

**本科生毕业论文（设计）**

题目：面向图像表示模型的知识蒸馏算法研究

**学 院： 计算机科学与技术学院**

**专 业： 计算机科学与技术+**

**智能科学与技术**

**班 级： 2019计算机科学与技术**

**+智能科学与技术01**

**学 号： 201906062715**

**学生姓名： 王子平**

**指导老师： 梅建萍**

**提交日期： 2023年6月6日**

面向图像表示模型的知识蒸馏算法研究

# 摘 要

面向图像表示模型的知识蒸馏算法是指将基于自监督学习的图像表示模型通过知识蒸馏方法压缩为模型参数更少，性能尽可能匹配原模型的轻量化模型，目前已有的知识蒸馏方法主要针对深度分类模型，因而无法将现有知识蒸馏算法直接应用于图像表示模型。本文以基于对比学习方法的图像表示模型为蒸馏对象，利用教师模型的中间层特征图、图像表示向量、类别预测向量三类信息，设计面向图像表示模型的知识蒸馏算法，对教师模型进行知识蒸馏。本文提出的面向图像表示模型的知识蒸馏算法结合了基于特征图的知识蒸馏算法[1][17][19]与传统知识蒸馏算法[12]中的部分设计，并针对图像表示模型的特点对算法进行相应的调整。在本文的实验部分中，通过消融实验对各算法模块的效果进行了相应的探究。实验结果表明，本文设计的面向图像表示模型的知识蒸馏算法，可以得到与教师模型性能相近的学生模型。

**关键词** 知识蒸馏，对比学习，图像表示模型

**RESEARCH ON KNOWLEDGE DISTILLATION ALGORITHM FOR IMAGE REPRESENTATION MODELS**

# ABSTRACT

The knowledge distillation algorithm for image representation models refers to compressing self-supervised learning-based image representation models into lightweight models with fewer model parameters while maintaining performance as close as possible to the original model. Existing knowledge distillation methods mainly focus on deep classification models and cannot be directly applied to image representation models. In this paper, we propose a knowledge distillation algorithm specifically designed for image representation models, using contrastive learning-based image representation models as the distillation targets. We utilize three types of information from the teacher model, including intermediate feature maps, image representation vectors, and class prediction vectors, to design the knowledge distillation algorithm. The proposed algorithm combines the design of feature map-based knowledge distillation algorithms[1][17][19] and some aspects of traditional knowledge distillation algorithms[12], while making adjustments according to the characteristics of image representation models. In the experimental section of this paper, we conduct ablation experiments to explore the effects of each algorithm module, and the results demonstrate that the proposed knowledge distillation algorithm for image representation models can achieve student models with performance close to that of the teacher model.

**KEY WORDS:** knowledge distillation,contrastive learning,image representation model



目 录

[摘 要 2](#_Toc23874)

[ABSTRACT 3](#_Toc6905)

[1 绪论 5](#_Toc19682)

[2 研究背景与研究现状 6](#_Toc1069)

[2.1 研究背景 6](#_Toc3239)

[2.2 国内外研究历史及现状 7](#_Toc26583)

[2.2.1 对比学习 7](#_Toc14692)

[2.2.2 知识蒸馏 8](#_Toc6292)

[3 面向图像表示模型的知识蒸馏算法 10](#_Toc18413)

[3.1 算法整体框架 10](#_Toc14805)

[3.2 基于特征图的知识蒸馏 11](#_Toc28725)

[3.3 基于图像表示的知识蒸馏 12](#_Toc17659)

[3.4 传统知识蒸馏 12](#_Toc19050)

[3.5 集成的蒸馏损失函数 13](#_Toc11394)

[4 实验 14](#_Toc30203)

[4.1 评价方法 14](#_Toc13693)

[4.1.1 线性评估 14](#_Toc8609)

[4.1.2 实验具体评估方法说明 14](#_Toc8682)

[4.2 面向图像表示模型的跨模型知识蒸馏 14](#_Toc17637)

[4.2.1 模型的选择 14](#_Toc23204)

[4.2.2 知识蒸馏实验 18](#_Toc7873)

[4.2.3 学生模型的性能对比研究 20](#_Toc32265)

[4.3 消融研究 21](#_Toc27207)

[4.3.1 类别预测信息的消融研究 21](#_Toc8319)

[4.3.2 SimSiam方法中predictor结构维度变化研究 22](#_Toc367)

[5 结论 23](#_Toc13957)

[6 参考文献 24](#_Toc10116)

[7 致 谢 26](#_Toc7157)

# 

# 1 绪论

深度学习作为人工智能中的一种技术，是实现人工智能的重要手段之一。目前，深度学习在理论研究和应用开发方面取得了重大进展，已经在自然语言处理、计算机视觉等领域取得了较多成果，并且还在不断发展中。通过深度学习，计算机可以从大量的数据中自动发现规律和特征，进而实现各种智能应用。随着数字经济的迅速发展，大数据时代为深度学习提供了海量数据，为深度神经网络提供了支撑条件，但同时这也带来了对于大型模型和高性能计算能力的需求增长。

但是，深度学习的诸多技术仍然无法实现广泛应用，随着深度学习模型规模的不断扩大和参数量的迅速增长，如何将这些模型部署在边缘设备（如嵌入式设备、移动设备）并进行推理变得极为困难。在移动计算领域，通常采用的设备是计算能力、存储容量以及能耗都受到很大限制的移动设备，这些设备的算力无法支撑大规模模型进行推理需要实现的复杂运算，这一限制已经成为深度学习模型发展为广泛应用的基础设施的一个重要瓶颈。

模型压缩技术是指通过一系列优化方法，减少深度学习模型的参数量和复杂度，从而提高其在移动设备等资源受限环境下的推理速度和效率。其主要方法有参数剪枝、量化、知识蒸馏、层次融合。知识蒸馏作为一种重要的模型压缩技术，近年来成为研究热点并展现出广阔的前景。但目前的知识蒸馏技术基本针对深度分类模型，而深度分类模型是一种基于有监督学习方法的模型，依赖有标签数据集。尽管有监督学习在诸多任务中表现出色，但由于特殊场景数据获取困难、标注成本高、数据隐私等问题，有监督学习模型的发展受到了限制。

近年来，基于无监督学习的深度网络模型逐渐兴起并成为重点研究方向，自监督学习可以从无标签的数据中学习数据的特征，目前主流的自监督学习可以分为生成式方法和判别式方法。对比学习是一种自监督学习中的判别式方法，目前已经在图像表示领域得到了广泛的应用，可用于人脸识别、图像检索和视觉跟踪等任务。对比学习可以得到能够提取更具鉴别性和泛化性的特征表示的特征提取器。将知识蒸馏算法应用于基于无监督学习的对比学习方法，以将性能较好的重型网络压缩为轻量级网络，使得模型更易于在资源受限环境下部署和推理。

本文从中间层特征图和图像表示的角度出发，研究面向图像表示的知识蒸馏算法，旨在提出能够对基于对比学习的图像表示模型进行有效蒸馏的知识蒸馏算法，实现图像表示模型的轻量化，为基于自监督学习方法的神经网络模型的落地应用提供技术支撑。

# 2 研究背景与研究现状

## 2.1 研究背景

随着近几年自监督学习技术的发展，预训练表示模型成为获取特征表示的重要工具。主流的自监督学习方法分为生成式方法与判别式方法，生成式方法方面Transformer[21]，BERT[7]，GPT-3[9]等大规模预训练模型取得了巨大成功，在图像表示方面的自监督学习的兴起得益于手工设计的一些网络前置任务，如相关区域预测[8]，拼图游戏[16]，着色[23]和旋转预测[3][10]，这些网络前置任务依赖于某种特设的启发式规则，限制了所学习的图像表征的泛化性能，对比学习是一种判别式自监督方法，其目标在于训练得到一个特征提取提器以获得图像表示。对比学习大多基于类孪生神经网络，同一图像经过两种不同的数据变换后得到一对正样本对，不同图像变换后的图像认为是负样本对，对比学习的核心思想就是拉近正对样例距离的同时使得负对样例之间距离尽可能远，从而使得图像在特征空间中得到合理的知识表示，继而提升图像表示在下游任务中的表现。具体而言，对比学习相关技术包括样本对构造、图像变换、网络架构、损失函数。

在样本对构造方面，在一个批次的训练在特征数据中，可能会出现假负样本与困难样本过少的情况，假负样本是指同一类别的数据空间中的相似度较高但在同一批次的训练过程中会被认为是负样本，故需要剔除假负样本以提升模型性能，困难样本是指同一类别的数据在特征空间中相似度不高或不同类别的数据在特征空间中相似度过高，前者被称为困难正样本对，后者被称为困难负样本对，困难样本对比例的增加可以提升特征提取器的性能。

在图像增强方面，图像变换如裁剪、放缩及其组合可以被用于构造正样本对，但由于变换的随机性，可能会产生无效的样本对，对应不同的变换需要相应的方法尽可能避免无效样本对的生成，另外图像语义增强，如修改图像的色彩是一种有效的增强方式，图像增强对于对比学习来说非常重要，适当的增强方式及其组合能使模型性能得到极大提升。

