doi:10.11959/j.issn.1000-436x.2017145

基于混合 maxout 单元的卷积神经网络性能优化

赵慧珍，刘付显，李龙跃，罗畅
（空军工程大学防空反导学院，陕西 西安 710051）

摘 要：针对深度卷积神经网络中 maxout 单元非最大特征无法传递、特征图像子空间池化表达能力不足的局限性，提出混合 maxout (mixout, mixed maxout)单元。首先，计算相同输入在不同卷积变换下所形成的特征图像子空间的指数概率分布；其次，根据概率分布计算特征图像子空间的期望；最后，利用伯努利分布对子空间的最大值与期望值加权，均衡单元模型。分别构建基于 mixout 单元的简单模型和网中网模型进行实验，结果表明 mixout 单元模型性能较好。

关键词：深度学习；卷积神经网络；maxout 单元；激活函数
中图分类号：TP391.3 **文献标识码**：A

Improving deep convolutional neural networks with mixed maxout units

ZHAO Hui-zhen, LIU Fu-xian, LI Long-yue, LUO Chang
(School of Air and Missile Defense, Air Force Engineering University, Xi'an 710051, China)

Abstract: The maxout units have the problem of not delivering non-max features，resulting in the insufficient of pooling operation over a subspace that is composed of several linear feature mappings, when they are applied in deep convolutional neural networks. The mixed maxout (mixout) units were proposed to deal with this constrain. Firstly, the exponential probability of the feature mappings getting from different linear transformations was computed. Then, the averaging of a subspace of different feature mappings by the exponential probability was computed. Finally, the output was randomly sampled from the max feature and the mean value by the Bernoulli distribution, leading to the better utilizing of model averaging ability of dropout. The simple models and network in network models was built to evaluate the performance of mixout units. The results show that mixout units based models have better performance.

Key words: deep learning, convolutional neural network, maxout units, activation function

1 引言

近年来，均衡随机单元进而规则化深度卷积神经网络（CNN, convolutional neural network）[1,2]成为改善深度学习过拟合问题的有效手段[3]。2012 年，Hinton 等[4]提出的 dropout 是首个通过均衡随机单元以达到 CNN 模型规则化的方法，通过在全连接层利用伯努利分布对连接加权，减少分类对任意单元的过度依赖，从而改善过拟合现象；Krizhevsky 等[5]验证了 dropout 在不同规模数据集上的适用性；Wang 等[6]利用高斯近似法进行快速 dropout 训练；Ba 等[7]利用与深度模型共享参数的二置信网络计算每个隐含层的 dropout 概率，改进模型学习效果；Tompson 等[8]将 dropout 拓展应用到整个特征空间，形成空间 dropout 方法；Wan 等[9]在 dropout 的基础上提出 dropconnect 方法，与 dropout 随机将神经元输出置零不同，dropconnect 随机将权重矩阵元素置零。Dropout 能够训练大量共享参数的单元模型，且均衡这些单元模型对整个模型输出的影响，有效改善过拟合现象，提高模型特征学习能力。然而，dropout 在后向传播中的更新针对的是不同训练子集上的不同模型，因此，能够在参数共享条件下将

收稿日期：2016-09-27；**修回日期**：2017-03-02
基金项目：国家自然科学基金资助项目（No.61601499）
Foundation Item: The National Natural Science Foundation of China (No.61601499)

总体分块训练的模型是适用于 dropout 的理想模型[10]；而传统模型在激活函数的限制下，均为总体训练，所以 dropout 与传统 CNN 模型的结合不能较好地利用其均衡单元模型的能力。

为更好地利用 dropout 均衡单元模型的能力，Goodfellow 等[10]设计了 maxout 单元，同时提出了基于 maxout 单元的 CNN 网络模型。多个 maxout 单元能够组合成任意连续线性线段，从而对任意非线性凸函数进行逼近，因此可以更加准确地模仿生物神经元特性，避免传统非线性激活函数引起的单元非活性化，更加有效地利用 dropout 均衡单元模型的能力。此外，每个 maxout 单元对特征图像子空间进行了最大池化[11]操作，选取子空间的最大特征图像，用来代表图像子空间的信息。然而，maxout 单元只输出了子空间的最大特征值，忽略了非最大特征值的影响，造成 maxout 单元对特征图像子空间的池化能力不足。针对非最大特征值无法传递的问题，Springenberg 等[3]提出概率 maxout（probout, probabilistic maxout）单元，计算所有特征图的概率，再根据概率对子空间抽样，使非最大特征值有可能被选择，但该算法增加了模型的复杂度。

针对 maxout 单元只利用特征图像子空间的最大特征值、其他特征值无法传递和特征图像子空间池化表达能力不足的问题，本文提出 mixout 单元，首先根据指数概率分布计算特征图子空间的期望，再利用伯努利分布对输入特征图像的最大值与期望值加权，使 mixout 单元在保留 maxout 单元的分段线性优势的基础上提高对特征图像子空间的池化表达能力，从而更加充分地利用 dropout 均衡单元模型的能力。最后，构建了基于 mixout 单元的简单模型，以分析其池化能力；根据 CNN 中常见的网中网（NIN, network in network）结构，构建了结合 mixout 单元的 M-NIN 模型，利用 3 个标准数据集 CIFAR-10、CIFAR-100 和 SVHN，分析基于 mixout 单元的一般模型的分类能力。

