第47卷第10期 中国科学技术大学学报 Vol.47,No.10
2017年10月 JOURNAL OF UNIVERSITY OF SCIENCE AND TECHNOLOGY OF CHINA Oct. 2017

文章编号:0253-2778(2017)10-0799-09

集成最大汇合:最大汇合时只有最大值有用吗

张 皓[1,2],吴建鑫[1,2]

(1.计算机软件新技术国家重点实验室,江苏南京 210023;2.南京大学计算机科学与技术系,江苏南京 210023)

摘要:卷积神经网络中的汇合层基于局部相关性原理进行亚采样,在减少数据量的同时保留有用信息,从而有助于提升泛化能力.同时,汇合层可以有效提高感受野.经典的最大汇合采用赢者通吃策略,这有时会影响网络的泛化能力.为此提出集成最大汇合,用于替代传统卷积神经网络中的汇合层.在每个局部汇合区域,集成最大汇合以 p 的概率使输出最大的神经元失活,激活输出第二大的神经元.集成最大汇合可以看作多个基础潜在网络的集成,也可以理解为一种输入经历一定局部形变下的经典最大汇合过程.实验结果表明,相比经典汇合方法及其他相关汇合方法,集成最大汇合取得了更好的性能.DFN-MR 是近期主流结构 ResNet 的一个衍生,相比 ResNet,DFN-MR 有着更多的基础潜在网络数目,同时避免了极深网络.保持其他超参数不变,通过将 DFN-MR 中步长为2的卷积层改为集成最大汇合串联步长为1的卷积层的结构,可以使网络性能得到显著提高.

关键词:卷积神经网络;汇合层;网络集成;数据扩充

中图分类号:TP183 **文献标识码**:A doi:10.3969/j.issn.0253-2778.2017.10.001

引用格式:张皓,吴建鑫.集成最大汇合:最大汇合时只有最大值有用吗[J].中国科学技术大学学报,2017,47(10):799-807.

ZHANG Hao, WU Jianxin. Ensemble max-pooling: Is only the maximum activation useful when pooling[J]. Journal of University of Science and Technology of China, 2017,47(10):799-807.

Ensemble max-pooling: Is only the maximum activation useful when pooling

ZHANG Hao[1,2], WU Jianxin[1,2]

(1. *National Key Laboratory for Novel Software Technology, Nanjing* 210023, *China*;
2. *Department of Computer Science and Technology, Nanjing University, Nanjing* 210023, *China*)

Abstract: The pooling layer in convolutional neural networks performs subsampling on the basis of the local correlation principle, reducing the data size while keeping useful information in order to improve generalization, and effectively increase receptive fields simultaneously. The winner-take-all strategy is used in classical max-pooling, which will affect the generalization of the network sometimes. A simple and effective pooling method named ensemble max-pooling was introduced, which can replace the pooling layer in conventional convolutional neural networks. In each pooling region, ensemble max-pooling drops the neuron with maximum activation with probability p, and outputs the neuron with second largest activation. Ensemble max-pooling can be viewed as an ensemble of many basic underlying networks, and it

收稿日期:2017-05-22;**修回日期**:2017-06-24

作者简介:张皓,男,1990年生,博士生,研究方向:机器学习与计算机视觉.E-mail:zhangh@lamda.nju.edu.cn

通讯作者:吴建鑫,博士/教授.E-mail:wujx2001@nju.edu.cn

can also be viewed as the classical max-pooling with some local distortion of the input. The results achieved are better than classical pooling methods and other related pooling approaches. DFN-MR is derived from ResNet, compared with which it has more basic underlying networks and avoids very deep networks. By keeping other hyperparameters unchanged, and replacing each convolutional layer in DFN-MR with a tandem form, i.e., a combination of an ensemble max-pooling layer and a convolutional layer with stride 1, it is shown to deliver significant gains in performance.

Key words: convolutional neural network; pooling layer; network ensemble; data augmentation

0 引言

卷积神经网络(convolutional neural networks)在计算机视觉及其他一些领域取得了显著的成果.一个经典的卷积神经网络模型由卷积层、汇合层(pooling layer)、全连接层、非线性激活函数交替组合而成.其中,汇合层基于局部相关性原理进行亚采样,在减少数据量的同时保留有用信息,从而有助于提升泛化能力.同时,汇合层可以有效提高感受野(receptive field).

经典的最大汇合是一个确定性的过程.每个局部汇合区域数值最大的神经元被选为输出.更新时,只有数值最大的神经元得到后一层向前传播的梯度,而局部汇合区域内其他的神经元则被忽略.当训练数据有限时,这种赢者通吃(winner-take-all)的策略可能会使网络不能很好地泛化到测试集[1-3].

本文提出一种新的用于克服赢者通吃弊端的汇合方法,称为集成最大汇合.与经典最大汇合方法相比,集成最大汇合还额外利用了每个局部汇合区域的非最大值的信息.在每个局部汇合区域,集成最大汇合以概率 p 输出第二大的神经元.训练时,在每个局部汇合区域,输出是从伯努利分布中采样得到的.

