

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

损失函数3： 特征图之间的L2损失函数。它可以通过计算最深层分类器和每个浅层分类器的特征图之间的L2损失来获得。通过 L2 损失函数，模型将特征图中的隐含知识引入每个浅层分类器的瓶颈层，从而引导所有的分类器在其瓶颈层中的特征图拟合最深层分类器的特征图。

损失函数3由式（2.5）表示。它代表最深层分类器对浅层分类器的“提示”。BYOT模型将“提示”定义为教师模型隐藏层的输出，目的是指导学生模型的学习。它的工作原理是缩短浅层分类器的特征图和最深层分类器的特征图之间的距离。 然而，由于不同深度的分类器的特征图具有不同的大小，因此需要添加额外的神经网络层来对齐它们。BYOT没有使用卷积层，而是使用了瓶颈架构，该架构能够更有效地提升模型性能。式（2.5）中表示第个浅层分类器的特征，表示最深层分类器的特征。

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

需要说明的是，所有新添加的层（图2.2中ResBlock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，也就是式（2.3）（2.4）（2.5），神经网络整体的损失函数可以写成：

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

其中BYOT使用两个超参数和来平衡式（2.3）（2.4）（2.5）。

## 基于自注意力机制的自知识蒸馏

正如前文所提到的，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（2.3）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。为了解决这个问题，提出一种基于自注意力机制的自知识蒸馏方法来考虑各个浅层的不同重要性。



图3.1 SKDSA的框架图

SKDSA的框架图如图3.1所示，(1) 神经网络根据深度被分成几个浅层。(2) 在每一个浅层下方，添加上额外的瓶颈层和全连接层，再导向多个分类器。(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。

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。SKDSA利用由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。

### 自注意力机制

这里给出计算出自注意力机制权重的具体步骤。第一步，基于伪教师查询和浅层键来计算出权重系数，然后把它们加起来。记相应的权重为，则它们满足。每一个实例都为第个浅层保存关联系数。注意力权重由下式计算：

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

SKDSA使用softmax函数来衡量每一个浅层对于最深层的蒸馏重要性，其中代表最后一层表示所产生的查询，代表第个浅层产生的键。正如图3.2所示，SKDSA在浅层(学生层)和最深层(教师层)上加一个适配层，用来提取查询和键；然后通过查询和键计算出注意力分数。下式通过添加投影层计算出和：

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

投影层的目的是降维，把教师层的维数降到。同样的道理，为了计算注意力权重，SKDSA使用第个投影层来处理第个浅层的键特征，使每一个查询的键和查询的维数相符。此外，在计算注意力权重前，通过式（2.1）引入温度缩放机制。加下来通过式（3.1）得到查询和键的相似性，以此来计算注意力权重。

### SKDSA的损失函数

类似于BYOT方法，SKDSA方法也引入了三种损失函数。

（1）损失函数1：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。

SKDSA使用的第一种损失函数是真实标签与最深层分类器的交叉熵。它使用来自训练数据集的真实标签和最深层分类器的输出来计算。通过这种方法，SKDSA把隐含在数据集中的知识从真实标签引入到最深层分类器。值得一提的是，BYOT方法计算的是所有分类器（包括所有浅层分类器和最深层分类器）的输出与真实标签的交叉熵损失再统计它们的和，而SKDSA仅仅计算最深层分类器与真实标签的交叉熵损失。

第一种损失函数由式（3.3）表示。它由最深层的分类器输出和真实标签计算得出：

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

（2）损失函数2：相对熵。

在很多情况下，人们需要比较两个概率分布。具体来说，假设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。

人们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，它以Solomon Kullback和Richard Leibler方法的两位作者命名，因此也被称为KL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。两个分布和之间相对熵通常表示为。它的计算方式为：

|  |  |  |
| --- | --- | --- |
|  |  | （2.16） |

相对熵值的直观理解是，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。

SKDSA使用的第二种损失函数是学生softmax输出和教师softmax输出的相对熵。通过引入相对熵，自知识蒸馏框架将教师神经网络(最深的神经网络)的影响传递到每个浅层分类器。