2 maxout 单元与 dropout 方法

2.1 传统激活函数

传统激活函数有 sigmoid、tanh 等饱和函数和 ReLU、LReLU 等非饱和函数，本节选取了几种常用的函数做了简要介绍与分析，便于后续实验对比。

Sigmoid 是常用的饱和激活函数，具有对称性和零均值特性等优点，其定义为

$$h(x)=\frac{1}{1+\mathrm{e}^{x}} \tag{1}$$

其中，x 为激活函数层的输入。然而在输入 x 的值较大或较小时，输出 $h(x)$ 趋于饱和，不能驱动网络底层的参数更新，如图 1(a)所示。

线性修正单元（ReLU, rectified linear unit）是常用的非饱和激活函数，具有单侧抑制、相对宽阔的兴奋边界、稀疏激活性等优点[12]。ReLU 定义为

$$h(x)=\max(0,x) \tag{2}$$

如图 1(b)所示，ReLU 将负数输入置为 0 的处理使梯度不能在负数部分传播，限制了学习速率和学习效果。

Mass 等[13]提出弱修正线性单元（LReLU, leaky ReLU）以避免 ReLU 在负数处理部分的零梯度问题，LReLU 引入坡度因子的概念对 ReLU 负数部分进行修正，使函数在负数部分具有固定梯度。LReLU 定义为

$$h(x)=\max(\alpha x,x) \tag{3}$$

其中，α 为坡度因子，可在 (0,1) 内灵活取值，取值越小，对负数部分的修正越小，如图 1(c)所示。LReLU 的训练结果对坡度因子的取值较为敏感。

He 等[14]针对损失函数对不同坡度因子不可导的特点提出参数修正线性单元（PReLU, parametric ReLU），其函数表达与 LReLU 相同，如式(3)所示，但与 LReLU 不同的是，PReLU 坡度因子 α 通过后向传播学习得到，灵活可变，如图 1(c)所示。灵活变化的坡度因子使 PReLU 在小数据集上容易出现过拟合。

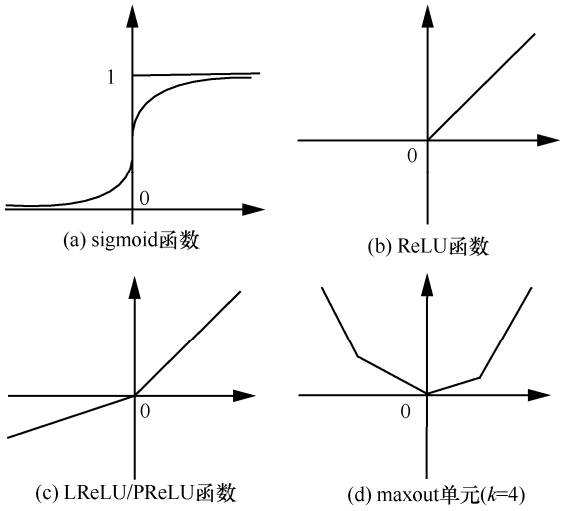


图 1 几种激活函数

在与规则化方法 dropout 结合的过程中，sigmoid 函数的饱和性限制了模型底层参数的更新，ReLU 系列模型的总体训练机制制约了 dropout 均衡单元模型的能力，而具有激活函数作用的 maxout 单元通过其连续分段线性及对 k 个特征图像子空间的最大池化作用，与 dropout 结合具有较好的表现[3,10]。

2.2 maxout 单元

设深度网络的第 l 层输入为 $\boldsymbol{x}\in\mathbb{R}^d$，则经过 maxout 单元处理后的输出 $h(\boldsymbol{x})$ 为 k 个特征图像组成的子空间的最大值，如式(4)所示。

$$h(\boldsymbol{x})=\max_{j\in[1,k]} z_j \tag{4}$$

其中，z_j 为输入 $\boldsymbol{x}$ 加权后与偏置项 $\boldsymbol{b}_j$ 之和，z_j 为

$$z_j=\boldsymbol{x}^{\mathrm{T}}\boldsymbol{W}_j+\boldsymbol{b}_j \tag{5}$$

其中，权值 $\boldsymbol{W}\in\mathbb{R}^{d\times m\times k}$ 与偏置项 $\boldsymbol{b}\in\mathbb{R}^{m\times k}$ 均由后向传播学习得到。当输入为一维数据且 $k=4$ 时，maxout 的函数图像如图 1(d)所示，maxout 单元以 4 段连续线段逼近二次函数。

Maxout 单元对于任意凸函数具有分段线性近似描述能力[10]，使其不仅能够在后向传递梯度时能够避免梯度消失/溢出等传统饱和激活函数遇到的问题，还能阻止 ReLU 函数负数处理部分引起的单元失活。

2.3 dropout 方法

均衡模型单元就是均衡每个单元对模型输出的影响，降低单元相互协同对模型输出的影响，从而提升模型学习的一般性。在训练过程中，若 2 个单元的协同作用对模型的输出有较大影响，模型会调整参数来学习这 2 个单元的共生关系，容易产生过拟合。CNN 等深度学习网络层数较多，参数量大，若模型单元不均衡，学习到的参数就不能反映数据集的真实信息。

Dropout 是一种通过均衡模型单元从而规则化模型的方法[4]。设深度网络含有 $L-1$ 个隐含层和一个 softmax 分类层，在训练阶段，对于每一个隐含层 l，设其输入为 $\boldsymbol{x}_l$，权重为 $\boldsymbol{W}_l$，偏置为 $\boldsymbol{b}_l$，激活函数为 f，则其输出 $h_l(\boldsymbol{x})$ 为

$$h_l(\boldsymbol{x})=f(\boldsymbol{W}_l^{\mathrm{T}}\boldsymbol{x}_l+\boldsymbol{b}_l),1\leqslant l\leqslant L-1 \tag{6}$$

隐含层 l 的输出并不直接作为 $l+1$ 层的输入，而是进行 dropout 处理，如式(7)所示。

$$\boldsymbol{x}_{l+1}=dropout(\boldsymbol{h}_l,p_{\mathrm{drop}}) \tag{7}$$

其中，p_{drop} 是 dropout 的概率，一般取 50%。式(7)表达的意思是，$\boldsymbol{h}_l$ 的值在以 p_{drop} 的概率置零后作为下一层的输入。

对于最后 softmax 分类层，输出为模型的最终分布，如式(8)所示。

$$D_L(\boldsymbol{x})=softmax(\boldsymbol{W}_L^{\mathrm{T}}\boldsymbol{x}_L+\boldsymbol{b}_L) \tag{8}$$

在测试阶段，隐藏单元的输出不再随机置零，而是将所有输出的权重减半，来抵消训练过程中 $\frac{1}{2}$ 的神经元输出置零带来的损失，这种模型称为平均模型。平均模型的损失约等于训练阶段模型 dropout 输出的损失。多次 dropout 训练输出分布的几何平均值越接近平均模型的值，dropout 越能够均衡单元模型[16]，模型的学习效果就越好。

