Learning Student Networks via Feature Embedding

通过特征嵌入学习学生网络

Hanting Chen, Yunhe Wang, Chang Xu, Chao Xu and Dacheng Tao, Fellow, IEEE

Hanting Chen, Yunhe Wang, Chang Xu, Chao Xu 和 Dacheng Tao，IEEE 会士

Abstract-Deep convolutional neural networks have been widely used in numerous applications, but their demanding storage and computational resource requirements prevent their applications on mobile devices. Knowledge distillation aims to optimize a portable student network by taking the knowledge from a well-trained heavy teacher network. Traditional teacher-student based methods used to rely on additional fully-connected layers to bridge intermediate layers of teacher and student networks, which brings in a large number of auxiliary parameters. In contrast, this paper aims to propagate information from teacher to student without introducing new variables which need to be optimized. We regard the teacher-student paradigm from a new perspective of feature embedding. By introducing the locality preserving loss, the student network is encouraged to generate the low-dimensional features which could inherit intrinsic properties of their corresponding high-dimensional features from teacher network. The resulting portable network thus can naturally maintain the performance as that of the teacher network. Theoretical analysis is provided to justify the lower computation complexity of the proposed method. Experiments on benchmark datasets and well-trained networks suggest that the proposed algorithm is superior to state-of-the-art teacher-student learning methods in terms of computational and storage complexity.

摘要 - 深度卷积神经网络已广泛应用于众多应用中，但其对存储和计算资源的高要求限制了其在移动设备上的应用。知识蒸馏旨在通过从训练良好的重型教师网络中提取知识来优化一个可移植的学生网络。传统的基于教师-学生的方法依赖于额外的全连接层来桥接教师和学生网络的中间层，这引入了大量的辅助参数。相比之下，本文旨在在不引入需要优化的新变量的情况下，将信息从教师传播到学生。我们从特征嵌入的新视角来看待教师-学生范式。通过引入局部保持损失，鼓励学生网络生成低维特征，这些特征可以继承其对应的高维特征的内在属性。由此产生的可移植网络可以自然地保持与教师网络相同的性能。提供了理论分析以证明所提方法的计算复杂性较低。在基准数据集和训练良好的网络上的实验表明，所提算法在计算和存储复杂性方面优于最先进的教师-学生学习方法。

Index Terms-deep learning, teacher-student learning, knowledge distillation.

关键词 - 深度学习，教师-学生学习，知识蒸馏。

I. INTRODUCTION

I. 引言

DEEP neural networks (DNNs) have provided state-of-the- art performance in various fields such as image classification [1], [2], semantic modeling [3], [4], visual quality evaluation [5], object detection [6], [7], and segmentation [8]. However, a neural network with considerable parameters requires heavy computation for both training and test, which is difficult to use on edge devices such as mobile phones and smart cameras. For example, a VGGNet [1] consisting of 16 convolutional layers has more than 500MB parameters and it requires more than $10^{10}\times$ floating number multiplications, which cannot be tolerated by portable devices. Therefore, how to compress and accelerate existing CNNs has become a research hotspot.

深度神经网络 (DNNs) 在图像分类 [1]、语义建模 [3]、视觉质量评估 [5]、目标检测 [6] 和分割 [8] 等多个领域提供了最先进的性能。然而，具有大量参数的神经网络在训练和测试时需要进行大量计算，这使得它们难以在移动电话和智能摄像头等边缘设备上使用。例如，包含 16 个卷积层的 VGGNet [1] 拥有超过 500MB 个参数，并且需要超过 $10^{10}\times$ 次浮点数乘法，这对于便携设备来说是不可接受的。因此，如何压缩和加速现有的卷积神经网络 (CNNs) 已成为研究热点。

Recently, a variety of CNN compression methods have been proposed to tackle the aforementioned issues such as quantization [9], [10], weight and feature approximation [11], encoding [12], approximation [13], and pruning [14], [15]. Wherein, weight pruning based methods achieve the highest compression performance since there are considerable subtle weights in most of pre-trained CNNs. In specific, Han et.al. [16] showed that over 70% subtle weights in AlexNet [2] can be removed without affecting its original top-5 accuracy. Wang et.al. [14] further pointed out that the redundancy can exist in both large and small weights and considerable redundancy also exists in modern CNNs such as ResNet [17].

最近，提出了多种 CNN 压缩方法来解决上述问题，例如量化 [9]、权重和特征近似 [11]、编码 [12]、近似 [13] 和剪枝 [14]、[15]。其中，基于权重剪枝的方法实现了最高的压缩性能，因为大多数预训练的 CNN 中存在大量微小权重。具体而言，Han 等人 [16] 表明，在不影响其原始前五名准确率的情况下，可以去除 AlexNet [2] 中超过 70% 的微小权重。Wang 等人 [14] 进一步指出，冗余可以存在于大权重和小权重中，现代 CNN(如 ResNet [17]) 中也存在相当大的冗余。

Although pruning based methods can provide very high compression and speed-up ratios, compressed CNNs by exploiting these approaches cannot be directly used in mainstream platforms (e.g. Tensorflow and Caffe) and hardwares (e.g. NVIDIA GPU cards) since they require special architectures and implementation tricks (e.g. sparse convolution and Huffman encoding). As deeper networks often have higher performance than that of shallow networks [18], the teacher-student learning paradigm [19], [20], [21], [22], [23], [24], [25], [26] has emerged to learn portable networks (student network) of deeper architecture yet fewer parameters and convolution filters from the original network (teacher network). Compared with other approaches, portable networks generated by the teacher-student paradigm are much more flexible since they are exactly regular neural networks which do not need any additional supports for implementing online inference.

尽管基于剪枝的方法可以提供非常高的压缩和加速比，但利用这些方法压缩的卷积神经网络 (CNN) 无法直接在主流平台 (例如 Tensorflow 和 Caffe) 和硬件 (例如 NVIDIA GPU 卡) 上使用，因为它们需要特殊的架构和实现技巧 (例如稀疏卷积和哈夫曼编码)。由于深层网络通常比浅层网络具有更高的性能 [18]，因此教师-学生学习范式 [19]，[20]，[21]，[22]，[23]，[24]，[25]，[26] 应运而生，以从原始网络 (教师网络) 学习具有更深架构但参数和卷积滤波器更少的可移植网络 (学生网络)。与其他方法相比，由教师-学生范式生成的可移植网络更加灵活，因为它们正是常规神经网络，不需要任何额外支持来实现在线推理。

However, besides directly making outputs of teacher and student networks similar, most of existing methods cannot directly inherit teacher information in other layers to the student network. Since the student network has a thinner architecture than its teacher's, the feature dimensionality (the number of filters) in the student network is much less than that of its teacher. Therefore, a lot of works [21, 122], [23], [24], [25], [26] proposed to use an intermediate fully-connected layer to connect hint and guided layers to approximately inherit the teacher information. In addition, the fully-connected layer brings in a large number of auxiliary parameters, which have larger space and computational complexities and cannot be applied on large-scale CNNs in practice. Taking Student 4 in Table I as an example, the memory usage for storing the fully-connected layer is about 135MB , which is much larger than those of the student network(9MB)and the teacher network(35MB).

然而，除了直接使教师和学生网络的输出相似外，大多数现有方法无法直接将教师信息从其他层继承到学生网络。由于学生网络的架构比教师网络更为精简，学生网络中的特征维度 (滤波器数量) 远少于教师网络。因此，许多研究 [21, 122]，[23]，[24]，[25]，[26] 提出了使用中间全连接层来连接提示层和引导层，以近似继承教师信息。此外，全连接层引入了大量辅助参数，这些参数具有更大的空间和计算复杂性，实际上无法应用于大规模的卷积神经网络。以表 I 中的学生 4 为例，存储全连接层的内存使用量约为 135MB ，这远大于学生网络 (9MB) 和教师网络 (35MB) 的内存使用量。

Given high-dimensional features from teacher network and low-dimensional features from student networks, it is natural to regard the information propagation between these two networks as a feature embedding task. Therefore, we propose a manifold learning based method for learning portable CNNs. In summary, the proposed approach makes the following contributions:

给定来自教师网络的高维特征和来自学生网络的低维特征，自然可以将这两个网络之间的信息传播视为特征嵌入任务。因此，我们提出了一种基于流形学习的方法来学习可移植的卷积神经网络 (CNN)。总之，所提出的方法做出了以下贡献:

- We propose to bridge teacher and student networks via a
- 我们提议通过一个

Hanting Chen and Chao Xu are with the Key Laboratory of Machine Perception (Ministry of Education) and Coopertative Medianet Innovation Center, School of EECS, Peking University, Beijing 100871, P.R. China. Email:htchen@pku.edu.cn, xuchao@cis.pku.edu.cn

韩廷辰和徐超在北京大学电子工程与计算机科学学院机器感知重点实验室 (教育部) 和合作媒体网络创新中心工作，地址: 北京市 100871，电子邮件:htchen@pku.edu.cn，xuchao@cis.pku.edu.cn

Hangting Chen and Yunhe Wang are with the Noah's Ark Laboratory, Huawei Technologies Co., Ltd, HuaWei Building, No. 3 Xinxi Road, Shang-Di Information Industri Base, Hai-Dian District, Beijing 100085, P.R. China. E-mail:htchen@pku.edu.cn, yunhe.wang@huawei.com

韩廷辰和王云鹤在华为技术有限公司的诺亚方舟实验室工作，地址: 北京市海淀区上地信息产业基地新西路 3 号华为大厦，邮政

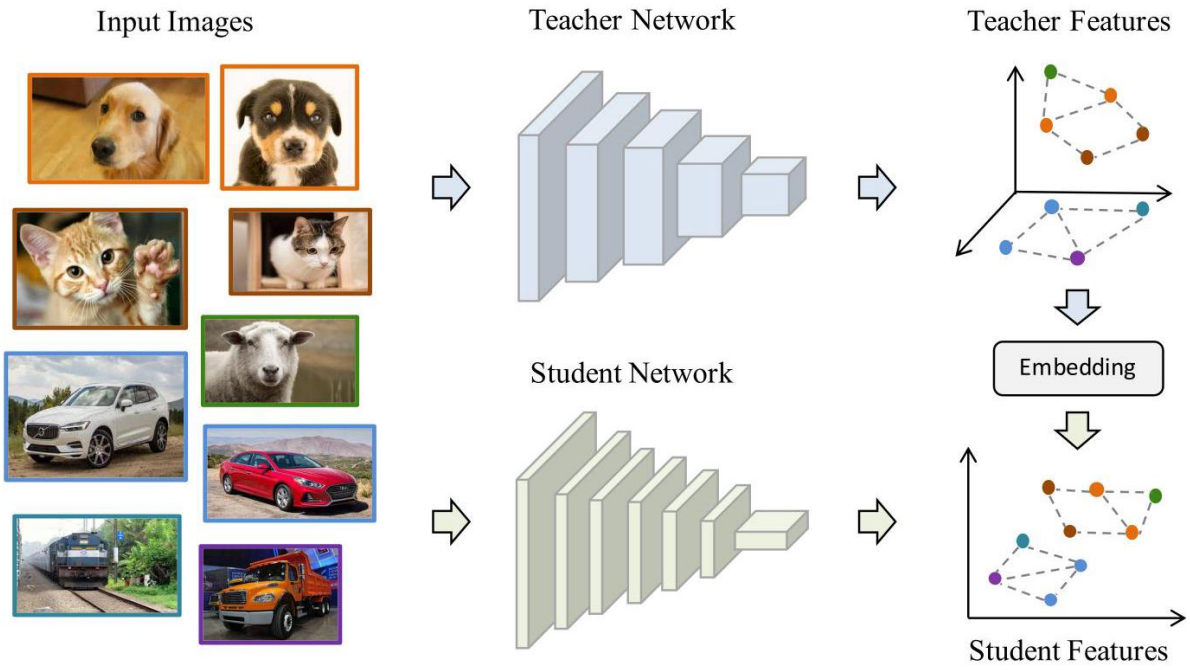


Fig. 1: The diagram of the proposed method for learning portable deep neural networks. The top line is the teacher network and the bottom line is the student network. By applying the proposed method with a locality preserving loss, the student network embeds features in a low-dimensional space and maintains the relationship between samples in the high-dimensional space.