在网络架构方面，对比学习主要采用类孪生网络架构，根据两个分支的网络架构与网络参数更新方式的不同设定，可分为同步对称网络、同步非对称网络、异步对称网络、异步非对称网络，按两个分支的网络架构相同与否分为对称与非对称，按两个分支的网络参数更新方式是否同步可分为同步与异步。不同的网络架构结合了特定的结构设计以取得较好的模型性能。

对比学习研究中常见的一个问题是坍塌解，或称崩溃解，指对于任意的输入模型的输出结果为同一特征向量或仅占特征空间中的一个子空间，大量的负样本一定程度上可以避免崩溃解，在仅使用正样本的模型中，可以通过一些特定的结构如非对称网络、动量更新、交叉梯度等方式防止崩溃解。

面向深度分类模型的知识蒸馏可以通过传递抽象层次较高的标签知识上使浅层学生网络学习到有效知识，该方法[12]以计算类别概率分布之间的KL散度实现，损失函数包含如下两个部分：1）教师网络模型与学生网络模型输出的类别预测的概率分布间的KL散度；2）学生网络模型输出的类别预测的概率分布与真实标签的概率分布之间KL散度。而对于基于对比学习方法的图像表示模型，如SimSiam[5]对比学习方法，图像表示向量是主干网络（编码器）的输出，而作为训练过程中的辅助结构预测器的输出在一定程度上可以被认为是类别预测信息；此外，中间层特征图和图像表示向量蕴含丰富的知识，解决了标签知识抽象层次高、信息单一的问题，故在算法设计上引入了已有的基于特征图的知识蒸馏算法，可以有效应对异构的网络架构下匹配教师网络与学生网络的特征图。

近年来，在知识蒸馏领域，从传统的知识蒸馏方法[12]到基于特征图的知识蒸馏方法[1][17][19][24]，已经有了丰富的研究成果，本设计结合已有方法并针对图像表示模型做出相应的调整，以基于对比学习的图像表示模型为知识转移对象，探究能够有效压缩模型并尽可能保持模型性能的知识蒸馏算法。

## 2.2 国内外研究历史及现状

### 2.2.1 对比学习

对比学习的思想可以追溯到2006年的研究[13]，该研究提出两幅图像进行对比学习的思想，2020年2月，陈挺等人提出了SimCLR方法[4]，该方法需要大量的负对样例与较大的训练批量大小，该框架的实例化模型在Image Net数据集上训练出的图像表示以7%的提升超越了以前的SOTA，甚至接近有监督模型的效果，使得对比学习受到近年来成为研究热点，何恺明等人提出的MoCo方法[14]将负对样例保存在一个队列中，并提出动量编码器。BYOL方法[11]沿用了动量编码器并采用非对称网络架构，使用在线网络和目标网络通过互相学习的方式获得良好的图像表示，并指出大量的负对样例不是必要的，SimSiam方法[5]提出以上对比学习方法的成果其实得益于孪生网络的架构，并认为相比于动量编码器，提前停止梯度下降是关键，通过交叉梯度反向传递的方式取得了成功。

总体而言，对比学习这种判别式自监督学习方法已经有了丰富的研究成果，本设计的目的在于研究对比学习模型的中间层知识表示，并以知识蒸馏方法实现模型压缩和知识迁移。

### 2.2.2 知识蒸馏

近期关于将中间层知识进行迁移的知识蒸馏研究兴起于基于特征图的知识迁移方法的成功，一般认为中间层包含了层级化的知识[7]，相比于高度抽象的输入层表示知识有更丰富的信息。在中间层语义研究上，近期的研究[22]表明可以将预训练模型以块（Block）为拆分单位进行拆分，衡量块之间的相似度并划分为等价类，在此基础上将模块重组视为最优化问题以少接近零训练的方式实现模型重组，这也说明中间层蕴含了分层级的抽象语义信息。

知识蒸馏的损失函数与知识的重构方式决定了教师网络的知识如何被提取并迁移到学生网络中，近年来众多基于中间层知识的迁移方法被提出，这些方法或引入新颖的结构，或从数学理论角度探究“知识”的有效转移。Romero等人于2015提出的Fitnets方法[17]在LKD损失函数基础上引入了LHT损失函数并从教师网络和学生网络中分别切片出Hints块（Block）和Guided块（Block）解决异构教师-学生网络结构匹配问题，以教师网络的Hints块向学生网络的Guided块传递信息（LHT），再令学生网络与Guided块之间进行蒸馏（LKD），该方法将中间层知识迁移引入，随后众多不同迁移中间层的方法相继涌现，2017年提出的AT方法[24]将特征图（feature map）以基于激活（activation）方式映射到注意力图（attention map）或和基于梯度方式映射到雅可比行列式，以注意力图或雅可比行列式的形式将中间层知识传递给学生网络，以缩小多个教师-学生块之间的LAT损失函数为目标使得学生网络学习到注意力信息。2019年提出的VID方法[1]利用了教师网络与学生网络中间层之间的变分互信息以最大化利用互信息获得性能提升。2019年提出的SP方法[19]受到相似的输入在教师网络中间层激活后得到的表示也相似这一现象的启发，提出通过保留此种同一批输入数据所得输出的相似性进行知识迁移，即以最大化教师网络与学生网络之间相似矩阵的相似性。上述方法都是基于特征图设计出知识传递形式。2020年提出的CRD[20]认为此前的一些方法对输出维度之间的独立性假设是限制了模型性能的一个原因，基于对比学习近年来取得的成功，该方法选择将对比学习的思想引入到知识蒸馏中以此来获取数据在表示空间中的高阶依赖，拉近同一样本在学生网络和教师网络分别得到的表示，排斥样本在学生网络得到的表示与样本在教师网咯得到的表示，构造了新的目标函数提高了教师模型与学生模型之间互信息的下限。

但以上方法在匹配教师网络与学生网络之间的层或块（多个层）时都采取了一种硬匹配机制，层与层之间或块与块之间以人工指定方式（随机匹配或一一匹配）构成固定的搭配，但是不同网络的不同层所包含的抽象信息层次可能并不匹配，这将使得知识迁移产生负面的效果，出于此想法，近期提出的SemCKD方法[6]将学生网络的每一层的特征图映射到与教师网络各中间层相匹配的特征图尺度上，并利用注意力机制计算教师网络中间层输出的特征图到学生网络中间层输出的特征图的注意力分数，以此形成教师网络层与学生网络层之间的软匹配，在软匹配机制基础上将教师网络与学生网络的中间层的特征图转换到相同尺度，以均方误差（MSE）衡量转换后的特征图之间的相似度，结合注意力分数的相似度之和构成损失函数LSemCKD，从而避免了中间层语义不匹配所导致的负正则化效应。

# 

# 3 面向图像表示模型的知识蒸馏算法

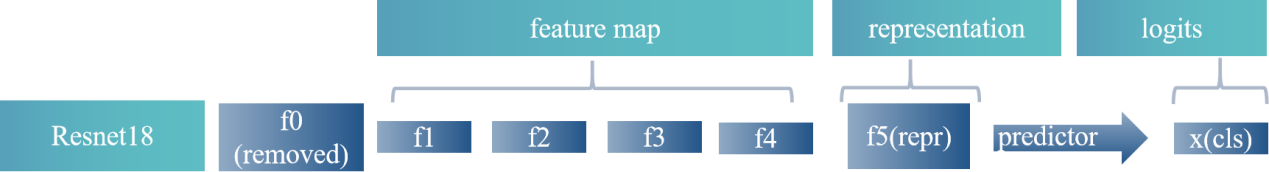
## 3.1 算法整体框架

在这一节中，我们介绍面向图像表示模型的知识蒸馏算法的整体框架，并提供一些必要的符号说明。

图3-3展示了整个蒸馏算法的框架，本系统的知识蒸馏算法主要针对图像表示模型，我们选择SimSiam[5]对比学习框架训练的Resnet50与辅助训练的predictor结构共同作为教师模型，该模型结构主要由三部分组成，如图3-1所示，包含主干网络的编码器Resnet50,替换了Resnet50中全连接层的projector以及预测类别的predictor，由上述三部分可以分别得到各自的输出，故该教师模型可以得到三类信息：(1)中间层特征图(2)图像表示向量(3)类别预测向量。在学生模型上，选择参数量更少的轻量级模型Resnet18，为了进行知识蒸馏，也将学生模型Resnet18的输出分为三类信息，Resnet18的模型结构如图3-2所示，共可获得七项输出f0、f1、f2、f3、f4、f5、x,如图3.3所示将f1、f2、f3、f4作为中间层特征图，将由f4经过下采样与铺平得到的f5作为图像表示向量，将由f5进行全连接后得到的输出x作为类别预测向量。将教师模型与学生模型之间的三类信息各自计算损失函数，并赋予不同的权重，共同构成知识蒸馏的损失函数。