集成最大汇合可以看作多个基础潜在网络的集成.训练时的每次迭代会在各局部汇合区域选不同的神经元,这相当于改变网络的连接结构,定义出一个新的基础潜在网络;当多层集成最大汇合层堆叠时,可能的基础潜在网络数目以指数级别增加.集成最大汇合也可以理解为一种输入经历一定局部形变下的经典最大汇合.当多层集成最大汇合层堆叠时,这样的局部形变数目以指数级别增加.

近期,基于跳跃连接的 ResNet 网络结构[4]成为当前主流的网络结构之一,ResNet 可以看作多个不同深度的基础潜在网络的集成[5],每个基础潜在网络是一个单分支的网络.DFN-MR(deeply fused networks-merge and run)是基于 ResNet 设计得到的,其在保证基础潜在网络数目足够多的同时去掉了对总体性能贡献不大的极深的基础潜在网络,取得了比 ResNet 更好的效果[6].ResNet 和 DFN-MR 中都没有使用汇合层,它们用步长为 2 的卷积层来降低空间大小.本文向 DFN-MR 中引入集成最大汇合,并通过实验检验其效果.

1 符号表示及经典汇合方法

本文使用 $\{x_1,x_2,\cdots,x_n\}$ 表示汇合层中一个局部汇合区域的输入,使用 $a\in R$ 表示一个局部汇合区域的输出.当我们使用经典的 2×2 汇合层时,$n=4$,即每 4 个输入,我们汇合得到一个输出.

我们假设存在一个对 $\{1,2,\cdots,n\}$ 的排列 π,使得

$$x_{\pi(1)}\geqslant x_{\pi(2)}\geqslant x_{\pi(3)}\geqslant\cdots\geqslant x_{\pi(n)} \tag{1}$$

经典的汇合方法包括最大汇合和平均汇合.最大汇合是在每个汇合区域内,选择数值最大的输入神经元作为输出,即

$$a=\max_{1\leqslant i\leqslant n}x_i=x_{\pi(1)} \tag{2}$$

式中,最后一步是因为式(1)的假设.平均汇合是在每个汇合区域内,计算局部汇合区域输入的平均值作为输出,即

$$a=\frac{1}{n}\sum_{i=1}^{n}x_i \tag{3}$$

平均汇合认为,局部汇合区域的所有神经元都对输出有着相同的贡献.

2 汇合层的作用

集成最大汇合是设计用来在经典的卷积神经网络结构中替代汇合层的,因此在提出集成最大汇合操作之前,在本节我们将回顾汇合层在经典卷积神经网络中的作用.

经典的卷积神经网络使用多个卷积层和汇合层对输入信号进行加工.卷积层使用多个卷积滤波器提取输入的多个局部特征,其输出数值大小衡量输入与卷积滤波器描述特征的匹配程度.卷积层的输出在经过非线性激活函数后,作为汇合层的输入.汇

合层基于局部相关性原理对输入进行亚采样，从而在减少数据量的同时保留有用信息.汇合层的作用体现在以下两方面.

一方面，汇合层可以减少特征映射的空间大小，并降低提取特征对局部位置的敏感程度.在卷积层发现一个特征后，它的精确位置不及它和其他特征的相对位置的关系重要.汇合层利用了这一特性来进行亚采样，以减少特征映射的空间大小.当多个卷积层和汇合层堆叠时，汇合层会不断地减小特征映射的空间大小，提取得到的特征对局部位置更加不敏感，网络的参数数量和计算量也会下降，这在一定程度上缓解了过拟合，从而有助于提升泛化能力.

另一方面，汇合层可以有效提高感受野.经过一个 3×3 卷积层感受野增加 2，而经过一个 2×2 汇合层感受野乘以 2，因此深层的、尤其是含有多个汇合层的卷积神经网络可以有效提高感受野.相比于传统计算机视觉领域使用的较小感受野的特征，深度卷积神经网络可以自动学习大感受野的特征，而使用大感受野的特征是深度卷积神经网络在图像识别领域取得优异性能的重要原因之一.

3　相关工作

3.1　随机汇合方法汇合层

经典的汇合方法是一个确定性的过程，即当局部汇合区域的输入给定后，其输出是一个定值.近年来，有若干相关工作是将汇合转化为一个随机过程.

随机汇合[1]的输出是根据各 x_i 的相对数值大小采样得到，即

$$a = x_i \text{ w.p. } p_i = \frac{x_i}{\sum_{j=1}^{n} x_j}, \forall i \tag{4}$$

式中，w.p.表示“以概率”.

最大汇合失活[7]和随机最大汇合[2]的思想基本一致，都是在汇合操作前，先在每个局部汇合区域对输入作失活(dropout)处理，即以概率 p 对输入神经元置零，再做汇合操作，其输出可以等价表示为

$$a = \begin{cases} x_{\pi(i)}, & \text{w.p. } p_i = p^{i-1}(1-p) \\ 0, & \text{w.p. } p^n \end{cases} \tag{5}$$

当 $x_{\pi(1)}, x_{\pi(2)}, \cdots, x_{\pi(i-1)}$ 全部失活时，输出是 $x_{\pi(i)}$.

