深度神经网络训练中适用于小批次的归一化算法

王 岩 吴晓富
(南京邮电大学通信与信息工程学院 南京 210003)

摘 要 近年来,批归一化(Batch Normalization,BN)算法已成为深度网络训练不可或缺的一部分。BN 通过计算批次中示例的均值和方差来对输入进行归一化,从而缓解深度神经网络训练中的梯度爆炸或者消失的问题。但是,由于算法与批次大小有关,BN 算法用于小批次时会因为不准确的估计导致性能下降。批重归一化(Batch ReNormalization,BRN)用指数移动平均(Exponential Moving Average,EMA)后的值对输入进行归一化操作,减小了归一化算法对批次的依赖。本文基于图像分类任务研究了在输入是小批次时归一化技术的应用,提出了通过改变 EMA 初值并对估计值加以修正来得到更准确的参数估计的批归一化算法。实验结果表明,所提算法与标准的 BN 和 BRN 算法相比,收敛速度更快,准确率有一定的改善。

关键词 图像分类,归一化算法,小批次,指数移动平均
中图法分类号 TP183 **文献标识码** A

Novel Normalization Algorithm for Training of Deep Neural Networks with Small Batch Sizes

WANG Yan WU Xiao-fu
(School of Telecommunication and Information Engineering, Nanjing University of Posts and Telecommunications, Nanjing 210003, China)

Abstract Batch Normalization (BN) algorithm has become a key ingredient of the standard toolkit for training deep neural networks. BN normalizes the input with the mean and variance computed over batches to mitigate the possible gradient explosion or disappearance during training of deep neural networks. However, the performance of BN algorithm often degrades when it is applied to small batch sizes due to inaccurate estimates of mean and variance. Batch ReNormalization (BRN) normalizes the input with the values of exponentialmoving average (EMA), reducing the dependency of the normalization algorithm on batches. This paper proposed a novel normalization algorithm with improved estimate on the moving mean and varianceby changing the initial value of EMA and adding corrections to the estimates. The experimental results show that the proposed algorithm has better performance in convergence speed and accuracy than both the standard BN and BRN algorithms.

Keywords Image classification, Normalization algorithm, Small batches, Exponential moving average

1 引言

深度学习在语音信号识别、图像分类、人脸识别和目标检测、自动驾驶等任务上取得了长足的进展[1-4]。然而,标准的深度神经网络很难训练,即使具有非饱和激活功能(如 ReLUs[5]),仍然可能会由于雅可比式乘以每层的输入激活进而发生梯度消失或爆炸,甚至在 Alex-Net[5] 中间激活可以相差几个数量级,这会导致产生大小差距较大的权重参数,对于较大的参数,权重的稍微变化就会产生相对大的影响,训练时会重点训练较大参数而忽略较小的参数。因此,权重初始化的选取,以及学习率和各种形式的归一化,对优化性能至关重要。

在当前的神经网络中,归一化层已经是网络结构中不可或缺的一部分。最常用于卷积神经网络(Convolutional Neural Network,CNN)的是文献[6]提出的批归一化(Batch Normalization,BN)算法,其通过计算当前批次的均值和标准差来对激活进行归一化操作。批重归一化(Batch ReNormalization,BRN)[7] 通过修正训练和测试中归一化的不同来解决小批次问题。批归一化算法的其他变体,有层归一化算法[8](Layer Normalization,LN)。LN 对深度网络中某一层的所有神经元的输入求均值和方差,来对激活进行归一化操作。LN 可以在 RNN 上取得不错的效果,但是用于 CNN 时往往会带来性能的退化。Group Normalization[9](GN)主要是针对 BN 对小批次效果差而提出的,其将深度网络某一层的所有神经元分成若干组,然后在每个组内做归一化,这样与批次大小无关,不受其约束。Weight Normalization[10](WN)对神经网络的权重 W 进行解耦合,同时结合初始化方法,可以在 CNN 上取得不错的效果,并可以加快训练速度。文献[6-12]都证明了归一化算法可以克服由初始化不良而引起的训练困难问题,且算法有助于更深层次模型中的梯度流动。

