DOI:10.7544/issn1000-1239.2016.********
卷(期):起止页,年

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（南京大学）南京 210046）

[2]（南京大学计算机科学与技术系 南京 210046）

（wujx@lamda.nju.edu.cn）

Ensemble Max-Pooling: Is Only the Maximum Activation Useful When Pooling

Zhang Hao[1,2] and Wu Jianxin[1,2]

[1]（*State Key Laboratory for Novel Software Technology*（*Nanjing University*）, *Nanjing* 210046）

[2]（*Department of Computer Science and Technology*, *Nanjing University*, *Nanjing* 210046）

Abstract The pooling layer in convolutional neural networks is doing subsampling based on the local correlation principle. It can reduce the data size while keeping useful information in order to improve generalization. Meanwhile, it can effectively increase receptive fields. The winner-take-all strategy is used in classical max-pooling, which will affect the generalization of the network sometimes. We introduce a simple and effective pooling method named ensemble max-pooling, 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 output the neuron with second largest activation. Ensemble max-pooling can be viewed as an ensemble of many basic underlying networks, and it can also be viewed as the classical max-pooling with some local distortion of the input. We achieve better results than classical pooling methods and other related pooling approaches. DFN-MR is derived from ResNet, and DFN-MR has more basic underlying networks and it 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 networks; pooling layer; network ensemble; data augmentation

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．我们提出集成最大汇合，一种简单而有效的汇合方法，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，而激活输出第二大的神经元．集成最大汇合可以看作是多个基础潜在网络的集成，也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．通过实验比较，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

中图法分类号 TP183

卷积神经网络(convolutional neural networks)近期在计算机视觉及其他一些领域取得了显著的成果．其中汇合层(pooling layer)一方面可以减少特征映射(feature map)的空间大小并降低被提取得到的特征

收稿日期：2017-02-25；**修回日期**：
通信作者：吴建鑫（wujx@lamda.nju.edu.cn）
第一作者为学生.

对局部位置的敏感程度，另一方面可以有效提高感受野(receptive field)．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．深度卷积神经网络相比传统方法能取得优异性能的原因之一在于其能学习到大感受野的特征.

经典的最大汇合(max-pooling)是一个赢者通吃(winner-take-all)的过程，每个局部汇合区域数值最大的神经元被输出．在更新时，只有这个数值最大的神经元得到后一层向前传播的梯度．当训练数据有限时，网络将可能不能很好地泛化(generalize)到测试集[1][2][3].

在本文，我们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．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．集成最大汇合可以看作是多个基础潜在网络的集成，也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.

ResNet[4]可以看作是多个不同深度的基础潜在网络的集成[5]．DFN-MR(deeply fused networks-merge and run)是基于 ResNet 设计得到的，并取得了比 ResNet 更好的效果[6]．ResNet 和 DFN-MR 都没有使用汇合层，它们用步长(stride)为 2 的卷积层来降低空间大小．在本文，我们将向 DFN-MR 中引入集成最大汇合，并通过实验观察其效果.

本文其余内容如下安排：第 1 节给出本文使用的符号表示以及两种经典汇合方法；第 2 节回顾相关工作；第 3 节提出集成最大汇合操作；第 4 节将集成最大汇合引入 DFN-MR；第 5 节在 CIFAR-10 和 ImageNet 两个数据集上面进行实验；第 6 节总结全文.

1 符号表示及经典汇合方法

在本文，我们使用$\{x_1, x_2, \ldots, x_n\}$表示汇合层中一个局部汇合区域的输入，$a \in R$ 表示该区域的输出．我们假设存在一个对$\{1, 2, ..., n\}$的排列 π，使得

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

经典的汇合方法包括最大汇合和平均汇合(average pooling)．最大汇合是在每个汇合区域内，输出数值最大的神经元:

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

其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，输出神经元的均值:

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

2 相关工作

2.1 随机汇合方法

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之后，其输出是一个定值．近年来，有若干相关工作是将汇合变成一个随机性的过程.

随机汇合(stochastic pooling)[1]的输出是根据各 x_i 的相对数值大小采样得到，即

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

其中 w.p.表示“以概率”(with probability).

