Modeling and Simulation 建模与仿真, 2024, 13(2), 1455-1465
Published Online March 2024 in Hans. https://www.hanspub.org/journal/mos
https://doi.org/10.12677/mos.2024.132137

Hans汉斯

一种阶段重置的知识蒸馏方法研究与仿真

陈骏立*，孙占全

上海理工大学光电信息与计算机工程学院，上海

收稿日期：2024年1月14日；录用日期：2024年3月15日；发布日期：2024年3月22日

摘　要

知识蒸馏是一种将知识从教师网络传递到学生网络的模型压缩方法。目前的知识蒸馏方法存在教师网络和学生网络之间的语义信息不一致的问题，具体而言，师生模型之间的前向推理距离不一致导致语义信息不一致，最终损耗蒸馏性能。为了解决这个问题，本文探索一种新的阶段重置知识蒸馏方法。该方法设计了以阶段为单位的知识蒸馏，师生网络相同阶段共享输出，降低了由学生与教师推理路径长度差异过大造成的特征语义不匹配的影响，从而提升学生网络的性能。最后，本文用提出的方法在公共数据集上进行仿真实验，并与最新的方法进行比较，实验结果表明本文提出的方法更具优势。

关键词

神经网络，分类模型，模型压缩，知识蒸馏，阶段重置

Design and Simulation of Stage Reset Knowledge Distillation Method

Junli Chen*, Zhanquan Sun

School of Optical-Electrical and Computer Engineering, University of Shanghai for Science and Technology, Shanghai

Received: Jan. 14th, 2024; accepted: Mar. 15th, 2024; published: Mar. 22nd, 2024

Abstract

Knowledge distillation is a compression technique used to transfer knowledge from a teacher network to a student network. However, the current knowledge distillation methods suffer from an issue of inconsistent semantic information between the teacher and student networks. This inconsistency arises due to variations in forward reasoning distance between the teacher-student model, resulting in a loss of distillation performance. To address this problem, this study introduces

*通讯作者。

文章引用: 陈骏立, 孙占全. 一种阶段重置的知识蒸馏方法研究与仿真[J]. 建模与仿真, 2024, 13(2): 1455-1465.
DOI: 10.12677/mos.2024.132137

a novel approach called "stage reset knowledge distillation." This method incorporates stage-based knowledge distillation, where the output is shared within the same stage of the teacher-student network, which reduced the influence of feature semantic mismatch caused by the large difference in reasoning path length between students and teachers, thus enhancing the performance of the student network. Experimental evaluations on a public dataset are conducted to validate the proposed method's efficacy. Comparative analysis against state-of-the-art techniques demonstrates the superior advantages offered by the proposed method.

Keywords

Neural Network, Classification Model, Model Compression, Knowledge Distillation, Stage Reset

1. 引言

随着人工智能技术迅速发展，卷积神经网络(CNNs)在图像分类[1]、目标检测[2]和语义分割[3]等广泛的计算机视觉应用中取得了前所未有的进展。目前这些性能最好的神经网络通常深度大、参数大、复杂度高。随着对资源受限设备的实时响应需求不断增加，越来越复杂的神经网络已经难以适应计算受限的设备上的应用，如移动设备和嵌入式系统。因此迫切需要新的解决方案，在不降低神经网络良好性能的情况下，降低模型的复杂性。针对这个问题，目前已经有了不少训练紧凑神经网络的技术，包括设计新的架构[4]，网络修剪[5]，量化[6]和知识蒸馏[7]。在这些方法中，知识蒸馏已经被证明是一种非常有效的模型压缩方法。

知识蒸馏的主要思想是将知识从大模型(教师模型)转移到小模型(学生模型)，让学生网络的性能接近教师网络的性能，用小模型来代替大模型，从而实现模型的压缩。在知识蒸馏方法中，首先对强大的教师网络进行预训练，将教师模型输出作为学生网络学习的监督信号，让学生网络的输出与教师网络输出相似。除了基于输出的知识蒸馏外，近年来很多研究[8] [9] [10]从特征层中提炼和转移知识，让学生的特征及特征变换与教师相似。

然而一个重要的问题往往被忽略从而限制了学生性能进一步提高。在卷积网络中卷积层所学到的知识是分层的，更深的中间特征层所对应的知识更抽象。学生网络和教师网络由于容量之间存在差距，导致学生网络的特征表达能力往往不如教师网络。师生之间的能力差距阻碍了学生模仿老师的确切特征。Mirzadeh 等人[11]发现一个参数更多、精度更高的教师比一个参数更少的教师教出来的学生更差。师生网络模型容量差距过大时，基于特征的知识蒸馏在传递信息时会出现语义信息不匹配问题，学生网络很难从教师网络中学习有效知识，导致学生网络模型出现负优化。有一些工作尝试解决这个问题，文章[12]提出通过师生特征的注意力相似程度，匹配师生知识传递路径，改变了传统知识蒸馏的手工设定知识传递路径。文章[13]提出计算师生层语义信息，从而绑定师生的知识传递路径。文章[14]指出教师过深的特征不适合同一阶段学生学习，提出利用教师浅层特征指导学生的深层特征。这些研究工作主要通过改变和匹配知识传递路径来实现师生语义信息匹配。而如何减少学生和学生之间语义信息差异却很少被研究。对于同构网络，师生网络模型的差异主要是网络深度不一致。较浅的学生推理过程相对于教师更加简单，在知识蒸馏过程中，学生网络无法总是完全拟合教师更复杂的特征。