利用 dropout 进行训练，相当于每次从大量数据中按照一定概率随机抽选了一部分进行下一步实验，多次迭代或实验后，每个单元被抽取的概率一致，基本不存在多个单元每次训练均同时出现的情况，避免了多个单元的协同作用，减少了对任意单元的过度依赖，使模型能够学习到输入数据的一般特征。

2.4 dropout 下 maxout 单元优势的分析

基于 maxout 单元的深度学习网络模型对 dropout 的充分利用来源于 2 个方面，一是 maxout 单元的连续线性分段特性，二是 maxout 单元对特征图子空间的池化作用。

1) Maxout 单元分段线性的作用。由 2.2 节可知，maxout 单元具有连续分段线性的特性。假设网络模型含有一层输入层和一层 softmax 分类层，softmax 层用来计算模型学习到的分布概率。输入层含有 N 组数据，利用 50%的概率对 2 层之间的连接进行 dropout 处理，则网络的连接共有 2^N 种可能方式，2^N 种网络模型可以学习到 2^N 种分布 $P_i(i\in(1,2^N))$。若对第一层数据进行线性处理，则利用 dropout 处理的分布 P_i 差异不大，故 P_i 的几何平均值的平方与平均模型输出 P_m 相等；若对第一层的数据进行非线性处理，则 P_i 之间差异相对较大，依据算术—几何平均值定理，P_i 的几何平均值的平方必小于平均模型的输出 P_m。根据归纳法，当网络模型加深时，上述结论依然成立。所以在利用 dropout 方法时，maxout 的连续分段线性能够使多次输出的几何平

均值最接近平均模型的输出，模型的学习效果较好。

2) Maxout 单元的池化作用。经过 maxout 单元处理后的输出 $h(x)$ 取 k 个特征图 z_j 组成的子空间的最大值，dropout 选择的连接具有一定的随机性，当连接改变时，特征图像子空间的组成会发生变化，对子空间进行池化，可以使输出相对于改变具有稳定性，从而表达特征。增加子空间的池化能力，一定程度上可以增加输出对图像的特征表达能力和细节描述能力。

然而，maxout 单元在 k 个特征图像组成的子空间中选取最大值，存在其他特征值无法前向传递，特征图像子空间池化表达能力不足的问题。

3 改进的 mixout 单元

由第 2 节分析可知，基于 maxout 单元的 CNN 模型对 dropout 均衡模型能力的充分应用来源于其分段线性逼近任意函数的能力以及对特征图像子空间的池化能力。基于 maxout 单元的模型处理图像的示意如图 2(b)所示。图 2(a)中实线方框内的图像子块作为输入，经过 2 种卷积核进行线性变换后，提取到同一输入与不同特征相匹配的值，分别为 0.8 与 0.3，同一输入的不同特征值组成了特征图子像空间，maxout 单元选取子空间的最大值，即 0.8 作为输出。在描述图像子块时，maxout 单元只描述了图像与特征 1 相匹配的部分，特征 2 被忽略。故 maxout 单元在 k 个特征图像组成的子空间中选取最大值传输，会造成其他特征值无法前向传递，特征图像子空间细节描述不足、池化表达能力不够的问题。

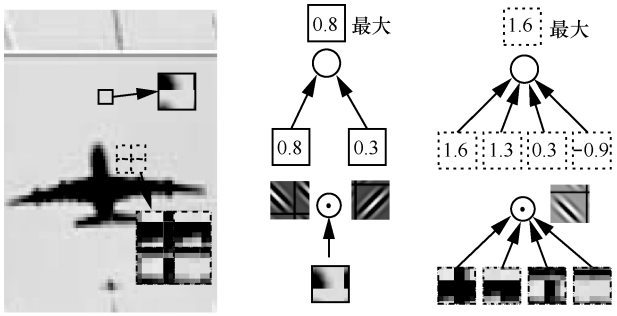


图 2 空间最大池化及 maxout 单元处理图像的示意

在采样阶段中的空间最大池化对图像的处理流程示意如图 2(c)所示。图 2(a)中虚线方框内的 4 个图像子块作为输入，经过同一种卷积核进行线性变换，提取到 4 个不同位置与相同特征相匹配的值 1.6、1.3、0.3、−0.9，对特征值进行最大池化操作，输出最大值 1.6。

同一输入在不同的卷积变换下提取了不同特征，形成了特征图像子空间，maxout 单元针对特征图像子空间进行池化；而下采样阶段的池化针对的是不同输入的特征图像，即不同的输入在相同卷积核的变换下，提取了多个输入的同一特征，而组成空间子空间，池化在空间子空间上进行[11]。

若在保持 maxout 单元分段线性逼近任意函数能力的基础上，提高对特征子空间的池化能力，CNN 模型将会对 dropout 均衡模型能力应用的更充分，特征学习能力更强，分类效果更好。

Yu 等[17]提出的混合池化是一种下采样阶段中的空间池化方法，利用伯努利分布对最大池化值与平均池化值进行抽样，使空间非最大值以 50%的概率对输出产生影响，提高了神经网络模型性能。在图 3(a)中将图 2(a)中虚线方框内的 4 个图像子块作为输入，经过同一种卷积核进行线性变换，提取到 4 个不同位置与相同特征相匹配的值 1.6、1.3、0.3、−0.9，对特征最大值 1.6 与期望值 0.6 进行伯努利抽样作为输出。

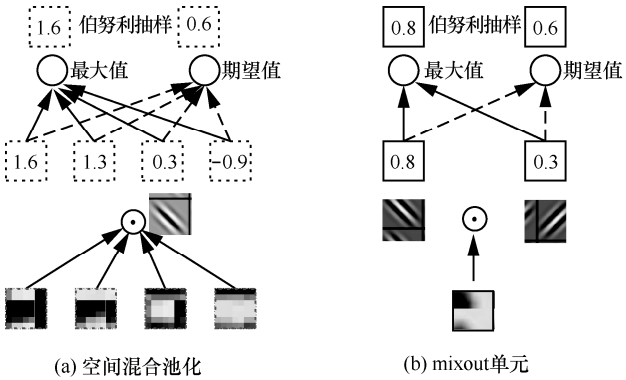


图 3 空间混合池化及 mixout 单元处理图像的示意

若将混合池化应用于 maxout 单元中，则一定程度上可以提高其池化能力。为充分利用不同卷积核提取的多个特征，本文结合混合池化的特点，提出 mixout 单元。

3.1 mixout 单元

设深度网络的第 l 层输入为 $\boldsymbol{x}\in\mathbb{R}^d$，$\boldsymbol{x}$ 可为首层输入或第 $l-1$ 层的输出，利用卷积核对输入进行线性变换，得到 $\boldsymbol{z}$，如式(9)所示。

$$\boldsymbol{z}=\boldsymbol{x}^{\mathrm{T}}\boldsymbol{W}+\boldsymbol{b} \tag{9}$$

