

一种利用知识迁移的卷积神经网络训练策略

罗可, 周安众[†], 罗潇

(长沙理工大学 计算机与通信工程学院, 长沙 410114)

摘要: 针对深层卷积神经网络在有限标记样本下训练时存在的过拟合和梯度弥散问题, 提出一种从源模型中迁移知识训练一个深层目标模型的策略. 迁移的知识包括样本的类别分布和源模型的低层特征, 类别分布提供了样本的类间相关信息, 扩展了训练集的监督信息, 可以缓解样本不足的问题; 低层特征包含样本的局部特征, 在相关任务的迁移过程中具有一般性, 可以使目标模型跳出局部最小值区域. 利用这两部分知识对目标模型进行预训练, 能够使模型收敛到较好的位置, 之后再用真实标记样本进行微调. 实验结果表明, 所提方法能够增强模型的抗过拟合能力, 并提升预测精度.

关键词: 卷积神经网络; 知识迁移; 过拟合; 梯度弥散; 预训练; 微调

中图分类号: TP181

文献标志码: A

Convolutional neural network training strategy using knowledge transfer

LUO Ke, ZHOU An-zhong[†], LUO Xiao

(College of Computer and Communication Engineering, Changsha University of Science and Technology, Changsha 410114, China)

Abstract: To overcome the overfitting and gradient vanishing of deep convolutional neural networks trained under limited labeled samples, a strategy is proposed to transfer knowledge from a source model to a deep target model. The transferred knowledge includes class distribution of the samples and low-level features of the source model. The class distribution provides class-related information about the samples, which extends the supervised informations of the training set to alleviate the problem of inadequate samples. The low-level feature contains the local characteristics of the samples, which is general in the process of transfer knowledge, and can make the target model jump out of the local minimum value area. Then, the two parts of knowledge are applied to the pre-training target model to make the model converge to a better position, and real labeled samples are used for fine-tuning. The experimental results show that the proposed method can both improve the anti overfitting ability of the model and prediction accuracy.

Keywords: convolutional neural network; knowledge transfer; overfitting; gradient vanishing; pre-training; fine-tuning

0 引言

深度神经网络提供了一个由低层到高层的逐层特征提取框架, 在计算机视觉领域取得了极大的成功. 尤其在图像分类任务中, 采用深层结构的卷积神经网络(Convolutional neural network, CNN)^[1]的效果大大超越了传统方法. 然而, CNN训练时对大量标记样本的过度依赖一直是它的一个缺陷. 因为CNN拥有大量待优化的参数, 训练时若样本不足, 极易产生过拟合现象. 为此, 研究者们提出了众多正则化方法, 如L2正则化^[2]、Dropout^[3]等, 试图通过对参数加以限制达到优化网络的目的, 虽然取得了一定成效, 但是

不能从根本上解决样本数量缺少的问题. 为了得到更多样本, Bucila等^[4]通过筛选标记样本的特征来合成具有同类特征分布的训练样本, 但此方法在大数据集中的代价过大. 在特定领域中, CNN的卷积层提取了样本相似的特征, 这些特征在不同的任务中具有一般性, 只要改变用于分类的全连接层, 并保留特征提取模块的参数, 最后使用少量样本微调(Fine-tuning)即可用于其他相关任务^[5]. 该方法虽然降低了对样本的需求, 但限制了模型结构, 不同的任务只能更改作为分类器的全连接层, 作为特征提取部分的卷积层无法修改.

收稿日期: 2017-09-10; 修回日期: 2017-12-18.

基金项目: 国家自然科学基金项目(11671125, 71371065, 51707013).

责任编辑: 柴利.

作者简介: 罗可(1961—), 男, 教授, 博士, 从事数据挖掘、计算机应用等研究; 周安众(1986—), 男, 硕士生, 从事数据挖掘、人工智能的研究.

[†]通讯作者. E-mail: sprite4@163.com.

另一方面, CNN的深层结构使其在训练时存在梯度弥散问题^[6]. 早期的神经网络使用 sigmoid 激活函数, 存在饱和区域, 当使用随机梯度下降 (Stochastic gradient descent, SGD) 算法进行训练时, 误差从输出向输入反向传播层层递减, 靠近输出的隐藏层训练得比较好, 而靠近输入的隐藏层几乎得不到训练. Hinton等^[7]提出的深度信念网络 (Deep belief network, DBN) 改变了这一状况, 该网络先进行逐层无监督预训练, 学习每一层输入到输出的非线性变换, 然后进行有监督的微调, 但是逐层预训练的方法在层数变得更多时不再适用. 基于有监督学习的 CNN 取而代之, CNN 有效的一个原因是 ReLU (Rectified linear unit) 激活函数^[8]的使用, 其模拟了大脑稀疏性的特点, 且不存在饱和区域, 缓解了梯度弥散问题, 但是在更深层的 CNN 中, ReLU 函数的作用有限. Szegedy等^[9]通过在中间层添加 softmax 分类层以弥补梯度的损失, 相当于为中间层提供了额外的监督信号, 在深层 CNN 的训练问题中获得了较好的效果.