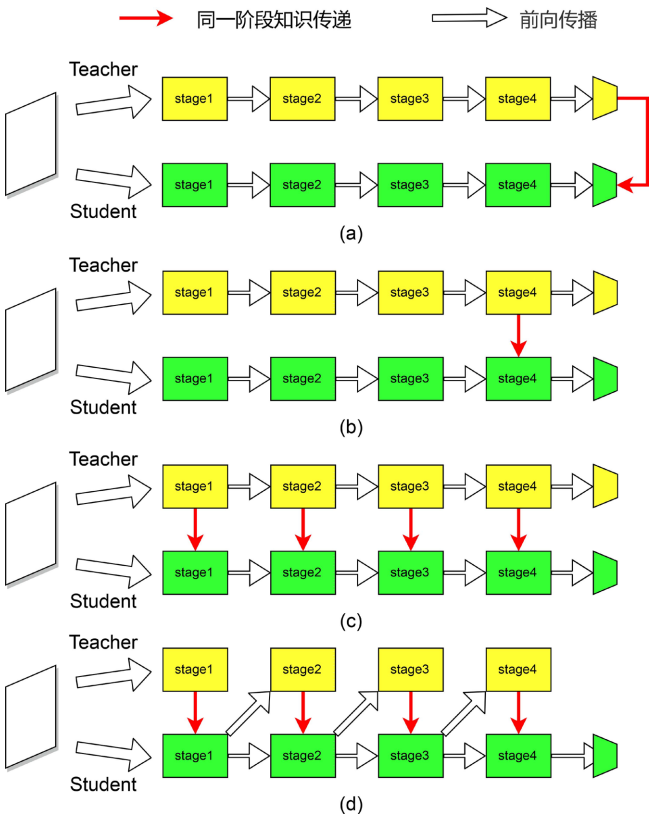


Figure 1. (a)~(c) Previous knowledge distillation frameworks. They feed instances from the input side, transferring knowledge at the same stage. (d) Our proposed staged distillation gives the same input to the student and teacher at the same stage

图 1. (a)~(c) 以前的知识蒸馏框架，从输入端输入实例，在同一阶段传递知识。(d) 我们提出的阶段重置蒸馏法在同一阶段为学生和教师提供相同的输入

为解决上述问题，本文提出一种新的蒸馏方法减少师生语义差距，让学生网络容易学习教师网络包含的信息。为了能够理解我们的想法，我们首先展示前人是如何处理这些知识转移的路径。如图 1 所示，(a)~(c)表示之前的蒸馏方法，它们将数据同时输入到学生和教师网络，在学生和教师相同的阶段进行知识蒸馏。例如(b)总是使用第四阶段的信息指导学生。这个过程看起来直观，但有趣的是，在师生模型容量差距巨大时，最后一个阶段的学生向老师学习是困难的。对于(c)的多阶段蒸馏，在早期的阶段学生并不能够跟上教师的节奏，而紧接进入下一个阶段的推理，当到达最后一个阶段时，学生已经难以模仿教师。为解决师生容量差距导致蒸馏效果不佳的问题，本文提出阶段重置蒸馏法 SRKD (Stage Reset Knowledge Distillation)，如图(d)所示。本文提出的方法以阶段为单位向教师学习，同一阶段的教师和学生共享相同的输入的。该方法设计的巧妙在于每个阶段的输入都被重置，对于同一阶段的教师和学生的输入保持一致，间接减少学生和教师推理长度差距。实验结果表明该方法在图像分类方面优于其他比较方法。

2. 阶段重置模型设计

如图 2 所示，是本文提出的阶段重置蒸馏的总体框架，教师网络是一个参数较大、并且经过预训练网络。在知识蒸馏阶段，只有学生网络参与训练和测试。

假设有一个教师模型和学生模型，分别用 f^T 和 f^S 表示。模型是在训练数据集 $D=\{(x_1,y_1),(x_2,y_2),\cdots,(x_N,y_N)\}$ 上进行优化的，其中 N 是训练样本的总数。真实的标签会监督学生模型，计算预测值与真实值标签之间的距离。一般交叉熵损失函数计算如下：