目前小批次训练普遍存在以下两个问题。

(1)对于越小的批次,虽然单个批次训练会更快,但想达

本文受国家自然科学基金项目(61372123,61401228,61671253),南京邮电大学科学研究基金项目(NY213002)资助。
王 岩(1995-),女,硕士,主要研究方向为图像分类,E-mail:hefeiwangyande@126.com;吴晓富(1975-),男,博士,主要研究方向为机器学习(人工智能信号处理)与计算机视觉,E-mail:xfuwu@njupt.edu.cn(通信作者)。

到与大批次同样的性能就要求更多的训练步数,因此越小的批次训练的时间会越长,时间成本和资源成本较高。

(2)越小的批次会带来性能的下降。

本文提出了一个新的基于修正移动均值和方差估计的归一化算法,并在图像分类任务上对其进行了测试。新的算法可以得到更准确的估计。不同于 BRN 只在归一化操作时划分小批次,而梯度更新是在大的批次上进行。本文研究的是在小批次上进行归一化操作和梯度更新,最后在小批次上进行测试。本文提出的算法相对其他归一化算法,收敛性能更佳,在全卷积网络 ALL-CNN[13] 和网中网(Network in Network,NiN)[14] 结构上可以较快地达到 60%的分类准确率,且取得了稍高的分类准确率。

2 相关工作

2.1 归一化算法

2015 年,Ioffe&Szegedy 提出的批归一化算法[6] 证明,在训练过程中将神经元的输入归一化到零均值和单位方差可以大大减少深度神经网络的训练时间。为了避免协变量偏移,即其中一层中的权重梯度高度依赖于先前的层输出,在训练时,批归一化算法根据它们在当前输入批次所有样本的均值和方差来对输入进行缩放;测试时,批归一化算法用训练时数据的均值和方差的无偏估计来对输入进行归一化操作。

BN 算法主要包含以下 3 个步骤。

(1)计算统计值:在当前批次上计算归一化所需统计值的均值和方差,输入 $x\in R^{m*d}$,m 指批次大小,d 指输入的特征图大小。

$$E(x^k)=\frac{1}{m}\sum_{i=1}^{m}x_i^k \tag{1}$$

$$Var[x^k]\leftarrow\frac{1}{m}\sum_{i=1}^{m}(x_i^k-E[x^k])^2 \tag{2}$$

其中,x_i^k 指 i 个样本的第 k 个特征图。

(2)归一化操作:把输入向量中的每个元素当成独立随机变量单独进行归一化,向量中各变量独立了,也就没有协方差矩阵了。这种归一化在各变量相关的情况下依然能加速收敛,仅仅使用了下面的公式进行预处理,也就是近似白化预处理。对于 d 维输入数据 $x=(x^{(1)}\cdots x^{(d)})$,归一化每一维:

$$\hat{x}=\frac{x^k-E[x^{(k)}]}{\sqrt{Var[x^{(k)}]}} \tag{3}$$

(3)线性变换:只对输入进行归一化可能会改变输入本来所能表现的特性或者分布,如在 sigmoid 函数中加入批归一化算法后可能会使得输入从非线性变成线性。为了解决这个问题,可以用可学习的参数增益 γ 和偏置 β 去拟合原先的分布。

$$y^{(k)}=\gamma^{(k)}\cdot\hat{x}^{(k)}+\beta^{(k)} \tag{4}$$

其中,当 $\gamma^{(k)}=Var[x^{(k)}]$,$\beta^{(k)}=E[x^{(k)}]$时,理论上可以得到与输入相同的分布。在实验中,一般初始化增益 $\gamma=1$,偏置 $\beta=0$,这里加入线性变换是为了让因训练而“刻意”加入的 BN 能够有可能还原最初的输入。

由式(1)、式(2)可知,批量归一化中的训练性能很大程度上取决于所获得的统计数据的质量,即取决于批次的大小。因此,批归一化在批次较小的情况下更难应用,例如在线学习。虽然分类任务通常可以使用相对较大的批次,但其他应用(如使用卷积网络的图像分割)使用较小的批次会导致性能退化。

