Journal of Computer Applications
计算机应用,2020,40(8):2219－2224
ISSN 1001-9081
CODEN JYIIDU
2020-08-10
http://www.joca.cn

文章编号:1001-9081(2020)08-2219-06
DOI:10.11772/j.issn.1001-9081.2019122139

基于自注意力机制的多域卷积神经网络的视觉追踪

李生武,张选德*
(陕西科技大学 电子信息与人工智能学院,西安 710021)
(*通信作者电子邮箱 zhangxuande@sust.edu.cn)

摘　要:为了解决多域卷积神经网络(MDNet)在目标快速移动和外观剧烈变化时发生的模型漂移问题,提出了自注意力多域卷积神经网络(SAMDNet),通过引入自注意力机制从通道和空间两个维度来提升追踪网络的性能。首先,利用空间注意力模块将所有位置上的特征的加权总和选择性地聚合到特征图中的所有位置上,使得相似的特征彼此相关;然后,利用通道注意力模块整合所有特征图来选择性地强调互相关联的通道的重要性;最后,融合得到最终的特征图。此外,针对MDNet算法因训练数据中存在较多相似但属性不同的序列所造成的网络模型分类不准的问题,构造了复合损失函数。该复合损失函数由分类损失函数和实例判别损失函数组成,首先,用分类损失函数来统计分类的损失值;然后,利用实例判别损失函数来提高目标在当前视频序列中的权重,抑制其在其他序列中的权重;最后,融合两项损失作为模型的最终损失。在目前广泛采用的测试基准数据集OTB50和OTB2015上进行实验,结果表明所提出的算法在成功率指标上相比2015年视觉目标跟踪挑战(VOT2015)的冠军算法MDNet分别提高了1.6个百分点和1.4个百分点,在精确率和成功率指标上优于连续域卷积相关滤波(CCOT)算法,在OTB50上的精确率指标优于高效卷积操作(ECO)算法,验证了该算法的有效性。

关键词:多域卷积神经网络;视觉追踪;自注意力机制;实例判别损失;深度学习
中图分类号:TP391.4　　**文献标志码:**A

Multi-domain convolutional neural network based on self-attention mechanism for visual tracking

LI Shengwu, ZHANG Xuande*
(*School of Electronic Information and Artificial Intelligence, Shaanxi University of Science and Technology, Xi'an Shaanxi* 710021, *China*)

Abstract: In order to solve the model drift problem of Multi-Domain convolutional neural Network (MDNet) when the target moves rapidly and the appearance changes drastically, a Multi-Domain convolutional neural Network based on Self-Attention (SAMDNet) was proposed to improve the performance of the tracking network from the dimensions of channel and space by introducing the self-attention mechanism. First, the spatial attention module was used to selectively aggregate the weighted sum of features at all positions to all positions in the feature map, so that the similar features were related to each other. Then, the channel attention module was used to selectively emphasize the importance of interconnected channels by aggregating all feature maps. Finally, the final feature map was obtained by fusion. In addition, in order to solve the problem of inaccurate classification of the network model caused by the existence of many similar sequences with different attributes in training data of MDNet algorithm, a composite loss function was constructed. The composite loss function was composed of a classification loss function and an instance discriminant loss function. First of all, the classification loss function was used to calculate the classification loss value. Second, the instance discriminant loss function was used to increase the weight of the target in the current video sequence and suppress its weight in other sequences. Lastly, the two losses were fused as the final loss of the model. The experiments were conducted on two widely used testing benchmark datasets OTB50 and OTB2015. Experimental results show that the proposed algorithm improves success rate index by 1.6 percentage points and 1.4 percentage points respectively compared with the champion algorithm MDNet of the 2015 Visual-Object-Tracking challenge (VOT2015). The results also show that the precision rate and success rate of the proposed algorithm exceed those of the Continuous Convolution Operators for Visual Tracking (CCOT) algorithm, and the precision rate index of it on OTB50 is also superior to the Efficient Convolution Operators (ECO) algorithm, which verifies the effectiveness of the proposed algorithm.

