计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
Computer Systems & Applications,2021,30(12):248−254 [doi: 10.15888/j.cnki.csa.008202]

E-mail: csa@iscas.ac.cn
http://www.c-s-a.org.cn
Tel: +86-10-62661041

基于关系型蒸馏的分步神经网络压缩方法①

刘　昊, 张晓滨

(西安工程大学 计算机学院, 西安 710048)
通讯作者: 张晓滨, E-mail: xiaobinzhangcn@126.com

摘　要: 针对关系型知识蒸馏方法中教师网络与学生网络的层数差距过大导致蒸馏效果下降的问题, 提出一种基于关系型蒸馏的分步神经网络压缩方法. 该方法的要点在于, 在教师网络和学生网络之间增加一个中间网络分步进行关系型蒸馏, 同时在每一次蒸馏过程中都增加额外的单体信息来进一步优化和增强学生模型的学习能力, 实现神经网络压缩. 实验结果表明, 本文的方法在 CIFAR-10 和 CIFAR-100 图像分类数据集上的分类准确度相较于原始的关系型知识蒸馏方法均有 0.2% 左右的提升.

关键词: 模型压缩; 知识蒸馏; 关系型知识蒸馏; 神经网络; 神经网络压缩

引用格式: 刘昊,张晓滨.基于关系型蒸馏的分步神经网络压缩方法.计算机系统应用,2021,30(12):248–254. http://www.c-s-a.org.cn/1003-3254/8202.html

Compression Method for Stepwise Neural Network Based on Relational Distillation

LIU Hao, ZHANG Xiao-Bin

(School of Computer Science, Xi'an Polytechnic University, Xi'an 710048, China)

Abstract: This study aims at the problem that the distillation effect decreases when the gap between the teacher network and the student network in relational knowledge distillation is too large. A stepwise neural network compression method based on relational distillation is proposed. The key point of this method is to add an intermediate network between the teacher network and the student network for relational distillation step by step. Moreover, in each distillation process, additional monomer information is added to further optimize and enhance the learning ability of the student model. The experimental results show that the classification accuracy of the proposed method on CIFAR-10 and CIFAR-100 image classification datasets is improved by about 0.2% compared with that of the original relational knowledge distillation method.

Key words: model compression; knowledge distillation; relational knowledge distillation; neural network; neural network compression

近些年来, 深度神经网络获得了越来越多的关注, 且在各种应用领域都取得了很好的成果, 例如计算机视觉[1,2]、自然语言处理[3] 等. 但是深层的神经网络往往受限于其模型大和复杂度高的特性, 很难直接部署到计算和存储能力有限的设备, 例如移动设备和嵌入式传感器等[4]. 因此, 在保证一定的准确度和精度的条件下, 合理的对深度神经网络进行模型压缩和优化成为了一个很有研究价值的问题. 知识蒸馏就是模型压缩方法中的一种非常典型的方法.

知识蒸馏的概念最早由 Bucila 等人[5] 在 2006 年提出, 由 Hindon 等人[6] 在 2015 年重拾并普及. 该方法的核心思想在于, 将复杂的、学习能力强的教师网络

① 基金项目: 陕西省自然科学基金 (2019JQ-849)
Foundation item: Natural Science Foundation of Shaanxi Province (2019JQ-849)
收稿时间: 2021-02-24; 修改时间: 2021-03-15; 采用时间: 2021-03-26