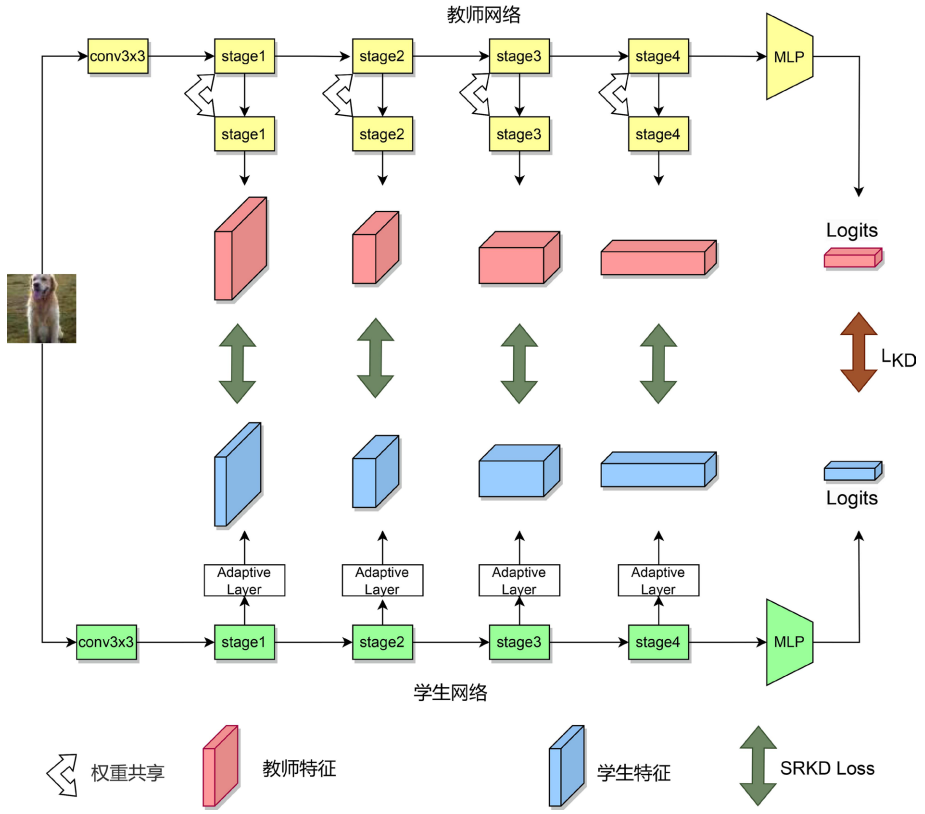


Figure 2. Framework of stage reset knowledge distillation
图 2. 阶段重置蒸馏框架图

$$L_{CE}=\frac{1}{N}\sum_{i=1}^{N}-\left\{y_i\cdot\log\left[\sigma\left(f^S\left(x_i\right)\right)\right]+\left(1-y_i\right)\cdot\log\left[1-\sigma\left(f^S\left(x_i\right)\right)\right]\right\} \tag{1}$$

其中 $f^S\left(x_i\right)$ 是输入 x_i 实例经过模型的 logit (softmax 之前)输出，$\sigma(\cdot)$ 为 softmax 函数。为了让学生网络的输出与教师网络的输出更加相似，试图减少学生和教师嵌入之间的分歧。采用 Kullback-Leibler (KL)散度最小化它们的距离，定义如下：

$$L_{KD}=\frac{1}{N}\tau^2\sum_{i=1}^{N}D_{KL}\left(\sigma\left(\frac{f^T\left(x_i\right)}{\tau}\right),\sigma\left(\frac{f^S\left(x_i\right)}{\tau}\right)\right) \tag{2}$$

其中 $f^T\left(x_i\right)$，$f^S\left(x_i\right)$ 分别教师网络和教师网络中倒数第二层(softmax 之前)的输出；τ 为温度因子，它被用作与目标软化程度相关的超参数；$\sigma(\cdot)$ 为 softmax 函数。D_{KL} 是度量 Kullback-Leibler 算子，衡量两个输出之间的距离。

如图 2 所示，stage 1 到 stage 4 表示模型的特征提取的各个阶段。假设学生网络为 f^S，学生网络由 $(S_1^S,S_2^S,...,S_n^S,S_c^S)$ 多个不同的阶段组合而成，S_n^S 表示学生模型第 n 阶段；S_c^S 表示学生网络的 MLP 多层感知机。因此学生网络 f^S 可以表示为：

$$f^S = S_C^S \bullet S_N^S \bullet S_{N-1}^S \bullet \cdots \bullet S_1^S \tag{3}$$

其中，将$\bullet$视为嵌套函数$f \bullet g(x) = f(g(x))$。每一个阶段都是一次下采样阶段，一个阶段通常由多个卷积层堆叠组成结构。不同阶段输出的特征空间和维度均不一致。给定一个输入x，通过前向传播可以计算每个阶段输出的特征。学生模型在各个阶段输出特征可以表示为：

$$F_i^s = S_i^S \bullet S_{i-1}^S \bullet \cdots \bullet S_1^S(x) \tag{4}$$

其中S_i^S是第i阶段的模块，F_i^s表示学生模型第i个阶段输出特征。上一个阶段的输出作为下一个阶段的输入。因此，第i个阶段输出特征的计算方式也可表示为：

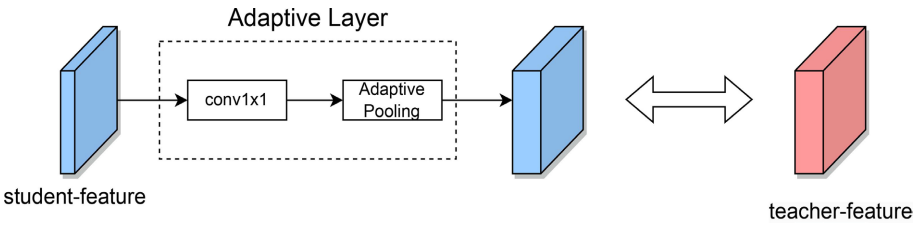


Figure 3. Adaptation layer
图 3. 自适应层

$$F_i^S = S_i^S\left(F_{i-1}^S\right) \tag{5}$$

其中S_i^S表示第i阶段模块；F_{i-1}^S表示$i-1$阶段模块的输出特征。由于教师网络每一个阶段的输入与学生网络同一阶段输入保持一致，因此教师该阶段过程可表示为：