Key words: Multi-Domain convolutional neural Network (MDNet); visual tracking; self-attention mechanism; instance discriminant loss; deep learning

收稿日期:2019-12-23;**修回日期:**2020-03-15;**录用日期:**2020-03-18。　**基金项目:**国家自然科学基金资助项目(61871260)。
作者简介:李生武(1994—),男,甘肃武威人,硕士研究生,主要研究方向:视觉跟踪、深度学习;　张选德(1979—),男,宁夏固原人,教授,博士,主要研究方向:图像质量评价、图像恢复。

0 引言

视觉追踪在无人驾驶、行人检测、智能交通等视觉应用中扮演着越来越重要的角色。基于相关滤波的追踪器可以利用一个循环矩阵在傅里叶域中完成快速计算，实现快速的目标跟踪，基于此，出现了很多高速且简易的追踪器[1-3]。随着卷积神经网络(Convolutional Neural Network，CNN)的不断发展，基于其强大的特征表示能力和高效的特征提取方式，CNN的应用领域不断扩展，各种针对特定问题设计的CNN模型不断被建立并成功地应用到各种图像任务中。在视觉追踪领域内，随着众多性能顶尖的以CNN为基础的深度视觉追踪算法的出现，它在追踪任务中应用的有效性已经得到了充分的验证。

一些基于CNN的模型算法就是利用其优秀的特征表示能力，致力于强化目标的表示，例如对抗学习跟踪(Visual Tracking Via Adversarial Learning，VITAL)[4]借鉴了生成对抗网络(Generative Adversarial Net，GAN)[5]的思想，在CNN中引入了对抗特征生成器来获取更加鲁棒的特征表示；结构感知网络(Structure-Aware Network，SANet)[6]通过将基于CNN的跟踪器和循环神经网络(Recurrent Neural Network，RNN)[7]结合起来，以获得目标的独特结构信息，提升模型对含有相似语义物体的鉴别能力；树结构卷积神经网络(CNNs in a Tree structure，TCNN)[8]通过构建树形结构的多个CNN来提高跟踪模型的可靠性，其中模型沿树中的路径在线更新，因此可以获得更鲁棒的模型。

多域卷积神经网络(Multi-Domain CNN，MDNet)[9]是VOT2015[10]的冠军算法，也是CNN在深度视觉追踪中应用最成功的代表性算法之一。MDNet的提出是受到用于目标检测的R-CNN[11]网络启发，该算法在ImageNet-Vid[12]数据集上离线训练好模型后，利用测试视频的首帧图像微调模型作为初始化，然后生成候选区域，最后确定预测框。

尽管基于CNN框架的MDNet算法在视觉追踪领域里取得了巨大的成功，但是受制于CNN框架内部存在的一些问题没有解决，如在追踪中的主要目的是在复杂环境中聚焦于目标的状态信息，而对其他外部信息不感兴趣。而一般的基于CNN框架的模型在处理数据时，等价地处理每一个特征图和特征子空间，没有重点关注的区域，尤其是在目标发生了剧烈的形态变化时，缺乏动态响应机制，这就限制了模型的多样表征能力。此外，针对在多域训练中，其中一个域对应一个用于训练的视频序列，会出现当前视频训练序列中的追踪目标在其他数据集中属于背景的情况，这种情况会导致网络鉴别目标的二义性，影响网络的识别精确率。本文通过在MDNet算法中引入自注意力(self-attention)机制和复合损失函数来解决上述问题。自注意力机制是由Vaswani等[13]首次提出的，通过利用该机制来获取输入的全局依赖性并且应用于机器翻译中。与此同时自注意力机制也开始逐渐应用到视觉图像处理领域中，如Zhang等[14]利用自注意力机制来学习一个良好的图片生成器；Wang等[15]将自注意力机制融入网络结构中用于图像分类，充分验证了在空间维度上非局部操作在图像和视频处理中的有效性。不同于这些工作，本文针对MDNet算法中存在的问题，通过引入自注意力机制提出了自注意力多域卷积神经网络(Multi-Domain convolutional neural Network based on Self-Attention，SAMDNet)视觉追踪算法，配合精心设计的复合损失函数，得到了显著的性能提升，最后通过充分的实验，验证了所提模型的有效性，并在多个广泛使用的测试基准上与数个先进的算法相比表现出众。