学到的知识迁移到较小的学生网络, 从而提升学生网络的精度. 但是在原始的知识蒸馏方法中, 学生网络只是学习了教师网络模型的输出, 并没有考虑教师网络中的其他信息. 随后 Romero 等人[7] 很快提出在知识蒸馏方法基础上, 引导学生模型先学习教师网络隐含层的特征表达来作为一个预训练模型, 再通过知识蒸馏得到目标学生网络, 以使得学生网络可以学习到更丰富的信息. Zagoruyko 等人[8] 在 2016年提出使用教师网络中的注意力图进一步训练学生网络, 通过定义注意力机制, 让学生网络拟合教师网络的注意力映射, 从而提升学生网络的效果, 达到知识蒸馏的目的. Yim 等人[9] 在 2017 年提出让学生网络去学习教师网络中层与层之间的关系而不是输出, 层间关系是用关系矩阵来定义的, 相当于让学生网络拟合了教师网络的学习过程, 这个思路也得到了不错的蒸馏效果. Chen 等人[10] 提出将原始蒸馏和文献 [7] 中提出的隐层蒸馏结合起来针对物体检测任务做知识蒸馏. Huang 等人[11] 提出将知识蒸馏的迁移过程看作是特征分布拟合过程, 并采用域适应中常用的最大平均差异来优化. Chen 等人[12] 提出使用跨样本相似性作为新的知识来让学生网络进行学习, 在人员检测和图像检索任务中都收到了不错的效果. Heo 等人[13] 提出迁移两种知识, 包括 ReLU 之后的特征响应大小以及每一个神经元的激活状态让学生网络进行学习. Heo 等人[14] 在 AAAI 上也提出了基于激活边界的知识蒸馏. Saputra 等人[15] 的工作对用于回归任务的网络进行知识蒸馏有一定的实践指导价值. Mishra 等人[16] 利用知识蒸馏技术提升低精度网络的分类性能. Wonpyo 等人[17] 在 2019 年提出关系型知识蒸馏方法, 让学生网络学习由教师网络的多个输出组成的结构性信息, 包括二元的距离结构信息和三元的角度结构信息, 在蒸馏的效果上相较之前的方法有了更明显的提升.

上述方法都是在教师网络知识迁移类型上来进行优化, 并没有考虑到教师网络和学生网络层数差距过大时, 学生网络的拟合能力受到限制从而导致蒸馏效果下降的问题. 对于这个问题, Cho 等人[18] 提出提前终止教师网络的训练来使学生网络更好的拟合教师网络, 这个方法在一定程度上降低了师生差距过大导致的影响, 但也损失了部分教师网络的信息. Jin 等人[19] 的工作受课程学习启发, 提出路由约束学习, 让学生网络学习教师网络训练过程中不同阶段的状态, 由易到难不断学习和优化. Mirzadeh 等人[20] 于 2019 年引出了“助教”的概念, 即在教师网络和学生网络中间使用一个中间的蒸馏网络进行过渡, 一定程度上避免了教师网络和学生网络层数差距过大的问题. 但是这个方法仅考虑了原始的知识蒸馏方法的优化, 蒸馏得到的学生模型效果并不是十分理想.

本文针对教师网络和学生网络层数差距过大影响蒸馏效果的问题, 基于关系型知识蒸馏方法, 提出了基于关系型蒸馏的分步神经网络压缩方法. 首先训练深层的神经网络作为教师网络, 选取中间网络来学习教师网络的单体输出与关系型输出作为过渡, 随后让目标学生网络学习中间网络的单体输出与关系型知识. 通过分步的关系型蒸馏来缓解教师网络和学生网络的层数差距, 同时在每一次关系型蒸馏过程中增加单体输出来丰富学生网络的信息, 实现高层到低层的神经网络压缩.

1 关系型分步知识蒸馏

本文以关系型知识蒸馏作为基础, 通过分步的蒸馏方法有效的避免了教师网络和学生网络层数差距过大时蒸馏效果下降的问题, 以使得较低层数的学生网络也可以很好的学习到较高层数的教师网络的知识, 从而获得更高质量的学生网络.

1.1 关系型蒸馏

传统的知识蒸馏方法通常只考虑到教师网络的输出表现或在此基础上的改进, 很少考虑到教师网络的结构信息. 相比之下, 关系型蒸馏引入了教师模型的多个输出组成结构单元来让学生网络进行学习. 由于这些信息更能体现出教师模型的结构化特征, 关系型蒸馏使得学生模型可以得到更好的指导.

关系型蒸馏的整体流程如图 1.

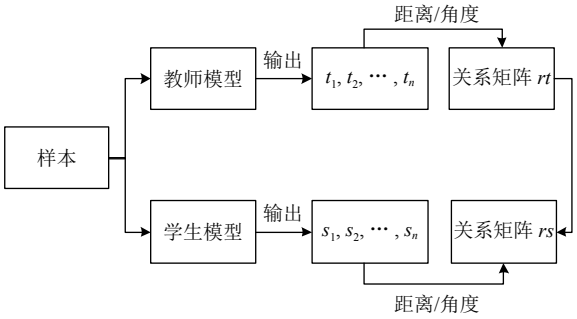

图 1 关系型蒸馏流程

步骤 1. 确定教师网络和目标学生网络的整体结构, 并用正常的训练方式训练教师网络;

步骤 2. 以教师网络的多个输出组成的结构性知识作为监督来训练学生网络;

步骤 3. 训练完成的学生网络即为目标模型.