第二种损失函数由式(3.4)表示。它是和的相对熵，目的是让添加的浅层分类器能够模拟最深层的预测结果，其中代表第个分类器经过softmax层的输出，代表最深层分类器经过softmax层的输出，代表蒸馏温度。

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

（3）损失函数3：特征图之间的L2损失函数。

L2 损失函数代表最小二乘误差。L2损失函数用于最小化误差，误差是真实值和预测值之间的所有平方差之和。

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

SKDSA使用的第三种损失函数是特征图之间的L2损失函数。它可以通过计算最深层分类器和每个浅层分类器的特征图之间的L2损失来获得。通过 L2 损失函数，模型将特征图中的隐含知识引入每个浅层分类器的瓶颈层，从而引导所有的分类器在其瓶颈层中的特征图拟合最深层分类器的特征图。

损失函数3由式（2.5）表示。它代表最深层分类器对浅层分类器的“提示”。BYOT模型将“提示”定义为教师模型隐藏层的输出，目的是指导学生模型的学习。它的工作原理是缩短浅层分类器的特征图和最深层分类器的特征图之间的距离。 然而，由于不同深度的分类器的特征图具有不同的大小，因此需要添加额外的神经网络层来对齐它们。SKDSA没有使用卷积层，而是使用了瓶颈架构，该架构能够更有效地提升模型性能。式（2.5）中表示第个浅层分类器的特征，表示最深层分类器的特征。

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

SKDSA的总损失函数可以写成：

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

其中代表式（2.2）中提到的知识蒸馏损失系数(kd loss)，代表基于自注意力机制的自知识蒸馏(SKDSA)损失函数。此外，通过修正式（2.6）中的特征损失函数，使损失函数正则化。SKDSA的损失函数可以写成：

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

### 网络模型

由于涉及非常多的网络层数，SKDSA模型使用了残差网络[37]的模型。在第四章的实验中，我们还测试了残差网络的变种宽残差网络[38]和密集残差网络[39]。

残差神经网络[37]可以说是过去几年计算机视觉和深度学习领域中最具开创性的工作。残差神经网络使训练多达数百甚至数千层的神经网络成为可能，并且仍然取得了令人信服的实验结果。残差神经网络利用其强大的表示能力，在图像分类以外的许多计算机视觉应用也取得了成功，例如对象检测和人脸识别。

根据通用逼近定理，给定足够的容量，可以知道单层的前馈网络足以表示任何函数。但是，当网络层数越大时，网络越容易过度拟合数据。因此，研究界有一个共同的趋势，即神经网络架构需要更层数更多同时避免过拟合。

然而，增加网络深度并不能通过简单地增加网络层数来实现。由于臭名昭著的梯度消失问题，层数较多的深度网络很难训练——因为梯度被反向传播到更早的层，重复乘法可能会使梯度无限小。结果如图2.10所示，随着层数的增加，神经网络的性能变得饱和甚至开始迅速下降。



图2.14 层数的增加导致更坏的性能[37]

残差网络的核心思想是引入“恒等跳跃连接”，跳过一层或多层，如下图所示：



图2.15 一个残差块



图2.16 残差网络的结构

He等人认为，堆叠层不应该降低网络性能，因为可以简单地在当前网络上堆叠恒等映射（不做任何事情的层），并且最终的网络架构将执行相同的操作。这表明层数更多的模型不应产生比层数较少的模型更高的训练误差。他们假设让堆叠层适合残差映射比让它们直接适合所需的底层映射更容易。上面的残差块使它精确地做到这一点。

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[38]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。

### 算法步骤

在推断阶段，SKDSA把神经网络最深层的输出作为模型预测的结果。SKDSA的算法步骤如表3.1所示。



图3.2 对SKDSA中自注意力模块的说明

表3.1 SKDSA算法步骤

| 基于自注意力机制的自知识蒸馏。 |
| --- |
| **输入**： 训练集; 模型参数随即初始化。  **输出**： 模型参数经过充分的训练和更新。  1： while 模型参数还没有收敛：  2： 从训练集中挑出一批数据记为。  3： 将放入特征译码器进行前向传播，获得浅层的表示和最深层表示。  4： 根据式（3.5）计算和。  5： 计算的注意力分数和学生键。  6： 利用式（3.7）中损失函数的梯度反向传播来更新参数，其中注意力分数作为权重。  7： end while |