图 1: 所提出的学习可移植深度神经网络的方法示意图。上方是教师网络，下方是学生网络。通过应用具有局部保持损失的所提出的方法，学生网络在低维空间中嵌入特征，并保持高维空间中样本之间的关系。

feature embedding task, so that student networks with fewer filters can generate low-dimensional features to preserve relationships between examples.

特征嵌入任务，以便具有较少滤波器的学生网络能够生成低维特征，以保持示例之间的关系。

- We introduce the locality preserving loss into the teacher-student learning paradigm and provide theoretical analysis to justify the lower computation complexity of the proposed algorithm.
- 我们将局部保持损失引入教师-学生学习范式，并提供理论分析以证明所提算法的计算复杂度较低。
- Experiments on benchmarks demonstrate that the proposed method can efficiently learn portable networks with state-of-the-art performance.
- 基准实验表明，所提方法能够高效地学习具有最先进性能的可移植网络。

This paper is organized as follows. Section II investigates related works on network compression algorithms. Section III proposes the student network learning method by feature embedding. Section IV analyzes the computational and space complexity of the proposed method. Section V shows the experimental results of the proposed method on several benchmark datasets and Section VI concludes this paper.

本文组织如下。第二节探讨了网络压缩算法的相关工作。第三节提出了通过特征嵌入的学生网络学习方法。第四节分析了所提方法的计算和空间复杂度。第五节展示了所提方法在多个基准数据集上的实验结果，第六节对本文进行总结。

编码 100085，电子邮件:htchen@pku.edu.cn，yunhe.wang@huawei.com

Chang Xu and Dacheng Tao are with the UBTech Sydney Artificial Intelligence Centre and the School of Computer Science in the Faculty of Engineering and Information Technologies at The University of Sydney, J12 Cleveland St, Darlington NSW 2008, Australia. E-mail: c.xu@sydney.edu.au, dacheng.tao@sydney.edu.au.

徐畅和陶大程在悉尼大学工程与信息技术学院计算机科学系和 UBTech 悉尼人工智能中心工作，地址: 澳大利亚新南威尔士州达灵顿克利夫兰街 J12，邮政编码 2008，电子邮件:c.xu@sydney.edu.au，dacheng.tao@sydney.edu.au。

II. RELATED WORKS

II. 相关工作

Since CNNs require heavy computation and storage, it is difficult to adapt a neural network to real-world applications directly. Recently, various related works have been proposed to reduce the complexity of CNNs. Compression methods can be grouped into network trimming, layer decomposition and knowledge distillation based on their techniques.

由于卷积神经网络 (CNN) 需要大量的计算和存储，直接将神经网络应用于现实世界的应用中是困难的。最近，提出了各种相关工作以减少 CNN 的复杂性。压缩方法可以根据其技术分为网络修剪、层分解和知识蒸馏。

A. Network Trimming

A. 网络修剪

Network Trimming aims to remove redundant neurons in CNNs to accelerate and compress the original network. Gong et.al. [27] suggested vector quantization to represent similar connections using a cluster center. Denton et.al. [11] exploited the singular value decomposition approach and decomposed the weight matrices of fully connect layers. Considering that 32-bit floating numbers are overfined for parameters of CNNs, Courbariaux et.al. [28] and Rastegari et.al. [29] explored binarized neural networks, whose weights are - 1/1 or -1/0/1. Han et.al. [15] utilized pruning, quantization and Huffman coding to achieve a higher compression ratio. In addition, Wang et.al. [14] introduced the discrete cosine transform (DCT) bases and converted convolution filters into the frequency domain, thereby producing a much higher compression ratio and speed improvement. Subsequently, Wang et.al. [30] compressed feature map of CNN in the frequency domain and accelerated the calculation of convolution directly. Sun et.al. [31] introduced least absolute shrinkage and selection operator to design a selection method for efficient networks. Wang et.al. [32] proposed a novel pruning algorithm with better generalization and pruning efficiency by utilizing Group Lasso. Huang and Yu [33] compressed the weight matrices during training by reshaping them into a high-dimensional tensor with a low-rank approximation.

网络裁剪旨在去除卷积神经网络中的冗余神经元，以加速和压缩原始网络。Gong 等人 [27] 建议使用向量量化来表示相似的连接，采用聚类中心。Denton 等人 [11] 利用奇异值分解方法对全连接层的权重矩阵进行了分解。考虑到 32 位浮点数对于卷积神经网络的参数来说过于精细，Courbariaux 等人 [28] 和 Rastegari 等人 [29] 探索了二值神经网络，其权重为 - 1/1 或 -1/0/1。Han 等人 [15] 利用剪枝、量化和霍夫曼编码实现了更高的压缩比。此外，Wang 等人 [14] 引入了离散余弦变换 (DCT) 基，并将卷积滤波器转换到频域，从而实现了更高的压缩比和速度提升。随后，Wang 等人 [30] 在频域压缩了卷积神经网络的特征图，并直接加速了卷积计算。Sun 等人 [31] 引入了最小绝对收缩和选择算子，设计了一种高效网络的选择方法。Wang 等人 [32] 提出了一个新颖的剪枝算法，通过利用组套索实现了更好的泛化能力和剪枝效率。Huang 和 Yu [33] 在训练过程中通过将权重矩阵重塑为低秩近似的高维张量来压缩权重矩阵。

Although aforementioned algorithms achieve satisfactory results in CNN compression, architectures of networks compressed by these methods would be significantly different from the that of ordinary networks, which means that special implementations are required for high-speed inference and development costs are increased.

尽管上述算法在卷积神经网络压缩方面取得了令人满意的结果，但这些方法压缩后的网络架构与普通网络的架构会显著不同，这意味着需要特殊的实现以实现高速推理，并且开发成本增加。

B. Layer Decomposition

B. 层分解

Traditional layers in DNNs (e.g. convolutional layers, fully-connect layers) often result in huge computational cost. Therefore, a number of works aim to design lightweight layers to obtain efficient networks. Jin et.al. [34] presented fully factorized convolutions to accelerate the feedforward of deep networks. Wang et.al. [35] factorized the convolutional layer by considering spatial convolution. SqueezeNet [36] achieved an accuracy similar with AlexNet yet with 50× fewer parameters by utilizing a bottleneck architecture. Pang et.al. [37] proposed sparse shallow MLP to construct a deep network with few parameters and high

accuracy. Since the convolution calculation in CNNs is time-consuming, a various of algorithms focused on redesigning the convolutional layers. MobileNets [38] introduced depth-wise separable convolutions that largely reduced the computation cost of convolutional layers. ShuffleNet [39] combined pointwise group convolution and channel shuffle to decrease the computational complexity while maintaining accuracy. Wu et.al. [40] presented a parameter-free "shift" operation to alternate vanilla convolutions. Moreover, Sandler et.al. [41] improved MobileNets by introducing an inverted residual structure, which consists of thin and linear bottleneck layers. Ma et.al. [42] proposed a new metric beyond FLOPs to evaluate the speed of CNNs, which leads to ShuffleNet V2. Unlike the methods to design a low-cost convolutional layer, Wang et.al. [43] reused the filters in CNNs by exploiting versatile filters and achieved a better performance.

深度神经网络中的传统层 (例如卷积层、全连接层) 往往导致巨大的计算成本。因此，许多研究旨在设计轻量级层以获得高效的网络。Jin 等人 [34] 提出了完全分解卷积以加速深度网络的前馈。Wang 等人 [35] 通过考虑空间卷积对卷积层进行了分解。SqueezeNet [36] 通过利用瓶颈架构实现了与 AlexNet 相似的准确性，但参数数量却少了 50× 。Pang 等人 [37] 提出了稀疏浅层多层感知器，以构建参数少且准确性高的深度网络。由于卷积神经网络中的卷积计算耗时，许多算法专注于重新设计卷积层。MobileNets [38] 引入了深度可分离卷积，大大减少了卷积层的计算成本。ShuffleNet [39] 结合了逐点组卷积和通道洗牌，以降低计算复杂性，同时保持准确性。Wu 等人 [40] 提出了无参数的 "移位" 操作，以替代普通卷积。此外，Sandler 等人 [41] 通过引入反向残差结构改进了 MobileNets，该结构由细长的线性瓶颈层组成。Ma 等人 [42] 提出了一个超越 FLOPs 的新指标来评估卷积神经网络的速度，这导致了 ShuffleNet V2 的出现。与设计低成本卷积层的方法不同，Wang 等人 [43] 通过利用多功能滤波器重用卷积神经网络中的滤波器，从而实现了更好的性能。

C. Knowledge distillation

C. 知识蒸馏

Different from directly compressing the heavy networks, some works investigate the intrinsic information of original networks to learn smaller networks. Knowledge Transfer, first pioneered by Hinton et.al. [19], aims to improve the training of a student network by borrowing knowledge from another powerful teacher network, while the student network has fewer parameters. It uses a softened version of the final output of a teacher network called softened target to instruct the student network. Besides the outputs, features of intermediate layers in teacher networks also contain useful information which can guide the learning of student networks. Therefore, Romero et.al. [21] minimized the difference between features of a hint layer in the student network and a guide layer in its teacher network, which enables the student network to receive sufficient information from teacher network. McClure and Kriegeskorte [20] proposed the pairwise distance of samples as a useful knowledge to transfer, which increases the robustness of student network. Motivated by ensemble learning methods, You et.al. [22] simultaneously utilized multiple teacher networks to learn a better student network. Moreover, several algorithms have been developed to investigate the restriction between teacher and student. Zagoruyk and Komodakis [24] exploited attention mechanism and proposed to transfer attention maps that are the summaries of full activations. Huang and Wang [25] treat the knowledge transfer as a distribution matching problem and used the Maximum Mean Discrepancy (MMD) metric to minimize the difference between features from teacher and student networks. Wang et.al. [26] exploited generative adversarial network to make feature distribution of teacher and student networks similar.

与直接压缩重型网络不同，一些研究探讨了原始网络的内在信息，以学习更小的网络。知识迁移最早由 Hinton 等人提出，旨在通过借用来自另一个强大教师网络的知识来改善学生网络的训练，而学生网络的参数更少。它使用教师网络最终输出的软化版本，称为软化目标，以指导学生网络。除了输出外，教师网络中间层的特征也包含有用的信息，可以指导学生网络的学习。因此，Romero 等人最小化了学生网络中提示层的特征与其教师网络中指导层特征之间的差异，从而使学生网络能够从教师网络中接收足够的信息。McClure 和 Kriegeskorte 提出样本的成对距离作为有用的知识进行迁移，这增加了学生网络的鲁棒性。受到集成学习方法的启发，You 等人同时利用多个教师网络来学习更好的学生网络。此外，已经开发了几种算法来研究教师与学生之间的限制。Zagoruyk 和 Komodakis 利用注意力机制，提出迁移注意力图，即完整激活的摘要。Huang 和 Wang 将知识迁移视为一个分布匹配问题，并使用最大均值差异 (MMD) 度量来最小化教师网络和学生网络特征之间的差异。Wang 等人利用生成对抗网络使教师和学生网络的特征分布相似。

Compared to network trimming approaches, teacher-student learning paradigms can be applied to mainstream hardware without special requirements, and can be combined with layer decomposition meth-

ods easily. Existing knowledge distillation algorithms can learn efficient networks under the guidance of the teacher networks. However, these methods usually exploited a fully-connect layer to bridge the gap between the high-dimensional features from teacher network and low-dimensional features from student network, which introduced a large number of additional parameters. For example, a $19GB$ fully-connect layer is required to connect $7 \times 7 \times 2048$ dimensional intermediate features in the teacher network (e.g. ResNet-101) and $7 \times 7 \times 1024$ dimensional intermediate features in the student network (e.g. Inception-BN). Given these huge space and computational complexities, a flexible and effective algorithm to transfer knowledge between features of the teacher and student is urgently required.

与网络裁剪方法相比，教师-学生学习范式可以在主流硬件上应用，而无需特殊要求，并且可以轻松与层分解方法结合。现有的知识蒸馏算法可以在教师网络的指导下学习高效的网络。然而，这些方法通常利用全连接层来弥补教师网络中的高维特征与学生网络中的低维特征之间的差距，这引入了大量额外的参数。例如，需要一个 $19GB$ 全连接层来连接教师网络 (例如 ResNet-101) 中的 $7 \times 7 \times 2048$ 维中间特征和学生网络 (例如 Inception-BN) 中的 $7 \times 7 \times 1024$ 维中间特征。鉴于这些巨大的空间和计算复杂性，迫切需要一种灵活有效的算法来在教师和学生的特征之间转移知识。

III. STUDENT NETWORK EMBEDDING

III. 学生网络嵌入

This section first reviews related works on the teacher-student learning paradigm, and then presents the student network embedding approach by introducing a locality preserving loss.