其中, $t_1,\cdots,t_n$代表教师模型的一个批次中的多个输出, $s_1,\cdots,s_n$代表一个批次中学生模型的多个输出, 关系矩阵 rt 和 rs 分别为教师和学生模型的输出经过距离/角度关系函数变换得到的关系矩阵.

从图 1 中可以看出, 模型整体是通过使用教师网络中距离或角度信息的关系矩阵作为监督来训练学生模型从而实现知识蒸馏, 因此, 损失函数可以定义如下:

$$L_{rkd}=\sum_{(x_1,\cdots,x_n)\in X^N} l(\varphi(t_1,\cdots,t_n),\varphi(s_1,\cdots,s_n)) \tag{1}$$

其中, $x_1,\cdots,x_n$代表一个批次样本中的 n 元组, $t_1,\cdots,t_n$代表教师模型的多个输出, $s_1,\cdots,s_n$代表学生模型的多个输出, φ为给定 n 元组的关系函数, l 是教师模型的结构信息与学生模型的结构信息的损失函数. 考虑到效率和运算成本, φ关系函数的选取目前只给出两种情况, 一种是距离关系函数φ_d, 此时 n 取值为 2; 一种是角度关系函数φ_a, 此时 n 取值为 3.

当选取距离关系函数时, 将每个批次的 m 个样本分成二元组(t_i,t_j)分别计算欧几里得距离, 距离计算方法如下所示:

$$\varphi_d\left(t_i,t_j\right)=\frac{1}{\mu}\left\|t_i-t_j\right\|_2,\ i\neq j \tag{2}$$

其中, μ是距离的正则化参数, 为了更好的关联其他二元组的距离数据, μ定义如下:

$$\mu=\frac{1}{\left|X^2\right|}\sum_{(x_1,x_2)\in X^2} t_i-t_{j_2},\ i\neq j\neq k \tag{3}$$

通过两两计算距离, 可以得到一个 $m\times m$ 的距离矩阵作为关系型的距离结构信息, 矩阵中包含了批次中每一个样本到其他所有样本的距离关系. 学生模型通过学习这个关系型距离结构信息来实现蒸馏学习, 此时损失函数如下:

$$L_{rkd-d}=\sum_{(x_i,x_j)\in X^N} l_\delta(\phi_d\left(t_i,t_j\right),\phi_d\left(s_i,s_j\right)) \tag{4}$$

其中, $\left(x_i,x_j\right)$为一个批次中选取的二元组, l_δ为 Smooth L1 损失.

当选取角度关系函数时, 将每个批次的 m 个样本分成若干三元组$\left(t_i,t_j,t_k\right)$分别计算角度, 角度计算方法如下所示:

$$\begin{cases}\varphi_{\mathrm{d}}\left(t_i,t_j,t_k\right)=\cos\angle t_it_jt_k=<\mathrm{e}^{ij},\mathrm{e}^{kj}>,i\neq j\neq k\\ \text{where } \mathrm{e}^{ij}=\dfrac{t_i-t_j}{t_i-t_{j_2}},\ \mathrm{e}^{kj}=\dfrac{t_k-t_j}{t_k-t_{j_2}}\end{cases} \tag{5}$$

通过每一个三元组计算角度, 得到一个 $m\times m\times m$ 的角度矩阵作为关系型的角度结构信息, 由于角度信息是一个更高阶的属性, 因此可以更高效的传递关系信息. 学生模型通过学习这个关系型角度结构信息来实现蒸馏学习, 此时损失函数如下:

$$L_{rkd-a}=\sum_{(x_i,x_j,x_k)\in X^N} l_\delta\left(\phi_a\left(t_i,t_j,t_k\right),\phi_a\left(s_i,s_j,s_k\right)\right) \tag{6}$$

其中, $\left(x_i,x_j,x_k\right)$为一个批次中选取的三元组.

距离结构信息和角度结构信息可以通过给定权重, 共同作为监督来训练学生网络, 让学生网络学习到更丰富的结构型信息. 但是如上一小节所述, 关系型蒸馏方法同样存在教师网络和学生网络层数差距过大时损失蒸馏效果的问题, 因此本文引入了分步蒸馏的思想来进行进一步的改进.