## SKDSA和袋装法的等价性证明

这一部分将证明SKDSA和集成学习中袋装法[40]的关系。

由于需要证明SKDSA方法和袋装法的等价性，所以需要先介绍袋装法的概念。

在训练模型时，无论处理的是分类问题还是回归问题，都可以获得一个函数，该函数接受输入，返回输出，并根据训练数据集进行定义。由于训练数据集的理论差异（因为数据集是来自真实未知潜在分布的观察样本），拟合模型也会有所不同：如果观察到另一个数据集，可以获得不同的模型。

袋装法的思路很简单：拟合几个独立的模型并“平均”它们的预测，以获得具有较低方差的模型。然而，在实践中，不可能拟合完全独立的模型，因为它需要太多的数据。因此，可以依靠引导样本的“近似属性”（代表性和独立性）来拟合几乎独立的模型。

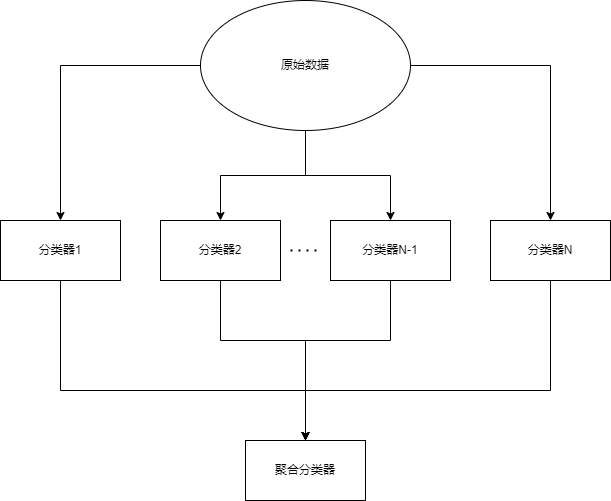


图2.13 袋装法示意图

袋装法的过程如图2.8所示。首先，创建多个 bootstrap 样本，以便每个新的 bootstrap 样本将充当另一个（几乎）独立的数据集，该数据集来自真实分布。然后，可以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。粗略地说，由于 bootstrap 样本是近似独立且同分布（i.i.d.）的，因此学习的基本模型也是如此。然后，“平均”弱学习器输出不会改变预期答案，但会减少其方差（就像平均 i.i.d. 随机变量保留预期值但减少方差一样）。

因此，假设有 L 个大小为 B 的引导样本（L 个独立数据集的近似值），表示为，它们可以拟合 L 个几乎独立的弱学习器。然后将它们聚合到某种平均过程中，以获得具有较低方差的集成模型。

有几种可能的方法可以聚合并行拟合的多个模型。对于回归问题，可以对所有单个模型的输出取平均以获得集成模型的输出。对于分类问题，每个模型输出的类可以看作是一个投票，获得多数票的类由集成模型返回（这称为硬投票）。仍然对于分类问题，还可以考虑所有模型返回的每个类的概率，对这些概率进行平均，并保持具有最高平均概率的类（这称为软投票）。如果可以使用任何相关权重，则平均值或投票可以是简单的或加权的。投票结果可以通过直接相加或者加权得出。

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。它还可以减少方差并有助于避免过度拟合。虽然它通常应用于决策树方法，但它可以与任何类型的方法一起使用。袋装法是模型平均方法的一个特例。

袋装法的过程是这样的。 (1) 假设训练数据集中有 N 个观测值和 M 个特征。 从训练数据集中随机抽取一个样本进行替换。(2) 随机选择 M 个特征的子集，并使用给出最佳分割的任何特征迭代地分割节点。(3) 让这棵树长到最大。(4) 上述步骤重复 n 次，并根据 n 棵树的预测聚合给出预测。

