

DOI:10.13232/j.cnki.jnju.2023.05.005

一种基于孪生网络的图片匹配算法

严镭宇¹, 李 伟^{1*}, 陈玉明¹, 黄 宏¹, 王文杰¹, 宋宇萍²

(1. 厦门理工学院计算机与信息工程学院, 厦门, 361024; 2. 厦门大学数学科学学院, 厦门, 361005)

摘 要:随着卷积神经网络(Convolutional Neural Networks, CNN)的不断改进, 基于CNN的图片匹配成为图像处理的关键, 然而, 许多基于CNN的图像相似度检测算法对图像特征的表达力较差, 且曼哈顿距离或欧式距离的计算方式导致在计算损失函数时模型不一定能很好地收敛. 针对此问题, 提出一种基于孪生网络和注意力机制的方法(CSNET)来提升图像匹配的性能, 主要步骤如下: 使用将激活函数改进为Mish函数的VGG16网络作为主干网络提取图像的特征, 在模型的卷积层加入注意力机制模块(Convolutional Block Attention Module, CBAM), 这提高了模型的特征提取能力和鲁棒性, 保证训练可以收敛. 对模型输入图片对的特征向量的欧氏距离, 再利用网络全连接层输出的相似度分数来度量被检图片是否相似. 将提出的CSNET与其他图片匹配方法在Omniglot和SigComp2011等数据集上进行比较, 实验结果表明, CSNET能有效提高图像相似度匹配的准确性.

关键词: 图像匹配, 孪生网络, 度量学习, 注意力机制, 卷积神经网络

中图分类号: TP183

文献标志码: A

A Siamese network-based image matching algorithm

Yan Rongyu¹, Li Wei^{1*}, Chen Yuming¹, Huang Hong¹, Wang Wenjie¹, Song Yuping²

(1. School of Computer and Information Engineering, Xiamen University of Technology, Xiamen, 361024, China;

2. School of Mathematical Sciences, Xiamen University, Xiamen, 361005, China)

Abstract: With the continuous improvement of Convolutional Neural Networks (CNNs), image matching based on CNNs has become a key to image processing. The Siamese Network metric for measuring the distance between two images has achieved a good result. However, many CNN-based image similarity detection systems have poor representation of image features, and the calculation of Manhattan or Euclidean distance leads to models that do not always converge well when calculating the loss function. To address this problem, this paper proposes a method based on the Siamese Network and attention mechanism to improve the performance of image matching (CSNET). The outline steps are as follows: the VGG16 network is used as the backbone network to extract the features of the images, the attention mechanism module is added after the convolutional layer of the model to enhance the feature extraction capability, and the Euclidean distance of the feature vectors of the input image pairs is used to measure whether the examined images are similar using the output of the fully connected layer of the network. The method in the paper is compared with other image-matching methods on the Omniglot dataset and SigComp2011 datasets. Experimental results show that the proposed CSNET effectively improves the accuracy of image similarity matching.

Key words: image matching, Siamese Network, metric learning, attention mechanism, CNN

基金项目: 福建省科学技术厅自然科学基金(2022J011233), 国家留学基金管理委员会访问学者项目(202308350042), 国家自然科学基金(61976183)

收稿日期: 2023-07-17

* 通讯联系人, E-mail: drweili@hotmail.com

随着计算机视觉在人工智能创新应用的飞速发展,在一个巨大且非结构化的图片集里检验图片的相似性成为计算机视觉领域的一个关键问题.近年来,人们提出了一系列图片匹配方法,主要分为两类.

第一类是基于手工设计特征的图片匹配方法,旨在预测图片对是否为正值(相似)(即两幅图片是否代表同一场景对).这类方法如视觉词汇袋(Bow)在预测一组较小的候选图片对方面有很好的效果^[1],也有一些方法使用Bow模型的判别性学习来预测输入数据集中的图片相似度.但需要训练的图片较多时,该方法的效果不佳,且人工成本过高.