$$F_i^T = S_i^T\left(F_{i-1}^S\right) \tag{6}$$

其中S_i^T为教师的第i阶段的模块，F_{i-1}^S为学生第$i-1$阶段编码器输出的特征，F_i^T为第i阶段通过教师编码器输出的特征。由于学生的一个阶段都需要向同一阶段教师学习。那么对于学生的一个阶段重置蒸馏的损失(Single Stage Reset KD)可表示为：

$$L_{SSRKD} = D\left(M_i^S\left(F_i^S\right), M_i^T\left(F_i^T\right)\right) \tag{7}$$

对于上式，F_i^S和F_i^T分别是学生和教师经过第i阶段的输出；为了确保学生和教师能够比较，在学生每个阶段输出特征之后加上一个自适应层。如图 3 所示的自适应层由 1 × 1 卷积层和自适应池化层组成的特征转化层。即M_i^S和M_i^T分别是学生和教师第i阶段的自适应特征转化层。D为 L2 损失，来最小化师生特征间的差距。本文的阶段重置蒸馏法(Stage Reset KD)在学生的每个阶段应用，因此重置蒸馏损失可以表示成：

$$L_{SRKD} = \sum_{i=1}^{n} D\left(M_i^S\left(F_i^S\right), M_i^T\left(F_i^T\right)\right) \tag{8}$$