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：

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

因为SKDSA使用注意力机制去将各个浅层的信息到最深层，SKDSA模型的最终预测表示可以写成：

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

其中满足。令式(3.5)中的(因为神经网络有层)，，，再令，

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

把式（3.6）和（3.7）代入(3.5)，可以得到：

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

由此证明了SKDSA模型的输出可以被视为不同模型的集成，这提升了SKDSA模型的泛化能力。而SKDSA模型中的注意力系数又恰恰对应于集成学习中的投票权重。

## 本章小结

这一章从理论上介绍了知识蒸馏、自知识蒸馏以及由此导出SKDSA方法。接下来还从理论上证明了SKDSA方法和袋装法的等价性。

# 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。接下来为了更进一步测试SKDSA的性能还进行了消融研究。

## 实验准备

### 实验环境

实验环境为外租的AutoDL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

1. CIFAR-10[41]和CIFAR-100[41]数据集

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。

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. Tiny ImageNet数据集[42]

ImageNet[43]项目是一个大型视觉数据库，用于视觉对象识别软件研究。该项目已经对超过 1400 万张图像进行了手工标注。ImageNet 包含 20，000 多个类别，每个典型类别由数百张图像组成。第三方图像 URL 的注释数据库可直接从 ImageNet 免费获得，但实际图像不属于 ImageNet。自 2010 年以来，ImageNet 项目每年举办一次软件竞赛，即 ImageNet 大规模视觉识别挑战赛 (ILSVRC)，软件程序将竞争正确分类和检测对象和场景。挑战赛使用一千个非重叠类的“修剪”列表。

ImageNet 大规模视觉识别挑战赛(ILSVRC)始于 2010 年，现已成为图像识别的标准基准。Tiny ImageNet Challenge 是一个类似的挑战，数据集较小而且图像类别较少。它包含 200 个图像类，训练集包含100000个图像，验证集包含10，000个图像，测试集包含10，000个图像。所有图像的大小都是64×64。

由于实验室的计算资源有限，SKDSA难以在数据集ImageNet上进行测试；只能在Tiny ImageNet数据集上进行了测试，它是ImageNet的一个节省资源的替代品。

1. Caltech-UCSD Birds 200数据集[44]

Caltech-UCSD Birds 200 (CUB-200)是一个图像数据集，包含 200 种鸟类（主要是北美）的6033张照片。

1. Stanford 40 Actions数据集

Stanford 40 Actions[45]数据集包含人类执行 40 项动作的图像。数据集总共有 9532 张图像，每个动作类有 180-300 张图像。

1. Stanford Dogs数据集

Stanford Dogs[46]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. MIT Indoor Scene Recognition数据集[47]

室内场景识别是高层次视觉中一个具有挑战性的开放性问题。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。

该数据库包含 67 个室内类别，共 15620 张图像。图片的数量因类别而异，但每个类别至少有 100 张图片。

### 超参数设置

所有的神经网络都是从零开始训练。优化方法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方法，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

第一名命中率被用来作为衡量的指标。第一名命中率的意思是测试集中预测概率最大的标签是正确标签的概率。实验使用第一名命中率来衡量模型的泛化能力，实验的优化的目标是尽可能提高验证集中的第一名命中率。

### 对比模型

将SKDSA的结果和五种已有的知识蒸馏方法相比较：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(SLA-SD)[9]，通过自知识蒸馏预测类别(CS-KD)[7]。实验同样把SKDSA和两种自注意力机制方法相比较：密集而隐含的注意力神经网络(DIANet)[35]，自注意力神经网络(SAN)[36]。

* **BYOT** [6]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]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]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]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]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**[35]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**[36]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]和SLA-SD[9]使用了其官方代码，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线模型的超参数。

### 网络架构

SKDSA在当前流行的卷积神经网络架构残差网络[37]，宽残差网络[38]，残差密集网络[39]上进行了实验。其中包含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。优化方法选择随机梯度下降方法（Stochastic gradient descent, SGD）。