本节首先回顾与教师-学生学习范式相关的工作，然后通过引入局部保持损失来提出学生网络嵌入方法。

A. Teacher-Student Interactions

A. 教师-学生互动

To learning student networks with portable architectures, Hinton et.al. [19] first proposed the Knowledge Distillation (KD) approach, which utilizes a softened output of a teacher network to transform information to a smaller network. McClure and Kriegeskorte [20] further proposed to minimize the pairwise distance of samples after employing the student network and the teacher network.

为了学习具有可移植架构的学生网络，Hinton 等人 [19] 首先提出了知识蒸馏 (KD) 方法，该方法利用教师网络的软输出将信息转移到更小的网络。McClure 和 Kriegeskorte [20] 进一步提出在使用学生网络和教师网络后最小化样本的成对距离。

Let $\mathcal{N}_T$ and $\mathcal{N}_S$ denote the original pre-trained convolutional neural network (teacher network) and the desired portable network (student network), respectively. The goal of knowledge distillation is utilizing $\mathcal{N}_T$ to enhance the performance of $\mathcal{N}_S$. Denote the softmax output of the teacher network $\mathcal{N}_T$ as $P_T = \mathrm{softmax}(a_T)$, where a_T denotes activations input into the softmax layer. Similarly, $P_S = \mathrm{softmax}(a_S)$ and a_S are softmax output and activations of the student network $\mathcal{N}_S$, respectively. Hinton et.al. [19] introduced the soften softmax output $\tau(P_T)$ and $\tau(P_S)$, which can be calculated as:

令 $\mathcal{N}_T$ 和 $\mathcal{N}_S$ 分别表示原始的预训练卷积神经网络 (教师网络) 和期望的便携网络 (学生网络)。知识蒸馏的目标是利用 $\mathcal{N}_T$ 来增强 $\mathcal{N}_S$ 的性能。将教师网络 $\mathcal{N}_T$ 的 softmax 输出表示为 $P_T = \mathrm{softmax}(a_T)$ ，其中 a_T 表示输入到 softmax 层的激活值。类似地，$P_S = \mathrm{softmax}(a_S)$ 和 a_S 分别是学生网络 $\mathcal{N}_S$ 的 softmax 输出和激活值。Hinton 等人 [19] 引入了软化的 softmax 输出 $\tau(P_T)$ 和 $\tau(P_S)$ ，可以计算为:

$$\tau(P_T) = \mathrm{softmax}\left(\frac{a_T}{\tau}\right),\ \tau(P_S) = \mathrm{softmax}\left(\frac{a_S}{\tau}\right). \tag{1}$$

Comparing with original one-hot like output, the soften output can transfer more information since it contains relationship between different classes. By matching the soften outputs of $\mathcal{N}_T$ and $\mathcal{N}_S$, student networks could inherits useful information from the teacher network. The student network is then learned using the following loss function:

与原始的类似 one-hot 输出相比，软化输出可以传递更多信息，因为它包含不同类别之间的关系。通过匹配 $\mathcal{N}_T$ 和 $\mathcal{N}_S$ 的软化输出，学生网络可以从教师网络中继承有用的信息。然后，使用以下损失函数来学习学生网络:

$$\mathcal{L}_{KD}(\mathcal{N}_S) = \mathcal{H}(\mathrm{y}, P_S) + \lambda \mathcal{H}(\tau(P_T), \tau(P_S)) \tag{2}$$

where $\mathcal{H}$ is the cross-entropy loss, y is the ground-truth label and λ is a Trade-off parameter. The first term denotes the classical classification objective

其中 $\mathcal{H}$ 是交叉熵损失，y 是真实标签，λ 是权衡参数。第一项表示经典的分类目标。

However, since architectures of teacher and student networks are significantly different, the constraint on the final output layer cannot be easily achieved. In addition, given the fact that features in the hidden layer also contain useful information, Romero et.al. [21] presented a more flexible FitNet approach by introducing an intermediate hidden layer to connect teacher and student networks, which achieves higher performance than that of KD method. FitNet is trained in a two-stage fashion following the student-teacher paradigm. Specifically, a fully-connected layer is first added after the guided layer in the student network. Let f_S and f_T denote the features generated by guided layer and hint layer in the teacher network. The loss function used in the first stage can be formulated as:

然而，由于教师网络和学生网络的架构显著不同，因此对最终输出层的约束不能轻易实现。此外，考虑到隐藏层中的特征也包含有用的信息，Romero 等人 [21] 提出了一个更灵活的 FitNet 方法，通过引入中间隐藏层来连接教师和学生网络，从而实现比 KD 方法更高的性能。FitNet 采用学生-教师范式以两阶段的方式进行训练。具体而言，在学生网络的引导层之后首先添加一个全连接层。令 f_S 和 f_T 分别表示教师网络中由引导层和提示层生成的特征。第一阶段使用的损失函数可以表述为:

$$\mathcal{L}_{HT}\left(\mathcal{N}_S\right)=\frac{1}{2}\left\|r\left(f_S\right)-f_T\right\|^2, \tag{3}$$

where r is the fully connect layer added to match f_S and f_T . Then, the student network $\mathcal{N}_S$ is further tuned using knowledge distillation as illustrated in Eq. 2 in the second stage. Since the feature dimensionality of the intermediate hint layer is much higher than that of the softmax layer, FitNet can transfer more useful information thus yield a student network with higher performance.

在这里，r 是添加的全连接层，以匹配 f_S 和 f_T 。然后，学生网络 $\mathcal{N}_S$ 在第二阶段进一步通过知识蒸馏进行调整，如公式 2 所示。由于中间提示层的特征维度远高于 softmax 层的特征维度，因此 FitNet 可以传递更多有用的信息，从而产生性能更高的学生网络。

In addition, lots of works have been proposed to further enhance the accuracy of the student network by introducing different assumptions. For example, Yim et.al. [23] introduced the FSP (Flow of Solution Procedure) matrix to transfer the relationship between convolutional layers. You et.al. [22] simultaneously utilized multiple teacher networks for learning a more accurate student network. Zagoruyko et.al. [24] transferred useful information from the teacher network using the attention maps. Huang et.al. [25] minimized the Maximum Mean Discrepancy (MMD) loss between feature maps from teacher and student networks. Wang et.al. [26] utilized the generative adversarial network to make feature distributions of both teacher and student networks similar. However, there are still two important issues to be addressed: 1) Eq 3 independently considers each individual data point without investigating their connections; 2) the introduced fully-connected layer r increases the cost for training the student network.

此外，许多研究工作已被提出，以通过引入不同的假设来进一步提高学生网络的准确性。例如，Yim 等人 [23] 引入了 FSP(解决方案过程流) 矩阵，以传递卷积层之间的关系。You 等人 [22] 同时利用多个教师网络来学习更准确的学生网络。Zagoruyko 等人 [24] 使用注意力图从教师网络中传递有用信息。Huang 等人 [25] 最小化教师和学生网络之间特征图的最大均值差异 (MMD) 损失。Wang 等人 [26] 利用生成对抗网络使教师和学生网络的特征分布相似。然而，仍然存在两个重要问题需要解决:1) 公式 3 独立考虑每个数据点，而没有研究它们之间的联系；2) 引入的全连接层 r 增加了训练学生网络的成本。

B. Locality Preserving Loss

B. 保持局部性的损失

As mentioned above, the feature dimensionality of the student network is lower than that of the teacher network (e.g. from 6912 to 5120), thus we propose to regard the portable network learning as a low-dimensional embedding procedure, which aims to learn effective low-dimensional features. Considering that input images with similar content should lie on the neighbor area in both high-dimensional and low-dimensional spaces, we propose to exploit the manifold learning approach to address the teacher-student learning paradigm.

如上所述，学生网络的特征维度低于教师网络的特征维度 (例如，从 6912 降到 5120)，因此我们建议将可移植网络学习视为一种低维嵌入过程，旨在学习有效的低维特征。考虑到具有相似内容的输入图像在高维和低维空间中应位于邻近区域，我们建议利用流形学习方法来解决教师-学生学习范式。

Lots of nonlinear manifold learning methods have been proposed for obtaining accurate low-dimensional representations. For instance, locally linear embedding (LLE [44]) attempts to represent the manifold

locally by reconstructing each input point as a weighted combination of its neighbors. Isomap [45] preserves geometric distances by returning an embedding where the distances between points is approximately equal to the shortest path distance, and laplacian eigenmaps (LE [46]) builds a graph incorporating neighborhood information of the dataset to compute a low-dimensional representation of the data set that optimally preserves local neighborhood information in a certain sense. However, these nonlinear methods are not applicable for large scale problems, due to their enormous computation and storage resource costs. In contrast, locality preserving projections (LPP [47]) is a linear alternative to those nonlinear methods and can be easily embedded into the learning of convolutional neural networks.

提出了许多非线性流形学习方法，以获得准确的低维表示。例如，局部线性嵌入 (LLE [44]) 试图通过将每个输入点重构为其邻居的加权组合来局部表示流形。等距映射 (Isomap [45]) 通过返回一个嵌入，使得点之间的距离大致等于最短路径距离，从而保持几何距离，而拉普拉斯特征映射 (LE [46]) 构建了一个图，结合数据集的邻域信息，以计算数据集的低维表示，从而在某种意义上最佳地保留局部邻域信息。然而，由于其巨大的计算和存储资源成本，这些非线性方法不适用于大规模问题。相比之下，保持局部性的投影 (LPP [47]) 是这些非线性方法的线性替代方案，并且可以很容易地嵌入到卷积神经网络的学习中。

Algorithm 1 Learning student network by exploiting the proposed locality preserving loss.

算法 1 通过利用提出的保持局部性损失来学习学生网络。

Input: A given teacher network $\mathcal{N}_T$ and its training set $\mathcal{X}$ with n instances, and the corresponding k -label set $\mathcal{Y}$, parameters: λ, γ , and τ .

输入: 给定的教师网络 $\mathcal{N}_T$ 及其训练集 $\mathcal{X}$ ，包含 n 个实例，以及相应的 k 标签集 $\mathcal{Y}$ ，参数: λ, γ 和 τ 。

1: Initialize the student network $\mathcal{N}_S$, whose number of parameters is significantly fewer than that in $\mathcal{N}_T$;

1: 初始化学生网络 $\mathcal{N}_S$ ，其参数数量显著少于 $\mathcal{N}_T$ 中的参数数量;

repeat

重复

Randomly select a batch $\left\{\left(\mathrm{x}^i, \mathrm{y}^i\right)\right\}_{i=1}^m$;

随机选择一个批次 $\left\{\left(\mathrm{x}^i, \mathrm{y}^i\right)\right\}_{i=1}^m$;

Employ the teacher network on the mini-batch:

在小批量上使用教师网络:

$\left[\tau\left(P_T\right), P_T, f_T\right] \leftarrow \mathcal{N}_T(\mathrm{x})$

Employ the student network on the mini-batch:

在小批量上使用学生网络:

$\left[\tau\left(P_S\right), P_S, f_S\right] \leftarrow \mathcal{N}_S(\mathrm{x})$;

Calculate the LP loss $\mathcal{L}_{LP} \leftarrow \frac{1}{m} \sum_{i,j} \alpha_{ij} \left\| f_S^i - f_S^j \right\|_2^2$;

计算 LP 损失 $\mathcal{L}_{LP} \leftarrow \frac{1}{m} \sum_{i,j} \alpha_{ij} \left\| f_S^i - f_S^j \right\|_2^2$;

Calculate the loss function $\mathcal{L}_{\text{Total}}$ (Fcn 8);

计算损失函数 $\mathcal{L}_{\text{Total}}$ (函数 8);

Update weights in $\mathcal{N}_S$ using gradient descent;

使用梯度下降更新 $\mathcal{N}_S$ 中的权重;

until convergence

直到收敛

Output: The student network $\mathcal{N}_S$.