至此，学生网络的优化包括三个损失，最终优化目标可以写成：

$$L = L_{CE} + L_{KD} + \lambda L_{SRKD} \tag{9}$$

其中L_{CE}为分类任务的交叉熵损失，L_{KD}为输出的蒸馏损失。L_{SRKD}为阶段重置蒸馏损失。λ为阶段重置蒸馏损失权重因子。

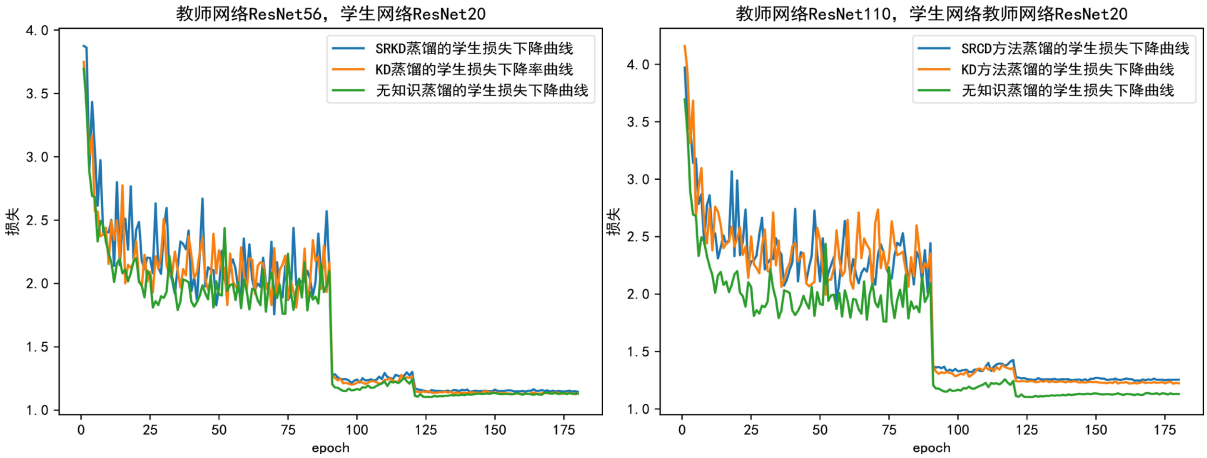
3. 仿真实验与结果分析

3.1. 数据集

本文采用了 3 个公共数据集对本文提出的方法进行仿真和对比实验。

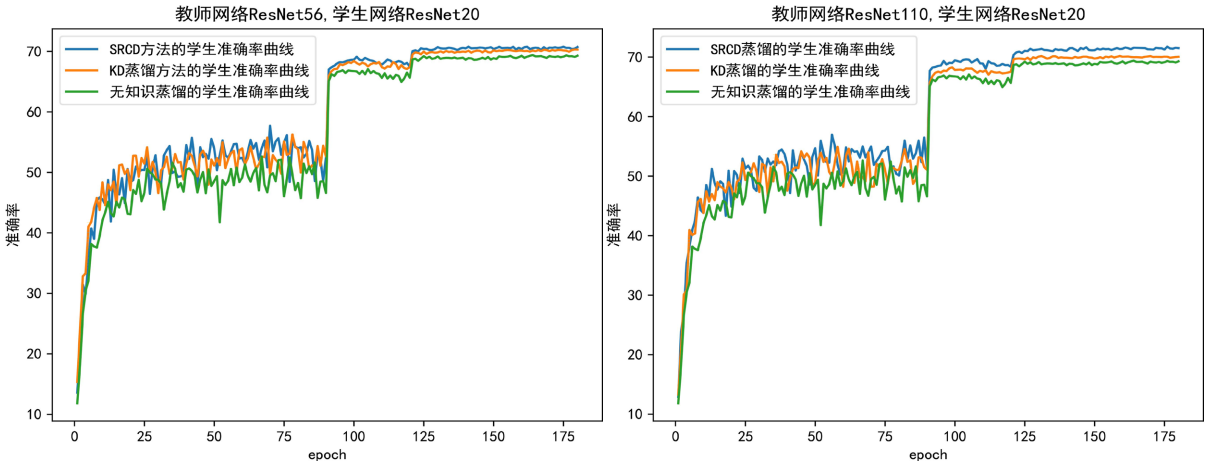
CIFAR-10：包含 50 K 训练图像和 10 K 测试图像，共 10 个类别，每张图像的大小为 32 × 32 × 3。在使用该数据集时，学生网络训练设置训练轮数 epoch 设为 180，批次大小 Batchsize 设为 64。学习率从 0.05 开始，在第 90、120、150 个 epoch 处除以 10。

CIFAR-100：包含 5 万张训练图像和 1 万张测试图像，共 100 个类别，大小与 CIFAR-10 相同。在使用该数据集时，训练设置参数与 CIFAR-10 数据集训练时一致。



(a) (b)

Figure 4. The loss curve and accuracy curve of the students network training stage
图 4. 学生网络训练阶段损失下降曲线



(a) (b)

Figure 5. The accuracy curve of the students network training stage
图 5. 学生网络训练阶段准确率曲线

TinyImageNet：ImageNet 的一个子集，它是一个更有挑战性的数据集，共有 200 个类。它有 10 万张训练图像和 1 万张验证图像。在预处理过程中，通过信道均值和标准差对图像进行归一化处理。在使用

该数据集时，学生网络的训练设置训练轮数 epoch 设为 100，批次大小 Batchsize 设为 64。学习率从 0.05 开始，在第 70、80、90 个 epoch 处除以 10。

3.2. 仿真实验

我们两个师生组进行了两个知识蒸馏方法以及无知识蒸馏方法下学生网络的仿真实验，如图 4、图 5 所示。仿真实验采用 CIFAR-100 作为训练集和验证集，经过预训练的 ResNet110、ResNet56 作为教师模型，ResNet20 作为学生模型。其中 ResNet110 教师网络的经过预训练的准确率为 74.31%，ResNet56 准确率为 72.32%。

其中图 4 是学生网络 ResNet20 在不使用蒸馏方法、使用 KD 蒸馏和使用 SRKD 蒸馏方法在训练的验证阶段损失下降变化曲线。由图可知，第 90 个 epoch、120 个 epoch 学习率下降时，学生网络模的损失都能够大幅下降。对相同的 ResNet20 学生网络进行知识蒸馏时，较小的 ResNet56 教师网络(如(a)所示)所产生的损失在最后阶段基本持平，更大的 ResNet110 教师网络(如(b)所示)在知识蒸馏时，学生与教师间的损失更大，这表明师生差距较大时，师生之间特征存在更大差异。

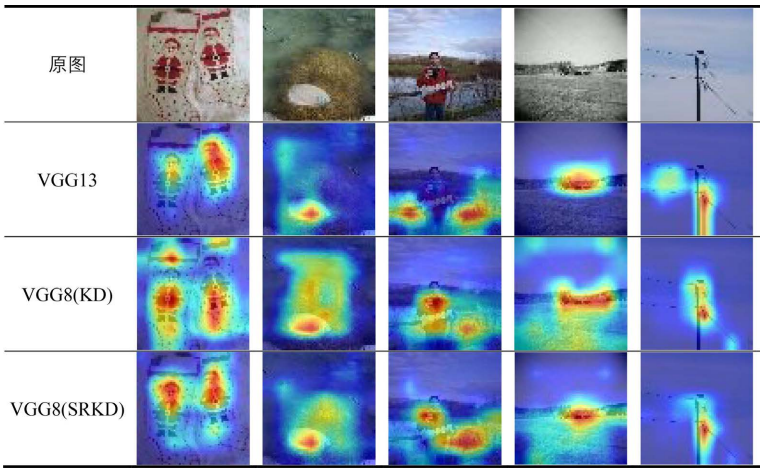


Figure 6. Visualization of attention
图 6. 注意力可视化图

如图 5 所示的是学生网络在不同教师蒸馏下的准确率变换曲线。从实验曲线可知，使用不同的教师时，学生网络在 SRKD 方法的蒸馏准确率高于其它两种方法，并且在教师网络更大的情况下，SRKD 的蒸馏效果比 KD 更加出色。