综上所述, 为了完成对深层模型的训练, 需要解决样本不足和梯度弥散两个问题. 本文提出一种利用知识迁移的训练策略, 从已训练的 CNN 中提取知识来对更深层的 CNN 进行预训练, 使模型收敛到较好的位置, 然后使用 SGD 算法微调, 使泛化性能获得较大提升.

1 问题描述

迁移学习^[10]的提出为学者们提供了一种新的解决问题的思路. CNN 从标记样本中获得监督信息, 自动学习出一个能够描述样本特征的非线性函数. 实际情况中, 满足深度网络需求的大数据集相当缺乏, 而为了拟合大量样本的特征, CNN 的层数在不断加深, 使得训练难度加大. 如果有一个已训练好的模型, 提取其中学到的知识并迁移到另一个 CNN 中, 不仅能减轻 CNN 对样本的依赖, 还能避免低层参数陷入局部最小值. 因此, 本文研究将一个已训练模型的知识迁移到另一个模型中的方法.

针对以上的设想, 有以下两个解决思路.

1) 在传统监督学习中, 模型利用了样本的离散类别信息, 但是很多时候类别之间有很大的相关性, 这个相关性并没有被标记出来, 从而损失了对训练很有帮助类间信息. 耿新等^[11]指出, 将样本的单标记转换成标记分布, 每个样本可以提供更多的监督信息, 相当于间接增加样本数量, 但标记分布一般需

要人工完成, 或者利用先验知识获得. 例如, 在人脸年龄估计中, 假设样本的标记年龄服从以真实年龄为中心的的高斯分布^[12]. 事实上, CNN 本身就包含类间信息, 只是在经过 softmax 输出后, 一定程度上被掩盖了. Hinton 对已训练的 CNN 使用知识提取 (Knowledge distillation, KD) 算法^[13], 可以得到包含样本类间信息的类别分布, 称为软目标 (Soft target), 用来训练简单的 CNN 模型. Tang 等^[14]的研究发现, KD 算法不仅可以训练简单模型, 还可以训练相关领域的复杂模型. 基于此, 本文提出一种采用软目标的预训练 (Pre-training with soft target, PST) 策略, 从源模型中提取软目标迁移到同领域的目标模型中, 可以从有限的样本中获得比单个标记更多的监督信息, 解决样本缺少的问题.

2) 深度学习思想旨在由低层到高层的特征组合来学习特征的层次结构. 然而, 在深度学习中使用目标函数是一个高度非凸函数, 存在多个局部最小值, 而且由于梯度弥散问题的存在, CNN 的前几层往往得不到很好的训练, 无法跳出局部最小值区域. 研究发现^[15-16], 以 CNN 的中间层作为学习目标来辅助训练深层模型, 可以缓解梯度弥散. 因此, 本文提出一种逐模块训练 (Module by module training, MMT) 策略, 目标模型相邻卷积层划分为一个模块, 以模块的方式学习源模型的低层特征, 类似于 DBN 的逐层预训练策略, 并将 MMT 与 PST 结合, 对样本类间信息和低层特征这两种知识同时进行迁移.

综合以上讨论, 本文利用知识迁移的训练策略主要有以下研究内容: 1) 确定源模型中哪些知识能够有效迁移, 即类别分布可以提供更多的监督信息, 低层特征可以指导目标模型学习, 使参数收敛到较好的位置. 2) 以预训练的方式进行知识的迁移, 然后再对模型微调, 不仅能利用源模型从训练中学到的知识, 同时也发挥了目标模型自身的性能. 3) 采用模块的方式将源模型转换成更深层的目标模型, 不仅提升了目标模型性能, 且更适合学习源模型迁移过来的知识.

2 相关基础

2.1 知识提取

CNN 使用 softmax 函数进行分类, 利用交叉熵损失函数最大化真实类别的概率, 而非真实类别的概率被最小化接近于 0, 相当于掩盖了这部分信息. Ba 等^[17]指出, CNN 中这部分掩盖的信息包含在 softmax 函数的输入信息中, 称为 logit. 通过让目标模型学习

源模型的logit特征,可以利用更多的样本类别信息.训练时的损失函数使用平方误差函数如下:

$$L_{\text{logit}}(w) = \frac{1}{2} \sum_t \|f(x^t, w) - z^t\|_2^2. \quad (1)$$

其中: x 是隐藏层到logit的输入, w 是模型的参数, f 是目标模型训练时输出的logit, z 是源模型输出的logit, t 是训练样本的数量.