输出: 学生网络 $\mathcal{N}_S$ 。

Specifically, given a labeled training set with n samples. $\left\{\left(\mathrm{x}^1, \mathrm{y}^1\right), \left(\mathrm{x}^2, \mathrm{y}^2\right), \cdots, \left(\mathrm{x}^n, \mathrm{y}^n\right)\right\}$, we denote features of x^i extracted by the teacher and student networks as f_T^i and f_S^i , respectively. Therefore, we propose to preserve the local relationship of the features generated by the student network like that of its teacher, which can be formulated as:

具体而言，给定一个带标签的训练集，包含 n 样本 $\left\{\left(\mathrm{x}^1, \mathrm{y}^1\right), \left(\mathrm{x}^2, \mathrm{y}^2\right), \cdots, \left(\mathrm{x}^n, \mathrm{y}^n\right)\right\}$ ，我们将教师和学生网络提取的 x^i 特征分别表示为 f_T^i 和 f_S^i 。因此，我们建议保留学生网络生成的特征的局部关系，使其与教师网络相似，这可以表述为:

$$\min_{W_S} \sum_{i,j} \alpha_{ij} \left\| f_S^i - f_S^j \right\|_2^2 \tag{4}$$

where W_S is the parameters of the student network before the guided layer and $\alpha_{i,j}$ describes the local relationship between the features generated by the selected hint layer of the teacher network. Specifically, $\alpha_{i,j}$ is defined as follows:

其中 W_S 是引导层之前学生网络的参数，而 $\alpha_{i,j}$ 描述了由教师网络选定提示层生成的特征之间的局部关系。具体而言，$\alpha_{i,j}$ 定义如下:

$$\alpha_{i,j}=\begin{cases}\exp\left(-\frac{\left\|f_T^i-f_T^j\right\|_2^2}{\sigma^2}\right) & \text{if } j\in N(i),\\ 0 & \text{otherwise,}\end{cases} \tag{5}$$

where $N(i)$ denotes the k nearest neighbor of the feature f_T^i of the i -th image x^i generated by teacher network, and σ is a normalized constant.

其中 $N(i)$ 表示由教师网络生成的第 i 张图像的特征 f_T^i 的 k 最近邻，而 σ 是一个归一化常数。

We can obtain a student network $\mathcal{N}_S$ preserving relationship between samples from a high-dimensional space to the target low-dimensional space by optimizing Eq. 4 However, to compute the k nearest neighbors, we need to take the entire training set in each iteration, which is inefficient. Therefore, we use the mini-batch strategy to learn the student network, and the k -nearest neighbors will only be discovered within the mini-batch. The locality preserving loss function is therefore reformulated as:

我们可以通过优化方程 4 来获得一个学生网络 $\mathcal{N}_S$ ，以保留从高维空间到目标低维空间的样本之间的关系。然而，为了计算 k 最近邻，我们需要在每次迭代中考虑整个训练集，这效率低下。因此，我们使用小批量策略来学习学生网络，而 k 最近邻仅在小批量内被发现。因此，局部保留损失函数被重新表述为:

$$\mathcal{L}_{LP}=\frac{1}{2m}\sum_{i,j}\alpha_{ij}\left\|f_S^i-f_S^j\right\|_2^2 \tag{6}$$

where m is the batch size of the student network.

其中 m 是学生网络的批量大小。

In addition, the ground-truth label data is also used for helping the training process of the student network. The entire objective function of the proposed network is then formulated as:

此外，真实标签数据也用于帮助学生网络的训练过程。然后，所提议网络的整个目标函数被表述为:

$$\mathcal{L}_{\text{Total}}=\frac{1}{2m}\left[\sum_i\mathcal{H}\left(\mathrm{y}^i,P_S^i\right)+\gamma\sum_{i,j}\alpha_{ij}\left\|f_S^i-f_S^j\right\|_2^2\right], \tag{7}$$

where γ is the weight parameter for seeking the trade-off of two different terms and P_S^i is the output of the classifier in the student network for the i -th sample x^i .

其中 γ 是用于寻求两个不同项之间权衡的权重参数，而 P_S^i 是学生网络中第 i 个样本 x^i 的分类器输出。

The first term in Fcn. 7 minimizes the cross entropy loss of classifier outputs to maintain the performance of the student network, and the second term embeds samples from the high-dimensional space to the low-dimensional space in the portable student network $\mathcal{N}_S$. Nevertheless, the knowledge distillation approach as discussed in Eq. 2 can be incorporated to further inherit more useful information from the teacher network. Therefore, we reformulate Eq. 7 as

Fcn. 7 中的第一个项最小化分类器输出的交叉熵损失，以维持学生网络的性能，第二个项将高维空间中的样本嵌入到便携式学生网络的低维空间中 $\mathcal{N}_S$ 。然而，如式 2 中讨论的知识蒸馏方法可以被纳入，以进一步从教师网络中继承更多有用的信息。因此，我们将式 7 重新表述为

$$\begin{aligned}\mathcal{L}_{\text{Total}}=&\frac{1}{m}\sum_i\left[\mathcal{H}\left(\mathrm{y}^i,P_T^i\right)+\lambda\mathcal{H}\left(\tau\left(P_T^i\right),\tau\left(P_T^i\right)\right)\right]\\&+\gamma\frac{1}{m}\sum_{i,j}\alpha_{ij}\left\|f_S^i-f_S^j\right\|_2^2\end{aligned} \tag{8}$$

Then, we use stochastic gradient descent (SGD) approach to optimize the student network. Since the proposed LP loss is exactly a linear operation, and the gradient of $\mathcal{L}_{LP}$ with respect to f_S^i can be easily calculated as:

然后，我们使用随机梯度下降 (SGD) 方法来优化学生网络。由于所提出的 LP 损失恰好是一个线性操作，并且 $\mathcal{L}_{LP}$ 关于 f_S^i 的梯度可以很容易地计算为:

$$\frac{\mathcal{L}_{LP}}{\partial f_S^i}=\frac{1}{m}\sum_{j:j\neq i}\alpha_{ij}\left(f_S^i-f_S^j\right). \tag{9}$$

The first term in Eq. 8, i.e. the classification loss, will affect all parameters in the student network, and parameters in $\mathcal{N}_S$ before the guided layer will be additionally updated by:

式 8 中的第一个项，即分类损失，将影响学生网络中的所有参数，而在引导层之前的 $\mathcal{N}_S$ 中的参数将额外更新为:

$$\frac{\partial \mathcal{L}_{LP}}{\partial W_S} = \sum_{i=1}^{m} \frac{\partial \mathcal{L}_{LP}}{\partial f_S^i} \cdot \frac{\partial f_S^i}{\partial W_S} \tag{10}$$

where $\frac{\partial f_S^i}{\partial W_S}$ is the gradient of the feature f_S^i . Alg1summarizes the detailed procedure of the proposed approach for learning student networks.

其中 $\frac{\partial f_S^i}{\partial W_S}$ 是特征 f_S^i 的梯度。Alg1 总结了所提出的方法学习学生网络的详细过程。

IV. ANALYSIS ON THE COMPLEXITY

IV. 复杂性分析

As mentioned above, traditional FitNet based methods introduce an additional fully-connected layer, which makes the training procedure of the student network very slow. In contrast, the proposed method does not need the fully-connected layer for connecting teacher and student networks. Thus, the space complexity and the computational complexity are much lower than those of FitNet based methods, as analyzed in Proposition 1

如上所述，传统的基于 FitNet 的方法引入了一个额外的全连接层，这使得学生网络的训练过程非常缓慢。相比之下，所提出的方法不需要全连接层来连接教师和学生网络。因此，空间复杂性和计算复杂性远低于基于 FitNet 的方法，如命题 1 中分析的那样。

Proposition 1. Denote dimensionalities of features generated by the teacher network and the student network as d_T and d_S , respectively. For a mini-batch with m samples, the computational complexity of the proposed scheme for inheriting information from the teacher network to the student network is $\mathcal{O}\left(m^2\left(d_S + d_T\right)\right)$.

命题 1. 设教师网络和学生网络生成的特征的维度分别为 d_T 和 d_S 。对于一个包含 m 个样本的小批量，所提出的方案将信息从教师网络继承到学生网络的计算复杂性为 $\mathcal{O}\left(m^2\left(d_S + d_T\right)\right)$ 。

Proof. The proposed method inherits information from the teacher network to the student network by calculating the locality preserving loss $\mathcal{L}_{LP}$, whose computational complexity is calculated through three steps.

证明。所提出的方法通过计算局部保持损失 $\mathcal{L}_{LP}$ 将信息从教师网络传递到学生网络，其计算复杂度通过三个步骤进行计算。

The first step is to calculate distances between features generated by the teacher network, i.e. $\left\|f_T^i - f_T^j\right\|_2^2$, whose computational complexity is $\mathcal{O}\left(m^2 d_T\right)$. The second step is to find the k nearest neighbors of each feature f_T^i to calculate α_{ij} in Eq. 6, whose computational complexity is $\mathcal{O}\left(km^2\right)$. Then, distances between features generated by the student network, i.e. $\left\|f_S^i - f_S^j\right\|_2^2$ will be calculated to obtain $\mathcal{L}_{LP}$, with the computational complexity $\mathcal{O}\left(m^2 d_S\right)$. Thus, the computational complexity of the proposed method is $\mathcal{O}\left(m^2\left(d_S + d_T + k\right)\right)$.

第一步是计算由教师网络生成的特征之间的距离，即 $\left\|f_T^i - f_T^j\right\|_2^2$ ，其计算复杂度为 $\mathcal{O}\left(m^2 d_T\right)$ 。第二步是找到每个特征 f_T^i 的 k 最近邻，以计算公式 6 中的 α_{ij} ，其计算复杂度为 $\mathcal{O}\left(km^2\right)$ 。然后，将计算由学生网络生成的特征之间的距离，即 $\left\|f_S^i - f_S^j\right\|_2^2$ ，以获得 $\mathcal{L}_{LP}$ ，其计算复杂度为 $\mathcal{O}\left(m^2 d_S\right)$ 。因此，所提出方法的计算复杂度为 $\mathcal{O}\left(m^2\left(d_S + d_T + k\right)\right)$ 。

Considering that, the dimensionality of features in a CNN is usually much larger than k and m , e.g. $k = 5, m = 128, d_T = 6 \times 6 \times 192 = 6912$, and $d_S = 8 \times 8 \times 80 = 5120$ in the Student 4 network as illustrated in Table I, the complexity $\mathcal{O}\left(km^2\right)$ could be ignored. Therefore, the computational complexity of the proposed method is $\mathcal{O}\left(m^2\left(d_S + d_T\right)\right)$.

考虑到 CNN 中特征的维度通常远大于 k 和 m ，例如 $k = 5, m = 128, d_T = 6 \times 6 \times 192 = 6912$ ，以及表 I 中所示的学生 4 网络中的 $d_S = 8 \times 8 \times 80 = 5120$ ，复杂度 $\mathcal{O}\left(km^2\right)$ 可以忽略。因此，所提出方法的计算复杂度为 $\mathcal{O}\left(m^2\left(d_S + d_T\right)\right)$ 。

In contrast, traditional methods [21], [26], [22], [24], [25] use a fully-connected layer to build the connection between teacher and student network have a $\mathcal{O}\left(md_S d_T\right)$ computational complexity. According to Proposition 1, the proposed teacher-student learning paradigm has a much smaller computational

complexity, which will accelerate the learning process for portable student networks. Taken the Student 4 network in Table 1 as an example, $md_Sd_T/m^2\left(d_S+d_T\right)\approx 23$, since m is much smaller than either d_S or d_T . In addition, considering $k=5, m=256, d_T=7\times 7\times 1024=50176$, and $d_S=7\times 7\times 2048=100352$ in Table VI, $md_Sd_T/m^2\left(d_S+\ d_T\right)\ \approx\ 131$. If the teacher network becomes more complex, d_T and $\mathcal{O}\left(md_Sd_T\right)$ will be increased significantly. In addition, the fully-connected layer introduces considerable parameters with a $\mathcal{O}\left(d_Sd_T\right)$ space complexity, to store such a fully-connected layer would require more than 100MB memory usage in practice. While the memory usage for storing parameter in the student network is only about 9MB . In contrast, the proposed method does not have any additional parameters. We will further illustrate this superiority in the experiment part.