此外，我们使用 Grad-CAM [15]对教师网络和学生网络进行注意力可视化，Grad-CAM 是一种用于在模型识别中可视化注意力图的通用工具。通过将空间注意力矩阵与原图相结合，我们可直观观察神经网络关注的空间位置，如图 6 所示。在图中，偏红的位置表示网络关注的重点区域，而偏蓝的位置表示关注较少的区域。本仿真实验主要探究学生注意力图与教师注意力图的相似性。我们采用了 VGG13 和 VGG8 作为教师网络和学生网络，在 Tiny-ImageNet 数据集训练。实验结果第一行的图像来自 Tiny-ImageNet 验证集的部分图片，VGG13 所在行的图像是 VGG13 教师网络对不同图像的注意力图。使用本文提出的 SRKD 蒸馏方法，学生网络的注意力图与教师网络相比注意力稍微分散。但与 KD 方法相比，SRKD 学生网络的注意力更接近教师网络，表明 SRKD 蒸馏法能让学生网络有效地从教师网络中提取知识。

3.3. 实验结果与分析

为了验证本文提出方法的效果，我们将之前的知识蒸馏方法与本文方法进行了对比。在 CIFAR-100 数据集上，我们对不同师生组进行了蒸馏，并记录了学生网络模型分类的 TOP-1 准确率。结果如表 1 所示，SRKD 方法在最终精度上有显著提高，优于其他比较方法。值得注意的是，SRKD 的平均准确率超出基线网络 2.43%，比 KD 高出 0.6%，比 FitNet 高出 2.17%，比 AT 高出 1.34%，比 SP 高出 1.33%，比 CC 高出 2.06%，比 RKD 高出 1.92%，比 PKT 高出 0.90%，比 NST 高出 1.81%。这验证了 SRKD 方法的有效性。在 WRN-40-2 和 WRN16-2 的师生对中，SRKD 的准确率达到 75.83%，这是所有方法中唯一超过教师网络的方法。另外，在 ResNet56 和 ResNet110 的教师网络对 ResNet20 的学生网络知识蒸馏，使用 FitNet、AT、CC、RKD、PKT 和 NST 的方法时，精度更高的教师蒸馏效果不如精度较低的教师网络效果好。但使用 SRKD 方法进行蒸馏时，ResNet110 对 ResNet20 蒸馏效果比 ResNet56 更好。这表明 SRKD 有效减少师生差距过大导致语义不一致的问题，并进一步提高了学生网络的准确性。

Table 1. Comparison results between mainstream methods and SRKD onCIFAR-100 dataset
表 1. 主流方法与 SRKD 在 CIFAR-100 数据集上的对比结果

Method	Network Architecture					
	WRN-40-2	WRN-40-2	ResNet56	ResNet110	ResNet110	VGG13
	WRN-16-2	WRN-40-1	ResNet20	ResNet20	ResNet32	VGG8
Teacher	75.61	75.61	72.32	74.31	74.31	74.64
Vanilla	73.26	71.98	69.06	69.06	71.14	70.36
KD [7]	74.92	73.54	70.66	70.67	73.08	72.98
FitNet [8]	73.58	72.24	69.21	68.99	71.06	71.02
AT [9]	74.08	72.77	70.55	70.22	72.32	71.43
SP	73.93	72.43	69.67	70.04	72.69	72.68
CC [16]	73.56	72.21	69.63	69.48	71.48	70.71
RKD [17]	73.55	72.22	69.61	69.25	71.82	71.48
PKT [18]	74.54	73.45	70.34	70.25	72.61	72.88
FSP	72.91	NA	69.95	70.11	71.89	70.20
NST [10]	73.68	72.24	69.6	69.53	71.96	71.53
SRKD	**75.83**	**73.94**	**71.84**	**72.08**	**72.37**	**73.36**

Table 2. Comparison results between mainstream methods and SRKD on Tiny-ImageNet data set
表 2. 主流方法与 SRKD 在 Tiny-ImageNet 数据集上的对比结果

Method	Vanilla	KD	FitNet	AT	M-FitNet	SRKD	Teacher
Top-1	44.89	46.05	45.97	46.33	45.91	**46.54**	48.98
Top-5	71.33	72.55	72.38	72.64	72.41	**72.78**	75.36

我们还在 TinyImageNet 数据集进行了实验，使用 ResNet110 和 ResNet20 作为教师和学生模型。结果如表 2 所示，实验结果的评估指标是 TOP-1 准确率和 TOP-5 准确率。结果表明，SRKD 优于其它主流方法，包括 KD、FitNet 和 AT。由于本文提出的 SRKD 方法是基于多阶段蒸馏的方法，为了公平性，我

们还比较了图 2(c)所示的 M-FitNet 多阶段蒸馏方法。然而，M-FitNet 的结果略低于单个阶段蒸馏的 FitNet。但 SRKD 蒸馏效果比 M-FitNet 和 FitNet 更好，这证实了我们的假设，SRKD 能够更好提炼教师的特征表达方式，有效减少学生向教师学习的难度，提升了知识蒸馏的效果。

Table 3. Comparison between the main method and SRKD combined with the main method on CIFAR-10 dataset
表 3. 在 CIFAR-10 数据集上，主流方法与 SRKD 结合的方法实验比较

Method		Vanilla	KD	KD+ SRKD	AT	AT+ SRKD	SP	SP+ SRKD	Teacher
T: Resnet20	Top-1	78.25	86.52	**86.81**	86.77	**87**	86.5	**86.74**	85.81
S: Resnet8	Top-5	98.69	99.25	**99.4**	99.31	**99.42**	99.14	**99.38**	99.22
T: Resnet32	Top-1	85.81	88.67	**88.97**	88.83	**89.06**	89.05	**89.09**	86.97
S: Resnet20	Top-5	99.22	99.41	**99.49**	99.48	**99.51**	99.52	**99.53**	99.37