1.2 基于关系型蒸馏的分步神经网络压缩方法

本文方法以关系型蒸馏方法为基础, 在教师网络和学生网络中间增加了一个中间网络作为过渡, 分步进行知识蒸馏, 同时在每一步的蒸馏过程中都学习了额外的单体信息. 单体信息是由教师或学生的输出首先除以一个软化系数 t[5], 再经过 Softmax 变换得到的, 软化系数 t 用来缓和教师网络的原输出, 取值越大, 输出的分布越缓和.

模型的整体训练流程如算法 1.

算法 1. 模型训练流程

输入: 多批次的图片样本$x_1,\cdots,x_n$

输出: 学生网络模型 S

步骤 1. 确定教师网络和学生网络的模型结构, 确定中间网络的层数及结构, 随后使用样本$x_1,\cdots,x_n$训练教师网络 T, 得到教师网络的输出信息$t_1,\cdots,t_n$

步骤 2. 根据式 (2) 和式 (5) 计算得到教师网络的二元距离关系矩阵 rdt 和三元角度关系矩阵 rat, 与教师网络软化的单体输出 $softt$ 协同作为监督训练中间网络 A, 得到中间网络的输出信息$a_1,\cdots,a_n$

步骤 3. 根据式 (2) 和式 (5) 计算得到中间网络的二元距离关系矩阵 rda 和三元角度关系矩阵 raa, 与中间网络软化的单体输出 $softa$ 协同作为监督训练学生网络 S

步骤 4. 训练好的学生网络 S 即为最终的目标模型

如图 2 所示, 模型整体经过了教师到中间网络和中间到学生网络的蒸馏过程. 以教师到中间网络为例, 中间网络模型在训练过程中会学习并拟合教师模型的 3 种知识, 包括教师模型软化的单个输出集合 *softt*; 教师模型输出的二元距离值组成的距离关系矩阵 *rdt*; 教师模型输出的三元角度值组成的角度关系矩阵 *rat*.

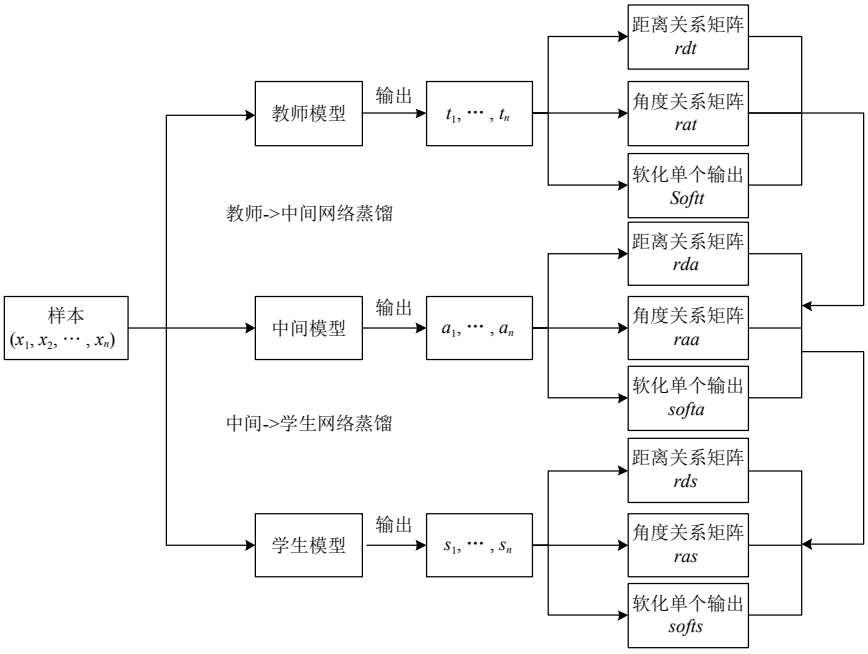

图 2 本文方法模型框架图

图 2 中, 软化输出通过交叉熵来计算损失值; 二元距离关系矩阵由教师网络的输出通过式 (2) 计算得到, 三元角度关系矩阵由教师网络的输出通过式 (5) 计算得到, 且二元距离矩阵和三元角度矩阵均使用了 Smooth L1 损失来计算损失值.

显然, 教师到中间网络蒸馏过程的整体的损失函数由 3 部分组成, 分别是单体损失、距离关系损失和角度关系损失, 具体可以表示为:

$$L_{zmrkd} = \lambda_{kd} L_{tkd} + \lambda_d L_{trkd-d} + \lambda_a L_{trkd-a} \tag{7}$$

其中, L_{zmrkd} 为教师向中间网络蒸馏的整体损失函数, L_{tkd} 为中间网络的单体输出损失, L_{trkd-d} 和 L_{trkd-a} 分别为 2.1 节中提到的中间网络的距离关系损失和角度关系损失, λ_{kd} 为单体输出权重, λ_d 为距离结构信息权重, λ_a 为角度结构信息权重.