3.2　ResNet 和 DFN-MR

近年来，有许多工作专注于设计更好的卷积神经网络结构，ResNet[4]是其中的一个代表性工作.跳跃连接的存在使得训练 ResNet 变得十分高效，并且可以使网络深度做到成百上千层[8].ResNet 结构可以看作多个不同深度的基础潜在网络的集成，其中那些较浅的基础潜在网络有助于避免训练时的梯度弥散[5].

文献[6]发现，ResNet 中用于集成的那些极深的基础潜在网络可能会影响较浅网络的求解空间与难度，去除这些极深的基础潜在网络不会对整体性能产生太大影响，有时反而能提升总体的性能，因此，文献[6]提出一种 DFN-MR 网络，其有足够多的基础潜在网络数目，同时避免了极深的网络出现，并通过实验验证了 DFN-MR 的性能优于 ResNet.

4　集成最大汇合

4.1　集成最大汇合操作

为避免经典最大汇合方法赢者通吃的弊端，本文提出一种新的具有随机性的最大汇合的方法，称为集成最大汇合，这种方法不仅利用了每个局部汇合区域最大值的信息，同时也利用了每个局部汇合区域非最大值的信息，具体过程如下：

在训练阶段，在每个局部汇合区域，集成最大汇合以 p 的概率将每个局部汇合区域第二大的神经元作为输出，即

$$a = \begin{cases} x_{\pi(1)}, \text{w.p. } 1-p \\ x_{\pi(2)}, \text{w.p. } p \end{cases} \tag{6}$$

式中，p 是一个可供调节的超参数.也就是说，在每个局部汇合区域，输出是从伯努利分布中采样得到的.对每个训练样本，每个局部汇合区域，采样是相互独立的.在误差反向传播过程中，与经典最大汇合的反向传播过程类似，这个被选为输出的神经元将得到后一层向前传播的梯度.

在测试阶段，如果继续使用式(6)中随机性的汇合方式，网络的输出会波动，这将使得网络的性能下降，因此，我们计算这个采样过程的期望的近似作为输出，输出 a 是对 $x_{\pi(1)}$ 和 $x_{\pi(2)}$ 的加权和.

$$a = (1-p)x_{\pi(1)} + px_{\pi(1)} \tag{7}$$

4.2　集成最大汇合与网络集成

集成最大汇合可以看作多个基础潜在网络的集成.训练时的每次迭代将在每个局部汇合区域对伯努利分布进行一次采样，并根据采样结果在各局部汇合区域选不同的神经元，这相当于改变网络的连接结构，定义出一个新的基础潜在网络.在测试阶段，通过使用加权而不是采样的方式，我们得到所有基础潜在网络的平均的近似估计.

当汇合层的输出是 $D \times H \times W$ 时，这样可能的基础潜在网络数目为

$$N = 2^{DHW} \tag{8}$$

式中，在每个局部汇合区域有取最大或第二大的神经元 2 种选择，而局部汇合区域的数目是 DHW，其数值大小随着模型大小由千到万不等.

当整个网络含有 L 层集成最大汇合层，其中第 l 层汇合层的输出是 $D_l \times H_l \times W_l$ 时，可能的基础潜在网络数目有

$$N = \prod_{l=1}^{L} 2^{D_l H_l W_l} = 2^{\sum_{l=1}^{L} D_l H_l W_l} \tag{9}$$

因此，多层集成最大汇合层的堆叠可大大提高基础潜在网络的数目.

4.3 集成最大汇合与数据扩充

集成最大汇合在每次迭代都会汇合得到一个新的特征映射，这相当于隐式地做了数据扩充.与经典的数据扩充直接作用在输入数据不同，集成最大汇合可以被理解成一种输入经历一定局部形变下的经典最大汇合过程，这样的数据扩充作用在中间层.这种局部形变类似于文献[10]中提出的弹性形变，并在 MNIST 数据集[10]上取得了非常好的效果.

由于各层对局部汇合区域中元素的采样是独立的，当多层集成最大汇合层堆叠时，这样的局部形变数目以指数级别增加.

4.4 集成最大汇合实现

集成最大汇合可直接从现有的深度学习框架的最大汇合层源代码中增加少量代码得到.在计算量上，集成最大汇合层只引入了少量的常数级计算，不影响整体渐进复杂度.实验发现，在 NVIDIA K-80 上，基于 Caffe[11]，采用网络结构单卡迭代 1 000 次计时取平均的方式，使用经典最大汇合的网络平均前向传播时间是 5.18031 ms，而使用集成最大汇合的网络平均前向传播时间是 5.28989 ms.相比经典最大汇合，集成最大汇合用时仅增加 1.9%.

5 集成最大汇合与 DFN-MR

经典的神经网络模型主要在宽度与深度方面进行不同程度的扩增，比如 AlexNet[12]，VGGNet[13]等经典网络通过宽度或深度增加的参数可以有效地提升其模型的表达能力.网络越深，其训练难度也随之相应增加，反而会导致性能的下降.ResNet[4]通过引入跳跃连接结构来试图解决极深网络在优化上带来的问题.