在此基础上, Hinton提出了更具普适性的KD算法,通过在softmax中引入一个温度参数 T ,产生更平滑的类别输出,即软目标,如下所示:

$$q_i(w) = \frac{\exp(z_i/T)}{\sum_j \exp(z_j/T)}. \quad (2)$$

得到源模型输出的软目标后,以目标模型训练时输出的软目标 q 和从源模型提取的软目标 p 的交叉熵作为损失函数

$$L = - \sum_i p_i \ln q_i. \quad (3)$$

对以上损失函数求导的公式为

$$\frac{\partial L}{\partial z_i} = \frac{1}{T} (q_i - p_i) = \frac{1}{T} \left[\frac{\exp(z_i/T)}{\sum_j \exp(z_j/T)} - \frac{\exp(v_i/T)}{\sum_j \exp(v_j/T)} \right]. \quad (4)$$

其中: z_i 为目标模型logit的某一维度,经过softmax输出的概率为 q_i , v_i 为源模型logit的对应维度,经过softmax输出的概率为 p_i . 当 T 取较大值时,上式可以近似为

$$\frac{\partial L}{\partial z_i} \approx \frac{1}{T} \left(\frac{1 + z_i/T}{N + \sum_j z_j/T} - \frac{1 + v_i/T}{N + \sum_j v_j/T} \right). \quad (5)$$

当logit的取值服从0均值分布的时候,即 $\sum_j z_j = \sum_j v_j = 0$, 上述公式可进一步简化为

$$\frac{\partial L}{\partial z_i} \approx \frac{1}{NT^2} (z_i - v_i). \quad (6)$$

在该情形下,目标函数变成了优化 z_i 与 v_i 之间的平方误差,与最小化公式(1)的损失函数是一致的.因此,对logit的学习实际上是KD算法的一个特殊情况,后者有更好的适用性.

2.2 特征选择

迁移学习中常用一种基于特征选择的迁移方法^[18],该方法首先识别出源领域与目标领域中共有的特征,然后利用这些特征进行知识迁移. CNN中较低层输出的特征正好具有这种特点,在相关的任务中

具有一般性. 文献[16]在训练深层CNN时,从目标模型中选择出特定卷积层,利用一个回归器去学习从源模型中选择出的卷积层的输出. 回归器在CNN中是一个额外添加的卷积层,能够使目标模型输出特征图的大小和数量与源模型输出特征图的大小和数量相一致,然后使用平方误差损失函数进行优化:

$$L(w_t) = \frac{1}{2} \|u_s(x, w_s) - r(v_t(x, w_t), w_r)\|_2^2. \quad (7)$$

其中: u_s 、 v_t 分别是源模型和目标模型从输入 x 到选择出的卷积层之间的非线性函数, w_s 、 w_t 分别是源模型与目标模型的参数, r 是回归器所表示的非线性函数.

3 训练策略

为了将知识进行迁移,首先基于源模型改进得到深层的目标模型,然后从源模型提取知识对目标模型进行两次预训练,最后用真实标记样本微调. 图1是整个流程的示意图. 为此,本文提出MMT和PST训练策略来完成知识的迁移.

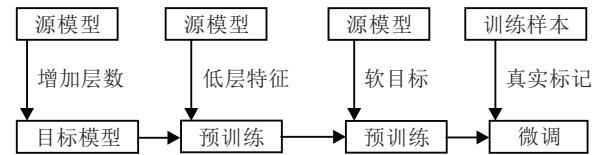


图1 利用知识迁移的训练策略示意图

3.1 MMT训练策略

训练深层CNN模型时,随着层数的增加,梯度弥散问题开始显现. 2.2节介绍的特征选择方法为本文提供了帮助,通过回归器让目标模型学习源模型中的特征,即卷积层输出的特征图,使其跳出局部最小值区域,可以收敛到较好的位置. 但是目标模型有更深的结构,与源模型的结构相差太大不利于知识迁移^[19],回归器的学习效果不理想. 为此,将目标模型基于源模型的结构进行加深,并采用模块化的方式学习源模型的特征,称为MMT策略. 首先将源模型以池化层为界划分为多个不同的模块,然后利用文献[9]的思想,将每个模块中的卷积核尺寸减小,并增加卷积层的层数,以此得到深层的目标模型,卷积层用于特征提取,池化层用于降维,如图2所示.

CNN中卷积层的计算公式为

$$x_j^l = f \left(\sum_{i \in M_j} x_i^{l-1} \otimes k_{ij}^l + b_j^l \right). \quad (8)$$

其中:等号左边 x_j^l 表示第 l 层输出的第 j 个特征图;等号右边表示与第 l 层的卷积核 k_{ij}^l 关联的所有 $l-1$ 层特征图 x_i^{l-1} 的卷积运算并求和,然后加上偏置参数 b_j^l ,最后通过 f 激活函数将输入转换为输出的特征图.