经过教师模型到中间网络的知识蒸馏, 可以得到一个蒸馏了教师网络中各种知识的较浅的中间网络, 该中间网络的输出将作为监督进一步训练学生网络. 与中间网络训练过程类似, 学生网络模型在训练过程中也会拟合中间网络的 3 种知识, 包括中间网络软化的单个输出集合; 中间网络输出的二元距离值组成的距离关系矩阵; 中间网络输出的三元角度值组成的角度关系矩阵.

中间网络到学生网络蒸馏过程的整体损失函数具体可表示为:

$$L_{smrkd} = \lambda_{kd} L_{skd} + \lambda_d L_{srkd-d} + \lambda_a L_{srkd-a} \tag{8}$$

其中, L_{smrkd} 为中间网络向学生网络蒸馏的整体损失函数, L_{skd} 为学生网络的单体输出损失, L_{srkd-d} 和 L_{srkd-a} 分别为学生网络的距离关系损失和角度关系损失.

模型整体通过增加中间网络分步进行知识蒸馏来缓和教师网络和学生网络之间的层数差距, 同时本文在每一次蒸馏过程中都使用了教师网络的 3 种知识协同对学生网络进行指导, 学生网络不仅可以迁移获得两种关系型信息, 还可以学习到教师网络单体输出带来的直接信息, 从而使得学生网络可以更好的拟合教师网络的学习能力.

2 实验分析

本文实验所采用的数据集为 CIFAR-10 和 CIFAR-100 图像分类数据集, 其中 CIFAR-10 数据集包括 60000 张尺寸为 32×32 的彩色图像, 图像分为 10 类, 每一个

分类中都有 6000 个图像, 包括 5000 个训练图像和 1000 个测试图像; CIFAR-100 和 CIFAR-10 类似, 但是有 100 个类, 每个类包含 600 个图像, 包括 500 个训练图像和 100 个测试图像. 为了更直观的体现深度变化, 采用 ResNet 作为实验模型的主体, 其中 ResNet-110 作为教师网络模型, ResNet-20 作为目标学生网络模型, ResNet-44 作为中间网络. 实验同时将本文的方法与原始的知识蒸馏 KD (Knowledge Distillation) 模型、文献 [7] 提到的 FT (FitNet) 模型、文献 [8] 提到的 AT (Attention Transfer) 模型、文献 [9] 提到的 FSP (Fast optimization, network minimization and transfer learning) 模型、文献 [11] 提到的 NST (Neural Selective Transfer) 模型、文献 [17] 提到的 RKD (Relational Knowledge Distillation) 模型以及文献 [20] 提到的 TAKD (Knowledge Distillation via Teacher Assistant) 模型在相同参数和模型设置下做了对比实验.

2.1 实验设置

本文的模型均基于 PyTorch 实现, 初始学习率设定为 0.1, batch-size 设定为 64, epochs 设定为 100. 蒸馏过程中的参数中, KD 模型软化系数 T 为 4.0, 权重 λ_{kd}为 16, AT 模型权重设定为 2.0, RKD 模型中距离结构信息权重λ_d为 25.0, 角度结构信息权重λ_a为 50.0. ResNet-20、ResNet-110、ResNet-44 的网络结构参照文献 [21] 中实现. 受限于机器性能, 本文给出的实验数据不代表模型最佳表现, 这里只给出实验对比结果来进行模型的对照.

2.2 实验结果与分析

实验采用的主体评价指标为图像分类准确率, 同时也给出了一部分模型参数量和召回率的实验数据对比. 为了达到更好的对照效果, 模型共用的参数如软化系数 T, 均采用相同的取值.

首先, 本文首先分别用 CIFAR-10 和 CIFAR-100 数据集训练教师网络 ResNet110 和学生网络 ResNet20, 以获得未经过蒸馏的模型的准确率, 随后使用本文的方法, 训练得到蒸馏过后的 ResNet20 作为目标学生模型. 所得结果如表 1 所示.

从表 1 可以看出, 本文的方法得出的模型结构与原始 ResNet20 相同, 但是在两个数据集上的准确率均有很大的提升.