从集成学习的角度来看，ResNet 可以被看作指数级别分支路径的混合[5].ResNet(见图 1(a))可以被等效地展开成一种多分支融合网络(见图 1(b))，不同分支可以在中间层进行信息融合(在 ResNet 中是加和的形式)，而这种多分支融合网络可以近似为很多基础潜在网络的集成(见图 1(c))，两者的区别在于基础潜在网络之间没有中间层的信息交互，它们只是共享对应层的网络参数.图 1 是仿照文献[14]制作而成.

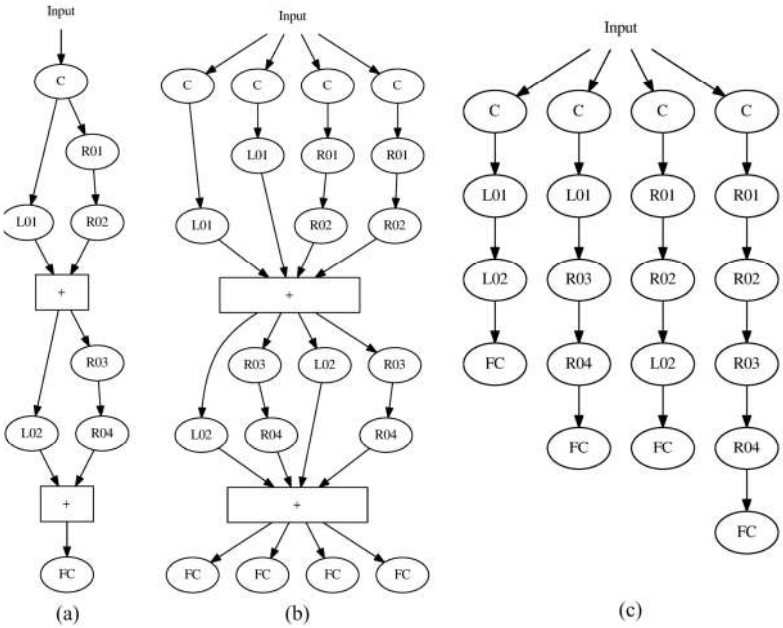

图 1 ResNet 指数级分支路径的混合

Fig.1 The mixing of exponential branching paths of ResNet

文献[6]发现极深的基础潜在网络除了能增加用于集成的网络数目外，对整体性能的贡献并不是最大的，反而会影响其他的组成网络，导致最终的结果受到影响.因此，一个好的网络应该满足“不减少基础潜在网络数目”和“避免极深基础潜在网络”这两个原则.

DFN-MR 和 ResNet 一样，都可以展开成一种多分支融合网络，可以近似为很多基础潜在网络的集成.相比 ResNet 结构，DFN-MR 的基础潜在网络数目更多，同时 DFN-MR 没有极深基础潜在网络对较浅网络的求解空间与难度的影响，更加容易收敛或训练，使其最终相比 ResNet 取得了更好的性能表现[6].

用于近似 ResNet 和 DFN-MR 的基础潜在网络是一种单分支的网络.ResNet 和 DFN-MR 中都没有使用汇合层，它们用步长为 2 的卷积层来降低空间大小.在单分支网络中，虽然汇合层和步长为 2 的卷积层都可以减小特征映射的大小，但汇合层可以有效提高感受野，从而在单分支网络中，汇合层的效果比步长为 2 的卷积层更好，因此，我们将 DFN-MR 中步长为 2 的卷积层改为 2×2 汇合层串联步长为 1 的卷积层的结构.

6　实验

我们在 CIFAR-10，CIFAR-100 和 ImageNet 三个数据集上进行实验.

6.1　CIFAR-10—汇合方法比较

CIFAR-10 数据集[15]包括 10 个类别，共有 50 000个训练数据和 10 000 个测试数据.每张图像大小 32×32×3，平均每个类别有 5 000 个训练数据.

实验目的是比较本文提出的集成最大汇合和几种相关工作，实验基于 Caffe[11]，使用的网络见表 1.实验使用带有 0.9 的动量的随机梯度下降方法进行优化，使用 0.003 的权值衰减.整个数据集减去训练集的均值，使用镜像的方式进行数据扩充，训练共进行 300 轮.初始学习率为 0.003，并在训练过程中不断递减，直到初始值的 1/100.实验结果见表 2，结果分以下 3 部分进行分析.