其中，$\boldsymbol{W}$ 为权重，$\boldsymbol{b}$ 为偏置项，参数 $\boldsymbol{W}$ 与 $\boldsymbol{b}$ 均由后向传播学习获得。

输入经过不同卷积核提取到不同特征，特征以图像为表现形式，故又称特征图像，多个不同的特征图像组成一个特征子空间，定义子空间第 i 个特征图像的指数概率 p_i 为

$$p_i = \frac{\mathrm{e}^{z_i}}{\sum_{1}^{k} \mathrm{e}^{z_j}} \tag{10}$$

其中，z_i 为利用第 i 个卷积核提取的特征图像的值。计算特征图像子空间在指数概率下的期望 $\tilde{\mathrm{E}}$ 为

$$\tilde{\mathrm{E}} = \sum_{1}^{k} p_i z_i \tag{11}$$

显然，指数概率下的期望 $\tilde{\mathrm{E}}$ 能够描述特征子空间的一般性。利用伯努利分布对特征图像子空间的最大值和平均值进行加权，输出 $h(x)$ 定义为

$$h(x) = \lambda \max_{i\in[1,k]} z_i + (1-\lambda)\tilde{\mathrm{E}} \tag{12}$$

其中，λ 服从伯努利二项分布，随机取 0 或 1，其定义为

$$\lambda \sim \mathrm{B}(0,1) \tag{13}$$

当 $\lambda = 1$ 时，特征子空间选取最大特征值，当 $\lambda = 0$ 时，特征子空间选择指数概率下的期望值。

3.2　mixout 单元的连续分段线性特征

定理 1　若网络中含有 2 层以上隐含层，且隐含层由 mixout 单元组成，则网络可以在紧集 $C \in \mathbb{R}^n$ 中以连续分段线性函数逼近任意连续凸函数 f。

证明　以 2 层隐含层网络为例说明，x 为输入，z_1 与 z_2 为输入加权后与偏置项的和，h_1 与 h_2 为隐含层，g 为输出，如图 4 所示。

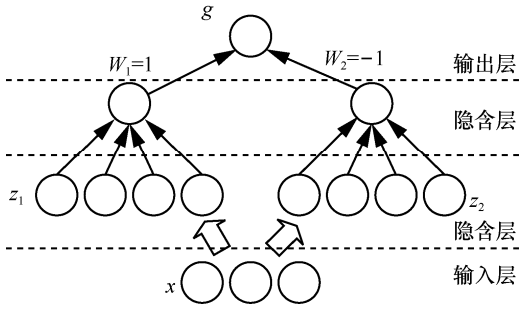


图 4　具有 2 层隐含层的网络

由文献[15]可知，对于任意 2 个正整数 m 与 n，存在 2 组 $n+1$ 维的实数参数向量 $[W_{1j}, b_{1j}]$ 与 $[W_{2j}, b_{2j}]$，$j \in [1,k]$，使 $g(v) = h_1(v) - h_2(v)$。则对任意分段线性函数可用 2 个凸分段线性函数之差表示，mixout 单元具有分段线性，2 个以上的 mixout 单元可以描述任意分段线性函数。

由 Stone-Weierstrass 近似定理可知[10]，设紧集 $C \in \mathbb{R}^n$，$f: C \to \mathbb{R}$ 为连续凸函数，ε 为任意正实数，则存在连续分段线性函数 g，对于所有 $v \in C$，有 $\left|f(v) - g(v)\right| < \varepsilon$。则在紧集 $C \in \mathbb{R}^n$ 中，存在连续分段线性函数，可以任意逼近 C 中的连续凸函数。

综上，2 个以上 mixout 单元可以描述任意分段线性函数，连续分段线性函数可以任意逼近连续凸函数。故若网络中含有 2 层以上隐含层，且隐含层由 mixout 单元组成，则网络能够逼近定义域 C 上的任意连续函数，且当 $\varepsilon \to 0$ 时，$k \to \infty$。

3.3　mixout 单元池化能力的分析

基于 mixout 单元的模型对图像的处理流程如图 3(b)所示。将图 2(a)中实线方框内的图像子块作为输入 x，利用式(9)对 x 进行 2 种不同的线性变换得到 $z_1 = 0.8$ 与 $z_2 = 0.3$，z_1 与 z_2 组成了特征图像子空间，利用式(10)计算每个特征图像的指数概率分别为 0.6、0.4，利用式(11)计算子空间期望值为 0.6，利用式(12)对最大值 0.8 与期望值 0.6 加权，得到输出。

显然，特征图像子空间的最大值描述了图像子块的最大特征值，期望值描述了图像子块非最大特征值的一般性。当模型进行多次迭代时，由于伯努利分布的随机性，mixout 的输出既可以描述图像子块的最大特征值，又可以描述子块的其他细节，从而提高 mixout 单元的池化能力。

综上，mixout 单元保留了 maxout 单元的分段线性，且提高了特征图像子空间的池化能力。由于基于 maxout 单元的 CNN 模型对 dropout 均衡模型能力的充分应用来源于其分段线性特征以及对特征图像子空间的池化能力，所以 mixout 单元能够提高对 dropout 均衡单元模型能力的应用。

4　实验分析

本文从 2 个方面对 mixout 单元进行实验分析。首先是对 mixout 单元池化能力的分析，构建简单的卷积神经网络模型，对 maxout、probout 及 mixout 单元所提取的特征进行定性与定量分析。其次，对基于 mixout 单元的一般模型的分析，构建基于 ReLU、LReLU、PReLU、maxout、probout 及 mixout 单元的网中网模型，并使用 dropout 方法对模型进

行规则化，对 3 种模型的学习分类能力进行定量分析。实验所用处理器为 Intel(R)Core(TM)i5-4590，CPU 主频为 3.3 GHz，显卡为 AMD 系列。实验所用软件平台为 MatConvNet。