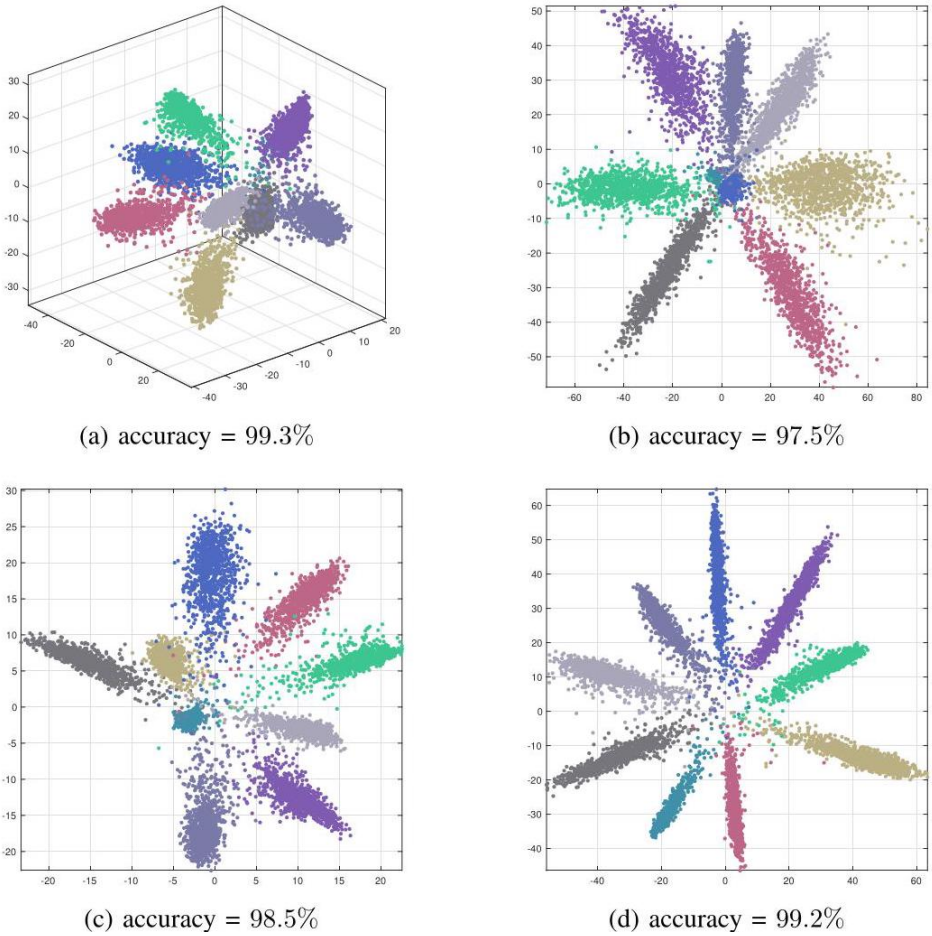
相比之下，传统方法 [21]、[26]、[22]、[24]、[25] 使用全连接层来建立教师网络和学生网络之间的连接，具有 $\mathcal{O}\left(md_Sd_T\right)$ 的计算复杂度。根据命题 1，所提出的教师-学生学习范式具有更小的计算复杂度，这将加速便携式学生网络的学习过程。以表 1 中的学生 4 网络为例，$md_Sd_T/m^2\left(d_S+d_T\right)\approx 23$ ，因为 m 远小于 d_S 或 d_T 。此外，考虑表 VI 中的 $k=5, m=256, d_T=7\times 7\times 1024=50176$ 和 $d_S=7\times 7\times 2048=100352$ ，$md_Sd_T/m^2\left(d_S+\ d_T\right)\ \approx\ 131$ 。如果教师网络变得更加复杂，d_T 和 $\mathcal{O}\left(md_Sd_T\right)$ 将显著增加。此外，全连接层引入了大量参数，具有 $\mathcal{O}\left(d_Sd_T\right)$ 的空间复杂度，实际存储这样的全连接层需要超过 100MB 的内存使用。而学生网络中存储参数的内存使用仅约为 9MB 。相比之下，所提出的方法没有任何额外的参数。我们将在实验部分进一步说明这一优势。

V. EXPERIMENTS

V. 实验

In this section, we implement experiments to validate the effectiveness of the method on three benchmark datasets, including MNIST, CIFAR-10, and CIFAR-100. Experimental results are further analyzed and discussed to investigate the benefits of the proposed method.

在本节中，我们实施实验以验证该方法在三个基准数据集上的有效性，包括 MNIST、CIFAR-10 和 CIFAR-100。实验结果进一步分析和讨论，以探讨所提出方法的优势。



(a) accuracy = 99.3%　(b) accuracy = 97.5%

(c) accuracy = 98.5%　(d) accuracy = 99.2%

Fig. 2: Visualization of features generated by different networks on the MNIST test set. (a) features of the original teacher network; (b) features of the student network learned using the standard back-propagation strategy; (c) features of the student network learned using FitNet; (d) features of the student network learned using the proposed method with locality preserving loss. Note that features of the same category in each figures are marked with the same color.

图 2: 在 MNIST 测试集上不同网络生成的特征可视化。(a) 原始教师网络的特征；(b) 使用标准反向传播策略学习的学生网络特征；(c) 使用 FitNet 学习的学生网络特征；(d) 使用具有局部保持损失的所提出方法学习的学生网络特征。请注意，每个图中同一类别的特征用相同颜色标记。

A. Validations on MNIST

A. 在 MNIST 上的验证

Visualization of Features. The locality preserving loss was introduced in Fcn. 8, which aims to learn the effective low-dimensional features from the pre-trained teacher network. In order to illustrate the superiority of the proposed method, we first trained a LeNet++ [48] as the teacher network, which has six convolutional layers and a fully-connected layers for extracting powerful 3D features. Numbers of neurons in each convolutional layer are32,32,64,64,128,128, and 3, respectively. This teacher network achieved a 99.3% test accuracy on the MNIST dataset, and the memory usage for storing all convolution filters of this teacher network is about

特征可视化。局部保持损失在 Fcn. 8 中被引入，旨在从预训练的教师网络中学习有效的低维特征。为了说明所提方法的优越性，我们首先训练了一个 LeNet++ [48] 作为教师网络，该网络具有六个卷积层和一个全连接层，用于提取强大的 3D 特征。每个卷积层中的神经元数量分别为 32、32、64、64、128、128 和 3。该教师网络在 MNIST 数据集上达到了 99.3% 的测试准确率，并且存储该教师网络所有卷积滤波器的内存使用量约为 2,982KB.

Then, a thinner student network with also six convolutional layers but half convolution filters per layer was initialized. As for the fully-connected layer, the number of neurons was set as 2 for seeking a low-dimensional embedding. Then, we trained this network using conventional back-propagation scheme, and the FitNet method on the MNIST dataset, respectively. Since the student network has fewer parameters, the accuracy of the network using conventional BP is only 97.5% , and the accuracy of the network using FitNet is about 98.5% .

然后，初始化了一个更薄的学生网络，该网络同样具有六个卷积层，但每层的卷积滤波器数量减半。至于全连接层，神经元的数量设置为 2，以寻求低维嵌入。然后，我们分别使用传统的反向传播方案和 FitNet 方法在 MNIST 数据集上训练了该网络。由于学生网络的参数较少，使用传统 BP 的网络准确率仅为 97.5% ，而使用 FitNet 的网络准确率约为 98.5% 。

We then trained a new student network with the same architecture by exploiting the proposed method as described in Fcn. 8, λ and τ were equal to 2 and 0.5, respectively, which refer to those in Hinton et.al. [19]. The learning rate η was set to be 0.01, empirically. The hyper-parameter k for searching neighbors was set as 5, and γ was set to be 1 . The accuracy of the resulting student network is 99.2% which is slightly lower than that of its teacher network but much higher than that of the student network straightforwardly learned using the conventional back-propagation method. In addition, the memory usage for convolution filters of this student network is about 734 KB , which only accounts for $\frac{1}{4}$ of that of the original teacher network.

然后，我们通过利用在 Fcn. 8 中描述的方法训练了一个具有相同架构的新学生网络，其中 λ 和 τ 分别等于 2 和 0.5，这与 Hinton 等人 [19] 的研究相对应。学习率 η 被经验性地设定为 0.01。用于搜索邻居的超参数 k 被设定为 5，而 γ 被设定为 1。所得到的学生网络的准确率为 99.2% ，略低于其教师网络，但远高于使用传统反向传播方法直接学习的学生网络。此外，该学生网络的卷积滤波器的内存使用量约为 734 KB ，仅占原教师网络的 $\frac{1}{4}$ 。

Moreover, features (i.e. input data of the softmax layer) of the above three networks were visualized in Fig. 2. Fig. 2 (a) shows that features of different categories extracted using the original teacher network are separated from each other in the 3-dimensional space, and can be easily distinguished by the following softmax layer. Since the student network has fewer parameters, features of the network trained using the conventional back-propagation are distorted as illustrated in Fig. 2 (b). Therefore, the performance of this network is lower than that of the teacher network. Fig. 2 (d) shows features of the student network learned using the proposed scheme with LP loss. It is clear that, features are separated by supervising of the proposed locality preserving loss, which can be seen as an excellent low-dimensional

embedding of features learned by the teacher network.

此外，上述三个网络的特征 (即 softmax 层的输入数据) 在图 2 中进行了可视化。图 2 (a) 显示，使用原教师网络提取的不同类别特征在三维空间中相互分离，并且可以通过后续的 softmax 层轻松区分。由于学生网络的参数较少，使用传统反向传播训练的网络特征如图 2 (b) 所示被扭曲。因此，该网络的性能低于教师网络。图 2 (d) 显示了使用提出的方案和 LP 损失学习的学生网络特征。显然，这些特征通过提出的局部保持损失进行监督，从而被分离，这可以视为教师网络学习的特征的优秀低维嵌入。

TABLE I: The performance of the proposed method on student networks with various architectures.

表 I: 所提出的方法在具有不同架构的学生网络上的性能。

Network	#layers	#params	#mult	speed-up	compression	Training time (min)		Additional params		Accuracy	
						FitNet	Ours	FitNet	Ours	FitNet	Ours
Teacher	5	9M	725M	×1	×1	150		-		90.21%	
Student 1	11	250K	∼ 30M	×13.17	×36	146	93	∼ 14M	0	89.03%	89.56%
Student 2	11	862K	108M	×4.56	×10.44	221	147	35M	0	91.01%	91.33%
Student 3	13	∼ 1.6M	392M	×1.40	×5.62	317	200	∼ 35M	0	91.14%	91.57%
Student 4	19	2.5M	382M	×1.58	×3.60	400	265	35M	0	91.55%	91.91%

网络	# 层	# 参数	# 倍数	加速	压缩	训练时间 (分钟)		额外参数		准确性	
						FitNet	我们的	FitNet	我们的	FitNet	我们的
教师	5	9M	725M	×1	×1	150		-		90.21%	
学生 1	11	250K	∼ 30M	×13.17	×36	146	93	∼ 14M	0	89.03%	89.56%
学生 2	11	862K	108M	×4.56	×10.44	221	147	35M	0	91.01%	91.33%
学生 3	13	∼ 1.6M	392M	×1.40	×5.62	317	200	∼ 35M	0	91.14%	91.57%
学生 4	19	2.5M	382M	×1.58	×3.60	400	265	35M	0	91.55%	91.91%

TABLE II: Different architectures on the CIFAR-10 dataset.