最大汇合只考虑最大的神经元；平均汇合认为所有神经元都对输出有着相同的贡献，即使有的神经元有着非常小的值.集成最大汇合是随机过程，避免了最大汇合赢者通吃的弊端；同时，由于集成最

表 1　用于在 CIFAR-10 数据集上几种汇合方法的网络结构

Tab.1　The network used to compare different pooling methods on CIFAR-10 data set

层	滤波器大小	通道数	步长	0-填补
conv1	5×5	32	1×1	2×2
pool1	3×3	—	2×2	0×0
conv2	5×5	32	1×1	2×2
pool2	3×3	—	2×2	0×0
conv3	5×5	64	1×1	2×2
pool3	3×3	—	2×2	0×0
fc4	—	64	—	—
fc5	—	10	—	—

表 2　不同汇合方法在表 1 网络，CIFAR-10 数据集上的实验结果

Tab.2　The results of different pooling methods on CIFAR-10 dataset，using the network described in Tab.1

模型	错误率	与最大汇合相比的相对改善
max-pooling	15.25%	0.00%
average pooling	16.20%	−6.23%
ensemble max-pooling ($p=0.05$)	15.51%	−1.70%
ensemble max-pooling ($p=0.1$)	15.14%	0.72%
ensemble max-pooling ($p=0.2$)	15.07%	1.18%
ensemble max-pooling ($p=0.3$)	14.28%	6.36%
ensemble max-pooling ($p=0.4$)	14.27%	6.43%
ensemble max-pooling ($p=0.5$)	14.62%	4.13%
ensemble max-pooling ($p=0.7$)	14.89%	2.36%
stochastic pooling	15.81%	−3.67%
max-pooling dropout/stochastic max-pooling ($p=0.05$)	15.22%	0.20%
max-pooling dropout/stochastic max-pooling ($p=0.1$)	15.17%	0.52%
max-pooling dropout/stochastic max-pooling ($p=0.2$)	14.97%	1.84%
max-pooling dropout/stochastic max-pooling ($p=0.3$)	16.59%	−6.96%
max-pooling dropout/stochastic max-pooling ($p=0.4$)	25.93%	−67.18%
max-pooling dropout/stochastic max-pooling ($p=0.5$)	not converge	—
max-pooling dropout/stochastic max-pooling ($p=0.7$)	not converge	—

大汇合只考虑最大和第二大神经元，缓和了平均汇合中数值小的神经元对输出的影响.实验中，我们发现，除 $p = 0.05$ 的集成最大汇合性能介于经典最大汇合和平均汇合之间，其余的均优于最大汇合及平均汇合.

集成最大汇合与随机汇合相比，尽管随机汇合没有引入额外的超参数 p，并且省去调超参数的工序，但是随机汇合中数值小的神经元有更大的概率被选中作为输出，其带来的信息损失使其性能劣于集成最大汇合.实验还发现，所有 p 取值的集成最大汇合均优于随机汇合.

为了进一步比较集成最大汇合与最大汇合失活/随机最大汇合，我们画出了两者的训练和测试损失，见图 2.对比集成最大汇合的训练和测试损失(即图中的对比线)，我们可以从偏差-方差分解[16]的角度来理解这个实验结果.随着 p 的增加，网络由高方差逐渐向高偏差移动.和集成最大汇合等效于多个基础潜在网络的集成的分析过程类似，最大汇合失活/随机最大汇合通过在各个局部汇合区域进行失活操作也可等效于多个网络集成.但是，在大的 p 的取值下，失活带来的信息损失会有害于而不是有利于训练过程，因此，最大汇合失活/随机最大汇合需要仔细调节失活概率 p 以达到最优的性能.在实验中我们发现，集成最大汇合可适用的 p 的动态范围比最大汇合失活/随机最大汇合更大，也就是说，在多个 p 的取值下集成最大汇合都能取得比最大汇合失活/随机最大汇合更好的性能，即集成最大汇合方法比最大汇合失活/随机最大汇合对 p 的取值更加健壮.

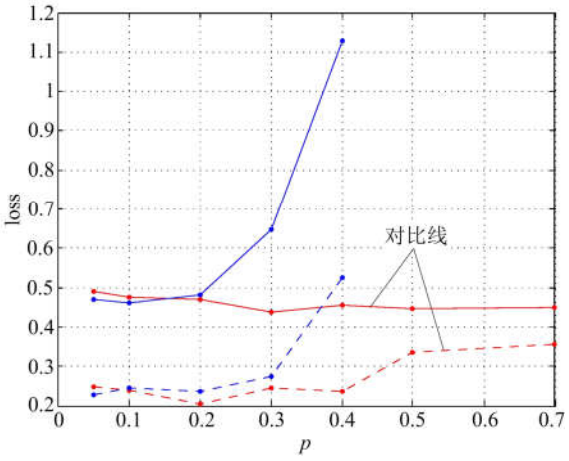

图 2 集成最大汇合与最大汇合失活/随机最大汇合的训练损失和测试损失

Fig.2 Training and testing loss of ensemble max-pooling and max-pooling dropout/stochastic max-pooling

6.2 CIFAR-10—DFN-MR

为了验证向 DFN-MR 中引入集成最大汇合带来的性能提升，本节实验基于 MXNet[18]，代码改写于文献[6].DFN-MR 共有卷积层 56 个，分为 3 组，各组卷积的通道数分别为 16，32，64.为了实验可比性，实验中，除了将 DFN-MR 网络中的步长为 2 的卷积层改为 2×2 汇合层串联步长为 1 的卷积层的结构外，其余超参数选择（如训练轮数，批量大小，学习率策略，权重初始化策略，权值衰减率等）与原文保持一致，即实验中使用的超参数选择可能并不是对我们的集成最大汇合模型最优.