第二类是基于深度学习的图片匹配方法.近年来,随着ImageNet等大规模标注数据集的出现,以感知机模型^[2]为基础的卷积神经网络(Convolutional Neural Networks, CNN)被广泛应用于深度学习并通过反向传播改进模型性能.深度卷积神经网络在图像处理中具有很强的特征提取能力,使用CNN已经成为主流的特征提取方法.随着各种CNN模型的提出,如AlexNet^[3],VGG-Net^[4],ResNet^[5],GoogleNet^[6],Efficientnet^[7]等,它们在图像分类、物体检测、图像检索等众多领域都取得了良好的效果.大量学者将长短时记忆网络(Long Short-Term Memory, LSTM)^[8]和CNN等神经网络方法应用于图像相似度检测,特别是将CNN作为主干网络的孪生网络在图片相似度领域有较好的效果,但它们大多存在默认特征图的每个通道和空间是同等重要的,导致重要信息表征能力的缺失和准确性不足等问题.

当有两个相同类型的输入数据时,欧氏距离和曼哈顿距离可以很好地计算两个向量之间的几何差异,但如果两个列表中的元素具有不同的数据类型或不同的含义与不同的特征,它们就不能发挥作用.以签名验证为例,签名板记录的数据来自不同的特征,如手的速度、字符的位置、字符的尺寸、签名所占的总空间、字符之间的距离等,这种情况需要一个能够处理这些不同的特征并将其投射到一个独特空间中的方法.为了解决此类问题,1994年Bromley et al^[9]提出了孪生网络.

孪生网络是一种基于相似性度量的小样本学

习方法,其结构简单,训练容易,两个主干特征提取网络之间的权值参数和偏置参数相同,但输入的是一组相似或不相似的图片对,已在人脸识别、语音处理和音频信号处理等领域取得了巨大的成效.然而,由于孪生卷积网络输出向量的距离极小,要求苛刻,结果可能会变得不准确,导致整个网络不收敛.

综上,传统图像匹配算法近年来的应用逐渐减少,基于CNN的图像匹配算法因为其精确度较高和人工成本较低的优点,应用逐渐增加,但仍存在一些缺点,具体表现:

(1)默认特征图的每个通道和空间是同等重要的,导致重要信息的表征能力的缺失;

(2)提取了大量与识别无关的全局特征,因此对算力要求很高,模型的泛化能力普遍不强;

(3)输出向量距离极小的强约束条件导致训练无法收敛;

(4)训练过程不够稳定,可能出现梯度消失的问题.

基于以上缺点,本文把用于相似度计算的孪生神经网络与注意力机制相结合,使用将激活函数替换为Mish的VGG16网络提取的图像特征和全连接网络来计算向量的距离,以此设计图像匹配框架.该方法的优点:

(1)使用全连接层计算特征向量的欧氏距离,使模型自适应地根据学习到的特征来调整度量方式,确保网络可收敛;

(2)在特征提取网络中加入注意力模块,有效地增强了CNN的表征能力,在一定程度上提高了图片匹配的准确性;

(3)将主干网络中的激活函数从ReLU函数替换为Mish函数,增强了网络的平滑性、非线性性和宽容性.

1 相关工作

注意力机制在图像领域的重大突破让图片匹配有了新的改进空间,而用孪生神经网络来度量学习图像对之间的差异,非常适合通过注意力机制模块来提升表征能力.

1.1 度量学习 度量学习将输入的样本对投射到一个嵌入空间,并使用如欧氏距离或余弦距离

等进行相似度的计算,其已在深度学习中得到了很好的应用. Schroff et al^[10]提出人脸相似性的三重测量来学习从面部图像到紧凑欧氏空间的映射. Ganin and Lempitsky^[11]用少数标准层和一个梯度反转层增强了一个深度架构,以实现领域适应行为. Sun and Saenko^[12]扩展了 CORAL 方法来学习深度神经网络中的非线性变换.

1.2 孪生网络 一般情况下用到的 CNN 只有一个输入,例如进行 Mnist 手写数字的分类或简单的目标检测. 为了检验输入数据对的相关性,学者们提出了皮尔逊相关系数^[13]、斯皮尔曼的 ρ 级系数^[14]等,但在两个结构为本体树的元素之间进行比较时,这些算法都无法处理结构复杂的数据. 为了解决这个问题, Bromley et al^[9]提出了孪生神经网络.