表 II:CIFAR-10 数据集上的不同架构。

Teacher	Student 1	Student 2	Student 3	Student 4
conv 3x3x96 pool 4x4	conv 3x3x16	conv 3x3x16	conv 3x3x32	conv 3x3x32
	conv 3x3x16	conv 3x3x32	conv 3x3x48	conv 3x3x32
	conv 3x3x16	conv 3x3x32	conv 3x3x64	conv 3x3x32
	pool 2x2	pool 2x2	conv 3x3x64	conv 3x3x48
			pool 2x2	conv 3x3x48 pool 2x2
conv 3x3x96 pool 4x4	conv 3x3x32	conv 3x3x48	conv 3x3x80	conv 3x3x80
	conv 3x3x32	conv 3x3x64	conv 3x3x80	conv 3x3x80
	conv 3x3x32	conv 3x3x80	conv 3x3x80	conv 3x3x80
	pool 2x2	pool 2x2	conv 3x3x80	conv 3x3x80
			pool 2x2	conv 3x3x80 conv 3x3x80 pool 2x2
conv 3x3x96 pool 4x4	conv 3x3x48	conv 3x3x96	conv 3x3x128	conv 3x3x128
	conv 3x3x48	conv 3x3x96	conv 3x3x128	conv 3x3x128
	conv 3x3x64	conv 3x3x128	conv 3x3x128	conv 3x3x128
	pool 8x8	pool 8x8	pool 8x8	conv 3x3x128 conv 3x3x128
				conv 3x3x128 pool 8x8
fc	fc	fc	fc	fc
softmax	softmax	softmax	softmax	softmax

教师	学生 1	学生 2	学生 3	学生 4
conv 3x3x96 pool 4x4	conv 3x3x16	conv 3x3x16	conv 3x3x32	conv 3x3x32
	conv 3x3x16	conv 3x3x32	conv 3x3x48	conv 3x3x32
	conv 3x3x16	conv 3x3x32	conv 3x3x64	conv 3x3x32
	pool 2x2	pool 2x2	conv 3x3x64	conv 3x3x48
			pool 2x2	conv 3x3x48 pool 2x2
conv 3x3x96 pool 4x4	conv 3x3x32	conv 3x3x48	conv 3x3x80	conv 3x3x80
	conv 3x3x32	conv 3x3x64	conv 3x3x80	conv 3x3x80
	conv 3x3x32	conv 3x3x80	conv 3x3x80	conv 3x3x80
	pool 2x2	pool 2x2	conv 3x3x80	conv 3x3x80
			pool 2x2	conv 3x3x80 conv 3x3x80 pool 2x2
conv 3x3x96 pool 4x4	conv 3x3x48	conv 3x3x96	conv 3x3x128	conv 3x3x128
	conv 3x3x48	conv 3x3x96	conv 3x3x128	conv 3x3x128
	conv 3x3x64	conv 3x3x128	conv 3x3x128	conv 3x3x128
	pool 8x8	pool 8x8	pool 8x8	conv 3x3x128 conv 3x3x128
				conv 3x3x128 pool 8x8
fc	fc	fc	fc	fc
softmax	softmax	softmax	softmax	softmax

TABLE III: Classification error on the MNIST dataset.
表 III:MNIST 数据集上的分类错误率。

Algorithm	#params	Misclass
Teacher	361K	0.55%
Standard back-propagation	$\sim$ 30K	1.90%
Knowledge Distillation 19	$\sim$ 30 K	0.65%
FitNet 21	$\sim$ 30K	0.51%
Student (Ours)	$\sim$ 30K	0.48%

算法	# 参数	错误分类
教师	361K	0.55%
标准反向传播	$\sim$ 30K	1.90%
知识蒸馏 19	$\sim$ 30 K	0.65%
FitNet 21	$\sim$ 30K	0.51%
学生 (我们的)	$\sim$ 30K	0.48%

Compression Results. In order to further illustrate the superiority of the proposed method, we followed the setting in Romero et.al. [21] and Wang et.al. [26] to conduct the student network learning experiment on the MNIST dataset. The teacher network consists of maxout convolutional layers as reported in Goodfellow et.al. [49] and the student network is twice as deep as the teacher network following Romero et.al. [21]. The hyper-parameter k of the proposed method for searching neighbors was set as 5, and γ was set to be 1, which were tuned on the last 5,000 images in the train set. The impact of parameters would be showed in the following experiment. The teacher network contains 3 maxout layers and a fully-connected layer, and the student network has 6 maxout layers and a fully-connected layer. Parameters in the student network is only about 8% of that in the teacher network. The guide layer of the teacher network is the 2nd layer while the hint layer of the student network is the 4th layer, respectively. The parameters of the networks were initialized randomly in $\mathrm{U}(-0.005, 0.005)$ and the networks were trained using stochastic gradient descent with RMSProp whose learning rate is 0.0005 and weight decay is 0.9 .

压缩结果。为了进一步说明所提方法的优越性，我们遵循 Romero 等人 [21] 和 Wang 等人 [26] 的设置，在 MNIST 数据集上进行学生网络学习实验。教师网络由 Goodfellow 等人 [49] 报告的 maxout 卷积层组成，学生网络的深度是教师网络的两倍，遵循 Romero 等人 [21] 的研究。所提方法用于搜索邻居的超参数 k 设置为 5，γ 设置为 1，这些参数是在训练集的最后 5,000 张图像上进行调整的。参数的影响将在以下实验中展示。教师网络包含 3 个 maxout 层和一个全连接层，学生网络则有 6 个 maxout 层和一个全连接层。学生网络中的参数仅约为教师网络中参数的 8% 。教师网络的引导层是第 2 层，而学生网络的提示层是第 4 层。网络的参数在 $\mathrm{U}(-0.005, 0.005)$ 中随机初始化，网络使用随机梯度下降法进行训练，采用 RMSProp 优化器，学习率为 0.0005，权重衰减为 0.9。

Table III reports the results of different networks on the MNIST dataset by exploiting the proposed method. In order to illustrate the advantage of the introduced locality preserving loss, the performance of student networks with the same architecture trained by using standard back-propagation, knowledge distillation [19], and FitNets [21] was also reported. It can be found in Table III, the student network trained using the standard back-propagation achieved a 1.90% error rate. The student network learned utilizing the knowledge distillation obtained a 0.65% misclassification error. The error rate of the student network trained by exploiting the FitNet approach is 0.51% , which outperforms both conventional back-propagation and knowledge distillation methods, and is slightly lower than that of the teacher network. In contrast, the student network using the proposed method achieves a 0.48% accuracy.

表 III 报告了利用所提出的方法在 MNIST 数据集上不同网络的结果。为了说明引入的局部保持损失的优势，还报告了使用标准反向传播、知识蒸馏 [19] 和 FitNets [21] 训练的相同架构的学生网络的性能。从表 III 中可以看出，使用标准反向传播训练的学生网络达到了 1.90% 的错误率。利用知识蒸馏学习的学生网络获得了 0.65% 的误分类错误。通过利用 FitNet 方法训练的学生网络的错误率为 0.51% ，其性能优于传统的反向传播和知识蒸馏方法，并且略低于教师网络的错误率。相比之下，使用所提出方法的学生网络达到了 0.48% 的准确率。

TABLE IV: 10-Class results of different networks on the CIFAR-10 datasets.
表 IV: 不同网络在 CIFAR-10 数据集上的 10 类结果。

Algorithm	plane	car	bird	cat	deer	dog	frog	horse	ship	truck
Teacher	90.1%	93.8%	86.0%	74.6%	93.5%	86.2%	95.2%	92.6%	95.3%	95.2%
FitNet [21]	90.7%	97.6%	91.0%	82.7%	93.8%	86.2%	92.7%	93.6%	94.6%	93.5%
Knowledge Distillation [19]	90.0%	95.2%	83.2%	84.4%	93.2%	87.1%	95.0%	91.6%	97.3%	93.7%
Multiple Teachers [22]	91.0%	96.8%	90.0%	83.1%	92.8%	87.1%	93.3%	94.2%	94.4%	93.8%
FSP Learning [23	90.9%	96.7%	90.4%	82.9%	93.3%	87.1%	95.6%	92.7%	96.0%	93.1%
Adversarial Learning 26	91.3%	96.5%	89.8%	84.2%	91.8%	87.5%	93.1%	95.3%	93.4%	93.8%
Student(Ours)	91.1%	97.0%	90.2%	83.6%	92.4%	87.2%	95.3%	93.2%	95.1%	94.1%

算法	平面	汽车	鸟	猫	鹿	狗	青蛙	马	船	卡车
教师	90.1%	93.8%	86.0%	74.6%	93.5%	86.2%	95.2%	92.6%	95.3%	95.2%
FitNet [21]	90.7%	97.6%	91.0%	82.7%	93.8%	86.2%	92.7%	93.6%	94.6%	93.5%
知识蒸馏 [19]	90.0%	95.2%	83.2%	84.4%	93.2%	87.1%	95.0%	91.6%	97.3%	93.7%
多教师 [22]	91.0%	96.8%	90.0%	83.1%	92.8%	87.1%	93.3%	94.2%	94.4%	93.8%
FSP 学习 [23]	90.9%	96.7%	90.4%	82.9%	93.3%	87.1%	95.6%	92.7%	96.0%	93.1%
对抗学习 [26]	91.3%	96.5%	89.8%	84.2%	91.8%	87.5%	93.1%	95.3%	93.4%	93.8%
学生 (我们的方法)	91.1%	97.0%	90.2%	83.6%	92.4%	87.2%	95.3%	93.2%	95.1%	94.1%

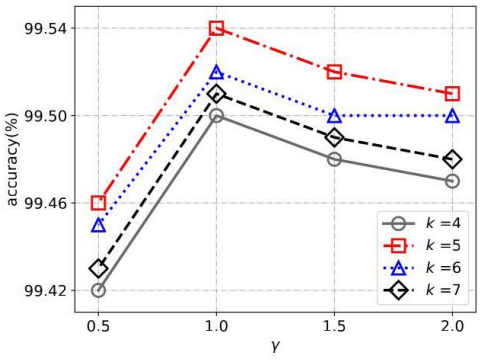


Fig. 3: The performance of the proposed method with different parameters α and β on the MNIST dataset.

图 3: 所提出的方法在 MNIST 数据集上不同参数 α 和 β 的性能。

Impact of parameters. As discussed above, the proposed method has two hyper-parameters: k and γ . We tested their impact on the accuracy of the student network on the validation set of MNIST dataset. It can be seen from Figure 3 that the student network trained utilizing the proposed method achieved the highest accuracy (99.54%) on the validation set when $k = 5$ and $\gamma = 1$. Therefore, we kept these hyper-parameters for the proposed method on the following experiments.

参数的影响。如上所述，所提出的方法有两个超参数: k 和 γ 。我们测试了它们对 MNIST 数据集验证集上学生网络准确性的影响。从图 3 可以看出，利用所提出的方法训练的学生网络在验证集上的准确率达到了最高值 (99.54%) ，当 $k = 5$ 和 $\gamma = 1$ 时。因此，我们在后续实验中保持了这些超参数。

B. Validations on CIFAR-10

B. 在 CIFAR-10 上的验证

After investigating the performance of the proposed method on the MNIST dataset, we then evaluated our method on the CIFAR-10 dataset, which consists of 32×32 pixel RGB color images with 10 categories. There are 50,000 training images and 10,000 testing images in this dataset. For fair comparison, images in the dataset were first processed using global contrast normalization (GCA) and ZCA whitening as suggested in Goodfellow et.al. [49] and Romero et.al. [21], and the last 10,000 training images were selected as the validation set for tuning the hyper-parameters. We followed Goodfellow et.al. [49] to train a teacher network, which consists of three convolutional layers and a fully-connected layer of 500 units with 5-linear-piece maxout activations. The teacher network has about 9M parameters with a 90.21% classification accuracy, and trained for 400 epochs using RMSProp with an initial learning rate of 0.005 and a weight decay of 0.9 . Note that the intermediate hidden layer of the teacher network and student network used in the proposed methods is same as that of FitNet [21] for a fair comparison.

在对所提出的方法在 MNIST 数据集上的性能进行调查后，我们接着在 CIFAR-10 数据集上评估了我们的方法，该数据集由 32×32 像素 RGB 彩色图像组成，共有 10 个类别。该数据集中有 50,000 张训练图像和 10,000 张测试图像。为了公平比较，数据集中的图像首先使用 Goodfellow 等人 [49] 和 Romero 等人 [21] 提出的全局对比度归一化 (GCA) 和 ZCA 白化进行处理，最后选择了 10,000 张训练图像作为验证集以调整超参数。我们遵循 Goodfellow 等人 [49] 的方法训练了一个教师网络，该网络由三个卷积层和一个包含 500 个单元的全连接层组成，使用 5 个线性分段的 maxout 激活函数。教师网络大约有 9M 个参数，分类准确率为 90.21% ，并使用 RMSProp 训练了 400 个周期，初始学习率为 0.005，权重衰减为 0.9。请注意，所提出方法中使用的教师网络和学生网络的中间隐藏层与 FitNet [21] 的相同，以便进行公平比较。

Trade off between compression/speed-up and accuracy. Since the CIFAR-10 dataset consists of more complex images, its samples cannot be easily distinguished by a casually designed neural network. Therefore, several student networks with different architectures, as reported in Romero et.al. [21], were established to further explore the compression performance and accuracy of the proposed method. Table II showed the detailed structures of these student networks. Numbers of parameters in these student networks are 250 K, 862 K, 1.6M , and 2.5M , respectively. The compression ratio and the speedup ratio of each student network can be directly calculated by comparing its numbers of parameters and the floating number multiplications to those of the teacher network, respectively. Detailed results were reported in Table I.