实验结果见表 3.其中关于 ResNet 的 6 项结果取自于原论文.对比 DFN-MR 和 DFN-MR(经典最大汇合)可以看出，将 DFN-MR 中步长为 2 的卷积层改为经典最大汇合和步长为 1 的卷积层带来较小的效果提升.对比 DFN-MR 和 DFN-MR 发现，向 DFN-MR 中引入集成最大汇合的系统性能有显著提升.由此可以看出，性能提升主要是由于集成最大汇合层利用了非最大值信息.

表 3 向 DFN-MR 中引入汇合层与相关工作在 CIFAR-10 数据集上的性能比较

Tab.3 The comparisons of introducing pooling into DFN-MR and other related works

模型	层数	错误率	与 DFN-MR 的相对改善
DFN-MR	56	5.50%	0.00%
DFN-MR (classical max-pooling)	56	5.44%	1.14%
DFN-MR (ensemble max-pooling, $p=0.05$)	56	5.13%	6.82%
DFN-MR (ensemble max-pooling, $p=0.1$)	56	4.62%	16.00%
DFN-MR (ensemble max-pooling, $p=0.2$)	56	5.06%	7.95%
DFN-MR (ensemble max-pooling, $p=0.5$)	56	4.69%	14.77%
ResNet[4]	110	6.61%	−20.19%
ResNet[17]	110	6.41%	−16.54%
ResNet (pre-activation)[8]	164	5.46%	0.72%
ResNet (pre-activation)[8]	1001	4.62%	16.00%
ResNet (stochastic Depth)[17]	110	5.23%	4.91%
ResNet (stochastic Depth)[17]	1202	4.91%	10.73%

注意到性能随 p 的变化不是单调的.从表 3 可以看出,错误率在 $p=0.1$ 时最小,而在 $p=0.2$时变大后又在 $p=0.5$ 时变得较小.这是因为深度神经网络面对的目标函数非凸,这使得网络性能和超参数的选取很多情况下不是单调关系.如文献[19]就观察到网络性能和神经网络中隐层结点个数的关系就不是单调的.从理论上讲,我们不可能得到最优的超参数选择,这是因为机器学习(包括深度学习在内)面临的问题通常是 NP 难甚至更难,而有效的学习算法必然是在多项式时间内运行完成,若能得到最优超参数取值,就意味着我们证明了"P=NP"[20-21].

从实际实验结果可以看出,对比 DFN-MR 和 ResNet 的几个相关工作,除了 1001 层和 1202 层的 ResNet 外,不同 p 取值下的 DFN-MR 的所有实验结果一致优于 ResNet.与 1001 层和 1202 层的 ResNet 相比,DFN-MR 的层数约为其 1/20,而最终错误率十分接近,说明了集成最大汇合的高效性.

6.3　CIFAR-100—汇合方法比较

CIFAR-100 数据集[15]包括 100 个类别,共有 50 000 个训练数据和 10 000 个测试数据.每张图像大小 32×32×3,平均每个类别有 500 个训练数据.

实验目的是为了在另外一个数据集上比较本文提出的集成最大汇合和几种相关工作,实验基于 Caffe,使用的网络见表 4.

表 4　用于在 CIFAR-100 数据集上比较几种汇合方法的网络结构

Tab.4　The network used to compare different pooling methods on CIFAR-100 data set

层	滤波器大小	通道数	步长	0-填补
conv1	3×3	16	1×1	1×1
conv2	3×3	16	1×1	1×1
pool2	2×2	—	2×2	0×0
conv3	3×3	32	1×1	1×1
conv4	3×3	32	1×1	1×1
pool42×2	—	2×2	0×0	
conv5	3×3	64	1×1	1×1
conv6	5×5	64	1×1	1×1
gap7	8×8	—	—	—
fc8	—	100	—	—

表 4 的网络结构并不是表 1 的简单加深,而是引入了近期卷积神经网络中较为广泛使用的一些设计理念[4,13],包括:①使用卷积核 3x3 的卷积层和卷积核 2×2 的汇合层;②在每个卷积层后,非线性激活函数前引入批量规范化层[22];③在最后一个卷积层后使用全局平均汇合提取最终特征而不是使用多个全连接层;④没有使用失活层[22];⑤使用文献[23]的方法进行参数初始化.

实验使用带有 0.9 的动量随机梯度下降方法进行优化,使用 0.001 的权值衰减.整个数据集减去训练集的均值,使用镜像的方式进行数据扩充,训练共进行 327 轮.初始学习率为 0.1,并在训练过程中不断递减,直到初始值的 1/100.实验结果见表 5,与表 2 的分析过程类似,即使是使用了不同的网络结构,集成最大汇合的结果也优于经典最大汇合和平均汇合以及其他相关工作.