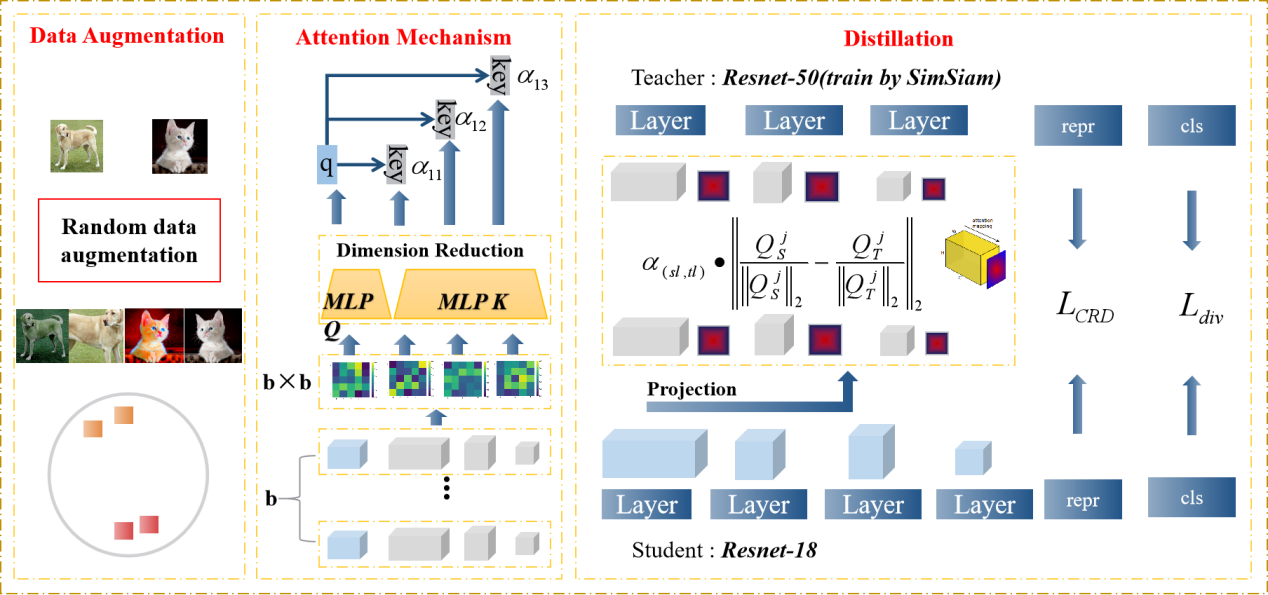
**图3-1 SimSiam的结构**



**图3-2 Resnet18的结构**

在算法框架中，主要由三部分组成：（1）首先进行数据增强操作，得到同一图像的两种不同数据增强后的图像副本，分别输入到教师模型和学生模型，相比直接将同一图像输入到两个模型中，数据增强操作能够提升模型的泛化能力；（2）在中间层特征图匹配问题上，引入论文[6]中提出的基于注意力机制的软匹配方式，计算层间匹配的注意力分数，用于对特征图相似度进行加权；（3）依次计算中间层特征图损失函数、图像表示损失函数、基于类别预测信息的损失函数，在特征图的相似度计算上，引入AT[24]方法中将特征图映射为注意力图的方法，以注意力图的形式传递中间层知识；在计算图像表示向量的损失函数时，引入CRD[20]方法中计算图像表示损失的方法，使用一种类似于InfoNCE损失函数但是比InfoNCE损失函数优化了互信息下限的损失函数；基于类别预测信息的损失函数采用通过教师模型的类别预测概率与学生模型的类别预测概率计算得到的KL散度，相比传统的知识蒸馏算法[12]中略去了学生模型的类别预测概率与独热编码标签的KL散度，这是考虑到本文提出的知识蒸馏算法是面向图像表示模型，而图像表示模型通常以自监督方式训练得到，并不具备标签信息。

在最初考虑算法框架时，在第三部分损失函数计算上只考虑计算中间层特征图损失与图像表示向量损失，但是实验结果表明仅使用此两种损失函数模型无法收敛，故在教师模型部分保留了以SimSiam方法[5]训练时的predictor结构，使得教师模型也能够输出类别预测信息。



**图3-3 面向图像表示模型的知识蒸馏算法框架**

## 3.2 基于特征图的知识蒸馏

在第二章中，已经介绍过一些基于中间层特征图的知识蒸馏相关研究，如Fitnets[17]、VID[1]、SP[19]、AT[24]，这些方法以不同的形式重构教师模型的中间层特征图，并以某种结构传递教师模型的特征图的信息。相关研究[7]表明，用于图像识别的模型得到的中间层特征具有不同抽象程度的层级性质，这使得教师模型与学生模型的中间层特征图在匹配时可能出现抽象程度不一致的情况，即中间层语义不匹配，继而引发知识蒸馏过程中出现负正则化效应。SemCKD方法[6]提出的基于注意力机制的软匹配方法有效缓解了该问题，该方法不再以手工指定方式匹配中间层，而是计算每一对特征图之间的相似度，并对相似度以注意力分数进行加权，而注意力分数是基于每一批次的输入得到的相似度矩阵通过注意力机制得到，其中相似度矩阵的计算方式同SP[19]方法一致。在SemCKD方法[6]中相似度的计算只是简单采用了MSE，在此基础上，受此前AT方法[24]的启发，我们考虑将中间层特征图映射为注意力图，以注意力图的形式计算中间层相似度，再以注意力分数对中间层相似度进行加权，得到如下中间层特征图损失函数：



（3-1）

其中与SemCKD[6]的区别在于引入AT方法[24]传递知识的形式,计算时将特征图映射为注意力图，映射方式为，,并进行了标准化操作。

## 3.3 基于图像表示的知识蒸馏

本文研究面向图像表示模型的知识蒸馏算法，在3.1中介绍的算法框架中提到过，教师模型是使用对比学习方法训练得到的Resnet50，并保留了predictor结构，学生模型是相较于Resnet50参数量更少的Resnet18。由于教师模型的主干网络是一个图像表示编码器，故将图像表示向量单独作为一类信息，并引入CRD[20]中方法，基于图像表示向量对教师模型进行蒸馏，其目标如下：

 （3-2）

其中h表示为如下形式：

 （3-3）

相较于原始的CRD方法，我们对图像进行数据增强操作，将同一图像的两个不同增强样本分别输入到学生模型和教师模型之中，得到同一图像两种不同增强方式后的图像表示，并计算损失函数。

## 3.4 传统知识蒸馏

传统的知识蒸馏即由Hinton等人的工作[12]，利用模型的类别信息进行蒸馏，其损失函数如下：

 （3-4）

其中表示蒸馏的温度系数，是权重系数，是soft max函数，和分别是学生模型和教师模型输出的类别预测向量，是one-hot编码形式的标签。

虽然本文研究面向图像表示模型的知识蒸馏算法，但在实验过程中发现，若考虑到教师模型采用无监督的对比学习方法，所用图像数据集可能为缺失图像类别信息的数据集，这也能训练得到性能较好的图像编码器，但在完全不利用类别信息的情况下，只利用中间层特征图与图像表示向量进行知识蒸馏，学生模型难以有效学习到教师模型的知识，故最终还是引入了利用类别信息的传统知识蒸馏方法[12]中损失函数的后半部分，即计算教师模型与学生模型输出的类别预测概率分布的KL散度，作为面向图像表示模型知识蒸馏的正则项，故本文采用的知识蒸馏算法框架中利用到类别信息的损失函数公式如下：

 （3-5）

相比与公式（3-4），不使用传统知识蒸馏方法[12]中损失函数的前半部分是因为本文所采用的图像表示模型是以自监督的对比学习方法训练得到，在假设图片数据集并不具有标签信息的情况下，仍然能使用该方法对图像表示模型进行知识蒸馏。而之所以能够引入传统知识蒸馏方法[12]中损失函数后半部分的原因是，以SimSiam对比学习方法[5]训练得到的教师网络除了作为主干网络的编码器resnet50之外，模型结构中还包括predictor结构，predictor的输出正是可以作为类别预测信息，而学生模型只需认为最终的输出为类别预测信息即可，最终实验表明，引入该正则项是有效的。

## 3.5 集成的蒸馏损失函数

前面三个小节依次介绍了基于中间层特征图、图像表示向量、预测类别向量，实验结果说明仅采用前两类信息无法有效地对教师模型进行蒸馏，故需要预测类别向量作为正则化项辅助蒸馏过程，将由三类信息各自构造的损失函数进行加权得到集成的蒸馏损失函数，如下所示:

 （3-6）

# 4 实验

## 4.1 评价方法

### 4.1.1 线性评估

线性评估方法是一种常用的对预训练的模型进行模型性能评估的方式，在本文中，线性评估的具体方法为载入预训练得到的编码器resnet50或resnet18的模型结构参数，并冻结全连接层之前的所有参数并将全连接层参数进行初始化，再载入数据集并以交叉熵损失函数计算损失并更新全连接层的参数，最终以准确率（accuracy）作为评价指标评估模型性能。