## 实验结果与分析

### 在数据集CIFAR-100和Tiny-ImageNet上的实验结果

SKDSA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在数据集CIFAR-100和Tiny-ImageNet上进行实验，并且和其他自知识蒸馏方法和自注意力方法进行对比，每个实验都重复了3次，实验结果分别如表4.1、表4.2、表4.3所示。实验结果记录了最后epoch的第一名命中率的平均数。最优的结果用粗体标注。

表4.2 SKDSA使用残差网络框架在数据集CIFAR-100和Tiny-ImageNet上的实验结果

| 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** |

表4.3 SKDSA使用宽残差网络框架在数据集CIFAR-100和Tiny-ImageNet上的实验结果

| 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** |

表4.4 SKDSA使用残差密集网络框架在数据集CIFAR-100和Tiny-ImageNet上的实验结果

| 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4.1、表4.2、表4.3不难看出，在数据集CIFAR-100上，和其他神经网络结构相比，SKDSA将准确率提升了3.38%到5.71%; 在数据集TinyImageNet上，SKDSA将准确率提升了1.94%到5.42%。如果仅仅看自知识蒸馏方法，几乎所有的自知识蒸馏模型都可以提升分类准确率。和所有的基准线方法相比，SKDSA取得比其他自知识蒸馏方法更优异的实验性能，在图片分类任务中做到了当前第一。这个结果说明使用自注意力机制的自知识蒸馏机制，模型能够更有效地利用浅层的知识。此外，增加数据增强模块(比如实验中使用裁剪的数据增强方法)，可以在数据集CIFAR-100上提升 1.96%到2.23%，在数据集TinyImageNet上提升0.21%到2.59%，这说明SKDSA方法和数据增强兼容。

### 在细粒度数据集上的实验结果

SKDSA分别使用基于残差网络框架、残差密集网络框架的SKDSA在一些细粒度数据集（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方法和自注意力方法进行对比，每个实验都重复了3次，实验结果分别如表4.4和表4.5所示。实验结果记录了最后epoch的第一名命中率的平均数。最优的结果用粗体标注。

表4.5 SKDSA使用残差网络框架在细粒度数据集上的各种自知识蒸馏方法效果对比

| 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** |

表4.6 SKDSA使用残差密集网络框架在细粒度数据集上的各种自知识蒸馏方法效果对比

| 密集残差网络框架 | 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4.4和表 4.5中可以看到，在细粒度数据集上SKDSA方法比其他自知识蒸馏方法性能更优异。特别要说明的是，在CUB-200-2011数据集中和基准线相比，SKDSA方法将第一名命中率从52.72%提升到67.03%; 在Stanford Dogs数据集中和基准线相比，SKDSA方法将第一名命中率从64.75%提升到72.84%。结果说明在细粒度数据集上SKDSA方法比其他自知识蒸馏方法更优异的实现性能。

除了预测的准确率高，SKDSA模型的训练速度和推理速度也比前三个基准线方法（DDSGD, BYOT, CS-KD）略快，比后两个基准线方法快很多。尤其需要说明的是，SKDSA方法的训练时间仅仅是FRSKD训练速度的四分之一。

### 不同自注意力机制的实验结果

表4.7 基于自注意力机制的视觉模型在数据集CIFAR-100上的对比

| 结构 | SKDSA | SAN | DIANet |
| --- | --- | --- | --- |
| ResNet18 | **80.51** | 75.90 | 76.62 |
| ResNet34 | **80.81** | 76.40 | 77.1 |
| ResNet50 | **81.13** | 77.2 | 78.6 |

接下来把SKDSA方法和两种自注意力模型在数据集CIFAR-100上的表现相比。正如表4.3所示，SKDSA方法显著优于其他两种自注意力方法，这彰显了SKDSA方法的优越性。