表 5　不同汇合方法在表 3 网络,CIFAR-100 数据集上的实验结果

Tab.5　The results of different pooling methods on CIFAR-100 data set, using the network described in Tab.3

模型	错误率	与最大汇合相比的相对改善
max-pooling	49.57%	0.00%
average pooling	47.60%	3.97%
ensemble max-pooling ($p=0.05$)	47.27%	4.64%
ensemble max-pooling ($p=0.1$)	45.58%	8.05%
ensemble max-pooling ($p=0.2$)	44.02%	11.20%
ensemble max-pooling ($p=0.3$)	43.58%	12.08%
ensemble max-pooling ($p=0.4$)	43.75%	11.74%
ensemble max-pooling ($p=0.5$)	43.67%	11.90%
ensemble max-pooling ($p=0.7$)	43.28%	12.69%
stochastic pooling	46.60%	6.00%
max-pooling dropout/stochastic max-pooling ($p=0.05$)	46.91%	5.37%
max-pooling dropout/stochastic max-pooling ($p=0.1$)	46.94%	5.31%
max-pooling dropout/stochastic max-pooling ($p=0.2$)	52.90%	−6.72%
max-pooling dropout/stochastic max-pooling ($p=0.3$)	58.97%	−18.97%
max-pooling dropout/stochastic max-pooling ($p=0.4$)	67.24%	−35.65%
max-pooling dropout/stochastic max-pooling ($p=0.5$)	72.92%	−47.11%
max-pooling dropout/stochastic max-pooling ($p=0.7$)	81.78%	−64.98%

6.4 ImageNet——微调

ImageNet数据集[14]共有1 000个类别,共有1 280 000个训练数据和50 000个验证数据.

实验基于Caffe[11],对比的模型是CaffeNet[11].CaffeNet是在Caffe上的基于AlexNet[12]的实验,两者主要的区别在于CaffeNet调换了AlexNet中汇合层和归一层的顺序,减小了内存开销.CaffeNet模型共迭代训练310 000次,小批量大小是256,得到的前1错误率为43.096%,前5错误率为19.97%.

本小节是基于CaffeNet预训练模型上的微调结果.微调时共迭代训练200 000次,初始学习率为0.001,每过50 000次迭代,学习率除以10,保持其他超参数不变.

基于CaffeNet在ImageNet上的微调结果见表6.对比集成最大汇合和CaffeNet预训练模型可以看出,通过集成最大汇合的随机性,可以使模型跳出原有收敛得到的局部最优从而继续训练.所有p的取值下,集成最大汇合的实验结果均优于原始CaffeNet模型,这说明集成最大汇合对超参数p有着比较广的适应区间.

表6 基于CaffeNet在ImageNet上的微调结果

Tab.6 The fine tuning results on ImageNet based on CaffeNet

模型	错误率(前1/前5)	与CaffeNet预训练模型的相对改善(前1/前5)
CaffeNet pre-trained model	43.10%/19.97%	0.00%/0.00%
ensemble max-pooling ($p=0.05$)	42.11%/19.19%	2.30%/3.93%
ensemble max-pooling ($p=0.1$)	41.97%/19.14%	2.60%/4.18%
ensemble max-pooling ($p=0.15$)	42.05%/19.23%	2.43%/3.70%
ensemble max-pooling ($p=0.3$)	42.22%/19.25%	2.03%/3.61%
ensemble max-pooling ($p=0.5$)	42.25%/19.33%	1.95%/3.21%
ensemble max-pooling ($p=0.7$)	42.28%/19.40%	1.88%/2.87%

6.5 ImageNet——重新训练

为了保证实验的可比性,在重新训练时,除了将CaffeNet网络中的经典汇合层改为对应的集成最大汇合层外,其余超参数选择(如训练轮数,小批量大小,学习率策略,权重初始化策略,权值衰减率等)保持和训练CaffeNet时使用的一致,即实验中使用的超参数选择可能并不是对我们的模型最优.

实验结果见表7,由表7可以看出,所有结果一致优于CaffeNet,说明集成最大汇合的有效性和对p有较广的适应区间.

表7 基于CaffeNet在ImageNet上的重新训练结果

Tab.7 The training from scratch results on ImageNet based on CaffeNet

模型	错误率(前1/前5)	与CaffeNet预训练模型的相对改善(前1/前5)
CaffeNet pre-trained model	43.10%/19.97%	0.00%/0.00%
ensemble max-pooling ($p=0.05$)	42.69%/19.19%	0.94%/2.20%
ensemble max-pooling ($p=0.1$)	42.37%/19.14%	1.68%/2.50%
ensemble max-pooling ($p=0.15$)	42.61%/19.23%	1.13%/2.53%
ensemble max-pooling ($p=0.3$)	42.81%/19.69%	0.65%/1.41%
ensemble max-pooling ($p=0.5$)	42.71%/19.67%	0.89%/1.51%
ensemble max-pooling ($p=0.7$)	42.85%/19.81%	0.58%/0.80%

7 结论

本文提出了一种简单有效的汇合方法,称为集成最大汇合,可用于替代现有卷积神经网络中的汇合层.我们在CIFAR-10和ImageNet数据集上进行实验,相比经典汇合和其他相关汇合方法,集成最大汇合方法取得了很好的效果.通过向ResNet的衍生DFN-MR中引入集成最大汇合,网络的准确率有显著提高.