4.1 mixout 单元池化能力的实验

为了更好地分析 mixout 单元的池化能力，本文构建了只有一层 mixout 单元的简单卷积神经网络，以排除其他因素的干扰。简单卷积神经网络包含一层输入层，一层 mixout 层，一层输出层，如图 5 所示。设隐含层 mixout 单元的数量为 100 个，利用 dropout 方法处理输出层连接，概率为 50%。将网络中的 mixout 单元替换为 maxout 单元及 probout 单元，其他参数均保持不变，进行对比实验。

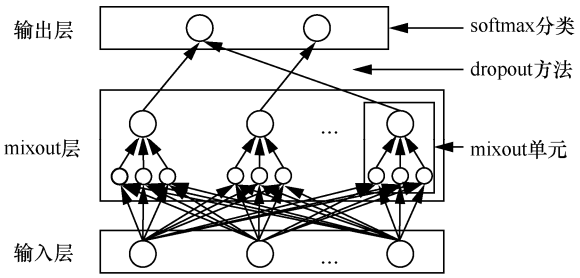
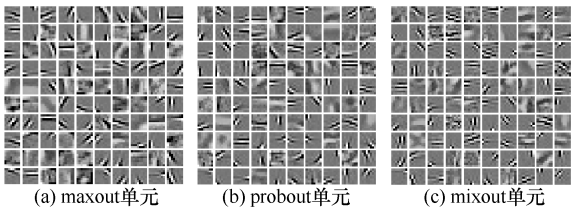


图 5 具有一层 mixout 单元的网络

利用数据库 CIFAR-10[19]进行训练。CIFAR-10 含有 60 000 张彩色图像，包括 50 000 张训练图像和 10 000 张测试图像，图像中的物体如动物、汽车等都处于图像中间，图像分辨率为 32 像素×32 像素，可分为 10 类，每类有 5 000 张训练图像和 1 000 张测试图像。将 3 种网络学习到的特征做可视化处理，结果如图 6 所示。



(a) maxout单元 (b) probout单元 (c) mixout单元

图 6 可视化处理网络学习到的特征

图 6(a)为基于 maxout 单元的网络学习到的特征，图 6(b)为 probout 单元学习到的特征，图 6(c)为 mixout 单元学习到的特征。由图 6 可知，与图 6(a)和图 6(b)相比，图 6(c)具有更好的边缘和细节信息。

KL 散度（Kullback-Leibler divergence）是一种用于描述 2 个分布之间差异性的工具[1]，其定义为

$$KL=\sum_{x\in\Omega}p(x)\ln\frac{p(x)}{q(x)} \tag{14}$$

其中，x 属于空间 Ω，p 与 q 分别是 x 在 Ω 中的 2 种概率分布，一般 p 表示真实分布，q 表示近似分布。KL 散度越大，表明二者之间差异越大；反之，表明近似分布 q 越接近真实分布 p。本文将平均模型的输出设为真实分布 p，将利用 dropout 方法处理的模型进行多次实验，取其分布的几何平均值为近似分布 q。基于 maxout、probout 和 mixout 单元模型的 KL 散度变化如图 7 所示。

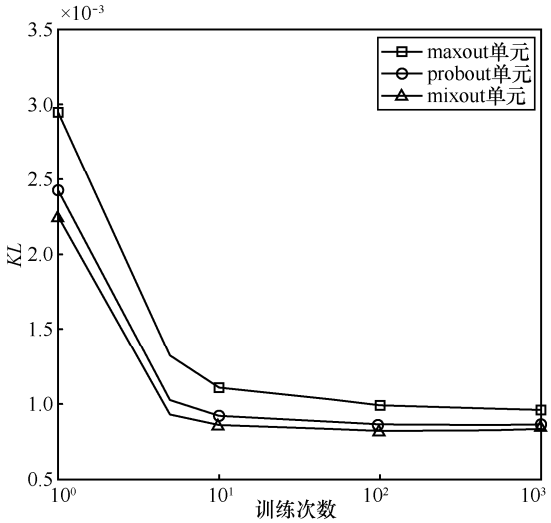


图 7 3 种模型的 KL 散度变化

由图 7 可知，随着实验次数的增加，3 种模型的 KL 散度均有下降趋势，说明多次实验所得分布的几何平均值越来越接近平均模型的分布，每个单元对模型的影响越来越均衡。mixout 单元的值始终最小，说明与其他几个函数相比，mixout 单元的分布更接近平均模型的分布，其每个单元对模型的影响更为均衡，对 dropout 均衡单元模型的能力利用得更加充分。

4.2 基于 mixout 单元的一般模型的分析

为进一步分析 mixout 单元的性能，本文对基于 mixout 单元的一般模型的特征学习能力进行实验。目前常用的 CNN 模型包括 AlexNet[3]、NIN[18]、ResNet[20]等，由于模型本身的结构和深度，不同模型的学习效果有些差异，相关文献对这些模型进行了详细分析，由于篇幅有限，此处对各个模型之间的差异不进行赘述。在现有模型中，NIN 网络以其特有的结构表现出优越的性能，因此，本文选取 NIN 网络作为基本框架，分别构建了基于 mixout

单元、maxout 单元、probout 单元、ReLU 函数 LReLU 函数以及 PReLU 函数的 NIN 模型，来进行实验分析。

4.2.1　实验数据集

实验数据集为 CIFAR-10[19]、CIFAR-100[19]与 SVHN[21]，其中，CIFAR-100 是 CIFAR-10 的扩展，含有 60 000 张彩色图像，可被分为 100 类，每类有 500 张训练图像和 100 张测试图像；SVHN 是从谷歌地图的街道图片上截取的房屋门牌号的集合，含有超过 600 000 张分辨率为 32 像素×32 像素的彩色图像，拍摄图像的角度与距离是随机的，因此数字具有不同的形变，但都处于图像的中间，与 MNIST 一致，SVHN 将图像按照数字 0~9 分为 10 类，而与 MNIST 不同的是，SVHN 每张图片可能不只含有一个数字，从而增加了分类的难度。