最大汇合失活(max-pooling dropout)[7]和随机最大汇合(stochastic max-pooling)[2]的思想十分接近，两者都先在每个局部汇合区域对输入做失活(dropout)[8]，即以概率 p 对各神经元置零，再进行汇合，其输出可以等价表示为

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

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

2.2 ResNet 和 DFN-MR

经典的神经网络主要在宽度与深度方面进行不同程度的扩增，但当网络变得越来越深时，其训练难度也随之增加．ResNet[4]通过引入跳跃连接(skip connection)来试图解决极深网络在优化上带来的问题，并且可以使网络深度做到成百上千层[9].

ResNet 结构可以被看作是指数多不同深度的基础潜在网络的集成[5]，其中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．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.

3 集成最大汇合

3.1 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，我们提出一种新的具有随机性的最大汇合方法，称为集成最大汇合，具体过程如下.

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

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

其中 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, 输出是基于从一个伯努利分布(Bernoulli distribution)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, 和经典最大汇合的反向传播过程类似, 这个被选择作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, 这将影响网络的性能. 因此, 我们计算这个采样过程的期望的近似作为输出, 即输出 a 是对 $x_{\pi(1)}$和 $x_{\pi(2)}$的加权和:

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

3.2 集成最大汇合与网络集成

集成最大汇合可以看作是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根据对伯努利分布的采样结果选取不同的神经元, 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

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

其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将大大提高基础潜在网络的数目.

3.3 集成最大汇合与数据扩充

集成最大汇合在每次迭代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(data augmentation). 和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 (elastic distortions), 而这种弹性形变在 MNIST 数据集[11]上取得了非常好的效果.

和前一小节的计算方法类似, 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.

3.4 集成最大汇合实现

集成最大汇合容易实现, 可以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. 实验中发现, 在 NVIDIA K-80 上, 基于 Caffe[12], 采用第 5.1 小节描述的网络结构单卡迭代 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, 用时仅增加 1.9%.

4 集成最大汇合与 DFN-MR

ResNet(见图 1(左))可以被等效地展开成一种多分支融合网络(见图 1(中)), 不同分支可以在中间层进行信息融合(在 ResNet 中是以加和的形式), 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(右)), 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[6]. 类似的, DFN-MR 也可以被近似为很多基础潜在网络的集成[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. 虽然两者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. 在实验部分, 我们将观察向 DFN-MR 中引入汇合层及集成最大汇合带来的性能提升.

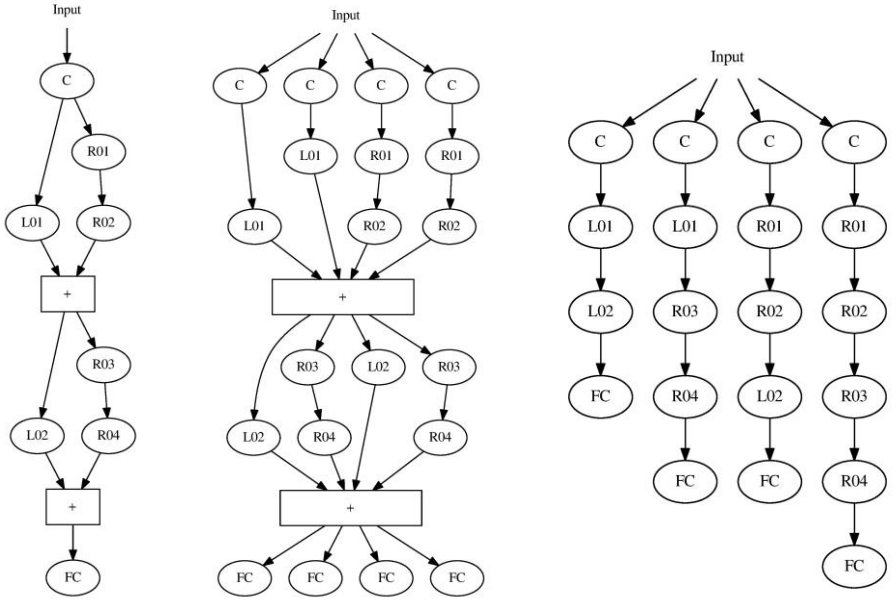

Fig. 1 ResNet (Left) can be expaned as a multi-branched fused network (Middle). This network resembles that of the ensemble of many basic underlying networks (Right). This figure is reproduced from the reference [6].