2.2 小批次归一化算法

上节已经提到了 BN 在批次很小或者不包含独立样本时(样本之间有联系时)准确率下降,不能用于在线学习等任务。Sergey Ioffe 假设其是由于在训练和测试之间进行了不同的操作,BN 训练时用小批量的均值而测试时采用训练时小批量均值的移动平均,对小的批次甚至单个样本来说,测试时也用训练时的统计特征进行归一化显然是不合理的。文献[7]提出了批重归一化算法来修正训练和测试中的不同,BRN 算法在训练和测试时通过指数移动平均(Exponential Moving Average,EMA)来对均值和标准差进行估计。EMA 通过平滑系数对不同时期的历史数据赋予不同的权重,以此来预测未知事物的未来趋势。归一化估计值不仅与当前批次的统计值有关,与前面训练步数的统计值也有关系,因此 BRN 算法减小了归一化算法对当前批次大小的依赖。

BRN 是用指数移动平均后的均值 μ 和标准差 σ 进行归一化,我们假设训练步数为 $1,\cdots,T$,μ_t 代表步数 t 的移动均值估计,μ_β^t代表步数为 t 时当前批次的统计均值,设初值 $\mu_0=0$,$\sigma_0=1$,μ_B 和 σ_B 分别代表当前批次的均值和标准差,衰减率 $\alpha=0.99$,指数移动平均在训练步长为 t 时计算值 μ_t:

$$\mu_t\leftarrow\alpha\mu_{t-1}+(1-\alpha)\mu_\beta^t \tag{5}$$

$$\sigma_t\leftarrow\alpha\sigma_{t-1}+(1-\alpha)\sigma_\beta^t \tag{6}$$

归一化操作引入参数 r 和 d,把 r 和 d 当作常量来修正均值和方差的估计。当 $r=1$ 和 $d=0$ 时,与之前 BN 做法一致,得到更稳定的数值分布。BRN 的归一化操作如下所示:

$$\frac{x_i-\mu}{\sigma}=\frac{x_i-\mu_\beta}{\sigma_\beta}\cdot r+d \tag{7}$$

$$r=\frac{\sigma_\beta}{\sigma},d=\frac{\mu_\beta-\mu}{\sigma} \tag{8}$$

3 基于小批次的归一化算法

3.1 改进的算法

在批重归一化中,用当前批次的均值和方差的指数移动平均来估计 μ 和 σ 的值,并未对 μ 和 σ 的初值选择做过多解释,初值一般设为 $\mu=0$,$\sigma=1$,衰减率 $\alpha=0.99$。

在这里,我们假设初值 $\mu_0=\sigma_0=0$,衰减率 $\alpha=0.99$,在训练步长为 t 时更的新方式为:

$$\mu_t\leftarrow\alpha\mu_{t-1}+(1-\alpha)\mu_\beta^t \tag{9}$$

由初值 $\mu_0=\sigma_0=0$ 可计算得到通项公式为:

$$\mu_t=(1-\alpha)\sum_{i=1}^{t}\alpha^{t-i}\mu_\beta^i \tag{10}$$

$$\begin{aligned}E[\mu_t]&=E[(1-\alpha)\sum_{i=1}^{t}\alpha^{t-i}\mu_\beta^i]\\&=E[\mu_\beta^t](1-\alpha)\sum_{i=1}^{t}\alpha^{t-i}\\&=E[\mu_\beta^t](1-\alpha^t)\end{aligned} \tag{11}$$

利用式(11)可以得到更新后的均值 μ 和 μ_β 之间的关系,为了使得 $E[\mu_t]=E[\mu_\beta^t]$得到更准确的估计,对移动均值 μ 做出如下修正:

$$\mu_t\leftarrow\frac{\mu_t}{(1-\alpha^t)} \tag{12}$$

同理:

$$\sigma_t\leftarrow\frac{\sigma_t}{(1-\alpha^t)} \tag{13}$$

本文提出的算法改进点如下:

(1)令初始化的 $\mu=\sigma=0$,对移动均值和方差的计算加入修正,更新后的值用于 r 和 d 的计算;