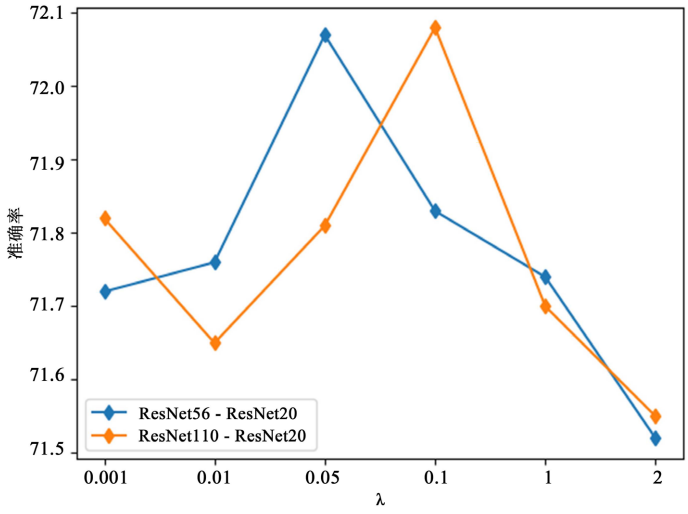


Figure 7. The influence of the parameters λ
图 7. 参数 λ 对准确率的影响

Table 4. Ablation experiments for the number of stage distillations
表 4. 分级蒸馏次数的烧蚀实验

Number	Stage3	Stage2	Stage1	Acc
0				71.21
1	✔			71.44
2	✔	✔		71.71
3	✔	✔	✔	**71.83**

此外，现有很多工作集中在基于特征的知识蒸馏的研究上，而 SRKD 方法可以与这些基于特征的知识蒸馏方法相结合。我们在数据集 CIFAR-10 上进行实验，将 SRKD 与 KD、AT 和 SP 的知识蒸馏方法相结合。实验中师生网络分别是 ResNet20、ResNet8 和 ResNet32、ResNet20。实验结果表 3 所示。实验采用 Top-1 和 Top-5 精度评估。对于 3 种基线方法，SRKD 有效提高了 KD、AT 和 SP 的性能，TOP-1 准确率分别提高了 0.29%、0.23%和 0.14%。因此，SRKD 方法可以与主流蒸馏方法结合，提高学生网络

的精度。

为了评估了超参数 λ 对于 SRKD 的影响。实验在 CIFAR-10 数据集上进行，使用 ResNet56、ResNet20 师生组和 ResNet110、ResNet20 师生组，在 0.001、0.01、0.05、0.1、1、2 等不同权重下进行实验。从图 7 的结果可以观察到，λ 在 0.05~0.1 的范围内精度最高，λ 大于 1 时 SRKD 的蒸馏效果会变差。因此，本节上述的所有实验的中 L_{SRKD} 损失权重因子 λ 设置为 0.1。

我们在 CIFAR-100 数据集中进行了消融实验(表 4)。实验采用 ResNet56 作为教师网络，ResNet20 作为学生网络。M-FitNet 作为基线实验(Number = 0)，我们首先将 SRKD 方法逐步引入不同的阶段来测试其效果。通常知识蒸馏方法以最后一个阶段特征作为蒸馏目标，我们首先将 SRKD 引入到最后一个阶段(Number = 1)，并观察到实验结果为 71.44%。接着，我们逐步增加第二和第一阶段的阶段重置(Number = 2、3)，蒸馏精度分别提升到 71.71%和 71.83%。在单独增加第三阶段的阶段重置蒸馏时，SRKD 方法的性能优于基线。逐步增加第二阶段的阶段重置学生网络提高精度最快，逐步增加第三阶段的阶段重置蒸馏，学生网络的精度最高。因此，通过增加阶段重置的次数，知识蒸馏性能逐渐提高。多阶段重置的知识提炼使教师网络特征更容易为学生所接受，能有效提升学生网络的性能。这些发现表明，SRKD 方法可以缓解语义不匹配的问题，并在知识蒸馏中起到积极的作用。

4. 结论

本文提出一种阶段重置知识蒸馏方法。具体来说，我们让学生以阶段为单位对齐教师网络的特征，同一阶段的师生保持相同的输入，同一阶段的教师的输出作为学生的学习目标，间接缩小了师生推理距离差距，从而解决由于师生模型容量差异导致的蒸馏效果不佳的问题。我们通过仿真实验展示了 SRKD 蒸馏方法的学生网络在训练阶段的损失和精度的变化过程，在多个数据集上使用不同结构的网络进行了广泛的对比实验和消融实验，验证了我们提出的方法的有效性。然而，SRKD 方法也存在一定的缺陷，在学生网络训练阶段，教师网络需要几乎两次推理，这显著增加了学生网络的训练时间。但与知识蒸馏的训练时间相比，我们更关注学生网络的推理时间和精度。综上所述，SRKD 提供了一种有效的阶段式蒸馏方式，并能与传统的知识方法相结合，有效提升学生网络性能。