综合考虑数据集的样本数与分类数可知，CIFAR-100 每类样本数有 600 张，数量较少，CIFAR-10 每类样本数有 6 000 张，数量一般，SVHN 每类样本数有 60 000 张，数量较多，实验选取这 3 类数据集，可以一定程度上验证模型的适用性。

4.2.2　M-NIN 模型及其参数设置

NIN 网络通常由卷积层、多层感知层与池化层组成[18]，其中，多层感知层通过非线性变换代替了 CNN 传统网络中的激活函数层。本文以文献[18]中的网络为基本框架，构建了基于 mixout 单元的 NIN 模型，即 M-NIN 模型，如图 8 所示。

由图 8 可知，M-NIN 模型含有 3 个 M-NIN 区块与一个 softmax 分类层，3 个 M-NIN 区块依次相连提取图像特征，最后通过 softmax 层进行分类，从而构成 M-NIN 网络。一个 M-NIN 区块由卷积阶段、多层感知阶段与池化阶段 3 个阶段组成，卷积阶段的操作有输入、卷积、逐层归一化、mixout 单元处理；mixout 多层感知（MMLP, mixout multi-layer perception）阶段包含相同的 2 个感知阶段，每个阶段包含一层感知层、一层逐层归一化层与一层 mixout 单元层处理；池化阶段包含一层池化层。

M-NIN 模型参数设置如表 1 所示。网络共有 3 个 M-NIN 区块和一个 softmax 分类层，3 个 M-NIN 区块中共有 3 个卷积层，均含有 192 个 mixout 单元，卷积核尺寸分别为 5×5、5×5、3×3，卷积为对不同位置的同一个非线性变换，故其子空间维数为 0，其边界补充像素为 2，滑动像素为 1；每个 M-NIN 区块中含有 2 层 MMLP，3 个 M-NIN 区块中的第 1 层 MMLP 含有的 mixout 单元数分别设置为 160、192、192，第 2 层 MMLP 含有的 mixout 单元数设置为 96、160、192，由于 mixout 单元是对相同输入的不同特征的操作，故其核尺寸均为 1×1，每个 mixout 单元作用于 5 个特征图像，故其子空间维数为 5，无边缘补充像素，滑动像素为 1；3 个 M-NIN 区块中共有 3 层池化层，分别选择最大池化、最大池化、平均池化，其核尺寸分别为 3×3、3×3、8×8，池化层针对的是同一个特征图像，故其子空间维数为 0，边界补充像素为 2，滑动像素分别设置为 2、2、1；每个区块最后连接均使用 dropout 方法进行随机选择，选择概率为 0.5；模型最后一层为用于分类的 softmax 层，针对数据集 CIFAR-10、CIFAR-100

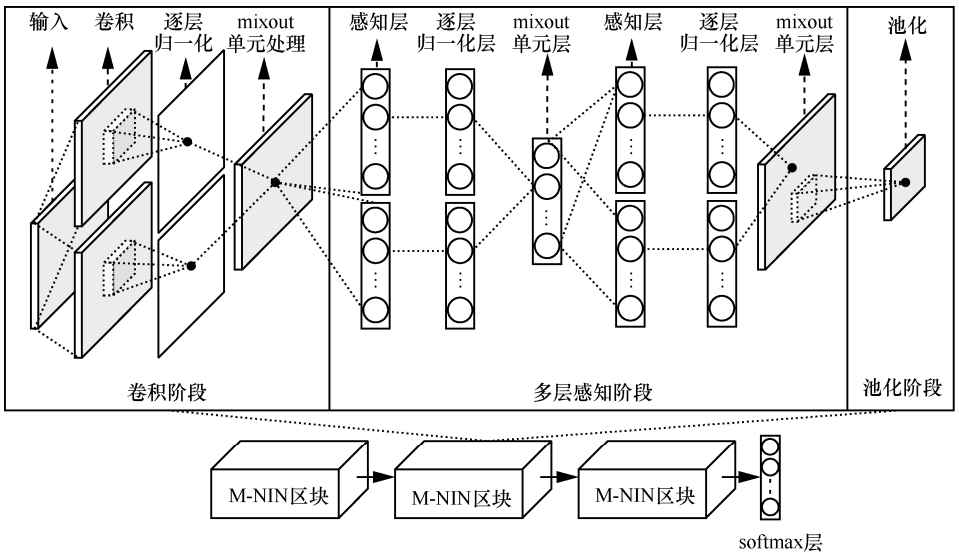


图 8　M-NIN 模型结构

表 1 M-NIN 模型参数设置

M-NIN 网络结构	层次	mixout 单元数	核尺寸	子空间维数	补充像素	滑动像素
M-NIN 区块 1	卷积层	192	5×5	0	2	1
	MMLP -1	160	1×1	5	0	1
	MMLP-2	96	1×1	5	0	1
	池化层	最大池化	3×3	0	2	2
M-NIN 区块 2	卷积层	192	5×5	0	2	1
	MMLP -1	192	1×1	5	0	1
	MMLP -2	192	1×1	5	0	1
	池化层	最大池化	3×3	0	2	2
M-NIN 区块 3	卷积层	192	3×3	0	2	1
	MMLP -1	192	1×1	5	0	1
	MMLP -2	160	1×1	5	0	1
	池化层	平均池化	8×8	0	2	1
softmax 层			分类 10/100			

与 SVHN，分别设置分类数为 10、100、10。

基于 maxout 单元、probout 单元的 NIN 网络与 M-NIN 网络类似，为保证实验的公平性，在训练基于 maxout 单元与 probout 单元的 NIN 网络时，只将 M-NIN 模型中的 mixout 单元分别更换为 maxout 单元、probout 单元，其他参数保持不变。M-NIN 模型参数设置参考了文献[3,10,17,22]。基于 ReLU、LReLU、PReLU 的 NIN 模型参数设置参考了文献[17, 23~25]，并根据本文具体实验进行了微调。