(2)在文献[8-9]中,训练和测试是经过相同的算法模块,所以本文提出不再区分训练和测试时的归一化方法,对训练和测试做相同的归一化。

3.2　算法流程

修正后的归一化算法的具体流程如算法 1 所示。

算法 1　修正后的归一化算法 renorm_debias

输入:x 代表一个小批量 $B=x_{1\cdots m}$;训练步数 t,初始值 $t=0$;μ_t,σ_t 代表 t 时刻的移动均值方差,初值 $\mu_0=0$,$\sigma_0=0$;移动平均更新率 $\alpha=0.99$;修正系数 $r_{max}=3$,$d_{max}=5$;可学习参数增益 γ 和偏置 β

输出:y_i

$t=t+1$

$$\mu_B^t \leftarrow \frac{1}{m}\sum_{i=1}^{m} x_i$$

$$\sigma_B^t \leftarrow \sqrt{\frac{1}{m}\sum_{i=1}^{m}(x_i-\mu_B)^2}$$

$$\mu_t \leftarrow \alpha\mu_{t-1}+(1-\alpha)\mu_B^t$$

$$\alpha^t=\alpha^{t-1}\times\alpha,\ \mu_t \leftarrow \mu_t/(1-\alpha^t)$$

$$\sigma_t \leftarrow \alpha\sigma_{t-1}+(1-\alpha)\sigma_B^t$$

$$\alpha^t=\alpha^{t-1}\times\alpha,\ \sigma_t \leftarrow \sigma_t/(1-\alpha^t)$$

$$r \leftarrow \text{stop gradient}\left(\text{clip}_{[1/r_{max},r_{max}]}\left(\frac{\sigma_B^t}{\sigma_t}\right)\right)$$

$$d \leftarrow \text{stop gradient}\left(\text{clip}_{[-d_{max},d_{max}]}\left(\frac{\mu_B^t-\mu_t}{\sigma_t}\right)\right)$$

$$\hat{x}_i \leftarrow \frac{x_i-\mu_B^t}{\sqrt{{\sigma_B^t}^2+\epsilon}}\cdot r+d$$

$$y_i \leftarrow \gamma\hat{x}_i+\beta$$

4　实验验证与性能分析

4.1　实验任务

图像分类任务是视觉领域的基础任务之一,图片分类的任务是对于一个给定的图片,预测其类别标签。

为了评估使用归一化算法进行有监督的图片分类任务的性能,我们使用 CIFAR-100 自然图像数据集[15]在卷积神经网络上进行测试。CIFAR-100 中包含了 50 000 个 32×32 的 RGB 图像用于训练,10 000 个图像用于验证将 CIFAR-100 中的所有图像标记为 100 个类。

4.2　实验参数设置

对于实验验证,我们关注两个深度神经网络(DNN),即 9 层的全卷积 ALL-CNN[13]和 3 层的网中网(Network in Network,NiN)[14]结构。所用 ALL-CNN 的详细网络结构如表 1 所列,NiN 结构与文献[14]相同。

表 1　全卷积网络的结构

网络层	细节描述
输入层	输入 24×24×3 的 RGB 图片
conv1	3×3 conv. BN. 96 ReLU, stride 1
conv2	3×3 conv. BN. 96 ReLU, stride 1
conv3	3×3 conv. BN. 96 ReLU, stride 2
conv4	3×3 conv. BN. 192 ReLU, stride 1
conv5	3×3 conv. BN. 192 ReLU, stride 1
conv6	3×3 conv. BN. 192 ReLU, stride 2
conv7	3×3 conv. BN. 192 ReLU, stride 1
conv8	3×3 conv. BN. 192 ReLU, stride 1
conv9	3×3 conv. BN. 10 ReLU, stride 1
global_pool	global average pooling(6 × 6)
softmax	10/100-way softmax

在 ALL-CNNs 和 NiN 的训练任务上采用动量优化算法和数据增强,选择交叉熵作为损失函数,动量优化器的动量设置为 0.9,随机初始化权重初始学习率为 0.1;对于 *BatchSize* 等于 128,10,5 时,每 64,26,26 个 epoch 学习率缩小为原来的 1/10,每个任务分别训练约 100 个 epoch。