图 1 ResNet(左)可以被展开成一种多分支融合网络(中)，这种多分支融合网络可以被近似为很多基础潜在网络的集成(右)．本图仿照文献[6]画成.

5 实验

我们在 CIFAR-10 和 ImageNet 一小一大两个数据集上进行实验.

5.1 CIFAR-10—汇合方法比较

CIFAR-10 数据集[14]包括 10 个类别，共有 50,000 个训练数据和 10,000 个测试数据.

本小节的实验目的是为了比较集成最大汇合和几种相关汇合方法，实现基于 Caffe[12]，使用的网络是 32C5-P3-32C5-P3-64C5-P3-64F-10F. 其中 32C5 表示通道(channel)数 32， 滤波器(filter)大小为 5 的卷积层；P3 代表滤波器大小为 3 的汇合层；64F 和 10F 分别代表输出神经元数 64 和 10 的全连接层. 在每个卷积和全连接层后接有 ReLU 层；在 64F 后接有 p=0.5 的失活层.

在实验中优化使用带有 0.9 动量(momentum)的随机梯度下降，训练共进行 300 轮(epoch)，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 1，结果分以下三部分进行分析.

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. 而集成最大汇合是随机性的过程，避免了最大汇合赢者通吃的弊端；集成最大汇合只考虑最大和第二大神经元，也缓和了平均汇合中数值小的神经元对输出的影响.

所有 p 取值的集成最大汇合均优于随机汇合，这是因为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.

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(loss)，见图 2. 最大汇合失活/随机最大汇合中在大的失活概率 p 的取值下，带来的信息损失会有害于训练过程，因此，最大汇合失活/随机最大汇合需要仔细调节 p 以取得最优的性能. 而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，集成最大汇合更加健壮(robust).

Table 1 The results of different pooling methods on CIFAR-10 dataset.

表 1 不同汇合方法在 CIFAR-10 数据集上的实验结果.

Model	*Error Rate*	***Relative** Improve Compared to Max-pooling*
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-pooing (p=0.05)	15.22%	0.20%
Max-pooling Dropout/Stochastic Max-pooing (p=0.1)	15.17%	0.52%
Max-pooling Dropout/Stochastic Max-pooing (p=0.2)	14.97%	1.84%
Max-pooling Dropout/Stochastic Max-pooing (p=0.3)	16.59%	-6.96%
Max-pooling Dropout/Stochastic Max-pooing (p=0.4)	25.93%	-67.18%
Max-pooling Dropout/Stochastic Max-pooing (p=0.5)	Not Converge	-
Max-pooling Dropout/Stochastic Max-pooing (p=0.7)	Not Converge	-