4.2.3 实验方法

实验按照文献[10]的方法训练模型。在处理 CIFAR-10 数据集时，首先对所有数据进行归一化与白化处理，再利用 40 000 张训练图像训练 M-NIN 模型；其他 1 0000 张训练图像用于交叉验证，记录验证错误率不再减少时的模型迭代次数；最后在整个训练集上训练模型直至达到该迭代次数。CIFAR-100 的训练方法与 CIFAR-10 一致。处理 SVHN 数据集时，采用 73 257 张训练图片和 20 032 张测试图片进行实验，此外，采用 531 131 张降低难度的图片作为额外训练数据。每类选取 400 张（共 4 000 张）训练图片用于模型训练，选取 200 张（共 2 000 张）额外训练图片用于交叉验证，最后在整个 598 388 张训练集上训练模型。利用上述方法及表 1 的参数分别训练基于 maxout 单元、probout 单元、与 mixout 单元的 NIN 模型；利用文献[25~27]的结构与参数分别训练基于 ReLU、LReLU、PReLU 的 NIN 模型。

4.2.4 实验结果与分析

实验结果如表 2 所示。整体而言，所有方法中，数据集 CIFAR-100 的误差均比数据集 CIFAR-10 与 SVHN 大，这是由于 CIFAR-100 每类样本数只有 600 张，数量较少，模型学习的样本不足，故数据集 CIFAR-100 上的误差整体较大；而数据集 SVHN 的每类样本数为 60 000 张，模型可以进行充分学习，故其整体分类错误率较低；基于 maxout 及其改进单元 probout、mixout 单元的模型均比基于 ReLU 及其改进函数的模型的错误率低，说明了在与模型 dropout 方法的结合中，maxout 单元、probout 单元及 mixout 单元具有更好的特征学习能力与图像分类能力；在 3 个数据集的分类中，M-NIN 模型的分类错误率均为最低，说明了 M-NIN 具有较好的特征学习能力。

表 2 不同模型在 3 个数据集中的错误率

模型	CIFAR-10	CIFAR-100	SVHN
ReLU 函数	12.45%	42.90%	2.78%
LReLU 函数	11.20%	42.04%	2.69%
PReLU 函数	11.79%	41.63%	2.51%
maxout 单元	11.68%	38.57%	2.47%
probout 单元	11.35%	38.14%	2.39%
mixout 单元	11.02%	37.39%	2.31%

为了充分分析 M-NIN 模型的分类能力，本文对实验数据进行进一步的处理。在 3 个数据集中，

分别计算基于 ReLU 函数、LReLU 函数、PReLU 函数、maxout 单元、probout 单元的模型与 M-NIN 模型的相对错误率，如图 9 所示。

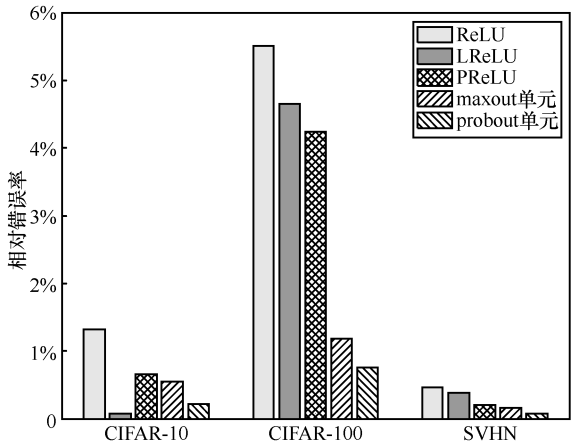


图 9　M-NIN 模型与其他模型的相对错误率

由图 9 可见，相比于数据集 CIFAR-10 与 SVHN，在 CIFAR-100 上，其他模型与 M-NIN 模型的错误率之差整体较大，说明了 M-NIN 模型在 CIFAR-100 上的性能提升较多，这是由于 CIFAR-100 的每类样本较少，mixout 单元能够更好地利用 dropout 均衡模型的能力规则化数据，减少过拟合，而数据量较大的训练集一定程度上有防止过拟合的作用，所以 M-NIN 模型在 CIFAR-100 上的错误率之差整体较大；相比于 maxout 单元，mixout 单元在 3 个数据集上的错误率降低了 5.5%、11.8%、1.6%，说明了基于 mixout 单元的模型有更好的特征学习能力与分类能力。值得注意的是，相比于 LReLU，PReLU 在小训练集 CIFAR-100 上的表现较好，这是因为 PReLU 灵活可变的坡度因子使其在训练过程中容易出现过拟合。

相比于 maxout 单元，mixout 单元增加了求取期望与伯努利抽样 2 个步骤，增加了算法复杂性，但综合整个模型的复杂性与参数计算量，增加的步骤并没有造成明显的时间消耗，以数据集 CIFAR-10 为例，记录 6 种模型前 10 次迭代的时间消耗如图 10 所示。需要强调的是，实验时间仅作为本文所有实验横向比较参考依据。

由图 10 可知，在每次迭代需要的时间上，M-NIN 模型与基于 maxout 单元、probout 单元的模型并没有明显区分开，这是因为 CNN 网络模型本身具有大量参数，步骤复杂，相比于整个模型的复杂度，probout 单元与 mixout 单元对 maxout 单元的改进并没有引入过多的计算量；此外，M-NIN 模型与基于 maxout 单元、probout 单元的模型每次迭代的时间消耗比基于 ReLU 及其改进函数的模型时间消耗要大，这是由于 ReLU 等是单纯的函数处理，而 maxout 单元等为相对复杂的单元变换，处理步骤相对较多。

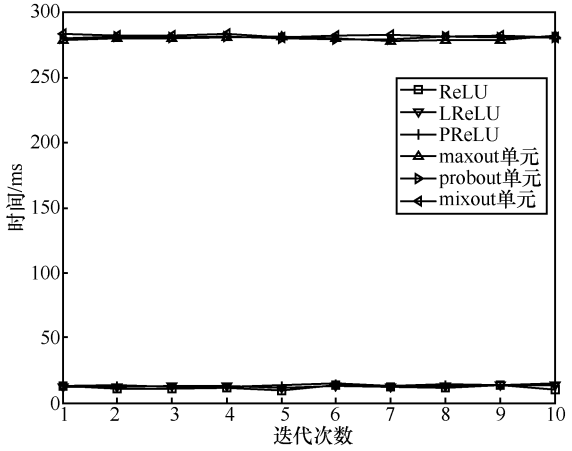


图 10　6 种模型的时间消耗比较

5　结束语

针对 maxout 单元只利用同一输入的最大特征图、其他特征无法传递和特征图子空间池化表达能力不足的问题，提出了 mixout 单元，在保留 maxout 单元分段线性优势的同时，提高特征图像子空间的池化表达能力，更加充分地利用 dropout 均衡单元模型的能力；构建了基于 mixout 单元的简单模型，验证了 mixout 单元较好地池化能力；构建了基于 mixout 单元的 NIN 模型，并利用标准数据集进行实验，证明了 M-NIN 模型较好的特征学习能力及分类能力。Mixout 单元对 dropout 的充分利用，使模型能够减少过拟合，然而对于较大数据集，大量的训练数据本身对于过拟合有改善作用，模型在大型训练集上的表现并不理想，需要进一步研究。