4.3　实验结果分析

首先,对批归一化算法在不同批次上的应用进行仿真实现,在批次即实验参数 *BatchSize* 为 128,10,5 时的仿真结果如图 1 和表 2 所示。从图中可以看到,当 *BatchSize*=5 时应用批归一化算法会比 *BatchSize*=10 的分类性能差,当批次由 10 降到 5 后会由于样本的减少而导致统计不够准确,这直接影响了最终的分类准确率。实际上,越小的输入批次会导致更长的训练时间和更低的分类性能。

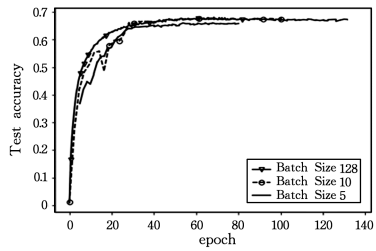


图 1　基于全卷积网络不同批次下应用批归一化算法的分类准确率

表 2　基于全卷积网络不同批次下应用归一化算法的分类准确率

归一化方法	批次大小	测试准确率/%
BN	128	67
BN	10	67.2
BN	5	65.8

在 2.2 节提到的小批次的 BRN 算法修正了批归一化中训练和测试之间的不同,采用移动均值和方差对输入进行归一化操作,从而减小 minibatch 相关性在 BN 模型中的依赖。本文对 BRN 算法进行了仿真,文献[7]的实验部分是对大的批次的输入进行分组,然后对组内样本进行归一化,在这里我们直接输入的就是小批次的数据,即输入样本数为 5,实验结果如图 2 和表 3 所示。

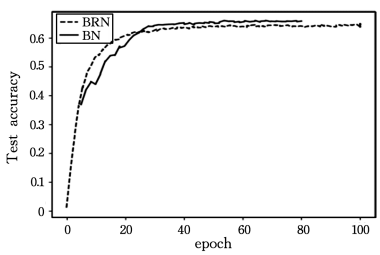


图 2　基于全卷积网络小批次下应用不同归一化算法的分类准确率

表 3　基于全卷积网络小批次下应用不同归一化算法的分类准确率

归一化方法	批次大小	测试准确率/%
BN	5	65.8
BRN	5	64.9

从图 2 可以看出,BRN 在 20epoch 左右时确实可以加速收敛,但是最终的效果却比批归一化低 1 个点。我们猜测这是由于任务的不同,BRN 在文献中虽然是针对小批次提出的,但是并没有对输入批次也会很小的情况做实验,文献中的

实验只在归一化时把大批次划分成小的数据进行归一化，梯度更新的操作是在大批次上完成的，而这里的实验是针对输入是小批次，同时对这个小的输入进行归一化操作和梯度更新。

本文对第 3 节提出的改进归一化算法进行仿真，实验结果如图 3 和表 4 所示，renorm_debias 是在 3.1 节提到的归一化算法的优化版本。优化后的算法大概在 15 个 epoch 达到 60%的分类准确率，而 BN 算法和 BRN 达到相同的准确率需要 23eopch 和 18 个 epoch，但是 BRN 有性能下降，而改进后的算法在准确率上较 BN 算法也略有优势。

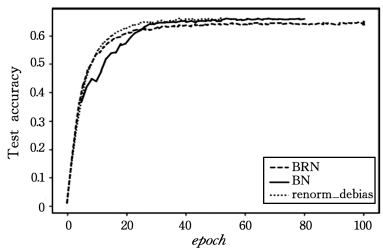


图 3 批次为 5 时在全卷积网络采用改进的归一化算法的测试准确率

表 4 批次为 5 时在全卷积网络采用改进的归一化算法的测试准确率

归一化方法	批次大小	测试准确率/%
BN	5	65.8
BRN	5	64.9
renorm_debias	5	66.2

本文同时在基于 NiN 的网络结构上对小批次的归一化算法进行了仿真实验，结果如图 4 所示。显然，改进后算法 renorm_debias 的损失函数下降得最快，损失值最低且准确率曲线比应用 BN 算法时更加平滑。renorm _debias 大概在 18 个 epoch 时达到 60%的分类准确率，而 BN 算法和 BRN 达到相同的准确率需要 26eopch 和 45 个 epoch，不过 BN 算法和改进算法的最终分类准确率都稳定在 64%左右，新的算法最终收益与 BN 算法差距不大。