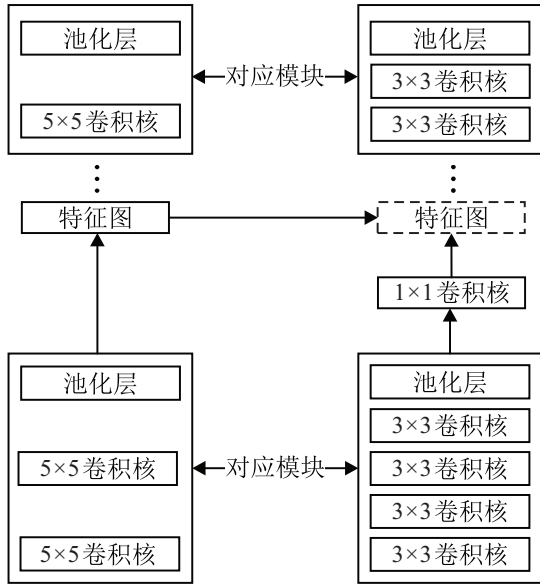


图2 源模型与目标模型对应模块学习示意图

源模型每个模块中包含了多个由式(8)组成的卷积运算,这种多层卷积运算被目标模型中更深层的卷积运算所代替,在深度学习中已证明^[17],深层的运算可以提取更加抽象的特征,有利于最后的分类。图2中,目标模型每个5×5卷积核的尺寸减小为3×3,并增加了层数,为了使参数数量保持基本不变,适当减少卷积核的数量,因此目标模型相较于源模型变得更深且更窄。卷积核尺寸和数量的变化使输出特征图的尺寸和数量也发生了变化,特征图大小的计算方法为

$$w' = (w + \text{pad} \times 2 - k_{\text{size}}) / \text{stride} + 1, \quad (9)$$

$$h' = (h + \text{pad} \times 2 - k_{\text{size}}) / \text{stride} + 1. \quad (10)$$

其中: w 、 h 分别是输入特征图的宽与高, k_{size} 是卷积核大小, stride 是卷积核步长, pad 是填充值, 输出特征图的宽和高分别是 w' 和 h' 。

根据式(9)和(10), 假设输入特征图大小为 28×28 , 填充值为1, 卷积核大小为 5×5 , 步长为1, 计算可得输出特征图大小为 26×26 。当用 3×3 的卷积核代替时, 选择填充值为0, 步长为1, 计算可得输出特征图大小仍然为 26×26 , 池化层同理。可见, 通过设置填充值和卷积核的步长可以保证目标模型每个模块输出的特征图与源模型中对应模块输出的特征图在尺寸上保持一致。最后根据文献[9], 采用 1×1 卷积核降维的思想, 在目标模型的模块后添加一个由 1×1 卷积核组成的卷积层, 用于增加模块输出的特征图数量, 使其与源模型中对应模块输出的特征图数量相同。

经过上述处理, 目标模型得到深层结构, 而且与源模型对应模块输出的特征图在尺寸和数量上保持一致, 减小了学习难度。

1×1 卷积核的加入不仅起到改变特征图数量的目的, 还具有更强的特征提取能力^[20], 使模块的学习能力得到增强。训练完之后需要从目标模型移除所有加入的 1×1 卷积核。与式(7)类似, 特征图的学习使用平方误差损失函数最小化对应模块输出的特征图之间的距离:

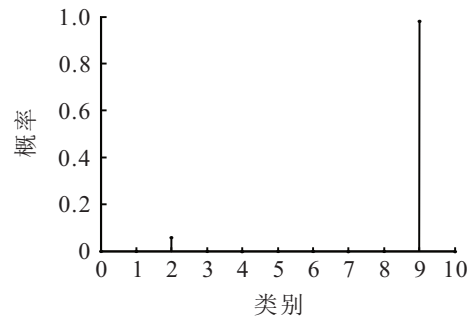
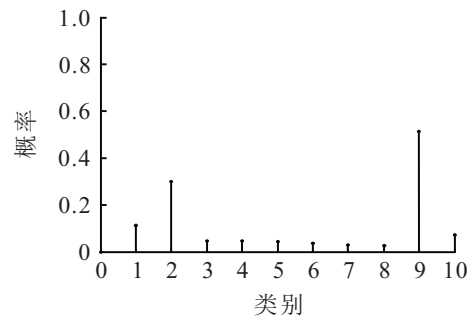
$$L_{\text{module}}(w) = \frac{1}{2} \|f(x, w_s) - r(g(x, w_t), w_r)\|_2^2. \quad (11)$$

其中: f 是源模型从输入到某个模块输出的非线性变换, g 是目标模型从输入到对应模块的非线性变换, r 是加入的 1×1 卷积核的变换。

训练时从最低层开始以逐个模块的方式学习, 且只需要训练模型较低层的几个模块。模型最后微调时, 靠近分类器的较高层由损失函数传递的梯度较大, 可以得到更多的调整。较低层得到的梯度较小, 但经过了逐模块的特征学习, 加之低层特征的一般性, 使其不需要有太大改变。同时, 以深层模块学习浅层模块的特征, 使特征的表达能力得到增强, 符合深度学习的思想。

3.2 PST训练策略

CNN 包含的样本类间信息可以通过在 softmax 函数中设置合适的温度参数 T , 以软目标的方法提取出来。 T 的取值越大, 输出的软目标越平滑, 如图3所示。当增大 T 的取值时, CNN 输出了样本的类别分布^[21], 这个分布是模型从样本中学习到的类间相关信息, 相当于扩展了类别标记, 提供了更多的监督信息。

(a) $T=1$ (b) $T=10$ 图3 T 取不同值时的softmax输出情况

在KD算法中,从源模型提取软目标后,与真实标记一起训练目标模型,真实标记用来修正软目标.损失函数如下所示:

$$L(w) = L_s(w) + \lambda L_h(w). \quad (12)$$

训练时,对式(12)右边的两个损失函数设置权重系数 λ .第1个损失函数是软目标的交叉熵,第2个损失函数是真实标签的交叉熵,且第2个损失函数的权重较低时得到的结果更好.