压缩/加速与准确性之间的权衡。由于 CIFAR-10 数据集包含更复杂的图像，因此其样本无法被随意设计的神经网络轻易区分。因此，建立了几种不同架构的学生网络，如 Romero 等人 [21] 所报告的，以进一步探索所提出方法的压缩性能和准确性。表 II 显示了这些学生网络的详细结构。这些学生网络中的参数数量分别为 250 K, 862 K, 1.6M 和 2.5M 。每个学生网络的压缩比和加速比可以通过将其参数数量和浮点数乘法与教师网络进行比较来直接计算。详细结果在表 I 中报告。

It can be found in Table I that, the student network (Student 1) with the highest compression and speed-up ratios has a lower classification accuracy. While, when we appropriately increase the number of parameters, student networks (Student 2-4) can achieve higher performance than that of the teacher network. Moreover, compared with the results of student networks learned using FitNet [21], the proposed method achieves better performance of all the four student networks.

从表 I 可以看出，具有最高压缩和加速比的学生网络 (学生 1) 具有较低的分类准确率。而当我们适当增加参数数量时，学生网络 (学生 2-4) 可以实现比教师网络更高的性能。此外，与使用 FitNet [21] 学习的学生网络的结果相比，所提出的方法在所有四个学生网络中都取得了更好的性能。

Complexity. As discussed in Proposition 1, the proposed method has significantly lower computational and space complexities. Therefore, we also report training time and additional parameters (i.e. weights in the fully-connected layer) of the proposed method and conventional FitNet based method in Table I. It is obvious that, the training time using the proposed method is much less than those of the FitNet and other FitNet based methohds, and we do not need any additional parameters for training student networks. Considering that there are lots of networks with heavy architectures [1], [6], the proposed method is a much more flexible approach with higher performance and efficiency.

复杂性。如命题 1 中所讨论的，所提出的方法在计算和空间复杂性上显著较低。因此，我们还在表 I 中报告了所提出的方法和传统 FitNet 基于方法的训练时间和额外参数 (即全连接层中的权重)。显然，使用所提出的方法的训练时间远低于 FitNet 和其他基于 FitNet 的方法，并且我们在训练学生网络时不需要任何额外参数。考虑到有许多具有复杂架构的网络 [1]，[6]，所提出的方法是一种更灵活的方式，具有更高的性能和效率。

Comparison with state-of-the-art methods. To illustrate the superiority of the proposed method, we compared it with other student-teacher learning methods using the same architecture (Student 4) on the CIFAR-10 dataset as summarized in Table V and Table IV. As a result, the student network utilizing the proposed method achieved a 91.91% accuracy. Comparison results show that our student network outperforms networks produced by state-of-the-art approaches, which proves that the proposed LP loss transfers more useful and intrinsic information from the teacher network. In addition, the student network with more portable architecture (Student 3) can achieve a 91.57% accuracy, which is slightly higher than that of the student network (Student 4) trained using FitNet.

与最先进方法的比较。为了说明所提出方法的优越性，我们将其与其他使用相同架构 (学生 4) 的学生-教师学习方法在 CIFAR-10 数据集上的结果进行了比较，如表 IV 和表 V 所总结的。结果表明，利用所提出方法的学生网络达到了 91.91% 的准确率。比较结果显示，我们的学生网络优于最先进方法生成的网络，这证明了所提出的 LP 损失从教师网络传递了更有用和内在的信息。此外，具有更便携架构的学生网络 (学生 3) 可以达到 91.57% 的准确率，这略高于使用 FitNet 训练的学生网络 (学生 4)。

TABLE V: Classification results of different networks on CIFAR-10 and CIFAR-100 datasets.

表 V: 在 CIFAR-10 和 CIFAR-100 数据集上不同网络的分类结果。

Algorithm	#layers	#params	CIFAR-10	CIFAR-100
Teacher	5	9M	90.21%	62.78%
Student(Ours)	19	2.5M	91.91%	65.13%
FitNet [21]	19	2.5M	91.55%	64.89%
Knowledge Distillation [19]	19	$\sim$ 2.5M	91.04%	63.07%
Multiple Teachers [22]	19	2.5M	91.66%	65.06%
FSP Learning [23]	19	2.5M	91.89%	64.65%
Adversarial Learning [26]	19	$\sim$ 2.5M	91.68%	65.11%

算法	# 层	# 参数	CIFAR-10	CIFAR-100
教师	5	9M	90.21%	62.78%
学生 (我们的)	19	2.5M	91.91%	65.13%
FitNet [21]	19	2.5M	91.55%	64.89%
知识蒸馏 [19]	19	$\sim$ 2.5M	91.04%	63.07%
多教师 [22]	19	2.5M	91.66%	65.06%
FSP 学习 [23]	19	2.5M	91.89%	64.65%
对抗学习 [26]	19	$\sim$ 2.5M	91.68%	65.11%

C. Validations on CIFAR-100

C. 在 CIFAR-100 上的验证

Moreover, we also conducted the validation on the CIFAR- 100 dataset. This dataset has 60,000 RGB color images of pixel 32×32 , which is the same as that of CIFAR-10. However, the CIFAR-100 dataset has 100 categories with only 600 images per class, which is a more challenging benchmark dataset for conducting the classification experiment. For example, the teacher network used in Table V on the CIFAR-100 dataset is only about 62% , which is much lower than that on the CIFAR-10 dataset. Similarly, images in this dataset were pre-processed using global contrast normalization and ZCA whitening. Random flipping, random crop and zero padding was also used for data augmentation as suggested in Romero et.al. [21].

此外，我们还对 CIFAR-100 数据集进行了验证。该数据集包含 60,000 张 RGB 彩色图像，像素 32×32 ，与 CIFAR-10 相同。然而，CIFAR-100 数据集有 100 个类别，每个类别仅有 600 张图像，这使其成为一个更具挑战性的基准数据集，用于进行分类实验。例如，在 CIFAR-100 数据集上使用的教师网络在表 V 中的性能仅约为 62% ，这远低于 CIFAR-10 数据集上的性能。同样，该数据集中的图像经过全局对比度归一化和 ZCA 白化的预处理。还使用了随机翻转、随机裁剪和零填充进行数据增强，正如 Romero 等人 [21] 所建议的。

We used the fourth student network (Student 4 in Table 1) to conduct the experiment on the CIFAR-100 dataset. Table V reports the classification results of student networks learned by exploiting different student-teacher learning paradigms with the same architecture. As a result, the student network learned by the proposed method obtained a 65.13% accuracy. It is clear that the student network learned using the proposed LP loss outperforms those of networks generated by the state-of-the-art methods, which demonstrated that the proposed method inherits useful information from the teacher network in a more effective and efficient way.

我们使用第四个学生网络 (表 1 中的学生 4) 在 CIFAR-100 数据集上进行实验。表 V 报告了通过利用不同的学生-教师学习范式学习的学生网络的分类结果，所有网络具有相同的架构。因此，采用所提出的方法学习的学生网络获得了 65.13% 的准确率。显然，使用所提出的 LP 损失学习的学生网络优于由最先进方法生成的网络，这表明所提出的方法以更有效和高效的方式从教师网络中继承了有用的信息。

D. Validations on ImageNet

D. 在 ImageNet 上的验证

We next conducted experiments on an extremely large image dataset, namely ImageNet ILSVRC 2012 [2], which has about 1.28M training images and 50 K validation images. We followed the settings in Huang and Wang [25] to implement the teacher-student learning paradigm. Wherein, we used ResNet-101 [17] and Inception-BN [50] as the teacher and student networks, respectively. The teacher network has about 128M parameters and student network has only about 32M parameters, which is a relatively portable

model. We used the scale and aspect ratio augmentation in Ioffe and Szegedy [50] and color augmentation in Krizhevsky et.al. [2] following Huang and Wang [25].

我们接下来在一个极大的图像数据集上进行了实验，即 ImageNet ILSVRC 2012 [2]，该数据集大约有 1.28M 个训练图像和 50 K 个验证图像。我们遵循 Huang 和 Wang [25] 的设置来实现教师-学生学习范式。在其中，我们分别使用 ResNet-101 [17] 和 Inception-BN [50] 作为教师网络和学生网络。教师网络大约有 128M 个参数，而学生网络只有大约 32M 个参数，这使其成为一个相对便携的模型。我们在 Ioffe 和 Szegedy [50] 中使用了尺度和长宽比增强，并在 Krizhevsky 等 [2] 中使用了颜色增强，遵循 Huang 和 Wang [25] 的方法。

These networks were optimized using Nesterov Accelerated Gradient (NAG), and the weight decay and the momentum were set as 10^{-4} and 0.9, respectively. We trained the networks for 100 epochs, and the initial learning rate was set as 0.1 and divided by 10 at the 30,60 and 90 epochs, respectively. The batch size was set as 256 and the hyper-parameters of the proposed method are the same as those in CIFAR experiments. In addition, λ and τ in knowledge distillation were equal to 2 and 0.5, respectively [25]. For FitNet, attention transfer and neuron selectivity transfer, the value of λ was set to $10^2, 10^3$ and 5, respectively, which refers to the settings in [25].

这些网络使用 Nesterov 加速梯度 (NAG) 进行优化，权重衰减和动量分别设置为 10^{-4} 和 0.9。我们训练网络 100 个周期，初始学习率设置为 0.1，并在第 30、60 和 90 个周期分别降低为 10。批量大小设置为 256，所提方法的超参数与 CIFAR 实验中的相同。此外，知识蒸馏中的 λ 和 τ 分别等于 2 和 0.5 [25]。对于 FitNet，注意力转移和神经元选择性转移，λ 的值分别设置为 $10^2, 10^3$ 和 5，这参考了 [25] 中的设置。

Table V shows the classification results of student networks on the ImageNet dataset by exploiting the proposed method and state-of-the-art learning methods. The top-1 accuracy and the top-5 accuracy of the teacher network are 77.32% and 93.42% , respectively. The student network without inheriting information (i.e. the standard BP) from the teacher achieved a 74.80% top-1 accuracy and a 92.18% top-5 accuracy.

表 V 显示了通过利用所提出的方法和最先进的学习方法在 ImageNet 数据集上学生网络的分类结果。教师网络的 top-1 准确率和 top-5 准确率分别为 77.32% 和 93.42% 。未从教师网络继承信息 (即标准反向传播) 的学生网络达到了 74.80% 的 top-1 准确率和 92.18% 的 top-5 准确率。

It can be found in Table V that, all the teacher-student learning methods achieve better results than that of the standard back-propagation, which demonstrates that there are abundant information in the teacher network. The student network learned using the proposed method obtained a 93.13% top- 5 accuracy, which is about 1% higher than that of the original student network. When compared to other methods, the student network learned through the proposed locality preserving loss achieved the highest accuracy. In addition, as discussed in Proposition 1, the proposed method has significantly lower computational and space complexities. Especially, deep neural networks on the ImageNet dataset have sophisticated architectures with billions of parameters, and the proposed method is more suitable for efficiently learning portable student networks. For example, the number of additional parameters for connecting teacher and student networks needed by FitNet based methods is about 5,035M. In contrast, the proposed method does not need any additional parameters but can obtain better results.

从表 V 可以看出，所有的教师-学生学习方法都取得了比标准反向传播更好的结果，这证明了教师网络中存在丰富的信息。使用所提出的方法学习的学生网络获得了 93.13% 的 top-5 准确率，比原始学生网络高出约 1% 。与其他方法相比，通过所提出的局部保持损失学习的学生网络达到了最高的准确率。此外，如命题 1 中所讨论的，所提出的方法在计算和空间复杂度上显著降低。特别是，ImageNet 数据集上的深度神经网络具有数十亿参数的复杂架构，而所提出的方法更适合高效学习可移植的学生网络。例如，基于 FitNet 方法连接教师和学生网络所需的额外参数数量约为 5,035M。相比之下，所提出的方法不需要任何额外参数，但可以获得更好的结果。

E. Intermediate Layer Selection

E. 中间层选择

Recalling Eq. 3, most of the teacher-student learning algorithms utilize the output features of intermediate layers of the teacher and student networks (i.e. the guide layer and hint layer). An important issue is how to select the hint layer from teacher network and the guided layer from the student network and how would the selection of such layers affect the performance. Zagoruyko and Komodakis [51] has discussed this problem and the experimental results showed that different selections of intermediate layers could only have a minor influence on the accuracy of student networks (less than 0.2% on the CIFAR-10 dataset).