### 4.1.2 实验具体评估方法说明

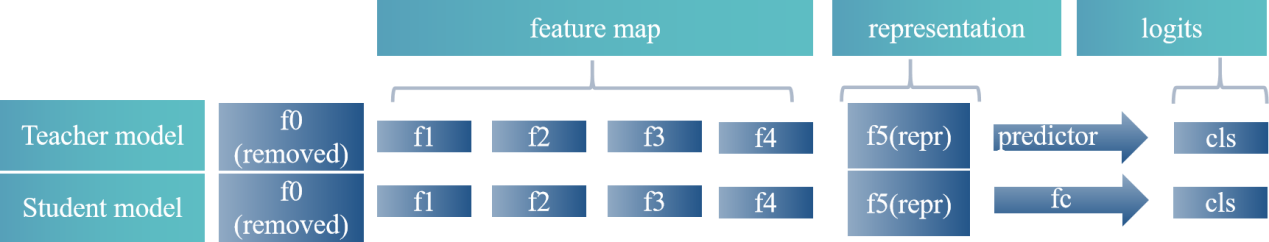
在本研究涉及的实验中，由于研究面向图像表示模型，故训练得到的教师模型和学生模型中，更受关注的是全连接层之前的模型参数，并采取线性评估方式对全连接层重新训练，得到模型在分类任务上的准确率。

在以对比学习SimSiam[5]方法训练教师模型和以本文设定的知识蒸馏算法训练学生模型时，无法在训练过程中进行线性评估得到每一轮次训练之后的模型准确率，而只能得到训练过程中损失值的变化曲线来观察训练过程，并在训练完成后以线性评估方式评价模型。

## 4.2 面向图像表示模型的跨模型知识蒸馏

### 4.2.1 模型的选择

在模型的选择上，需要考虑两个问题，一是教师模型应当是图像表示模型，二是学生模型应当比教师模型具有更少的参数量。本实验选取ResNet50作为教师模型，ResNet18作为学生模型。

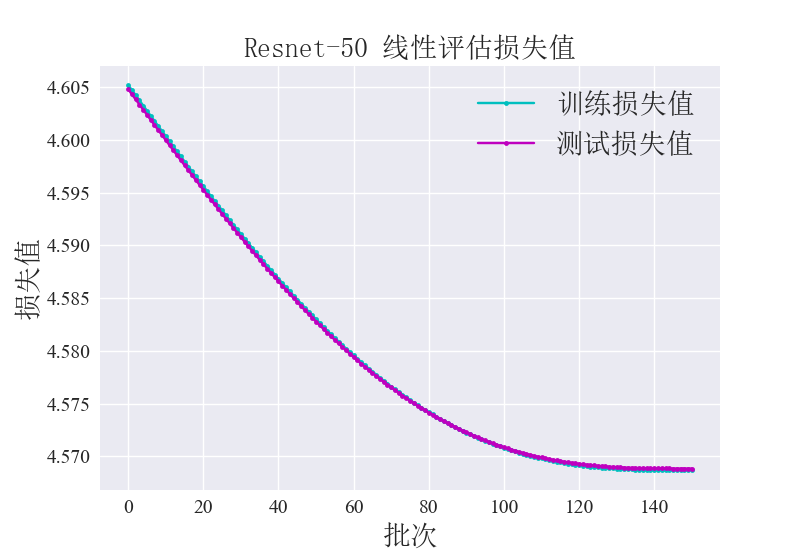
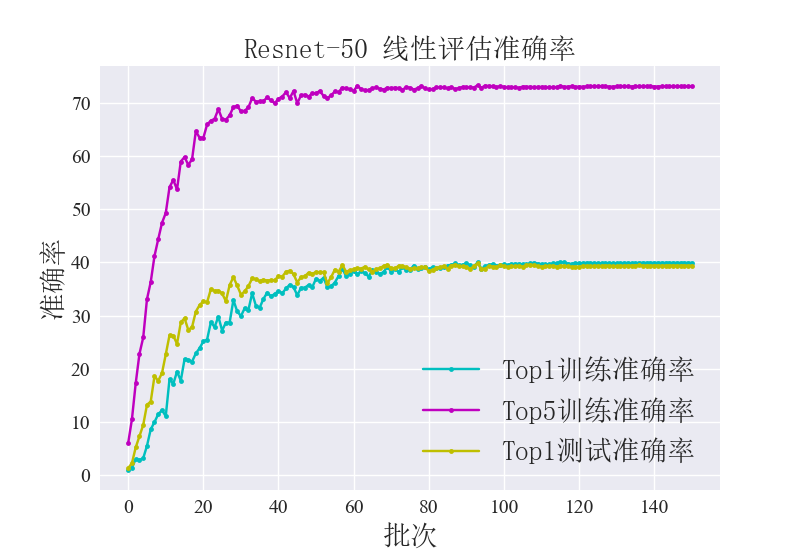


**图4-1教师模型和学生模型的选择**

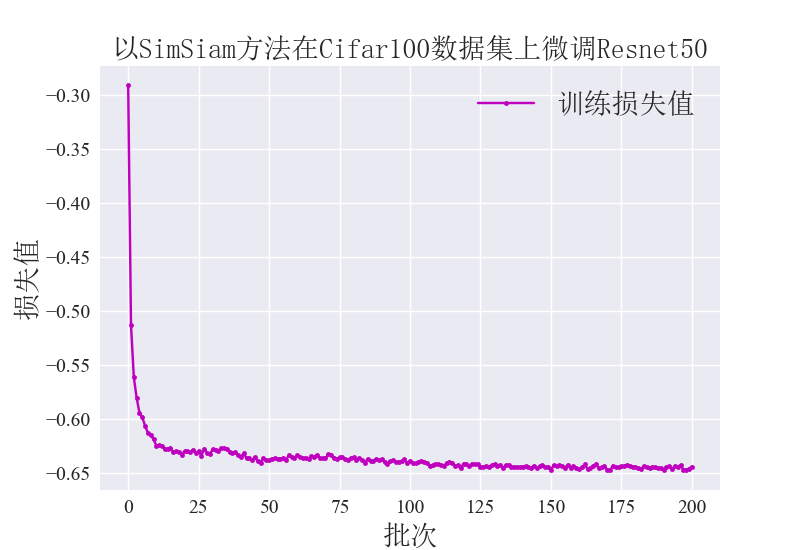
对于教师模型Resnet50，我们采用对比学习方法SimSiam[5]的训练得到的编码器。在模型参数的选择上，我们考虑了以下两种选择：（1）原论文中提供的在Image Net数据集上预训练的ResNet50模型，但该预训练模型文件中并不包含模型训练过程中的predictor结构的参数，并且由于本研究在评价模型时以cifar100数据集进行评价，我们进行了对比实验来观察该预训练模型以SimSiam方法[5]在cifar100数据集上进行迁移学习前后以线性评估方式得到的准确率的变化。（2）直接以对比学习方法SimSiam[5]在cifar100数据集上训练一个resnet50，并保留predictor结构的参数信息。

对于第一种选择，我们做了一组对比实验，对照组为直接将SimSiam论文提供的预训练模型文件在cifar100数据集上进行线性评估，实验组为先将预训练的ResNet50以SimSiam方法在cifar100上进行迁移学习训练100个轮次，再将模型在cifar100数据集上进行线性评估。