软目标形式的概率输出提取自源模型,并非样本的真实类别分布,源模型性能越好,软目标的误差也越小,这使得目标模型非常依赖于源模型的性能.为此,本文提出改进的策略,将软目标以预训练的方式迁移到目标模型,然后用真实标记样本进行微调.这样不仅迁移了源模型输出的软目标知识,又最大限度地利用了目标模型自身的深层特征提取能力.具体步骤如下.

Step 1:将源模型的softmax函数的参数 T 设定为较大值,从而得到输出的每个样本的软目标;

Step 2:将软目标作为样本的标签,通过SGD算法训练目标模型,此时目标模型的 T 值与源模型输出软目标时的 T 值保持一致;

Step 3:目标模型预训练完后,再将 T 设置为1,用真实标记样本微调模型.

预训练的策略有两个好处:一是利用了样本的类间相关信息,弥补了训练样本监督信息不足的缺点;二是对源模型的依赖更少,因为最终要通过真实标记样本微调,使目标模型更加可靠.

3.3 结合MMT和PST的训练策略

模型最后的训练策略结合了MMT和PST.首先从模型中较低层开始逐模块预训练,一般情况下训练最低层的两个模块;然后再利用软目标预训练整个目标模型;最后利用真实标记样本微调.具体步骤如下.

Step 1:将源模型的卷积核尺寸减小并增加层数,形成的深层CNN作为目标模型,并以池化层为界划分为不同模块;

Step 2:源模型与目标模型对应的低层模块进行逐模块地特征学习;

Step 3:将源模型softmax函数的温度参数 T 设为较大的值,并提取样本的软目标;

Step 4:目标模型softmax函数的温度参数 T 设置为与源模型相同的值,以Step3提取的软目标作为样本的标记对目标模型进行预训练;

Step 5:将目标模型softmax函数的温度参数 T 设置为1,以真实标记样本微调.

4 实验与结果分析

为了评估本文提出的训练策略对深度模型的性能影响,使用MNIST^[22]和CIFAR-10/100^[23]数据集进行验证.采用MatConvNet深度学习框架,GPU为GTX950,这样可以大大提高训练效率.使用均值为0,方差0.01的高斯分布对模型随机初始化,总共训练5个模型,采用多数投票机制,即选择大多数模型的分类值作为最后的结果.实验不仅比较了SGD、KD和本文方法的效果,还对比了不同深度的CNN对结果的影响.参数 T 的取值在MNIST上为20,在CIFAR-10/100上为10.

4.1 MNIST数据集上的实验结果与分析

MNIST数据集由大小为 28×28 的手写数字图片组成,包含了60 000张训练图片和10 000张测试图片.由于LeNet模型^[22]在MNIST数据集的分类任务中取得了较好的效果,本文采用的源模型基于LeNet改进,称为网络A;将sigmoid激活函数替换为ReLU激活函数,并增加了卷积核的数量,参数总数为43万.目标模型根据3.1节的方法改进而来,参数总数45万,称为网络B.模型训练时共迭代50次,学习率为0.05,MMT策略应用于模块1和模块2.表1所示是两个网络各层的卷积核尺寸和数量取值情况.

表1 网络A和网络B的各层卷积核尺寸和数量

模块	网络A	网络B
1	$5 \times 5 \times 20$	$5 \times 5 \times 20$
		$3 \times 3 \times 20$
		$3 \times 3 \times 40$
2	$5 \times 5 \times 20$	$3 \times 3 \times 80$
		$3 \times 3 \times 100$
3	$4 \times 4 \times 500$	$3 \times 3 \times 200$
		$2 \times 2 \times 500$

表2显示了在MNIST数据集上的实验结果.训练集大小采用随机选择部分训练样本的方式获得.网络B在只采用50%训练集以及使用KD算法训练后,测试效果接近于采用全部训练集的SGD算法,而SGD算法采用50%训练集时测试误差上升明显.可见,软目标弥补了样本数量不足的缺点,而本文方法取得了比KD算法更好的效果,在使用全部训练集的情况下优于Dropout和FitNet模型,且使用本文方法训练后,Dropout和FitNet模型的测试误差得到进一步降低,验证了本文所提出的训练策略可适用于多种不同的模型.