Moreover, using more than 1 guided layer/hint layer results in a slight improvement of performance (less than 0.1% on the CIFAR-10 dataset). Therefore, we do not study the selection of intermediate layers. Instead, we focus on the performance of the student networks under the same experimental settings.

回顾公式 3，大多数教师-学生学习算法利用教师网络和学生网络中间层的输出特征 (即引导层和提示层)。一个重要的问题是如何从教师网络中选择提示层，以及如何从学生网络中选择引导层，这些层的选择将如何影响性能。Zagoruyko 和 Komodakis [51] 讨论了这个问题，实验结果表明，不同的中间层选择对学生网络的准确性影响很小 (在 CIFAR-10 数据集上小于 0.2%)。此外，使用超过 1 个引导层/提示层会导致性能略有提升 (在 CIFAR-10 数据集上小于 0.1%)。因此，我们不研究中间层的选择，而是专注于在相同实验设置下学生网络的性能。

TABLE VI: Classification results of different networks on ImageNet datasets.

表 VI: 不同网络在 ImageNet 数据集上的分类结果。

Algorithm	Model	#params	Top-1	Top-5
Teacher	ResNet-101	128M	77.32%	93.42%
Student using LP loss (Ours)	Inception-BN	32M	75.91%	93.13%
Standard back-propagation	Inception-BN	32M	74.80%	92.18%
FitNet [21]	Inception-BN	32M	75.52%	92.73%
Knowledge Distillation [19]	Inception-BN	32M	75.44%	92.65%
Attention Transfer [24]	Inception-BN	32M	75.36%	92.74%
Neuron Selectivity Transfer [25]	Inception-BN	32M	75.66%	92.89%

算法	模型	# 参数	Top-1	Top-5
教师	ResNet-101	128M	77.32%	93.42%
使用 LP 损失的学生 (我们的)	Inception-BN	32M	75.91%	93.13%
标准反向传播	Inception-BN	32M	74.80%	92.18%
FitNet [21]	Inception-BN	32M	75.52%	92.73%
知识蒸馏 [19]	Inception-BN	32M	75.44%	92.65%
注意力转移 [24]	Inception-BN	32M	75.36%	92.74%
神经元选择性转移 [25]	Inception-BN	32M	75.66%	92.89%

VI. CONCLUSION

VI. 结论

Here we examine the deep neural network compression problem for learning portable networks from original teacher models. Besides features generated by the teacher network, the relationship between samples in the feature space is another important information for maintaining the performance of the student network. In this paper, we present a novel teacher-student learning paradigm by introducing the locality preserving loss. The neighbor relationship between samples represented by the teacher network is embedded into the student network. Therefore, the resulting portable network can achieve similar performance as that of the teacher network naturally. Experiments on benchmark datasets show that the proposed method can efficiently produce portable neural networks with higher performance, which is superior to the state-of-the-art approaches.

在这里，我们考察了从原始教师模型学习可移植网络的深度神经网络压缩问题。除了教师网络生成的特征外，特征空间中样本之间的关系是维持学生网络性能的另一个重要信息。在本文中，我们通过引入局部保持损失提出了一种新颖的教师-学生学习范式。由教师网络表示的样本之间的邻居关系被嵌入到学生网络中。因此，生成的可移植网络能够自然地实现与教师网络相似的性能。在基准数据集上的实验表明，所提出的方法能够有效地生成具有更高性能的可移植神经网络，优于最先进的方法。

REFERENCES

参考文献

[1] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in ICLR, 2015.

[2] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in NIPS, 2012, pp. 1097- 1105.

[3] K. Greff, R. K. Srivastava, J. Koutník, B. R. Steunebrink, and J. Schmid-huber, "Lstm: A search space odyssey," IEEE transactions on neural networks and learning systems, vol. 28, no. 10, pp. 2222-2232, 2017.

[4] Y. Yuan, L. Mou, and X. Lu, "Scene recognition by manifold regularized deep learning architecture," IEEE transactions on neural networks and learning systems, vol. 26, no. 10, pp. 2222-2233, 2015.

[5] W. Hou, X. Gao, D. Tao, and X. Li, "Blind image quality assessment via deep learning," IEEE transactions on neural networks and learning systems, vol. 26, no. 6, pp. 1275-1286, 2015.

[6] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," in NIPS, 2015, pp. 91-99.

[7] M. Gong, J. Zhao, J. Liu, Q. Miao, and L. Jiao, "Change detection in synthetic aperture radar images based on deep neural networks," IEEE transactions on neural networks and learning systems, vol. 27, no. 1, pp. 125-138, 2016.

[8] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in CVPR, 2015, pp. 3431-3440.

[9] Y. Gong, L. Liu, M. Yang, and L. Bourdev, "Compressing deep convolutional networks using vector quantization," arXiv preprint arXiv:1412.6115, 2014.

[10] J. Cheng, J. Wu, C. Leng, Y. Wang, and Q. Hu, "Quantized cnn: a unified approach to accelerate and compress convolutional networks," IEEE Transactions on Neural Networks and Learning Systems, 2017.

[11] E. L. Denton, W. Zaremba, J. Bruna, Y. LeCun, and R. Fergus, "Exploiting linear structure within convolutional networks for efficient evaluation," in NIPS, 2014.

[12] W. Chen, J. T. Wilson, S. Tyree, K. Q. Weinberger, and Y. Chen, "Compressing neural networks with the hashing trick," in ICML, 2015.

[13] P. Gysel, J. Pimentel, M. Motamedi, and S. Ghiasi, "Ristretto: A framework for empirical study of resource-efficient inference in convolutional neural networks," IEEE Transactions on Neural Networks and Learning Systems, 2018.

[14] Y. Wang, C. Xu, S. You, D. Tao, and C. Xu, "Cnnpack: packing convolutional neural networks in the frequency domain," in NIPS, 2016, pp. 253-261.

[15] S. Han, H. Mao, and W. J. Dally, "Deep compression: Compressing deep neural networks with pruning, trained quantization and huffman coding," arXiv preprint arXiv:1510.00149, 2015.

[16] S. Han, J. Pool, J. Tran, and W. Dally, "Learning both weights and connections for efficient neural network," in NIPS, 2015.

[17] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in CVPR, 2016, pp. 770-778.

[18] J. Ba and R. Caruana, "Do deep nets really need to be deep?" in NIPS, 2014.

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

[20] P. McClure and N. Kriegeskorte, "Representational distance learning for deep neural networks," Frontiers in computational neuroscience, vol. 10, 2016.

[21] A. Romero, N. Ballas, S. E. Kahou, A. Chassang, C. Gatta, and Y. Bengio, "Fitnets: Hints for thin deep nets," in ICLR, 2015.

[22] S. You, C. Xu, C. Xu, and D. Tao, "Learning from multiple teacher networks," in ACM SIGKDD, 2017.

[23] J. Yim, D. Joo, J. Bae, and J. Kim, "A gift from knowledge distillation: Fast optimization, network minimization and transfer learning," in CVPR, 2017.

[24] S. Zagoruyko and N. Komodakis, "Paying more attention to attention: Improving the performance of convolutional neural networks via attention transfer," in ICLR, 2017.

[25] Z. Huang and N. Wang, "Like what you like: Knowledge distill via neuron selectivity transfer," arXiv preprint arXiv:1707.01219, 2017.

[26] Y. Wang, C. Xu, C. Xu, and D. Tao, "Adversaial learning of portable student networks," in AAAI, 2018.

[27] Y. Gong, L. Liu, M. Yang, and L. Bourdev, "Compressing deep convolutional networks using vector quantization," arXiv preprint arXiv:1412.6115, 2014.

[28] M. Courbariaux, I. Hubara, D. Soudry, R. El-Yaniv, and Y. Ben-gio, "Binarized neural networks: Training deep neural networks with

weights and activations constrained to+ 1 or-1," arXiv preprint arXiv:1602.02830, 2016.

[29] M. Rastegari, V. Ordonez, J. Redmon, and A. Farhadi, "Xnor-net: Imagenet classification using binary convolutional neural networks," in ECCV. Springer, 2016, pp. 525-542.

[30] Y. Wang, C. Xu, C. Xu, and D. Tao, "Beyond filters: Compact feature map for portable deep model," in ICML, 2017, pp. 3703-3711.

[31] K. Sun, S.-H. Huang, D. S.-H. Wong, and S.-S. Jang, "Design and application of a variable selection method for multilayer perceptron neural network with lasso," IEEE transactions on neural networks and learning systems, vol. 28, no. 6, pp. 1386-1396, 2017.

[32] J. Wang, C. Xu, X. Yang, and J. M. Zurada, "A novel pruning algorithm for smoothing feedforward neural networks based on group lasso method," IEEE transactions on neural networks and learning systems, vol. 29, no. 5, pp. 2012-2024, 2018.

[33] H. Huang and H. Yu, "Ltnn: A layerwise tensorized compression of multilayer neural network," IEEE transactions on neural networks and learning systems, 2018.

[34] J. Jin, A. Dundar, and E. Culurciello, "Flattened convolutional neural networks for feedforward acceleration," arXiv preprint arXiv:1412.5474, 2014.

[35] M. Wang, B. Liu, and H. Foroosh, "Factorized convolutional neural networks." in ICCV Workshops, 2017, pp. 545-553.

[36] F. N. Iandola, S. Han, M. W. Moskewicz, K. Ashraf, W. J. Dally, and K. Keutzer, "Squeezenet: Alexnet-level accuracy with 50x fewer parameters and 1.5mb model size," arXiv preprint arXiv:1602.07360, 2016.

[37] Y. Pang, M. Sun, X. Jiang, and X. Li, "Convolution in convolution for network in network," IEEE transactions on neural networks and learning systems, vol. 29, no. 5, pp. 1587-1597, 2018.

[38] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam, "Mobilenets: Efficient convolutional neural networks for mobile vision applications," arXiv preprint arXiv:1704.04861, 2017.

[39] X. Zhang, X. Zhou, M. Lin, and J. Sun, "Shufflenet: An extremely efficient convolutional neural network for mobile devices," arXiv preprint arXiv:1707.01083, 2017.

[40] B. Wu, A. Wan, X. Yue, P. Jin, S. Zhao, N. Golmant, A. Gholaminejad, J. Gonzalez, and K. Keutzer, "Shift: A zero flop, zero parameter alternative to spatial convolutions," arXiv preprint arXiv:1711.08141, 2017.

[41] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "Mobilenetv2: Inverted residuals and linear bottlenecks," in CVPR, 2018, pp. 4510-4520.

[42] N. Ma, X. Zhang, H.-T. Zheng, and J. Sun, "Shufflenet v2: Practical guidelines for efficient cnn architecture design," arXiv preprint arXiv:1807.11164, 2018.

[43] C. X. C. X. D. T. Yunhe Wang, Chang Xu, "Learning versatile filters for efficient convolutional neural networks," in NIPS, 2018.

[44] S. T. Roweis and L. K. Saul, "Nonlinear dimensionality reduction by locally linear embedding," science, vol. 290, no. 5500, pp. 2323-2326, 2000.

[45] J. B. Tenenbaum, V. De Silva, and J. C. Langford, "A global geometric framework for nonlinear dimensionality reduction," science, vol. 290, no. 5500, pp. 2319-2323, 2000.

[46] M. Belkin and P. Niyogi, "Laplacian eigenmaps and spectral techniques for embedding and clustering," in NIPS, 2002, pp. 585-591.

[47] X. He and P. Niyogi, "Locality preserving projections," in NIPS, 2004, pp. 153-160.

[48] Y. Wen, K. Zhang, Z. Li, and Y. Qiao, "A discriminative feature learning approach for deep face recognition," in ECCV, 2016.

[49] I. J. Goodfellow, D. Warde-Farley, M. Mirza, A. Courville, and Y. Ben-gio, "Maxout networks," arXiv preprint arXiv:1302.4389, 2013.

[50] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in ICML, 2015, pp. 448-456.

[51] S. Zagoruyko and N. Komodakis, "Paying more attention to attention: Improving the performance of convolutional neural networks via attention transfer," arXiv preprint arXiv:1612.03928, 2016.