1 MDNet追踪算法

MDNet算法的目标是训练出一个MDNet使得该网络能够在任意域中也就是任意视频序列中将目标和背景区分出来。尽管不同的视频序列带有不同的目标和背景标签，然而在各种场景下目标和背景都存在一些共有的属性，例如对照明变化的鲁棒性、运动模糊、缩放比例变化等。在网络模型的离线训练阶段，MDNet通过其多域学习结构将不同的训练视频序列中的通用属性保留在网络的卷积层部分，最后的多分支全连接层对应不同的序列，负责在各自对应视频序列中甄别输入的候选框属于目标类还是背景类。在测试追踪阶段，新生成的单分支全连接层替换掉原来的多分支，构成追踪网络，追踪过程中综合利用网络更新策略、边界框修正策略和样本生成策略来完成对目标的分类和定位。

2 自注意力多域卷积神经网络

本文提出的自注意力多域卷积神经网络SAMDNet结构如图1所示，网络的输入为107×107大小的RGB图像，整个网络主要由两部分组成。其中一部分为网络主体结构，由3个卷积层(CONV1~CONV3)和2个全连接层(FC4、FC5)组成，卷积层和VGG(Visual Geometry Group)网络中对应的部分完全一样。此外，网络末端还有k个全连接层(FC6[1]~FC6[k])作为对应的k个域，也就是k个用于网络训练的视频序列，全连接层FC6中的每一个分支均使用Softmax交叉熵函数来计算目标和背景概率，另外也参与在离线训练阶段复合损失的计算。另一部分为引入的自注意力机制，自注意力机制的实现主要由空间注意力模块(Spatial Attention Module，PAM)和通道注意力模块(Channel Attention Module，CAM)具体完成，其中PAM位于卷积层CONV1和卷积层CONV2之间，CAM位于卷积层CONV3和全连接层FC4之间。

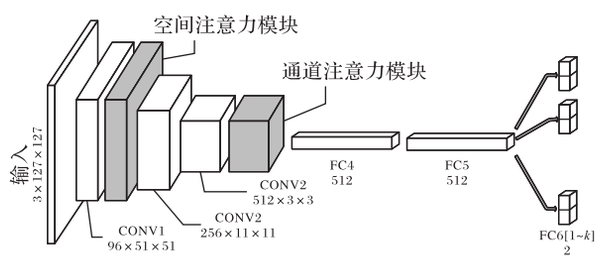


图1 自注意力多域卷积神经网络结构

Fig. 1 Network structure of multi-domain convolutional neural network based on self-attention

2.1 自注意力机制算法

注意力机制从本质上来讲类似于人类的视觉注意力机制，当人类在观察事物时，视觉系统快速扫描捕获的图像，获得了重点关注的目标区域，也就是视觉的注意力焦点，之后视觉系统在焦点附近分配更多的资源来获取更多的细节信息，同时其他的不相关或者无用的信息就被抑制了。而自注意力机制是对注意力机制的改进，它减少了对外部信息的影响，更加擅长去捕捉数据和特征的内部相关性。本文提出的算法就是利用自注意力机制的特性，动态地在空间和通道两个维度上去自学习注意力矩阵，而后将注意力融入模型进而驱动多域卷积神经网络聚焦于关注追踪目标内部特征，最后获得更为鲁棒的目标表示。

本文算法中的自注意力机制主要是通过空间注意力模块

和通道注意力模块实现:空间注意力模块将所有位置上的特征的加权总和选择性地聚合到特征图中的所有位置上,使得相似的特征彼此相关;通道注意力模块通过整合所有特征图来使得网络重点去关注那些互相关联的通道特征图,有选择地提取重要通道的特征信息。