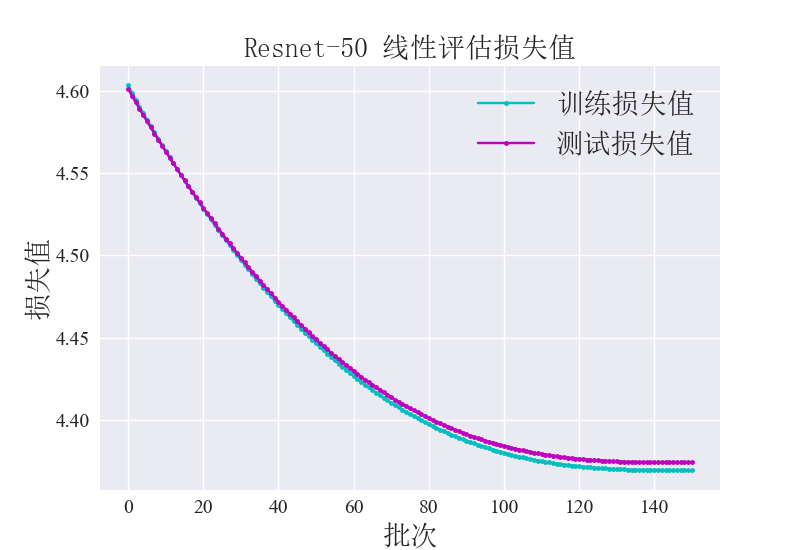
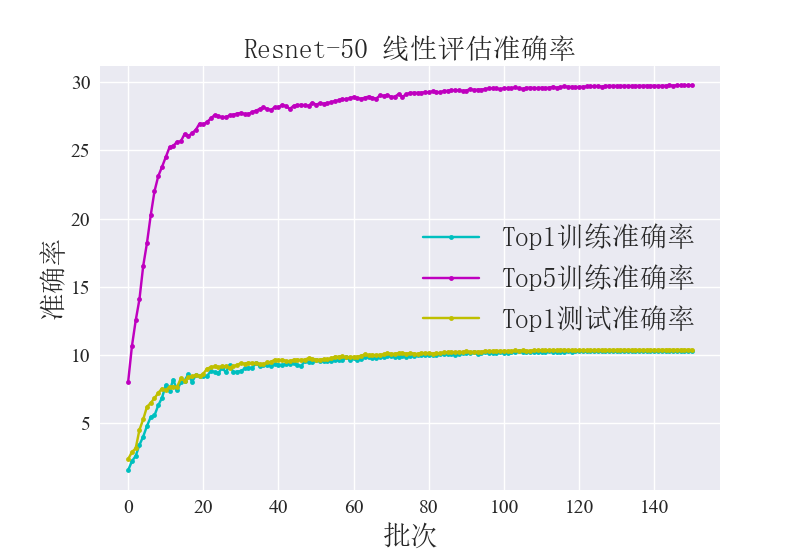
得到实验结果如图4-2所示，SimSiam[5]论文中提供预训练我们发现当预训练模型直接在cifar100数据集上进行线性评估时，能够达到40.01%的top1准确率，但是当经过迁移学习后，将模型在cifar100数据集上进行线性评估时，模型的top1准确率却只有17.67%，这表明模型性能反而下降了，这可能由多个原因导致，经分析，可能原因有如下几个：（1）进行迁移学习的批次大小，在cifar100数据集进行迁移学习时，批次大小仅设置了128，相比于在Image Net上以512的批次大小进行训练必然有所差距。（2）训练的轮次不够，模型参数尚未完全调整到适合cifar100数据集的程度，观察迁移学习过程中的loss曲线可以发现，在最初的epoch中，损失值曲线经历了剧烈的波动阶段，之后开始较平稳地下降，波动阶段可以视为模型从Image Net数据集迁移到cifar100数据集的适应过程，而后续的平稳下降阶段受epoch限制尚未达到收敛的地步。



**图4-2 预训练模型Resnet50在cifar100上迁移学习前线性评估结果**

****

**图4-3 图左和图右分别为为不冻结和冻结Resnet50编码器进行微调的结果**

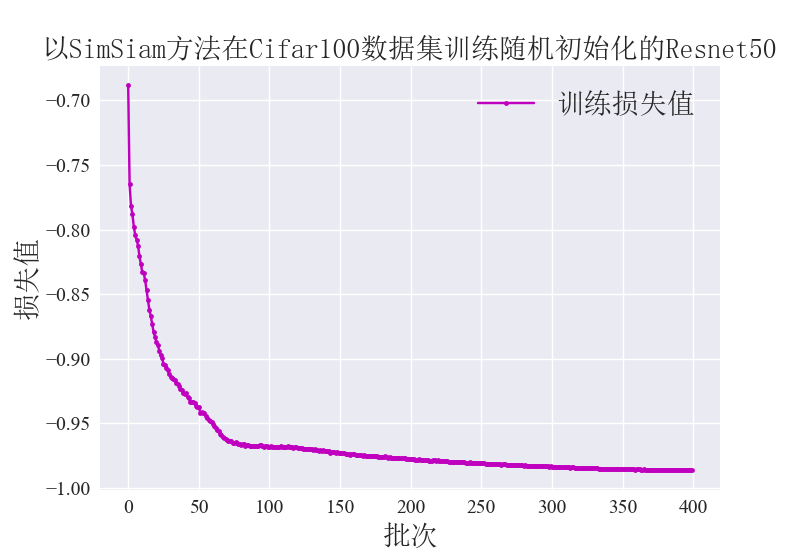
****

**图4-4 不冻结Resnet50进行迁移学习后Resnet50的线性评估结果**

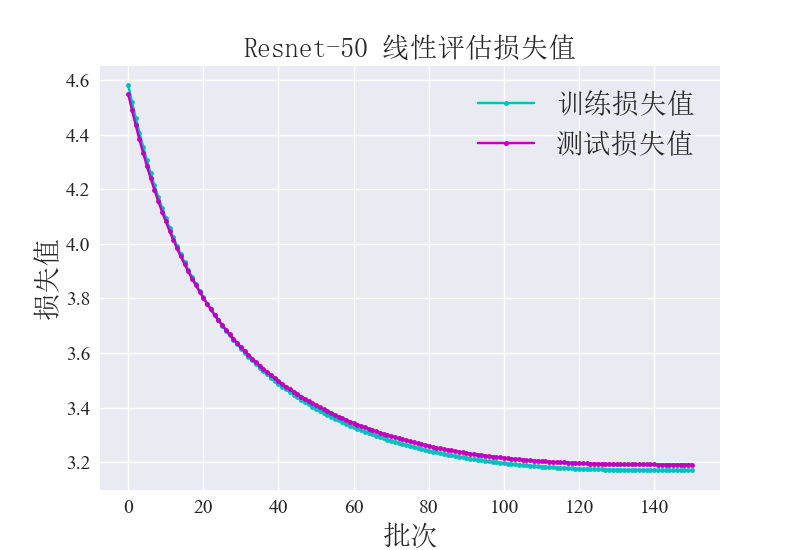
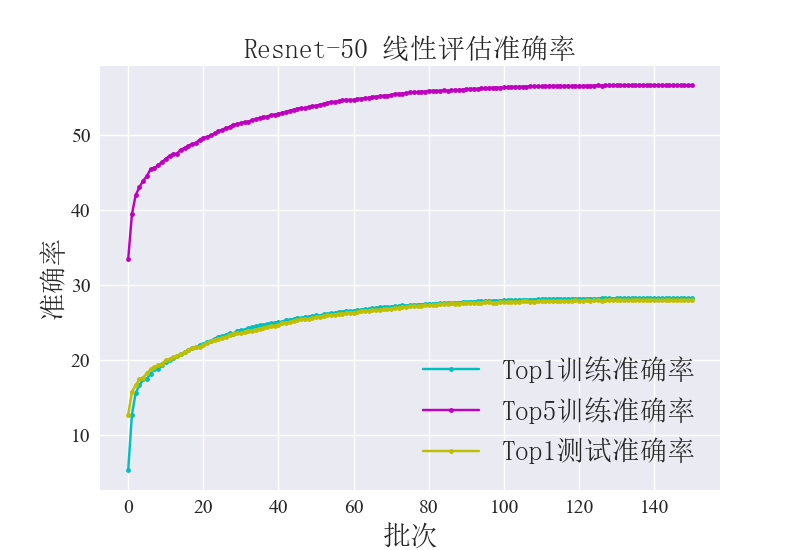
根据上述实验结果的分析我们得到结论：在进行迁移学习实验时，当批次大小和训练轮次受限的情况下，SimSiam[5]论文中提供的Resnet50预训练模型在cifar100数据集上进行迁移学习反而会使得模型性能降低，这是因为在进行迁移学习的过程中，编码器的参数是经过较充分的训练后得到的，而SimSiam方法[5]使用的架构中的predictor结构的参数是随机初始化的，这使得二者出现不匹配的情况，导致在进行迁移学习时出现了混乱的现象，观察图4-3左1图中的损失曲线可以发现，损失值出现了剧烈的波动，这正是上述原因所导致的结果。

在此基础上，我们考虑以下两种选择：一是将编码器Resnet50冻结再进行迁移学习，其目的在于训练predictor结构的参数，这里predictor的输出维度设置为100，使得我们可以将性能较好的预训练模型作为教师模型，在predictor训练充分的前提下，这是我们能够得到的性能最好的教师模型，但很可能出现predictor训练不充分导致类别预测信息质量较低，而用于计算KL散度的类别预测信息的质量是非常关键的；二是故我们考虑直接使用SimSiam方法[5]训练一个随机初始化的Resnet50，同步训练编码器和predictor结构，作为知识蒸馏实验中的教师模型。

采取上述第一种方法进行迁移学习得到的损失曲线如图4-3右1所示，训练实验设置中采取批次大小为128，训练200个轮次，实验结果表明损失值稳定下降并趋于收敛。



**图4-5 训练随机初始化的Resnet50的损失曲线**

****

**图4-6 以SimSiam方法训练随机初始化的Resnet50得到模型的线性评估结果**

**表4-1备选的教师模型线性评估结果对比**

|  |  |  |  |
| --- | --- | --- | --- |
|  | top1训练准确率 | Top5训练准确率 | top1测试准确率 |
| Resnet50(迁移学习前) | 40.10% | 73.99% | 40.01% |
| Resnet50(迁移学习后) | 10.27% | 29.79% | 10.36% |
| Resnet50(2048) | 34.44% | 66.15% | 33.93% |
| Resnet50(100) | 28.21% | 56.67% | 27.93% |