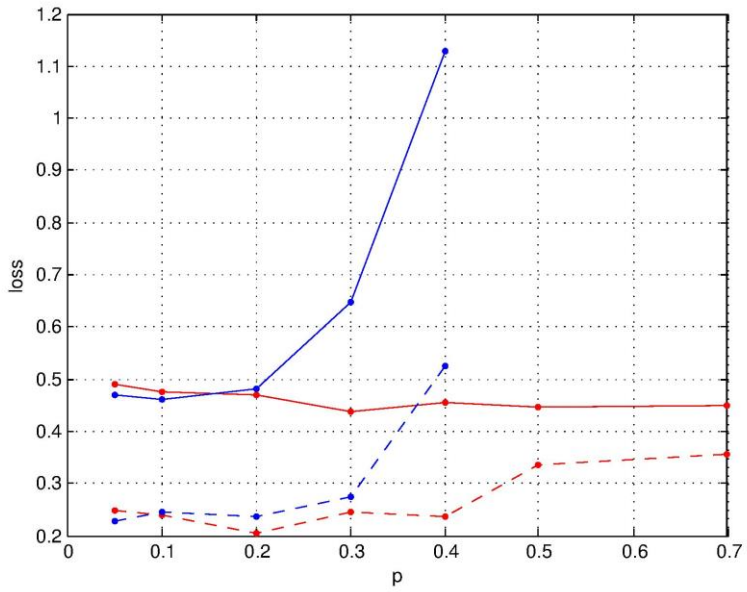

Fig. 2 Training and testing loss of ensemble max-pooling and max-pooling dropout/stochastic max-pooling. In which red lines represent ensemble max-pooling, and blue lines represent max-pooling dropout/stochastic max-pooling; dashed lines represent training loss, and solid lines represent testing loss.

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. 其中红线表示集成最大汇合, 蓝线表示最大汇合失活/随机最大汇合; 虚线表示训练损失, 实线表示测试损失.

Table 2 The comparisons of introducing pooling into DFN-MR and other related works.

表 2 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.

Model	*Layers*	*Error Rate*	***Relative** Improve Compared to DFN-MR*
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[16]	110	6.41%	-16.54%
ResNet (Pre-activation)[9]	164	5.46%	0.72%
ResNet (Pre-activation)[9]	1001	4.62%	16.00%
ResNet (Stochastic Depth)[16]	110	5.23%	4.91%
ResNet (Stochastic Depth)[16]	1202	4.91%	10.73%

5.2 CIFAR-10—DFN-MR

本小节的实验目的是观察向 DFN-MR 中引入集成最大汇合对性能带来的影响. 实现基于 MXNet[15], 代码改写于文献[6]的实现.[1] 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, 在实验中除了将 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 的卷积层的结构之外, 其余超参数选择(如训练轮数, 学习率策略, 权重初始化策略等)保持和原文一致, 也就是说, 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 2. 向 DFN-MR 中引入最大汇合带来较小的性能提升, 而引入集成最大汇合带来显著提升. 这主要是由于集成最大汇合层利用了非最大值信息.

DFN-MR (Ensemble Max-pooling)与 1001 层的 ResNet 和 1202 层的 ResNet (Stochastic Depth)最终错误率十分接近甚至更优, 而层数约为其 1/20, 并且一致优于其他 ResNet 结果, 说明了集成最大汇合的高效性.

5.3 ImageNet—微调

ImageNet 数据集[17]共有 1,000 个类别,共有 1,280,000 个训练数据和 50,000 个验证(validation)数据.

实现基于 Caffe[12], 对比模型是 CaffeNet[12]. CaffeNet 是在 Caffe 上的基于 AlexNet[13]的实现, 两者主要的区别在于 CaffeNet 调换了 AlexNet 中汇合层和归一(normalization)层的顺序, 从而减小内存开销. CaffeNet 模型共迭代训练 62 轮.

本小节是基于 CaffeNet 预训练模型上的微调(fine-tuning)结果. 微调时共迭代 40 轮, 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.

实验结果见表 3. 通过集成最大汇合的随机性, 可以使模型跳出原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 p 有着比较广的适应区间.

5.4 ImageNet—重新训练

本小节是基于 CaffeNet 重新训练(training from scratch)的结果. 为了实验可比性, 在重新训练时, 其余超参数选择保持和 CaffeNet 一致, 也就是说, 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 4. 所有结果一致优于 CaffeNet, 并且所有 p 取值下结果均优于 CaffeNet, 说明集成最大汇合的有效性和对 p 比较广的适应区间.

[1] https://github.com/zlmzju/fusenet.

Table 3 The fine-tuning results on ImageNet based on CaffeNet.

表 3 基于 CaffeNet 在 ImageNet 上的微调结果.

Model	*Error Rate (Top-1/Top-5)*	***Relative** Improve Compared to Pre-trained Model (Top-1/Top-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%

Table 4 The training from scratch results on ImageNet based on CaffeNet.

表 4 基于 CaffeNet 在 ImageNet 上的重新训练结果.