表2 MNIST数据集上的测试误差

模型(方法)	训练集大小/%	测试误差/%
网络A(SGD)	100	0.92
网络B(SGD)	100	0.82
网络B(SGD)	50	1.12
网络B(KD)	100	0.77
网络B(KD)	50	0.85
网络B(本文方法)	100	0.50
网络B(本文方法)	50	0.61
Dropout(文献[3])	100	0.79
Dropout(本文方法)	100	0.55
FitNet(文献[16])	100	0.51
FitNet(本文方法)	100	0.42

4.2 CIFAR-10/100数据集上的实验结果与分析

CIFAR-10 包含 60 000 张 32×32 的 RGB 图片,共 10 个类别. 训练数据 50 000 张图片(每类 5000 张),测试数据 10 000 张图片. CIFAR-100 数据集与 CIFAR-10 类似,样本数量相同,共 100 个类别(每类 500 张),测试数据 10 000 张. 由于类别较多,每个类别的样本很少, CIFAR-100 数据集所提供的类别信息很有限.

以文献[20]介绍的 NiN(Network in Network)网络结构作为源模型,记为网络 C,分为 3 个模块,每个模块包含一个卷积层和由 1×1 卷积核组成的多层感知器(MultiLayer perceptron, MLP)^[20],参数总数 97 万. 目标模型有网络 D 和网络 E,根据 3.1 节的方法改进而来. 网络 C 包含 MLP,虽然对性能提升有很大帮助,但是会增加大量的计算量,于是网络 D 去掉了 MLP,参数总数 99 万. 为了在 CIFAR-100 数据集上获得更好的效果,网络 E 使用了更深的结构,参数总数 97 万. MMT 策略应用于目标模型的第 1、第 2 模块. 训练时, CIFAR-10 数据集迭代 100 次, CIFAR-100 数据集迭代 300 次,学习率 0.5,每迭代 20 次下降 10 倍. 各网络的卷积核尺寸和数量情况如表 3 所示.

表3 网络C、D和E的各层卷积核尺寸和数量

模块	网络C	网络D	网络E
1	5×5×192 MLP	5×5×64 3×3×96 3×3×96	5×5×32
			3×3×32
			3×3×32
			3×3×64
2	5×5×192 MLP	3×3×128 3×3×128 3×3×128	3×3×96
			3×3×96
			3×3×96
			3×3×96
3	3×3×192 MLP	3×3×192 3×3×192	3×3×128
			3×3×128
			3×3×128
			2×2×128

表 4 显示了在 CIFAR-10/100 数据集上的实验结果. 网络 D 虽然有深层的结构,但去掉了 MLP 模块,在使用 SGD 算法时,测试误差要比网络 C 高. 因为 KD 算法更适合浅层结构,没有改善其性能,但在没有使用数据扩充的情况下与 SGD 算法接近,而由本文方法训练的网络 D 在同样没有数据扩充的情况下,测试误差明显降低,可见利用类别分布信息在一定程度上缓解了样本缺少的问题. 网络 E 的深层结构造成训练困难,使用 SGD 算法达不到可观的效果,而本文方法采用了 MMT 策略,更适合深层模型,效果提升明显. 在使用数据扩充的情况下,优于 Dropout 和 NiN 模型的结果,与采用了更深层结构的 FitNet 模型的结果接近. Dropout 因为本身的浅层结构,使用了本文方法后效果提升不明显;深层结构的 NiN 模型使用本文方法后精度有较大的提升;而 FitNet 模型的训练策略中也采用了低层模型的预训练,采用本文方法后相当于增加了一次样本扩充,效果有一定的提升.

表4 CIFAR-10/100数据集上的测试误差

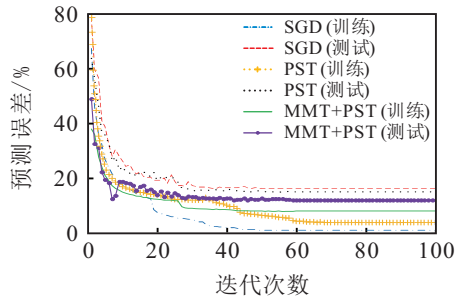
模型(方法)	数据扩充	CIFAR-10/%	CIFAR-100/%
网络C(SGD)	是	12.16	38.96
网络D(SGD)	是	14.29	40.58
网络D(KD)	否	14.41	40.60
网络D(本文方法)	否	11.50	37.31
网络E(SGD)	是	11.41	36.92
网络E(KD)	否	12.88	37.76
网络E(本文方法)	否	9.96	36.79
网络E(本文方法)	是	8.53	35.09
Dropout(文献[3])	是	11.68	37.20
Dropout(本文方法)	是	11.13	36.82
NiN(文献[20])	是	8.81	35.68
NiN(本文方法)	是	7.02	33.88
FitNet(文献[16])	是	8.39	35.04
FitNet(本文方法)	是	7.96	34.26

4.3 预训练的影响

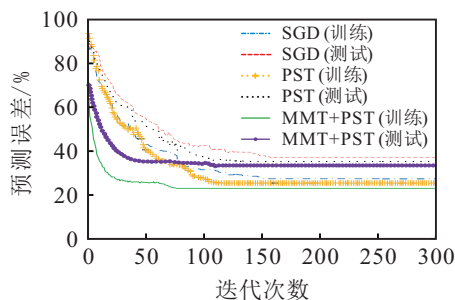
为了比较预训练策略对模型的影响,实验分别使用 PST 和 MMT+PST 策略预训练后,再以真实标记样本微调来观察模型的收敛情况,并与 SGD 算法进行了对比.