## 消融实验与分析

在消融研究中，每个实验都被重复了三次，实验结果记录第一名命中率在最后一个epoch的平均数。最优的实验结果用粗体字标注。实验结果如表4.7和表4.8所示。

表4.8 残差网络框架不同方法下，添加和不添加自注意力模块的实验结果对比

| ResNet18 | CIFAR-100 | TinyImageNet |
| --- | --- | --- |
| w/o SA | 77.80 | 57.20 |
| SKDSA | **80.51** | **60.02** |
| MixUp | 78.33 | 58.43 |
| MixUp+SA | **81.41** | **61.44** |
| SLA | 77.52 | 58.48 |
| SLA+SA | **82.81** | **63.02** |

表4.9 宽残差网络框架不同方法下，添加和不添加自注意力模块的实验结果对比

| WRN-16-2 | CIFAR-100 | TinyImageNet |
| --- | --- | --- |
| w/o SA | 71.71 | 51.92 |
| SKDSA | **74.80** | **53.21** |
| MixUp | 72.21 | 52.82 |
| MixUp+SA | **76.73** | **53.68** |
| SLA | 73.00 | 50.77 |
| SLA+SA | **76.83** | **53.89** |

### 自注意力机制的重要性

为了证明自注意力模块在利用浅层关系信息的重要性，实验使用一个去掉了自注意力模块的模型作为对照。表4.7和表4.8的第一行的w/o SA是去掉自注意力机制的SKDSA（也就是BYOT），第三行的MixUp代表另一种数据增强方法。可以看出，在去掉了自注意力模块之后，每个实验的分类准确性都大幅下降，这证实了SKDSA的每一部分都很有必要。

### 自注意力方法与数据增强方法的相容性

正如表4.7和表4.8所示，在给MixUp方法（一种数据增强方法）增加自注意力模块后，在架构依次为WRN-16-2，ResNet18，WRN-16-2，ResNet18的模型中，第一名命中率依次提高4.52%，3.08%，0.86%，3.01%。对于SLA方法，结果依次提高3.83%，5.29%，3.12%，4.54%。可以发现，通过给SKDSA方法加上增强策略，实验表现同样可以大幅提高。以上这些结果说明，SKDSA方法同样和数据增强方法兼容。



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

### 蒸馏温度分析

为了研究蒸馏温度对实验结果的影响，实验测试了SKDSA方法在不同蒸馏温度{1，2，3，4}下的表现，实验同样测试了给自注意力模块使用温度缩放策略的有效性。如图4.1所示，当蒸馏温度由1提升到2时，预测准确率有所提升，说明SKDSA模型可以从知识蒸馏中获益。图4.1中绿柱代表的是自注意力模块没有使用温度缩放的结果，黄柱代表的是自注意力模块使用温度4缩放的结果，由此可以看出使用温度缩放策略来做自注意力机制使结果更优。