Model	*Error Rate (Top-1/Top-5)*	***Relative** Improve Compared to Pre-trained Model (Top-1/Top-5)*
CaffeNet Pre-trained Model	43.10%/19.97%	0.00%/0.00%
Ensemble Max-pooling (p=0.05)	42.69%/19.53%	0.94%/2.20%
Ensemble Max-pooling (p=0.1)	**42.37%/19.47%**	**1.68%/2.50%**
Ensemble Max-pooling (p=0.15)	42.61%/19.52%	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%

6 总结

我们提出了一种简单而有效的汇合方法，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更好的效果. 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，我们将在更深网络和多个数据集下进行实验，同时对集成最大汇合和经典最大汇合的异同进行可视化.

参 考 文 献

[1] Zeiler M, Fergus R. Stochastic pooling for regularization of deep convolutional neural networks[C]//Proc of the Int Conf on Learning Representation, 2013

[2] Huang Yuchi, Sun Xiuyu, Lu Ming, et al. Channel-max,channel-drop and stochastic max-pooling[C]//Proc of the IEEE Conf on Computer Vision and Pattern Recognition Workshops, NJ: IEEE, 2015: 9–17

[3] Cai Meng, Shi Yongzhe, Liu Jia. Stochastic pooling maxout networks for low-resource speech recognition[C]//Proc of the 2014 IEEE Int Conf on In Acoustics, Speech and Signal Processing, NJ: IEEE, 2014: 3266–3270

[4] He Kaiming, Zhang Xiangyu, Ren Shaoqing, et al. Deep residual learning for image recognition[C]//Proc of the 2016 IEEE Conf on Computer Vision and Pattern Recognition, NJ: IEEE, 2016: 770–778

[5] Veit A, Wilber M, Belongie S. Residual networks are exponential ensembles of relatively shallow networks[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[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]//Proc of the Int Conf on Neural Information Processing, Berlin: Springer, 2015: 46–54

[8] Hinton G, Srivastava N, Krizhevsky A, et al. Improving neural networks by preventing co-adaptation of feature detectors[OL]. [2017-02-14] https://arxiv.org/abs/1207.0580

[9] He Kaiming, Zhang Xiangyu, Ren Shaoqing, et al. Identity mappings in deep residual networks[C]//Proc of the 14th European Conf on Computer Vision, Berlin: Springer, 2016: 630–645

[10] Simard P, Steinkraus D, Platt J, et al. Best practices for convolutional neural networks applied to visual document analysis[C]//Proc of the Int Conf on Document Analysis and Recognition, Washington: IEEE Computer Society, 2003: 958–962

[11] 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

[12] Jia Yangqing, Shelhamer E, Donahue J, et al. Caffe: Convolutional architecture for fast feature embedding[C]//Proc of the 22nd ACM international conference on Multimedia, New York: ACM, 2014: 675–678

[13] Krizhevsky A, Sutskever I, Hinton G. Imagenet classification with deep convolutional neural networks[C]//Advances in Neural Information Processing Systems, Massachusetts: MIT Press, 2012: 1097–1105

[14] Krizhevsky A, Hinton G. Learning multiple layers of features from tiny images. 2009

[15] Chen Tianqi, Li Mu, Li Yutian, et al. Mxnet: A flexible and efficient machine learning library for heterogeneous distributed systems[C]//Proc of Neural Information Processing Systems, Workshop on Machine Learning Systems, NIPSW, Massachusetts: MIT Press, 2015

[16] Huang Gao, Sun Yu, Liu Zhuang, et al. Deep networks with stochastic depth[C]//Proc of the 14th European Conference on Computer Vision, Berlin: Springer, 2016: 646–661

[17] Deng Jia, Dong Wei, Socher Richard, et al. Imagenet: A large-scale hierarchical image database[C]//Proc of In IEEE Conf on Computer Vision and Pattern Recognition, NJ: IEEE, 2009: 248–255

Zhang Hao, born in 1994. Master. His main research interests include computer vision and machine learning.

Wu Jianxin, born in 1978. PhD, professor, PhD supervisor. Recipient of the “Thousand Talents Program” for Distinguished Young Scholars. His main research interests include computer vision and machine learning.