该网络的运行原理如图 1 所示. 图像对中的 X_1 和 X_2 同时输入权值共享的特征提取网络,提取特征向量 $G_W(X_1)$ 和 $G_W(X_2)$,再将特征向量经过一个非线性函数来度量它们之间的距离. 由于孪生网络存在输出的欧式距离极小的强约束条件,得出的图像对特征向量的度量会有误差,导致孪生神经网络无法进行训练.

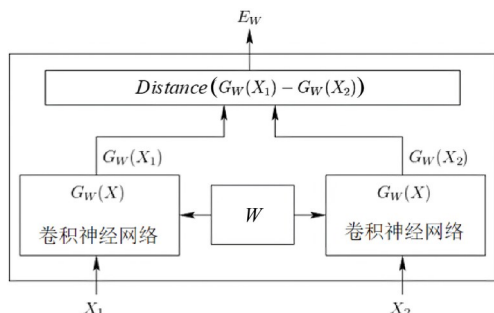


图 1 孪生神经网络的运行原理

Fig. 1 The operating principle of Siamese Network

1.3 注意力模块 注意力模型已经成为神经网络研究的一个重要领域,注意机制的灵感可以归结为人们对环境的生理感知,这是人类利用有限的处理资源从大量信息中快速选择高价值信息的一种手段,其大幅提高了知觉信息处理的效率和准确性. 事实证明,注意力机制提高了模型对有用特征赋予更多权重以改善特征提取的能力,在各种计算机视觉任务中取得了良好的效果^[15-16],

其中较有影响力的是 SENET 和基于 SENET 的 CBAM(Convolutional Block Attention Module)模块,本文模型也采用 CBAM 模块来提升性能.

2 算法设计

给图片匹配设计一个通用的神经网络,为了对输入该网络的图片对进行特征提取,本文提出一种以 VGG16 为主干网络并加入 CBAM 模块来提升性能的孪生神经网络(CSNET),如下所示.

算法 CSNET

输入: 一组图片对

输出: 该图片对的相似分数

1. 将图片对的两张图片分别输入网络的两个输入头;
2. 利用权值共享的 VGG16 网络提取图片对的多维特征信息;
3. 将这对特征信息平铺到一维上;
4. 求出这对一维向量相减的绝对值;
5. 将相减的向量输入一个两层的全连接网络;
6. 该全连接网络输出的分数经过一个 Sigmoid 函数后输出一个 0~1 的相似度分数;
7. 将该相似度分数与设定的阈值作比较,大于阈值判断为相似,小于阈值判断为不相似.

该算法首先通过卷积神经网络提取图片对的特征,再采用一个两层的全连接层来计算图片对之间的欧氏距离并输出匹配相似度,以此判断该图片对是否相似. 网络结构的设计如图 2 所示.

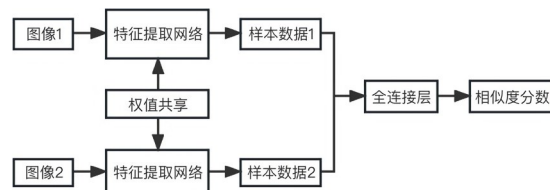


图 2 CSNET 的网络结构

Fig. 2 The structure of CSNET

2.1 主干网络 VGG16 的参数小,特征表征能力强,通过组合与堆叠小的卷积核比大的卷积核有更强的鲁棒性. 图片处理的具体流程如下:

(1) 一张原始图片被 resize 到指定大小,本文使用 105×105 的分辨率.

(2) conv1 包括两次 3×3 卷积网络,一次 2×2 最大池化,输出的特征层为 64 通道.

(3) conv2 包括两次 3×3 卷积网络,一次 2×2

最大池化,输出的特征层为128通道.

(4)conv3包括三次 3×3 卷积网络,一次 2×2 最大池化,输出的特征层为256通道.