表4-1中列举了备选具体的教师模型的的线性评估的结果，最终我们选用以SimSiam[5]方式在cifar100数据集上训练的随机初始化的Resnet50模型，该模型线性评估得到的准确率约为27.93%，相比批次大小为512在Image Net上训练的模型约40.01%的准确率有12.08%的差距，但在4.3.1小节中的消融研究中表明，我们需要保留SimSiam方法[5]中的predictor结构，以获取类别预测信息。

我们分析了以SimSiam方法[5]在cifar100数据集上训练的随机初始化的Resnet50模型的性能低于SimSiam原论文提供的在Image Net数据集上训练的预训练模型Resnet50的性能的原因，其可能原因有以下几点：（1）实验中批次大小仅设置为128，限制了模型的能力；（2）由于在后续实验中引入了类别预测信息用于知识蒸馏，故我们将SimSiam方法[5]中的predictor结构中的输出维度从2048改为了100，以匹配cifar100数据集的类别数，表4-1中的数据表明，将predictor结构的输出维度从2048降低到100确实使训练得到的编码器的性能降低了。

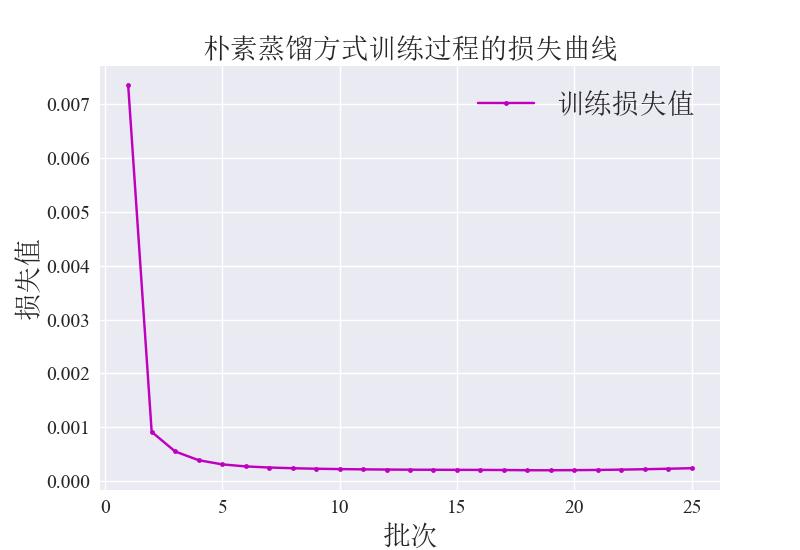
### 4.2.2 知识蒸馏实验

本实验通过两种方式（即优化蒸馏和朴素蒸馏）对图像表示模型进行知识蒸馏，朴素蒸馏方式直接将教师模型和学生模型的特征图与图像表示一一对应并采用MSE作为损失函数，在维度匹配上采用了下采样和高维向量嵌入低维的方式，将此蒸馏方法得到的学生模型线性评估得到的准确率作为本算法研究中知识蒸馏性能的基准。优化蒸馏结合了特征图信息、图像表示信息、类别预测信息以3.1小节中的算法进行蒸馏，并调节三类信息构成的损失函数所占权重。

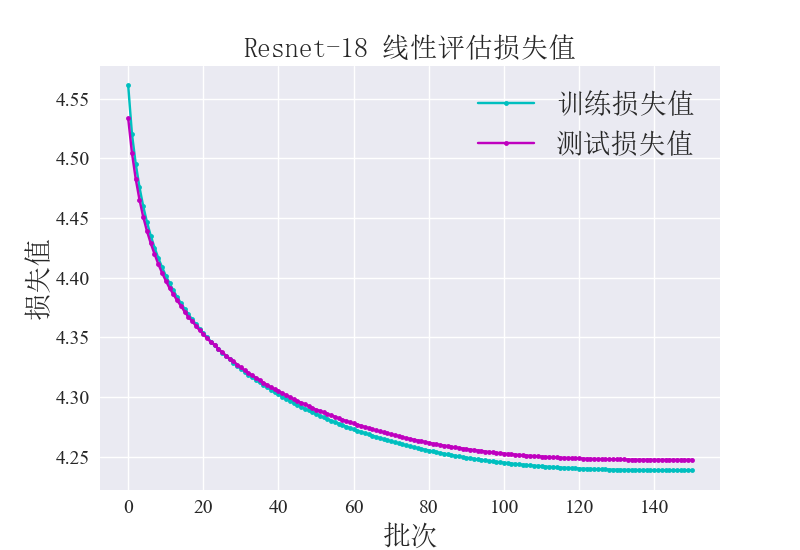
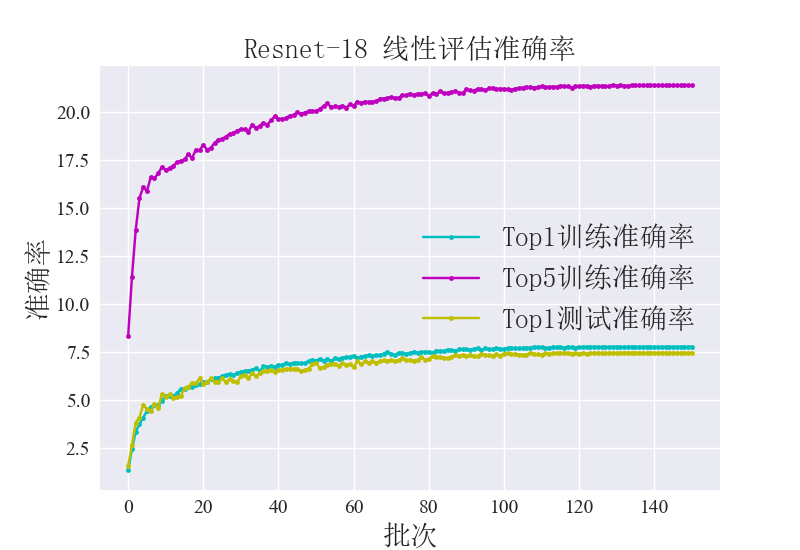
**表4-2知识蒸馏所得学生模型线性评估结果对比**

|  |  |  |  |
| --- | --- | --- | --- |
|  | top1训练准确率 | top5训练准确率 | top1测试准确率 |
| Resnet18(朴素蒸馏) | 7.72% | 21.39% | 7.45% |
| Resnet18(优化蒸馏) | 32.95% | 59.80% | 30.56% |
| Resnet50(教师模型) | 28.21% | 56.67% | 27.93% |

在具体的实验配置中，采用以SimSiam方法[5]训练的Resnet50及predictor结构作为教师模型，其线性评估准确率为27.93%。基于该教师模型，我们使用朴素蒸馏与优化蒸馏两种方法进行了对比实验。