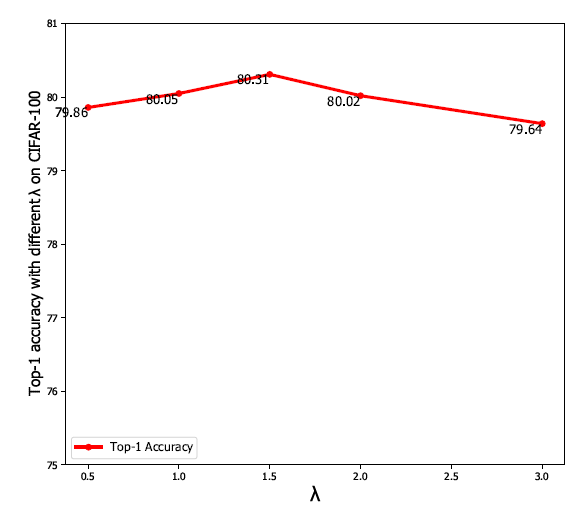


图4.2 使用不同值时的第一名命中率

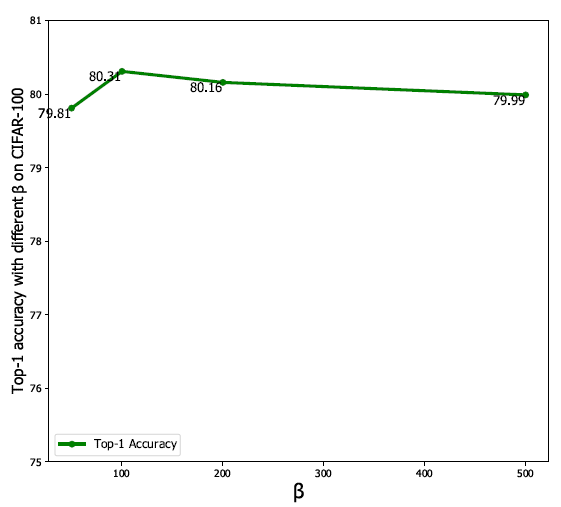


图4.3 使用不同时的第一名命中率

### 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实验结果的影响。测试的取值范围是{0.5，1，1.5，2，3}，取值范围是{50，100，200，500}。实验以残差网络作为骨干在数据集CIFAR-100和TinyImageNet上做了实验。图4.2说明了不同超参数下SKDSA方法的分类准确性。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方法在超参数的变化下表现稳定，表现最优的超参数是，。

## 本章小结

本章使用SKDSA在多个学术界和工业界的数据集(包括CIFAR-10，CIFAR-100，Caltech-UCSD Bird(CUB200)[44]，Stanford 40[45]，Stanford Dogs (Dogs)[46]，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算法（包括DDGSD，BYOT，CS-KD，SLA+SD，FRSKD等）作为对比。可以看出SKDSA方法在多个数据集上的性能相比现有的自知识蒸馏方法结果都有令人鼓舞的提升，证实了SKDSA算法的有效性。

另外，这一章还对SKDSA框架中的自注意力模块、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方法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

为了解决深度学习规模庞大难以训练的问题，在迁移学习的启发下，人们已经提出知识蒸馏将从一个较大模型（教师模型）学到的信息迁移到另一个较小模型（学生模型）。为了教师网络规模庞大难以训练的问题，人们又提出了自知识蒸馏技术。自知识蒸馏技术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。

过去的自知识蒸馏模型（比如BYOT）将作为学生模型的各个浅层一视同仁，忽略了作为教师模型的网络最深层对它们的不同影响。为了解决这个问题，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）。在SKDSA方法中，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还可以从理论上被证明等价于集成学习。

SKDSA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的各个部件都有必要。

## 主要创新点

本论文的主要创新点可以归纳如下：

由于过去的基于特征的自知识蒸馏方法忽视了不同层特征信息之间重要性的差别，这可能会导致大部分暗知识的损失。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方法的图片分类准确性比现有的其他自知识蒸馏方法更高。

SKDSA方法看似很简单，但有充分的根据。尽管过去人们对自知识蒸馏和自注意力机制都有了充分的研究，但SKDSA是第一个将它们结合起来的。SKDSA方法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。

SKDSA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方法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。

## 未来工作展望

SKDSA算法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的深度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和SLA方法的结合在一些数据集上已经取得了更优的效果。未来可以考虑将BYOT和SLA方法结合在一起。

第六，尝试把SKDSA应用到半监督任务和弱监督任务中。

SKDSA方法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方法。

虽然SKDSA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应用到语言模型上的方法策略。

# 参考文献

[1] 卢宏涛, 张秦川. 深度卷积神经网络在计算机视觉中的应用研究综述. 数据采集与处理. 2016, 31(1): 1-7

[2] 奚雪峰, 周国栋. 面向自然语言处理的深度学习研究. 自动化学报. 2016, 42(10): 1445-65

[3] 侯一民, 周慧琼, 王政一. 深度学习在语音识别中的研究进展综述. 计算机应用研究. 2017, 34(8): 2241-6

[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] X. Liao, L. He, Z. Yang, C. Zhang. Video-based person re-identification via 3d convolutional networks and non-local attention. In: Asian Conference on Computer Vision, 2018: 620–634

[33] 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

[34] 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

[35] 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

[36] 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

[37] 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

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

[39] 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

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

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

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

[45] 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

[46] 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)

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

[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