(5)conv4包括三次 3×3 卷积网络,一次 2×2 最大池化,输出的特征层为512通道.

(6)conv5包括三次 3×3 卷积网络,一次 2×2 最大池化,输出的特征层为512通道.

2.2 注意力模块 本文使用的注意力机制是CBAM模块,CBAM可以被添加到卷积层的任何位置.空间注意力和通道注意力有不同的功能,CBAM能将两者结合起来,更好地权衡图像中的主体部分.本研究中,注意力机制被应用于图片相似度检测,将一幅大小为 105×105 的图像输入网络模型,图像通过网络模型后得到一个 $3 \times 3 \times 512$ 的特征图.CBAM由一维通道注意力模块 $F_c \in RC \times 1 \times 1$ 和二维空间注意模块 $F_s \in R_1 \times H \times W$ 组成,两者串联排列.图像特征首先被输入通道注意模块,再生成通道特征模块,利用特征的通道关系生成通道特征 β' ;然后,生成的特征 β' 经过空间注意模块,得到最终的特征 β'' .通道关注模块保留了更多的图像纹理信息和详细的语义信息,空间关注模块通过关注图像的轮廓和空间结构来提取有效的空间特征.

2.3 全连接距离度量层 一般实验对相似的图像对标注为1,对不相似的图像对标注为0,但由于孪生网络输出的向量欧式距离极小的强约束条件,得出的图像对特征向量的度量会有误差,导致孪生神经网络无法进行训练,所以本文在主干网络提取特征之后用全连接层的方式让模型学习计算两个特征向量的距离.具体步骤:在获得主干特征提取网络之后,可以获取一个多维特征,将其平铺到一维上获得两个一维向量,将这两个一维向量相减再进行绝对值求和,相当于求取两个特征向量插值的L1范数,也相当于求取两个一维向量的距离;将该距离输入一个两层的全连接网络,输出0~1的分数,全连接层经反向传播的训练后输出的分数即为这两张图片的相似度分数.

2.4 Mish 激活函数 Mish (Maxout-Based Activation Function for Deep Neural Networks)是一种用于深度神经网络的激活函数,在一定程度上

扩展了传统的激活函数(如ReLU, sigmoid和tanh)的能力,能够更好地处理非线性关系和梯度消失问题.如下所示:

$$f(x) = x \times \tanh(s \ln(1 + \exp(x))) \quad (1)$$

和传统的激活函数相比,Mish有几个优点:

(1)平滑性:Mish函数是平滑的,连续可导,所以在使用Mish作为激活函数时,神经网络的训练更加稳定,还可以避免梯度消失的问题.

(2)非线性性:Mish函数具有更强的非线性特性,能够更好地捕捉输入数据中的复杂模式和特征,这对于复杂的计算机视觉和自然语言处理的任务非常有益.

(3)宽容性:Mish函数能够处理较大的输入值,不会出现梯度爆炸的问题,这使得神经网络可以更好地适应具有大范围输入值的情况,提高了网络的鲁棒性.实际应用中,Mish激活函数通常用作CNN和全连接神经网络的激活函数,增强网络的表达能力和性能.使用Mish函数,可以在一定程度上改善模型的收敛速度和精度,提升深度学习模型的性能.

2.5 训练过程 输入一对图像 x_1, x_2 以及它们是否相似的标签 y (图像对相似时 $y=1$,不相似时 $y=0$)进行训练.首先,将输入的 x_1, x_2 的图像像素调整到 $105 \times 105 \times 3$ 的固定大小,设定一系列的训练参数,选择Adam优化器,Epoch设置为100轮,Batch_size设定为32.用全连接神经网络输出图像对的相似度分数(0~1)后,再将该分数与输入图像对的类别标签 y ($y=1$ 为相似, $y=0$ 为不相似)进行交叉熵运算,得出该次训练的损失,最后用损失对该网络进行反向更新,调整特征提取层与全连接层的参数,使其能较准确地预测输入图像对的相似度.交叉熵的计算如下:

$$l = -(y_n \times \lg(z_n) + (1 - y_n) \times \lg(1 - z_n)) \quad (2)$$

其中, z_n 表示第 n 个图像对是正样本的概率, y_n 表示第 n 个图像对的标签.