同时列出 ResNet110、ResNet44、ResNet20 及本文模型的网络层数及参数量, 如表 2 所示.

从表 2 可以看到, 本文中目标学生网络的参数量仅有 0.27 M, 远小于教师网络甚至是中间网络的参数量, 且通过表 1 的数据来看, 本文模型相对于同结构的 ResNet20, 准确率有相当大程度的提升.

表 1 初始模型与本文模型准确率 (%)

模型	CIFAR-10	CIFAR-100
ResNet110	93.26	72.61
ResNet20	91.63	68.52
本文模型	92.74	69.27

表 2 网络层数及参数量对比

模型	网络层数	参数量 (M)
ResNet20	20	0.27
ResNet44	44	0.66
ResNet110	110	1.7
本文模型	20	0.27

为了说明在教师网络和学生网络的差距过大时加入中间网络的影响, 以及在分步蒸馏过程中加入单体输出信息协同结构信息来进一步监督的影响, 本文在学生网络设定为 ResNet20 的情况下, 分别进行了教师网络为 ResNet20、ResNet50 以及 ResNet110 的关系型蒸馏实验, 同时使用 TAKD 的思路进行了关系型蒸馏的实验, 所得结果如表 3 所示.

表 3 本文模型与低层教师关系型蒸馏准确率 (%)

教师模型	CIFAR10 (学生模型)	CIFAR100 (学生模型)
ResNet20	92.53	69.09
ResNet50	92.61	69.18
ResNet110	92.56	69.04
关系型TAKD	92.68	69.19
本文模型	92.74	69.27

从表 3 可以看出, 当教师模型分别设定为 ResNet20 和 ResNet50 时, 随着教师网络自己的模型效果提升, 学生模型的准确率也有了一定的提升; 当教师模型设定为 ResNet110 时, 教师模型准确率有很大提升, 学生模型的准确率相比教师模型为 ResNet50 却有下降, 甚至跟教师模型为 ResNet20 时的准确率相仿, 因此在教师网络和学生网络差距过大时实际上完全失去了知识蒸馏的意义. 使用 TAKD 的思路在 ResNet110 和 ResNet20 中加入中间网络作为缓冲后, 学生模型准确率有了一定的提升. 本文模型在两次蒸馏过程中加入单体输出作为监督后, 可以在图像分类中获得更高的准确率, 达到更好的效果.

本文还与其他文献中给出的知识蒸馏方法的准确率进行了对比, 同时增加了 CIFAR-10 数据集的召回率

对比. 使用的教师模型均为 ResNet110, 学生模型为 ResNet20. 受表格格式限制, 使用 C-10 代表 CIFAR-10 数据集, C-100 代表 CIFAR-100 数据集, acc 代表准确率数据, rec 代表召回率数据. 所得结果如表 4 所示.

表 4 本文模型与其他蒸馏模型准确率及召回率 (%)

模型	C-10 (acc)	C-100 (acc)	C-10 (rec)
KD	92.12	69.23	91.79
FT	92.28	69.31	92.33
AT	92.53	68.97	92.27
FSP	92.31	69.13	92.16
NST	92.36	69.18	92.18
TAKD	92.33	69.25	92.37
RKD	92.56	69.04	92.49
本文	92.74	69.27	92.67

从表 4 可以看出, 本文模型在 CIFAR-10 数据集上相较于其他主流模型, 准确度和召回率均有明显提升, CIFAR-100 数据集中, 模型表现跟 FT 模型相比虽有些许差距, 但总体来说也收到了不错的蒸馏效果. 特别是相较于关系型蒸馏 RKD 模型, 本文模型在两个数据集上均有 0.2% 左右的准确度提升.

以上实验表明, 当教师网络和学生网络层数差距过大时, 通过选取中间网络分步对教师模型进行包括单体输出和结构型输出的蒸馏, 可以在让学生网络学到更多关于教师网络的信息, 迁移更丰富的知识, 进而缓解教师模型和学生模型差距过大时蒸馏效果变差的问题, 获得相比于其他大部分主流蒸馏模型更好的效果.