如图 4 所示,网络 D 在 CIFAR-10 上以 SGD 算法训练时,训练误差接近 0,但测试误差仍然很高,这是典型的由于数据不足所产生的过拟合现象. 经过 PST 预训练后,训练误差有所升高,同时测试误差得到降低,因为 PST 提取了软目标,为模型的训练提供了更多的监督信息,缓解了过拟合现象. 结合 MMT 与 PST 预训练后,测试误差达到最优. 具有深层结构的网络 E 在 CIFAR-100 上以 SGD 训练时,因为低层参数难以训练,导致收敛缓慢且达不到较高的精度. PST 策略扩展了样本的监督信息,可以在一定程度上降低误

差,结合MMT与PST策略后,低层参数经过预训练已收敛到较好的位置,迭代不到100次就已达到较平稳的状态,误差也得到了进一步降低,可见对低层模块预训练使参数更容易跳出局部最小值区域,收敛到一个更好的解。



(a) 网络D在CIFAR-10上的收敛结果

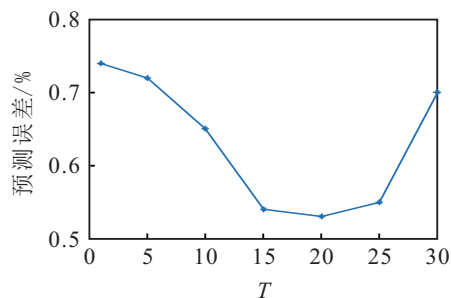


(b) 网络E在CIFAR-100上的收敛结果

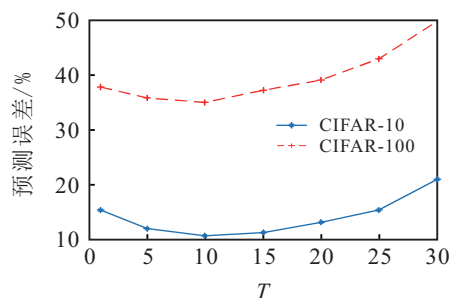
图4 不同预训练方法对结果的影响

4.4 温度参数的影响

为了观察参数 T 对训练的影响,图5显示了 T 的取值变化在MNIST和CIFAR10/100数据集上的测试误差。



(a) T 取不同值时在MNIST数据集上的结果



(b) T 取不同值时在CIFAR-10/100数据集上的结果

图5 T 取不同值时目标模型的性能变化

从图5可以看出,在MNIST数据集上, T 取20的效果最好,而在CIFAR-10/100数据集上, T 取10的效果最好。可见在不同的任务中, T 的最佳取值会有差别。因为softmax趋向于最大化真实类别的概率,错误类别的概率被抑制,如果源模型对某个数据集的分类准确率很高,则需要更大的 T 值才能提取出被掩盖的信息,也即样本类别之间的相关信息。 T 值的选取更多是依靠实验或是经验判断。

5 结论

本文提出了一种提取源模型的知识并迁移到目标模型的训练策略。该策略将已训练模型的知识代入到数据不足、难以训练的更深层的CNN模型,使预测精度获得较大提升。实验结果表明:通过预训练较低层的模块,可以使参数跳出局部最小值区域,收敛到一个较好的位置;从源模型提取的软目标扩展了样本的监督信息,一定程度上缓解了样本不足的问题;在不同的任务中,调节 T 的取值使找到的软目标越接近样本真实类别分布,对训练越有帮助。由于软目标是通过在softmax函数中设置一个 T 参数得到的,而softmax主要用于分类任务,在适用性上有所限制,今后的研究可以进一步寻找更通用的方法,扩大知识提取的适用范围。

参考文献(References)

- [1] Krizhevsky A, Sutskever I, Hinton G. ImageNet classification with deep convolutional neural networks[C]. Int Conf on Neural Information Processing Systems 2012. Harrah: Curran Associates Inc, 2012: 1097-1105.
- [2] Schmidhuber J. Deep Learning in neural networks: An overview[J]. Neural Networks, 2015, 61(1): 85-117.
- [3] Srivastava N, Hinton G, Krizhevsky A, et al. Dropout: A simple way to prevent neural networks from overfitting[J]. The J of Machine Learning Research, 2014, 15(1): 1929-1958.
- [4] Bucila C, Caruana R, Niculescu A. Model compression[C]. Proc of the 12th ACM SIGKDD Int Conf on Knowledge Discovery and Data Mining. Philadelphia: ACM, 2006: 535-541.
- [5] Girshick R, Donahue J, Darrell T, et al. Rich feature hierarchies for accurate object detection and semantic segmentation[C]. IEEE Conf on Computer Vision and Pattern Recognition. Columbus: IEEE Computer Society, 2014: 580-587.
- [6] Erhan D, Manzagol P A, Bengio Y, et al. The difficulty of training deep architectures and the effect of unsupervised pre-training[J]. Immunology of Fungal Infections, 2009,