基金项目

国防基础科研项目(JCKY2019413D001)、上海理工大学医工交叉项目(10-21-302-413)、国家自然基金项目(6217023627)。

参考文献

[1] Krizhevsky, A., Sutskever, I. and Hinton, G.E. (2012) ImageNet Classification with Deep Convolutional Neural Networks. *Communications of the ACM*, **60**, 84-90. https://doi.org/10.1145/3065386

[2] Redmon, J., Divvala, S., Girshick, R. and Farhadi, A. (2016) You Only Look Once: Unified, Real-Time Object Detection. 2016 *IEEE Conference on Computer Vision and Pattern Recognition*, Las Vegas, 27-30 June 2016, 779-788. https://doi.org/10.1109/CVPR.2016.91

[3] Long, J., Shelhamer, E. and Darrell, T. (2015) Fully Convolutional Networks for Semantic Segmentation. 2015 *IEEE Conference on Computer Vision and Pattern Recognition*, Boston, 7-12 June 2015, 3431-3440. https://doi.org/10.1109/CVPR.2015.7298965

[4] Cui, J., Chen, P., Li, R., *et al*. (2019) Fast and Practical Neural Architecture Search. 2019 *IEEE/CVF International Conference on Computer Vision*, Seoul, 27 October 2019-2 November 2019, 6509-6518. https://doi.org/10.1109/ICCV.2019.00661

[5] Luo, J.-H., Wu, J. and Lin, W. (2017) ThiNet: A Filter Level Pruning Method for Deep neural Network Compression. 2017 *IEEE International Conference on Computer Vision*, Venice, 22-29 October 2017, 5058-5066. https://doi.org/10.1109/ICCV.2017.541

[6] Jacob, B., Kligys, S., *et al.* (2018) Quantization and Training of Neural Networks for Efficient Integer-Arithmetic-Only Inference. 2018 *IEEE Conference on Computer Vision and Pattern Recognition*, Salt Lake City, 18-23 June 2018, 2704-2713. https://doi.org/10.1109/CVPR.2018.00286

[7] Hinton, G., Vinyals, O. and Dean, J. (2015) Distilling the Knowledge in a Neural Network. arXiv: 1503.02531.

[8] Romero, A., Ballas, N., Kahou, S.E., *et al*. (2015) FitNets: Hints for Thin Deep Nets. arXiv: 1412.6550.

[9] Zagoruyko, S. and Komodakis, N. (2016) Paying More Attention to Attention: Improving the Performance of Convolutional Neural Networks via Attention Transfer. arXiv: 1612.03928.

[10] Huang, Z. and Wang, N. (2017) Like What You Like: Knowledge Distill via Neuron Selectivity Transfer. arXiv: 1707.01219.

[11] Mirzadeh, S.I., Farajtabar, M., Li, A., Levine, N., Matsukawa, A. and Ghasemzadeh, H. (2020) Improved Knowledge Distillation via Teacher Assistant. *Proceedings of the AAAI Conference on Artificial Intelligence*, **34**, 5191-5198. https://doi.org/10.1609/aaai.v34i04.5963

[12] Ji, M., Heo, B. and Park, S. (2021) Show, Attend and Distill: Knowledge Distillation via Attention-Based Feature Matching. *Proceedings of the AAAI Conference on Artificial Intelligence*, **35**, 7945-7952. https://doi.org/10.1609/aaai.v35i9.16969

[13] Wang, C., Chen, D., Mei, J.-P., Zhang, Y., Feng, Y. and Chen, C. (2022) SemCKD: Semantic Calibration for Cross-Layer Knowledge Distillation. *IEEE Transactions on Knowledge and Data Engineering*, **35**, 6305-6319. https://doi.org/10.1109/TKDE.2022.3171571

[14] Chen, P., Liu, S., Zhao, H. and Jia, J. (2021) Distilling Knowledge via Knowledge Review. 2021 *IEEE/CVF Conference on Computer Vision and Pattern Recognitio*n, Nashville, 20-25 June 2021, 5008-5017. https://doi.org/10.1109/CVPR46437.2021.00497

[15] Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D. and Batra, D. (2017) Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization. 2017 *IEEE International Conference on Computer Vision*, Venice, 22-29 October 2017, 618-626. https://doi.org/10.1109/ICCV.2017.74

[16] Peng, B., Jin, X., *et al.* (2019) Correlation Congruence for Knowledge Distillation. *IEEE/CVF International Conference on Computer Vision*, Seoul, 27 October 2019-2 November 2019, 5006-5015. https://doi.org/10.1109/ICCV.2019.00511

[17] Park, W., Kim, D., Lu, Y. and Cho, M. (2019) Relational Knowledge Distillation. 2019 *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Long Beach, 15-20 June 2019, 3967-3976. https://doi.org/10.1109/CVPR.2019.00409

[18] Passalis, N. and Tefas, A. (2018) Learning Deep Representations with Probabilistic Knowledge Transfer. In: Ferrari, V., Hebert, M., Sminchisescu, C., Weiss, Y., Eds., *Computer Vision—ECCV* 2018. *Lecture Notes in Computer Science*, Springer, Cham, 268-284. https://doi.org/10.1007/978-3-030-01252-6_17