****

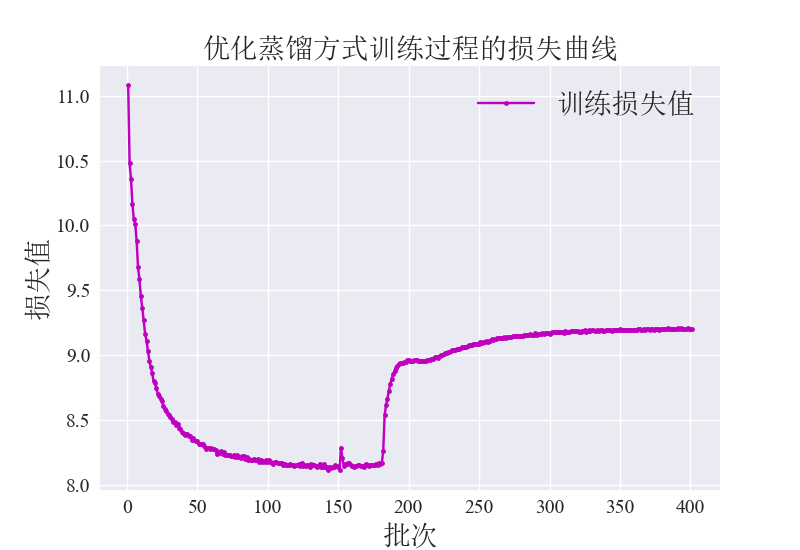
**图4-7 朴素蒸馏过程中的损失曲线**



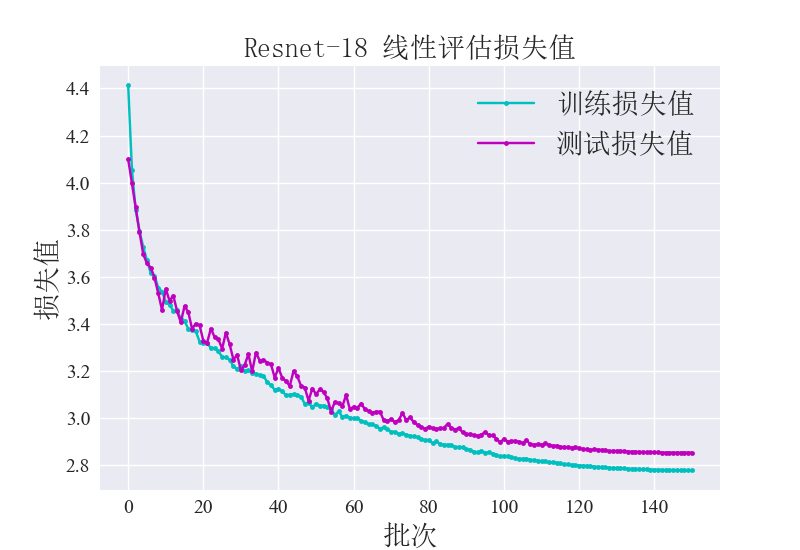
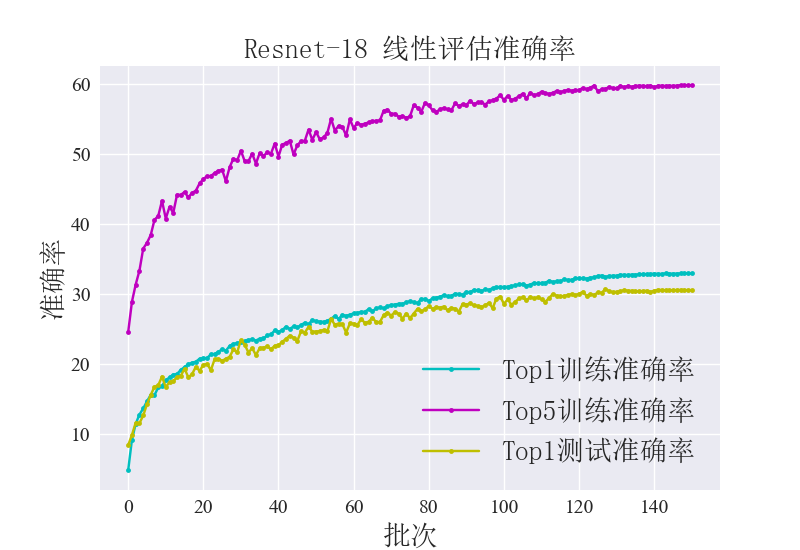
**图4-8 朴素蒸馏得到学生模型Resnet18的线性评估结果**

实验结果表明，采用朴素蒸馏方式得到学生模型的Top1准确率仅为7.45%，相较于教师模型的27.93%的Top1准确率差距很大。

实验设置批次大小为128时，进行优化蒸馏得到实验结果如下图所示：

****

**图4-9 优化蒸馏训练过程中的损失曲线**

****

**图4-10 优化蒸馏得到学生模型Resnet18的线性评估结果**

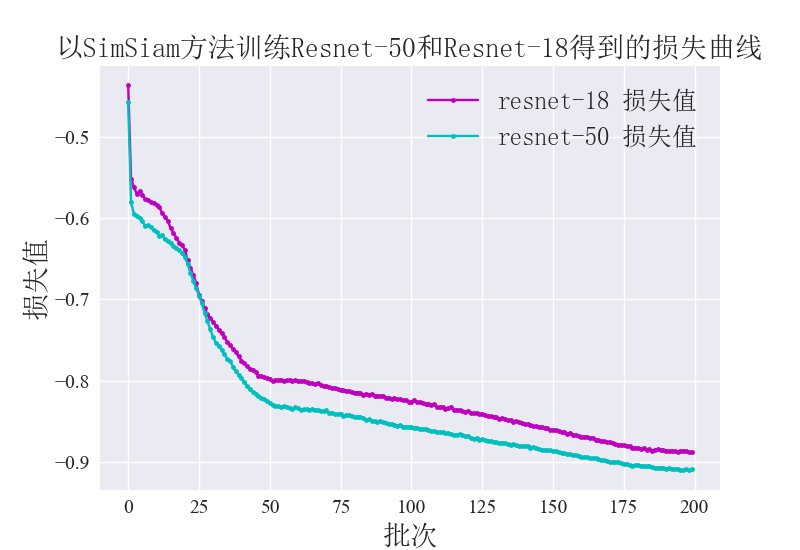
图4-9中引入后优化蒸馏过程中的损失曲线在训练的前180个epoch平稳降低，但之后损失值曲线出现了跳跃式的突变并逐渐收敛于一个稳定的数值，其突变的原因为在第150、180、210个轮次时对学习率进行了调整，之后从图4-10中对学生模型进行线性评估的结果来看，学生模型的Top1准确率出乎意料地达到了30.56%，略高于教师模型27.93%的Top1准确率，推断其原因有如下可能：（1）教师模型的准确率较低，这导致知识蒸馏算法设计合理的情况下，学生模型的表现超过了教师模型；（2）教师模型在进行线性评估时仅以主干网络编码器Resnet50进行线性评估，但是在进行优化蒸馏时学生模型额外学习到了predictor结构中的信息，这使得学生模型中的图像表示编码器（全连接层之前的结构）性能更优秀。但该结果已经能够说明该知识蒸馏算法是有效的。

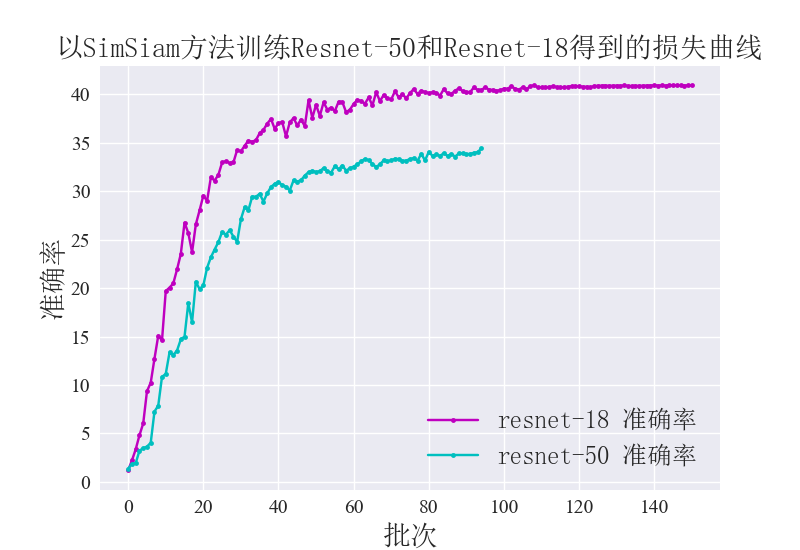
### 4.2.3 学生模型的性能对比研究

在上一小节中我们通过知识蒸馏方式得到了学生模型，并进行了性能评估，在这一小节中，我们尝试研究以下两种方式得到的学生模型性能的差别：（1）以SimSiam方法[5]直接训练一个ResNet18编码器；（2）通过知识蒸馏算法得到的学生模型Resnet18。

**表4-3 SimSiam方法训练resnet50和resnet18线性评估性能对比**

|  |  |  |  |
| --- | --- | --- | --- |
|  | Top1训练准确率 | Top5训练准确率 | Top1测试准确率 |
| Resnet18(2048) | 40.93% | 72.70% | 40.95% |
| Resnet50(2048) | 34.44% | 66.15% | 33.93% |



**图4-11 Resnet50和Resnet18通过SimSiam方式训练得到的损失曲线**

**图4-12 Resnet50和Resnet18线性评估结果**

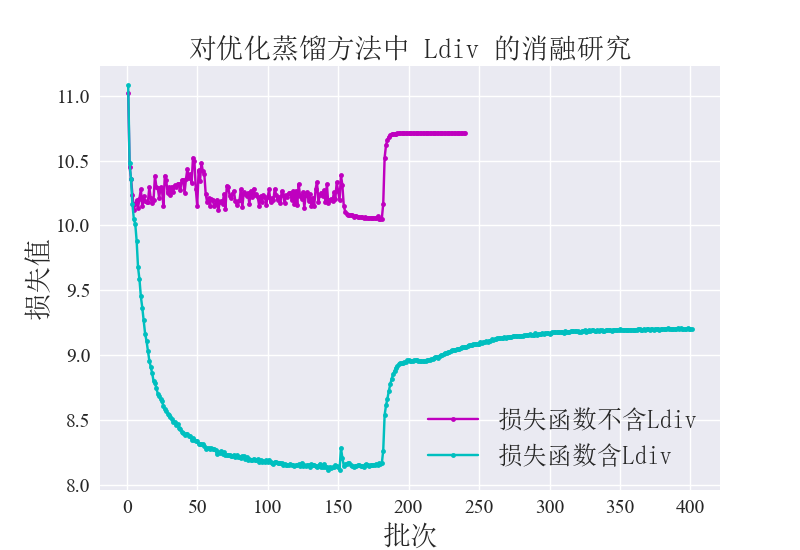
实验设置中批次大小设置为128，都训练200个轮次，最终对编码器进行线性评估的结果表明Resnet18的性能反而更优秀，分析其原因如下可能：两个模型训练时损失值曲线的变化基本一致，并且观察损失值曲线的变化趋势可以判断出在训练200个轮次的情况下模型尚未收敛，而Resnet18相比Resnet50具有更少的参数量，故在相同实验设置的情况下参数量更少的模型或许训练得更充分，因而具有更好的性能。