- 5(1): 153-160.
- [7] Hinton G E, Osindero S. A fast learning algorithm for deep belief nets[J]. *Neural Computation*, 2006, 18(7): 1527-1554.
- [8] Glorot X, Bordes A, Bengio Y. Deep sparse rectifier neural networks[C]. *Proc of the 14th Int Conf on Artificial Intelligence and Statistics*. Florida: JMLR Proceedings, 2011: 315-323.
- [9] Szegedy C, Liu W, Jia Y, et al. Going deeper with convolutions[C]. *Proc of the IEEE Conf on Computer Vision and Pattern Recognition*. Boston: IEEE, 2015: 1-9.
- [10] 许敏, 王士同, 顾鑫. TL-SVM: 一种迁移学习算法[J]. *控制与决策*, 2014, 29(1): 141-146.
(Xu M, Wang S T, Gu X. TL-SVM: A transfer learning algorithm[J]. *Control and Decision*, 2014, 29(1): 141-146.)
- [11] 耿新, 徐宁, 邵瑞枫. 面向标记分布学习的标记增强[J]. *计算机研究与发展*, 2017, 54(6): 1171-1184.
(Geng X, Xu N, Shao R F. Label enhancement for label distribution learning[J]. *J of Computer Research and Development*, 2017, 54(6): 1171-1184.)
- [12] Geng X, Yin C, Zhou Z H. Facial age estimation by learning from label distributions[J]. *IEEE Trans on Pattern Analysis & Machine Intelligence*, 2013, 35(10): 2401-2412.
- [13] Hinton G, Vinyals O, Dean J. Distilling the knowledge in a neural network[J]. *Computer Science*, 2015, 14(7): 38-39.
- [14] Tang Z, Wang D, Zhang Z. Recurrent neural network training with dark knowledge transfer[C]. *The IEEE Int Conf on Acoustics, Speech and Signal Processing*. Brisbane: IEEE, 2015: 5900-5904.
- [15] Gulcehre C, Bengio Y. Knowledge matters: Importance of prior information for optimization[J]. *J of Machine Learning Research*, 2016, 17(8): 1-32.
- [16] Romero A, Ballas N, Kahou S E, et al. Fitnets: Hints for thin deep nets[C]. *Int Conf Learning Representations 2015*. San Diego: Arxiv-prints, 2015: 1412-1550.
- [17] Ba J, Caruana R. Do deep nets really need to be deep[C]. *Advances in Neural Information Processing Systems*. Montreal: NIPS, 2013: 2654-2662.
- [18] 庄福振, 罗平, 何清, 等. 迁移学习研究进展[J]. *软件学报*, 2015, 26(1): 26-39.
(Zhuang F Z, Luo P, He Q, et al. Survey on transfer learning research[J]. *J of Software*, 2015, 26(1): 26-39.)
- [19] Soekhoe D, Putten D, Laat A. On the impact of data set size in transfer learning using deep neural networks[C]. *Advances in Intelligent Data Analysis XV*. Sweden: Springer Int Publishing, 2016: 50-60.
- [20] Lin M, Chen Q, Yan S. Network in network[C]. *Int Conf on Learning Representations*. Banff: Arxiv-prints, 2014: 1312-4400.
- [21] Geng X. Label distribution learning[J]. *IEEE Trans on Knowledge and Data Engineering*, 2016, 28(7): 1734-1748.
- [22] Lecun Y, Bottou L, Bengio Y, et al. Gradient-based learning applied to document recognition[J]. *Proc of the IEEE*, 1998, 86(11): 2278-2324.
- [23] Krizhevsky A. Learning multiple layers of features from tiny images[R]. Toronto: University of Toronto, 2009.

(责任编辑: 齐 霖)

下 期 要 目

- 不平衡数据分类方法综述 李艳霞, 等
- 求解存在运输空间约束多单元协作调度问题的拍卖算法 曾程宽, 等
- 考虑缓冲区故障的多产品生产系统的性能分析 杨喜娟, 等
- 时滞仿射线性参数变量系统的有记忆 H_∞ 状态反馈控制 潘晓真, 等
- 基于三支决策的主动学习方法 胡 峰, 等
- 一类互联非线性系统的分布式故障诊断观测器设计 夏静萍, 等
- 考虑执行器性能约束的刚体航天器鲁棒姿态跟踪控制 陈海涛, 等
- 平板导电结构缺陷脉冲涡流和超声复合检测方法 黄平捷, 等
- 基于改进烟花算法的随机装配线混流调度 刘俨后, 等
- 融合改进蚁狮算法和 T-S 模糊模型的噪声非线性系统辨识 赵小国, 等