2.1.1　空间注意力模块

空间注意力模块的结构如图2所示。网络模型的输入样本图像通过第一个卷积层(CONV1)后得到特征图 $\boldsymbol{F}\in\mathbf{R}^{C\times H\times W}$,其中上标 C、H、W 分别代表特征图通道数、特征图高度和特征图宽度,$\boldsymbol{F}$ 即空间注意力模块的输入。输入 $\boldsymbol{F}$ 分两次经过同一个一维卷积层后生成特征图 $\boldsymbol{A}$ 和 $\boldsymbol{B}$,且 $\{\boldsymbol{A},\boldsymbol{B}\}\in\mathbf{R}^{C\times H\times W}$,特征图 $\boldsymbol{C}$ 和 $\boldsymbol{F}$ 相同,$\boldsymbol{A}$ 和 $\boldsymbol{B}$ 矩阵重构后得到 $\{\boldsymbol{A}^*,\boldsymbol{B}^*\}\in\mathbf{R}^{N\times C}$,其中 N 为特征图中单个通道的像素个数,即 $N=H\times W$,之后将 $\boldsymbol{B}^*$ 和 $\boldsymbol{A}^*$ 的转置矩阵 $\boldsymbol{A}^{*\mathrm{T}}$ 作矩阵乘法,然后将得到的结果输入到Softmax层来计算注意力矩阵 $\boldsymbol{T}_{\mathrm{p}}\in\mathbf{R}^{N\times N}$,计算方式如式(1)所示。

$$x_{ji}=\exp(A_i\cdot B_j)\Big/\sum_{i=1}^{N}\exp(A_i\cdot B_j) \tag{1}$$

式中:x_{ji} 是用来度量位置 i^{th} 的像素对位置 j^{th} 像素的影响,x_{ji} 的值越大,位置为 i^{th} 和 j^{th} 的元素之间关联程度越高,越相似。与此同时特征图 $\boldsymbol{C}$ 进行矩阵重构得到 $\boldsymbol{C}^*\in\mathbf{R}^{C\times N}$,将 $\boldsymbol{C}^*$ 和注意力矩阵 $\boldsymbol{T}_{\mathrm{p}}$ 的转置矩阵 $\boldsymbol{T}_{\mathrm{p}}{}^{\mathrm{T}}$ 作矩阵乘法后得到的结果融入原输入的特征图 $\boldsymbol{F}$,最后得到空间注意力模块的输出 $\boldsymbol{E}\in\mathbf{R}^{C\times H\times W}$,计算方式如式(2)所示。

$$E_j=\alpha\sum_{i=1}^{N}(x_{ji}C_i)+F_j \tag{2}$$

α 为自学习参数,初始值为0。特征图 $\boldsymbol{E}$ 将作为之后网络组件的输入,此时特征图 $\boldsymbol{E}$ 中的每个位置都是来自原对应位置输入特征值加权和与来自所有位置的特征之和,因此每个位置的特征值都获得了一个全局的上下文信息,相似的特征会互相得到增益,有助于跟踪过程中目标的精确分类和定位。

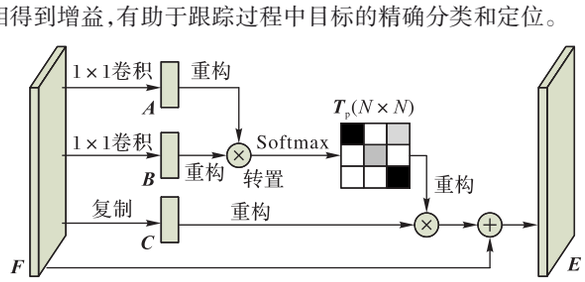


图2　空间注意力模块

Fig. 2　Spatial attention module

2.1.2　通道注意力模块

通道注意力模块结构如图3所示。和空间注意力不同的是,通道注意力 $\boldsymbol{T}_{\mathrm{c}}\in\mathbf{R}^{C\times C}$ 是从输入特征图 $\boldsymbol{F}\in\mathbf{R}^{C\times H\times W}$ 直接计算出来的,其中上标 C 为特征图的通道数。首先将 $\boldsymbol{F}$ 重构为 $\boldsymbol{F}^*\in\mathbf{R}^{N\times N}$,再将 $\boldsymbol{F}^*$ 与其转置矩阵 $\boldsymbol{F}^{*\mathrm{T}}$ 作矩阵乘法,输出的结果送入Softmax层中计算得到通道注意力矩阵 $\boldsymbol{T}_{\mathrm{c}}$,计算方式如式(3)所示。