3 实验分析

操作系统 Windows 10 专业版,CPU Intel i3 12100,GPU NVIDIA GeForce RTX 2060,显存6 GB,编程语言Python 3.7,深度学习框架PyTorch.

在预测图像对的相似性时会出现两种错误：第一种是把不同的图片误认为是同一对图片，对于这种情况，用假接受率(FAR)来表示可能性；第二种是把同一对图片误认为是不同的图片，对于这种情况，用假拒绝率(FRR)表示。本文使用的评估指标包括 FAR , FRR 和准确度(ACC)。如下所示：

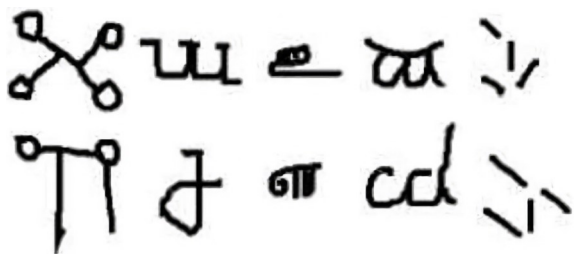
$$FAR = \frac{FP}{FP + TN} \times 100\% \quad (3)$$

$$FRR = \frac{FN}{FN + TP} \times 100\% \quad (4)$$

$$ACC = \frac{TP + TN}{FP + TP + TN + FN} \times 100\% \quad (5)$$

其中, TP (True Positive) 是相同的图片被判别为相同图片的样本数量, TN (True Negative) 是不同的图片被判别为不同图片的样本数量, FP (False Positive) 是不同图片被判别为相同图片的样本数量, FN (False Negative) 是相同图片被判别为不同图片的样本数量。

采用 Omniglot, SigComp2011 和 CEDAR 数据集, 其中 Omniglot 数据集包含来自 50 个国家的字母 (真实的和虚构的) 的 1623 个字符类和每个字符类的 20 个手写的图像, 如图 3 所示。模型通过输入数据集中随机生成一对图像对进行训练。



Angelic Armenian Grantha Tengwar ULOG

图 3 Omniglot 数据集的示例

Fig. 3 Examples of the Omniglot dataset

训练过程中的损失变化如图 4 所示。由图可见, 在训练到第 60 Epoch 后损失趋于平缓, 模型在 100 Epoch 内得到了有效的训练。

图 5 展示了模型输入两个相同的字母后输出的相似度分数 ($Similarity = 1.000$), 图 6 展示了模型输入两个不同的字母后输出的相似度分数 ($Similarity = 0.000$)。由图可见, 相似度越接近 1, 两张图片越相似, 这与实验预期是相符的。