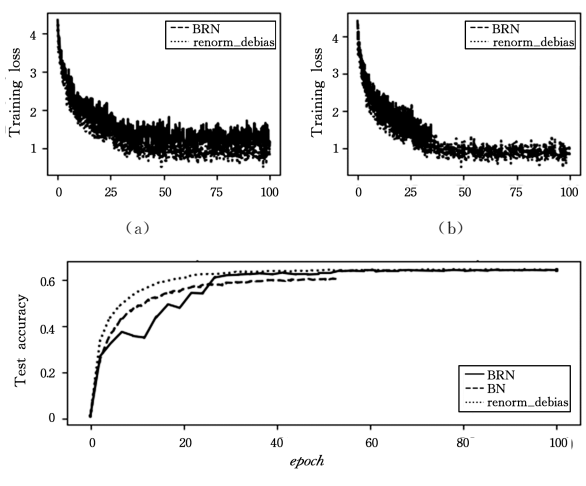


图 4 批次为 5 时在 NIN 网络中采用改进的归一化算法的训练损失和测试准确率

考虑到极端情况，我们也对批次等于 1 的情况进行了实验，本文中的算法相对标准 BN 算法在训练开始时就可以很快地提升测试集的分类准确率(见图 5)。

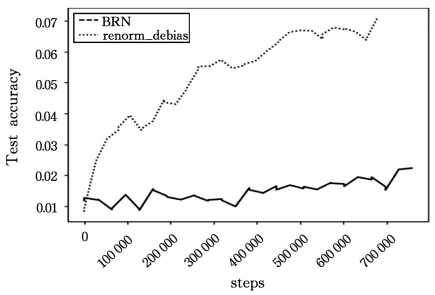
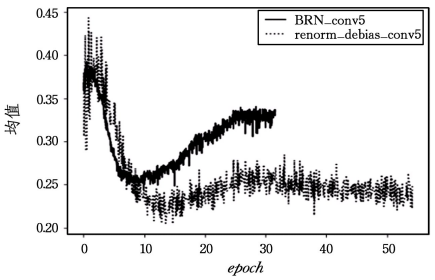
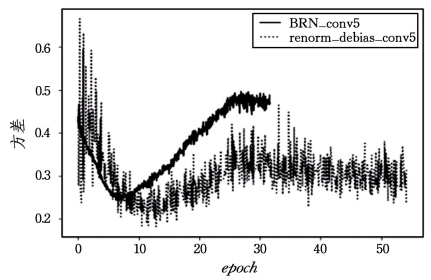


图 5 批次为 1 时在 NiN 网络上应用归一化算法的测试准确率

下面我们来观察改进前后算法经过激活后输出值的变化情况，即 relu(conv(wx)+b)的均值和方差分布。我们把九层全卷积网络中第五层卷积的均值和方差分布绘制在图 6 中，发现改进后的算法可以使得网络具有更接近 0 的均值和更小的方差。我们推测这可能是改进后的算法发挥了作用的原因，如在文献[16]中即是以零均值可以得到更自然的梯度从而加速收敛而提出了新的激活函数 elu。



(a)均值分布



(b)方差分布

图 6 基于 ALL-CNN 网络应用改进的归一化算法与标准 BN 算法的均值和方差

结束语 本文基于图像分类任务研究了归一化技术在小数据集上的应用，介绍了基于小批次上归一化算法的相关研究，分析了由于小批次带来的训练上的几点困难，在 BRN 基础上提出了新的算法，详细介绍了算法的出发点和算法的流程，并在基于卷积神经网络的图像分类任务上进行了小批次的实验仿真和结果分析。本文对归一化所需要的均值和方差的初值和计算方式进行了一些修改，同时也不再区分归一化算法在训练和测试时的不同操作。对比现有的算法，本文提出的算法在收敛性上更佳。

参 考 文 献

[1] LAWRENCE S,GILES C L,TSOI A C,et al. Face recognition: A convolutional neural-network approach[J]. IEEE transactions on neural networks,1997,8(1):98-113.

(下转第 308 页)

[13] PAN C. Multiple sclerosis identification by convolutional neural network with dropout and parametric ReLU[J]. Journal of Computational Science,2018,28:1-10.