3 结论与展望

本文针对关系型知识蒸馏方法中教师模型与学生模型的差距过大时蒸馏效果下降的问题, 选取中间网络, 分步对教师网络模型进行关系型蒸馏, 同时在每一次蒸馏过程中都在距离结构信息和角度结构信息这两种关系型信息之外, 额外迁移了单个输出信息来丰富学生网络的学习. 实验结果表明, 本文的模型在图像分类数据集上的表现相对于原始的关系型模型以及大部分前沿的知识蒸馏模型, 分类准确率更高, 效果更好. 但本文方法在中间网络的选取以及知识迁移方式上仍有改进空间, 将在后续的研究过程进一步深化和推进.

参考文献

1 Huang G, Liu Z, van der Maaten L, *et al*. Densely connected convolutional networks. 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Honolulu: IEEE, 2017. 2261–2269.

2 Hu J, Shen L, Sun G. Squeeze-and-excitation networks. 2018 IEEE/CVF IEEE Conference on Computer Vision and Pattern Recognition. Salt Lake City: IEEE, 2018. 7132–7141.

3 Devlin J, Chang MW, Lee K, *et al*. BERT: Pre-training of deep bidirectional transformers for language understanding. Proceedings of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT) 2019. Minneapolis: Association for Computational Linguistics, 2019. 4171–4186.

4 韩云飞, 蒋同海, 马玉鹏, 等. 深度神经网络的压缩研究. 计算机应用研究, 2018, 35(10): 2894–2897, 2903. [doi: 10.3969/j.issn.1001-3695.2018.10.003]

5 Buciluǎ C, Caruana R, Niculescu-Mizil A. Model compression. Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Philadelphia: ACM, 2006. 535–541.

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

7 Romero A, Ballas N, Kahou SE, *et al*. Fitnets: Hints for thin deep nets. arXiv: 1412.6440, 2014.

8 Zagoruyko S, Komodakis N. Paying more attention to attention: Improving the performance of convolutional neural networks via attention transfer. arXiv: 1612.03928, 2017.

9 Yim J, Joo D, Bae J, *et al*. A gift from knowledge distillation: Fast optimization, network minimization and transfer learning. 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Honolulu: IEEE, 2017. 7130–7138.

10 Chen GB, Choi W, Yu X, *et al*. Learning efficient object detection models with knowledge distillation. Proceedings of the 31st International Conference on Neural Information Processing Systems. Long Beach: Curran Associates Inc., 2017. 742–751.

11 Huang ZH, Wang NY. Like what you like: Knowledge distill via neuron selectivity transfer. arXiv: 1707.01219, 2017.

12 Chen YT, Wang NY, Zhang ZX. DarkRank: Accelerating deep metric learning via cross sample similarities transfer. Proceedings of the 32nd AAAI Conference on Artificial Intelligence. New Orleans: AAAI, 2018. 2852–2859.

13 Heo B, Kim J, Yun S, *et al*. A comprehensive overhaul of feature distillation. 2019 IEEE/CVF International Conference on Computer Vision (ICCV). Seoul: IEEE, 2019. 1921–1930.

14 Heo B, Lee M, Yun S, *et al*. Knowledge transfer via

distillation of activation boundaries formed by hidden neurons. Proceedings of the 33rd AAAI Conference on Artificial Intelligence. Hawaii: AAAI, 2019. 3779–3787.

15 Saputra MRU, Gusmao P, Almalioglu Y, *et al*. Distilling knowledge from a deep pose regressor network. 2019 IEEE/CVF International Conference on Computer Vision (ICCV). Seoul: IEEE, 2019. 263–272.

16 Mishra A, Marr D. Apprentice: Using knowledge distillation techniques to improve low-precision network accuracy. arXiv: 1711.05852, 2017.

17 Park W, Kim D, Lu Y, *et al*. Relational knowledge distillation. 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Long Beach: IEEE, 2019. 3962–3971.

18 Cho JH, Hariharan B. On the efficacy of knowledge distillation. 2019 IEEE/CVF International Conference on Computer Vision (ICCV). Seoul: IEEE, 2019. 4793–4801.

19 Jin X, Peng BY, Wu YC, *et al*. Knowledge distillation via route constrained optimization. 2019 IEEE/CVF International Conference on Computer Vision (ICCV). Seoul: IEEE, 2019. 1345–1354.

20 Mirzadeh SI, Farajtabar M, Li A, *et al*. Improved knowledge distillation via teacher assistant. Proceedings of the 34th AAAI Conference on Artificial Intelligence. Palo Alto: AAAI, 2020. 5191–5198.

21 He KM, Zhang XY, Ren SQ, *et al*. Deep residual learning for image recognition. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Las Vegas: IEEE, 2016. 770–778.