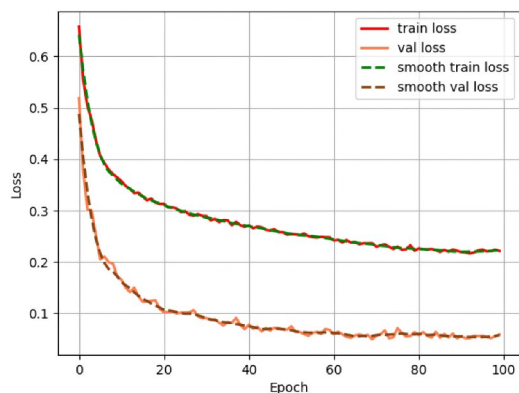


图 4 训练损失随迭代次数的变化曲线

Fig. 4 Variation of training loss over iterations

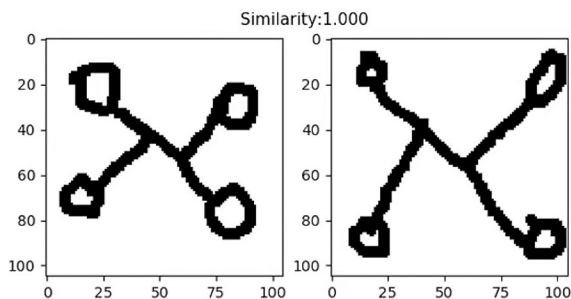


图 5 正样本的预测情况

Fig. 5 The prediction for positive samples

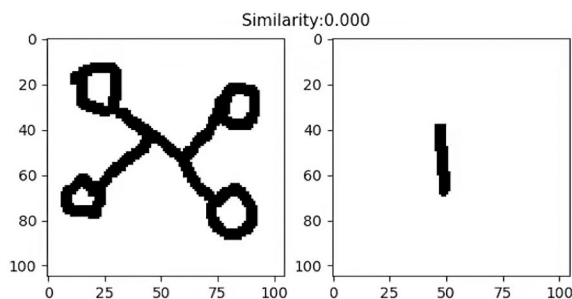


图 6 负样本的预测情况

Fig. 6 The prediction for negative samples

表 1 为 Omniglot 数据集上的检验结果, 由表可见, CSNET 的准确率为 98.6%, 优于没有加入注意力模块的 98.4%, 错误接受率与错误拒绝率也有一定程度的降低, 说明 CBAM 注意力机制模块提升了特征提取网络的表征能力和图片匹配的准确性。表 2 和表 3 为 SigComp2011 和 CEDAR 数据集上的检验结果, 由表可见, CSNET 的各项指标均优于各文献的性能, 该网络在没有全连接距离度量层的情况下训练无法收敛, 说明全连接距离度量层和简单的欧氏距离计算相比, 能更好

地进行度量学习.

表 1 Omniglot 任务的数据统计

Table 1 Data statistics of the Omniglot

方法	FRR	FAR	ACC
Siamese Network	1.3%	1.1%	98.4%
CSNET (OURS)	0.9%	1.3%	98.6%

表 2 SigComp2011 任务的数据统计

Table 2 Data statistics of the SigComp2011

方法	FRR	FAR	ACC
文献[17]的模型	0.53%	0	99.65%
CSNET (OURS)	0.48%	0	99.71%

表 3 CEDAR 任务的数据统计

Table 3 Data statistics of the CEDAR

方法	FRR	FAR	ACC
文献[18]的模型	10.46%	6.40%	91.57%
文献[19]的模型	16.37%	11.24%	86.10%
文献[20]的模型	6.78%	4.20%	95.66%
文献[21]的模型	0	4.20%	100%
CSNET (OURS)	0	0	100%

在 Omniglot 和 SigComp2011 数据集上进行交叉验证试验,测试 CSNET 的泛化能力,即使用在 Omniglot 数据集上训练好的 CSNET 在 SigComp2011 数据集上进行测试,再使用在 SigComp2011 数据集上训练好的 CSNET 在 Omniglot 数据集上进行测试,实验结果如表 4 所示.由表可见,CSNET 有较好的泛化能力.

表 4 CSNET 的跨数据集准确度检验

Table 4 Cross-dataset accuracy verification of CSNET

训练集	测试集	
	Omniglot	SigComp2011
Omniglot	98.60%	88.76%
SigComp2011	85.48%	99.71%

4 结论

本文提出一种基于孪生网络和注意力机制的图像匹配算法,通过卷积神经网络提取特征,并在网络中嵌入通道和空间注意力模块,使注意力模块能够有效地识别特征图中不同区域的重要性,并以加权的方式给予关键区域更多的权重,提高

度量学习的准确性.在 Omniglot, SigComp2011 和 CEDAR 数据集上进行了比较实验,从实验结果可以看出,本文提出的 CSNET 在加入注意力机制模块后,较大程度地提升了准确度.由于本文采用的 VGG16 网络有降低参数数量的优势,故在对更大的数据集进行验证时,可能会出现欠拟合的情况.今后将研究使用更复杂的 CNN,并使用与之更契合的注意力机制模块,进一步在海量数据集中提升图像匹配性能.