未来工作中,我们将在更深的网络和多个数据集下比较集成最大汇合方法的作用,同时对集成最

大汇合和经典最大汇合的异同进行可视化.此外,我们将把集成最大汇合与其他网络集成,数据扩充方法进行比较.

参考文献(References)

[1] ZEILER M D, FERGUS R. Stochastic pooling for regularization of deep convolutional neural networks [J]. Eprint, 2013: arXiv:1301.3557.

[2] HUANG Yuchi, SUN Xiuyu, LU Ming, et al. Channel-max, channel-drop and stochastic max-pooling [C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops. Boston, USA: IEEE, 2015: 9-17.

[3] CAI Meng, SHI Yongzhe, LIU Jia. Stochastic pooling maxout networks for low-resource speech recognition [C]// Proceedings of the 2014 IEEE International Conference on Acoustics, Speech and Signal Processing. Florence, Italy: IEEE, 2014: 3266-3270.

[4] HE Kaiming, ZHANG Xiangyu, REN Shaoqing, et al. Deep residual learning for image recognition [C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, USA: IEEE, 2016: 770-778.

[5] VEIT A, WILBER M, BELONGIE S. Residual networks are exponential ensembles of relatively shallow networks [EB/OL]. [2017-02-14] https://arxiv.org/abs/1605.06431v1.

[6] ZHAO Liming, WANG Jingdong, LI Xi, et al. On the connection of deep fusion to ensembling [EB/OL]. [2017-02-14] https://arxiv.org/abs/1611.07718.

[7] WU Haibing, GU Xiaodong. Max-pooling dropout for regularization of convolutional neural networks [C]// Proceedings of the International Conference on Neural Information Processing. Berlin: Springer, 2015: 46-54.

[8] HE Kaiming, ZHANG Xiangyu, REN Shaoqing, et al. Identity mappings in deep residual networks [C]// Proceedings of the 14th European Conference on Computer Vision. Berlin: Springer, 2016: 630-645.

[9] SIMARD P Y, STEINKRAUS D, PLATT J C, et al. Best practices for convolutional neural networks applied to visual document analysis [C]// Proceedings of the International Conference on Document Analysis and Recognition. Washington: IEEE Computer Society, 2003: 958-962.

[10] LECUN Y, BOTTOU L, BENGIO Y, et al. Gradient-based learning applied to document recognition [J]. Proceedings of the IEEE, 1998, 86(11): 2278-2324.

[11] JIA Y Q, SHELHAMER E, DONAHUE J, et al. Caffe: Convolutional architecture for fast feature embedding [C]// Proceedings of the 22nd ACM International Conference on Multimedia. Orlando, USA: ACM, 2014: 675-678.

[12] KRIZHEVSKY A, SUTSKEVER I, HINTON G E. ImageNet classification with deep convolutional neural networks [J]. International Conference on Neural Information Processing Systems, 2012, 25 (2): 1097-1105.

[13] SIMONYAN K, ZISSERMAN A. Very deep convolutional networks for large-scale image recognition [J]. Eprint, 2015: arXiv:1409.1556.

[14] DENG J, DONG W, SOCHER R, et al. ImageNet: A large-scale hierarchical image database [C]// Proceedings of IEEE Conference on Computer Vision and Pattern Recognition. Miami, USA: IEEE, 2009: 248-255.

[15] KRIZHEVSKY A, HINTON G. Learning multiple layers of features from tiny images[J]. Eprint, 2009: arXiv:1011.1669v3.

[16] GEMAN S, BIENENSTOCK E, DOURSAT R. Neural networks and the bias/variance dilemma [J]. Neural computation, 1992, 4(1): 1-58.

[17] HUANG Gao, SUN Yu, LIU Zhuang, et al. Deep networks with stochastic depth [C]// Proceedings of the 14th European Conference on Computer Vision. Berlin: Springer, 2016: 646-661.

[18] CHEN Tianqi, LI Mu, LI Yutian, et al. Mxnet: A flexible and efficient machine learning library for heterogeneous distributed systems [J]. Eprint, 2015: arXiv:1512.01274.

[19] DANIELS H, KAMP B, VERKOOIJEN W. Application of Neural Networks to House Pricing and Bond Rating [M]. Tilburg University, 1997.

[20] COBHAM A. The intrinsic computational difficulty of functions [J]. International Congress for Logic, 1969, 31(1): 43-52.

[21] EDMONDS J. Paths, trees, and flowers [J]. Canadian Journal of Mathematics, 2009, 17(3):361-379.

[22] IOFFE S, SZEGEDY C. Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift [C]// Proceedings of the 32nd International Conference on Machine Learning. Lille, France: ACM, 2015: 448-456.

[23] HE Kaiming, ZHANG Xiangyu, REN Shaoqing, et al. Delving deep into rectifiers: Surpassing human-level performance on ImageNet classification [C]// Proceedings of the 27th International Conference on Computer Vision. Santiago, USA: ACM, 2015: 1026-1034.