参考文献：

[1] 周昌令, 栾兴龙, 肖建国. 基于深度学习的域名查询行为向量空间嵌入[J]. 通信学报, 2016, 37(3): 165-174.
ZHOU C L, LUAN X L, XIAO J G. Vector space embedding of DNS query behaviors by deep learning[J]. Journal on Communications, 2016, 37(3): 165-174.

[2] 杨钊, 陶大鹏, 张树业, 等. 大数据下的基于深度神经网的相似汉字识别[J]. 通信学报, 2014, 35(9): 184-189.

YANG Z, TAO D P, ZHANG S Y, et al. Similar handwritten Chinese character recognition based on deep neural networks with big data[J]. Journal on Communications, 2014, 35(9): 184-189.

[3] SPRINGENBERG J T, RIEDMILLER M. Improving deep neural networks with probabilistic maxout units[J]. arXiv preprint arXiv:1312.6116, 2013.

[4] HINTON G E, SRIVASTAVA N, KRIZHEVSKY A, et al. Improving neural networks by preventing co-adaptation of feature detectors[J]. arXiv preprint arXiv:1207.0580, 2012.

[5] KRIZHEVSKY A, SUTSKEVER I, HINTON G E. ImageNet classification with deep convolutional neural networks[C]//The 26th Annual Conference on Neural Information Processing Systems. 2012: 1097-1105.

[6] WANG S I, MANNING C D. Fast dropout training[C]//The 30th International Conference on Machine Learning. 2013: 118-126.

[7] BA J, FREY B. Adaptive dropout for training deep neural networks[C]//The Advances in Neural Information Processing Systems. 2013: 3084-3092.

[8] TOMPSON J, GOROSHIN R, JAIN A, et al. Efficient object localization using convolutional networks[C]//The IEEE Conference on Computer Vision and Pattern Recognition. 2015: 648-656.

[9] WAN L, ZEILER M, ZHANG S, et al. Regularization of neural networks using dropconnect[C]//The 30th International Conference on Machine Learning. 2013: 1058-1066.

[10] GOODFELLOW I J, WARDE-FARLEY D, MIRZA M, et al. Maxout networks[C]//The 30th International Conference on Machine Learning. 2013: 1319-1327.

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

[12] NAIR V, HINTON G E. Rectified linear units improve restricted Boltzmann machines[C]//The 27th International Conference on Machine Learning. 2010: 807-814.

[13] MAAS A L, HANNUN A Y, NG A Y. Rectifier nonlinearities improve neural network acoustic models[C]//The 30th International Conference on Machine Learning. 2013, 30(1).

[14] HE K, ZHANG X, REN S, et al. Delving deep into rectifiers: surpassing human-level performance on ImageNet classification[C]// The IEEE International Conference on Computer Vision. 2015: 1026-1034.

[15] WANG S. General constructive representations for continuous piecewise-linear functions[J]. IEEE Transactions on Circuits and Systems I: Regular Papers, 2004, 51(9): 1889-1896.

[16] LI J, WANG X, XU B. Understanding the dropout strategy and analyzing its effectiveness on LVCSR[C]//38th IEEE International Conference on Acoustics, Speech, and Signal Processing. 2013: 7614-7618.

[17] YU D, WANG H, CHEN P, et al. Mixed pooling for convolutional neural networks[C]//International Conference on Rough Sets and Knowledge Technology. 2014:364-375.

[18] LIN M, CHEN Q, YAN S. Network in network[J]. arXiv Preprint arXiv:1312.4400, 2013.

[19] KRIZHEVSKY A, HINTON G E. Learning multiple layers of features from tiny images[R]. Computer Science Department, University of Toronto, Tech. 2009.

[20] HE K M, ZHANG X Y, REN S Q, et al. Deep residual learning for image recognition[J]. arxiv Prepraint arXiv: 1512.03385, 2015.

[21] NETZER Y, WANG T, COATES A, et al. Reading digits in natural images with unsupervised feature learning[R]. NIPS Workshop on Deep Learning and Unsupervised Feature Learning, 2011.

[22] CHANG J R, CHEN Y S. Batch-normalized maxout network in network[J]. arXiv Preprint arXiv:1511.02583, 2015.

[23] JIN X, XU C, FENG J, et al. Deep learning with s-shaped rectified linear activation units[J]. arXiv Preprint arXiv:1512.07030, 2015.

[24] LIAO Z, CARNEIRO G. On the importance of normalisation layers in deep learning with piecewise linear activation units[C]//The 2016 IEEE Winter Conference on Applications of Computer Vision. 2016: 1-8.

[25] XU B, WANG N, CHEN T, et al. Empirical evaluation of rectified activations in convolutional network[J]. arXiv Preprint arXiv: 1505.00853, 2015.

[26] SRIVASTAVA N, HINTON G, KRIZHEVSKY A, et al. Dropout: a simple way to prevent neural networks from overfitting[J]. The Journal of Machine Learning Research, 2014, 15(1): 1929-1958.

[27] CLEVERT D A, UNTERTHINER T, HOCHREITER S. Fast and accurate deep network learning by exponential linear units (ELUS) [J]. arXiv preprint arXiv:1511.07289, 2015.

作者简介：

赵慧珍（1990-），女，山东单县人，空军工程大学博士生，主要研究方向为深度学习、计算机视觉。

刘付显（1962-），男，山东曹县人，空军工程大学教授，主要研究方向为作战建模与仿真、指挥决策优化。

李龙跃（1988-），男，河南驻马店人，空军工程大学博士生，主要研究方向为作战建模与仿真、指挥决策优化。

罗畅（1988-），男，四川广安人，空军工程大学博士生，主要研究方向为深度学习、计算机视觉。