参考文献

- [1] Cao S, Snavely N. Learning to match images in large-scale collections//Fusiello A, Murino V, Cucchiara R. Computer vision - ECCV 2012. Workshops and demonstrations. Springer Berlin Heidelberg, 2012: 259—270.
- [2] Freund Y, Schapire R E. Large margin classification using the perceptron algorithm//Proceedings of the 11th Annual Conference on Computational Learning Theory. Madison, WI, USA: ACM, 1998: 209—217.
- [3] Krizhevsky A, Sutskever I, Hinton G E. Imagenet classification with deep convolutional neural networks. Communications of the ACM, 2017, 60(6): 84—90.
- [4] Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. 2014, arXiv:1409.1556.
- [5] He K M, Zhang X Y, Ren S Q, et al. Deep residual learning for image recognition//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, NV, USA: IEEE Press, 2016: 770—778.
- [6] Szegedy C, Liu W, Jia Y Q, et al. Going deeper with convolutions//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Boston, MA, USA: IEEE Press, 2015: 1—9.
- [7] Tan M X, Le Q. Efficientnet: Rethinking model scaling for convolutional neural networks//Proceedings of the 36th International Conference on Machine Learning: PMLR. Long Beach, CA, USA: PMLR Press, 2019: 6105—6114.
- [8] Hochreiter S, Schmidhuber J. Long short-term memory. Neural Computation, 1997, 9(8): 1735—

- 1780.
- [9] Bromley J, Guyon I, LeCun Y, et al. Signature verification using a “siamese” time delay neural network//Proceedings of the 6th International Conference on Neural Information Processing Systems. Denver, CO, USA: ACM, 1993: 737–744.
- [10] Schroff F, Kalenichenko D, Philbin J. FaceNet: A unified embedding for face recognition and clustering//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Boston, MA, USA: IEEE Press, 2015: 815–823.
- [11] Ganin Y, Lempitsky V. Unsupervised domain adaptation by backpropagation//Proceedings of the 32nd International Conference on Machine Learning. Lille, France: Omnipress, 2015: 1180–1189.
- [12] Sun B C, Saenko K. Deep CORAL: Correlation alignment for deep domain adaptation//Hua G, Jégou H. Computer vision - ECCV 2016 workshops. Amsterdam, The Netherlands: Springer, 2016: 443–450.
- [13] Benesty J, Chen J D, Huang Y T, et al. Pearson correlation coefficient//Cohen I, Huang Y T, Chen J D, et al. Noise reduction in speech processing. Springer Berlin Heidelberg, 2009: 1–4.
- [14] Spearman C. The proof and measurement of association between two things. By C. Spearman, 1904. The American Journal of Psychology, 1987, 100(3–4): 441–471.
- [15] Zhong Z L, Lin Z Q, Bidart R, et al. Squeeze-and-attention networks for semantic segmentation//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, WA, USA: IEEE Press, 2020: 13065–13074.
- [16] Yuan Y H, Chen X L, Wang J D. Object-contextual representations for semantic segmentation//The 16th European Conference on Computer Vision. Glasgow, UK: Springer, 2020: 173–190.
- [17] 朱琴. 基于孪生神经网络的合同手写签名认证研究. 硕士学位论文. 上海: 华东师范大学, 2022. (Zhu Q. Research on contract handwritten signature authentication based on Siamese Neural Network. Master Dissertation. Shanghai: East China Normal University, 2022.)
- [18] Chandra S, Maheskar S. Offline signature verification based on geometric feature extraction using artificial neural network//2016 3rd International Conference on Recent Advances in Information Technology. Dhanbad, India: IEEE, 2016: 410–414.
- [19] Kruthi C, Shet D C. Offline signature verification using support vector machine//2014 5th International Conference on Signal and Image Processing. Bangalore, India: IEEE Press, 2014: 3–8.
- [20] Xiao W H, Ding Y T. A two-stage siamese network model for offline handwritten signature verification. Symmetry, 2022, 14(6): 1216.
- [21] Li C, Lin F, Wang Z Y, et al. DeepHSV: User-independent offline signature verification using two-channel CNN//Proceedings of 2019 International Conference on Document Analysis and Recognition. Sydney, Australia: IEEE Press, 2019: 166–171.

(责任编辑 杨可盛)