## 4.3 消融研究

### 4.3.1 类别预测信息的消融研究

在4.2.2小节的知识蒸馏实验中，优化蒸馏的实验结果是引入了基于类别预测信息的后得到的，在本小节中我们对的重要性单独进行研究。

在知识蒸馏过程中，我们默认特征图信息和图像表示信息是必要的，图4-13和表4-4展示引入和不引入的两种实验设置下优化蒸馏的损失值曲线与线性评估结果，观察4-13可知，引入后损失曲线下降幅度更大并更平滑，线性评估结果也表明的引入对知识蒸馏的效果是关键的。

****

**图4-13 优化蒸馏训练过程中的损失曲线**

**表4-4 优化蒸馏中****的消融研究**

|  |  |  |  |
| --- | --- | --- | --- |
|  | Top1训练准确率 | Top5训练准确率 | Top1测试准确率 |
| Resnet18(引入) | 32.95% | 59.80% | 30.56% |
| Resnet18(不引入) | 11.62% | 28.34% | 12.32% |
| Resnet18(朴素蒸馏) | 7.72% | 21.39% | 7.45% |
| Resnet50(教师模型) | 28.21% | 56.67% | 27.93% |

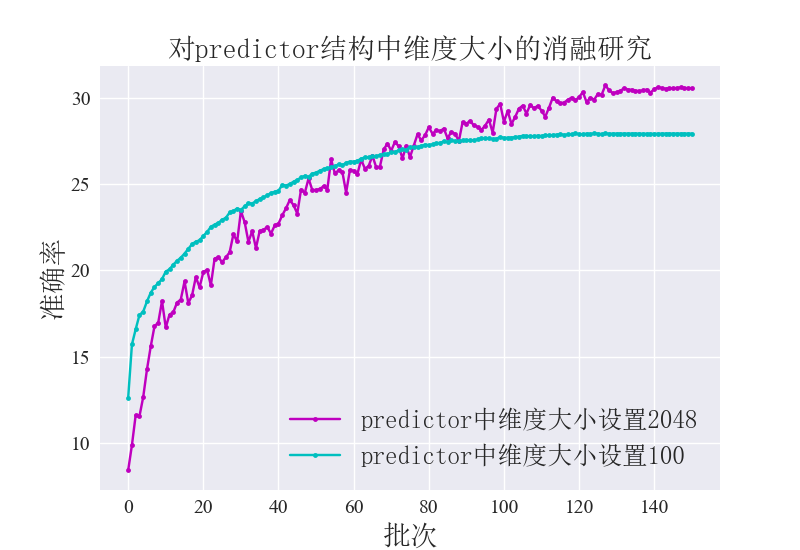
图4-13中绿色曲线是不引入的情况下仅利用优化蒸馏得到的损失值曲线，从损失值曲线的变化趋势可以看到，训练了240个批次但损失值在此期间并没有长期平稳下降的过程，只是不断地波动并最终稳定在一个数值，模型似乎达到了收敛状态，但在对学生模型进行线性评估时，却发现学生模型的top1准确率只有1%，该结果表明学生模型没有学到任何有效信息，引入是必要的。

### 4.3.2 SimSiam方法中predictor结构维度变化研究

在知识蒸馏实验中，为了能够利用类别预测信息，我们实际上修改了predictor的输出维度（从2048修改为100），实验结果表明该改动会导致模型性能的降低，下图4-4中的实验结果说明了该问题，当predictor的输出维度减少时，模型在线性评估时收敛地更快，性能上限降低了。

**表4-5 SimSiam方法中predictor结构输出维度的消融研究**

|  |  |  |  |
| --- | --- | --- | --- |
|  | Top1训练准确率 | Top5训练准确率 | Top1测试准确率 |
| Resnet50(2048) | 34.44% | 66.15% | 33.93% |
| Resnet50(100) | 28.21% | 56.67% | 27.93% |



**图4-14 predictor结构中设置不同维度的模型性能对比**

# 5 结论

本文通过对知识蒸馏和对比学习相关领域的调研，阅读相关文献对已有研究进行总结分析，并提出了一种面向图像表示模型的知识蒸馏算法。该算法以图像表示模型为目标，通过该蒸馏算法可以得到性能保持良好并具有更少参数量的轻量模型。在实验阶段，本文主要以对比学习SimSiam方法训练得到的图像表示模型为蒸馏对象，利用中间层特征图、图像表示向量、类别预测向量三类信息对教师模型进行知识蒸馏，并通过消融实验证明了类别预测信息的重要性以及中间层特征图与图像表示向量这两类信息的有效性。总体而言，本文为面向图像表示模型的知识蒸馏提供了一种可选择的方案。

# 6 参考文献

1. Ahn S, Hu S X, Damianou A, et al. Variational information distillation for knowledge transfer[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2019: 9163-9171.
2. Bengio Y, Courville A, Vincent P. Representation learning: A review and new perspectives[J]. IEEE transactions on pattern analysis and machine intelligence, 2013, 35(8): 1798-1828.
3. Chen T, Zhai X, Ritter M, et al. Self-supervised gans via auxiliary rotation loss[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2019: 12154-12163.
4. Chen T, Kornblith S, Norouzi M, et al. A simple framework for contrastive learning of visual representations[C]//International conference on machine learning. PMLR, 2020: 1597-1607.
5. Chen X, He K. Exploring simple siamese representation learning[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2021: 15750-15758.
6. Chen D,Mei J P, Zhang Y, et al. Cross-layer distillation with semantic calibration[C]//Proceedings of the AAAI Conference on Artificial Intelligence. 2021, 35(8): 7028-7036.
7. Devlin J, Chang M W, Lee K, et al. Bert: Pre-training of deep bidirectional transformers for language understanding[J]. arXiv preprint arXiv:1810.04805, 2018.
8. Doersch C, Gupta A, Efros A A. Unsupervised visual representation learning by context prediction[C]//Proceedings of the IEEE international conference on computer vision. 2015: 1422-1430.
9. Floridi L, Chiriatti M. GPT-3: Its nature, scope, limits, and consequences[J]. Minds and Machines, 2020, 30: 681-694.
10. Gidaris S, Singh P, Komodakis N. Unsupervised representation learning by predicting image rotations[J]. arXiv preprint arXiv:1803.07728, 2018.
11. Grill J B, Strub F, Altché F, et al. Bootstrap your own latent-a new approach to self-supervised learning[J]. Advances in neural information processing systems, 2020, 33: 21271-21284.
12. Hinton G, Vinyals O, Dean J. Distilling the knowledge in a neural network[J]. arXiv preprint arXiv:1503.02531, 2015.
13. Hadsell, R., Chopra, S., and LeCun, Y. Dimensionality reduction by learning an invariant mapping. In 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR’06), volume 2, pp. 1735–1742. IEEE, 2006.
14. He K, Fan H, Wu Y, et al. Momentum contrast for unsupervised visual representation learning[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2020: 9729-9738.
15. 23He K, Zhang X, Ren S, et al. Deep residual learning for image recognition[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2016: 770-778.
16. Noroozi M, Favaro P. Unsupervised learning of visual representations by solving jigsaw puzzles[C]//Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part VI. Cham: Springer International Publishing, 2016: 69-84.
17. Romero A, Ballas N, Kahou S E, et al. Fitnets: Hints for thin deep nets[J]. arXiv preprint arXiv:1412.6550, 2014.
18. Oord A, Li Y, Vinyals O. Representation learning with contrastive predictive coding[J]. arXiv preprint arXiv:1807.03748, 2018.
19. Tung F, Mori G. Similarity-preserving knowledge distillation[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. 2019: 1365-1374.
20. Tian Y, Krishnan D, Isola P. Contrastive representation distillation[J]. arXiv preprint arXiv:1910.10699, 2019.
21. Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need[J]. Advances in neural information processing systems, 2017, 30.
22. Yang X, Daquan Z, Liu S, et al. Deep model reassembly[J]. arXiv preprint arXiv:2210.17409, 2022.
23. Zhang R, Isola P, Efros A A. Colorful image colorization[C]//Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part III 14. Springer International Publishing, 2016: 649-666.
24. Zagoruyko S, Komodakis N. Paying more attention to attention: Improving the performance of convolutional neural networks via attention transfer[J]. arXiv preprint arXiv:1612.03928, 2016.
25. Zagoruyko S, Komodakis N. Wide residual networks[J]. arXiv preprint arXiv:1605.07146, 2016.

# 7 致 谢

首先，在此对我的母校浙江工业大学表示感谢，母校四年来对我的栽培为我打下了坚实的专业基础，拓展了我的知识视野，再者要感谢我的本科毕业设计指导教师梅建萍老师，梅老师为我给予了我细致入微的指导和建议。其次，我还要感谢我的家人和朋友们，他们一直支持和鼓励我，让我能够专注于我的学业和研究。我将承载着期盼，坚持终身学习的理念，不忘初心，砥砺前行，回报社会。