[14] YANG W B,YANG H C. Gesture recognition method based on convolutional neural network[J]. Journal of Anhui Polytechnic University,2018,33:41-46.

[15] WU J. Fruit classification by biogeography-based optimization and feedforward neural network [J]. Expert Systems, 2016, 33(3):239-253.

[16] LU S. Pathological Brain Detection in Magnetic Resonance Imaging Using Combined Features and Improved Extreme Learning Machines [J]. Journal of Medical Imaging and Health Informatics,2018,8:1486-1490.

[17] 刘艳虹,顾定倩,程黎,等.我国手语使用状况的调查研究[J].语言文字应用,2013,5(2):35-41.

[19] 徐鑫鑫,黄元元,胡作进.连续复杂手语中关键动作的提取算法[J].计算机科学,2018,45(S2):189-193.

[20] MELLISA P A,JEKLIN H,SAKKA N. Mammograms Classification Using Gray-level Co-occurrence Matrix and Radial Basis Function Neural Network [J]. Procedia Computer Science, 2015,59:83-91.

[21] LU H M. Facial Emotion Recognition Based on Biorthogonal Wavelet Entropy,Fuzzy Support Vector Machine,and Stratified Cross Validation [J]. IEEE Access,2016,4:8375-8385.

[22] 毛思晨.基于卷积网络和长短时记忆网络的中国手语词识别方法研究[D].合肥:中国科技大学,2018.

[23] ZHOU X X,SHENG H. Combination of stationary wavelet transform and kernel support vector machines for pathological brain detection [J]. Simulation,2016,92(9):827-837.

[24] ZHANG H Y,YUAN J Z. Survey on New Methods of Vision-based Hand Gesture Recognition[J]. Journal of Computational Science,2017,44:1-6.

(上接第 276 页)

[2] KRIZHEVSKY A,SUTSKEVER I,HINTON G E. Imagenet-classification with deep convolutional neural networks[C]// Advances in Neural Information Processing Systems. 2012:1097-1105.

[3] LECUN Y,BENGIO Y. Convolutional networks for images, speech,and time series[M]// The Handbook of Brain Theory and Neural Networks. MIT Press,1998.

[4] ABDEL-HAMID O,DENG L,YU D. Exploring convolutional neural network structures and optimization techniques for speech recognition[C]// INTERSPEECH 2013. Lyon,2013.

[5] KRIZHEVSKY A,SUTSKEVER I,HINTON G E. Imagenet classification with deep convolutional neural networks[C]// Advances in Neural Information Processing Systems. 2012:1097-1105.

[6] IOFFE S,SZEGEDY C. Batch normalization:accelerating deep network training by reducing internal covariate shift[C]// International Conference on International Conference on Machine Learning. JMLR. org,2015.

[7] IOFFE S. Batch renormalization:Towards reducing minibatch dependence in batch-normalized models[C]// Advances in Neural Information Processing Systems. 2017:1945-1953.

[8] BA J L,KIROS J R,HINTON G E. Layer normalization[J]. arXiv:1607.06450,2016.

[9] WU Y,HE K. Group normalization[C]// Proceedings of the European Conference on Computer Vision (ECCV). 2018:3-19.

[10] SALIMANS T,KINGMA D P. Weight normalization:A simple reparameterization to accelerate training of deep neural networks[C]// Advances in Neural Information Processing Systems. 2016:901-909.

[11] REN M,LIAO R,URTASUN R,et al. Normalizing the normalizers:Comparing and extending network normalization schemes [C]// ICLR. 2017.

[12] LIAO Q,KAWAGUCHI K,POGGIO T. Streaming Normalization:Towards Simpler and More Biologicallyplausible Normalizations for Online and Recurrent Learning[J]. arXiv:1610.06160v1,2016.

[13] SPRINGENBERG J T,DOSOVITSKIY A,BROX T,et al. Striving for simplicity:The all convolutional net[C]// ICLR. 2015.

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

[15] KRIZHEVSKY A,HINTON G. Learning multiple layers of features from tiny images:Technical Report:TR-2009[R]. University of Toronto,2009.

[16] CLEVERT D A,UNTERTHINER T,HOCHREITER S. Fast and accurate deep network learning by exponential linear units (elus)[C]// ICLR. 2016.