$$y_{ji}=\exp(F_i\cdot F_j)\Big/\sum_{i=1}^{C}\exp(F_i\cdot F_j) \tag{3}$$

其中:y_{ji} 是用来度量通道 i^{th} 对通道 j^{th} 的影响,之后将 $\boldsymbol{T}_{\mathrm{c}}$ 的转置矩阵 $\boldsymbol{T}_{\mathrm{c}}{}^{\mathrm{T}}$ 和 $\boldsymbol{F}^*$ 作矩阵乘法后再重构使乘积的维数为 $\mathbf{R}^{C\times H\times W}$。通道注意力模块的最终输出 $\boldsymbol{E}\in\mathbf{R}^{C\times H\times W}$ 的计算方式如下:

$$E_j=\beta\sum_{i=1}^{C}(y_{ji}F_i)+F_j \tag{4}$$

其中:β 为自学习参数,初始值设置为0。通过自适应的方式,在总体上控制通道注意力对每个通道的影响,最后将得到的通道注意力和原特征图进行融合得到了输出 $\boldsymbol{E}$,特征图 $\boldsymbol{E}$ 将作为之后网络组件的输入。

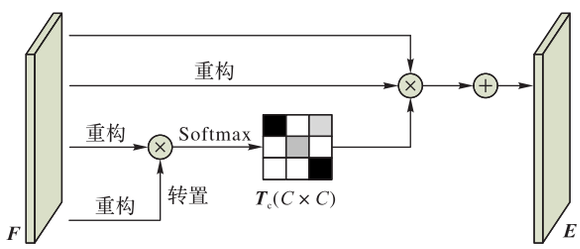


图3　通道注意力模块

Fig. 3　Channel attention module

2.2　用于离线训练的复合损失函数

在离线训练模型阶段,本文采用复合损失函数来统计误差,利用反向传播更新模型参数。复合损失函数由两部分损失项构成:一个为分类损失项,另一个为实例判别损失项。

离线训练时,多域卷积神经网络中的输出分数 $\boldsymbol{S}^d$ 定义方式如下所示:

$$\boldsymbol{S}^d=[\varphi^1(x^d;R),\varphi^2(x^d;R),\cdots,\varphi^D(x^d;R)]\in\mathbf{R}^{D\times 2} \tag{5}$$

其中:x^d 表示在第 d 个视频序列中的样本图片;R 为 d 视频序列中的目标真实框标记;D 为训练视频集中序列的总个数;$\varphi^d(;)$ 表示第 d 个序列对应的二分类分数。最后的全连接层($\mathrm{FC6}^1,\mathrm{FC6}^2,\cdots,\mathrm{FC6}^D$)的输出被同时送入用于二分类的Softmax函数 σ_{cls} 和用于实例判别的Softmax函数 σ_{ins},且完成一次迭代训练后得到 $\boldsymbol{S}\in\mathbf{R}^{p\times q}$ 的矩阵,p 为单次迭代中参与的样本数量,q 为样本图片属于的类别数量,i 和 j 分别是 $\boldsymbol{S}$ 的行和列索引。则 σ_{cls} 和 σ_{ins} 定义方式如式(6)和式(7)所示:

$$[\sigma_{\mathrm{cls}}(\boldsymbol{S}^d)]_{ij}=\exp(S_{ij}^d)\Big/\sum_{j=1}^{q}\exp(S_{ij}^d) \tag{6}$$

$$[\sigma_{\mathrm{ins}}(\boldsymbol{S}^d)]_{ij}=\exp(S_{ij}^d)\Big/\sum_{i=1}^{p}\exp(S_{ij}^d) \tag{7}$$

复合损失函数的定义方式如式(8)所示:

$$L=L_{\mathrm{cls}}+\gamma L_{\mathrm{ins}} \tag{8}$$

其中:L_{cls} 为二分类损失函数,比较目标和背景的分数来指引分类;L_{ins} 为实例判别损失函数,用来提高当前目标在当前序列中属于目标的正分数,抑制它在其他序列中属于目标的分数;γ 为一个超参数,作用是平衡两种损失的权重,本文根据多次实验设置的理想值为0.15。在离线训练网络的过程中,最内层的单次迭代只处理一个视频序列。假设当前序列为 $d(k)=k\bmod K$,则第 k 次迭代的二分类损失计算方式如式(9)所示:

$$L_{\mathrm{cls}}=-\frac{1}{N}\sum_{n=1}^{N}\sum_{q=1}^{2}[z_n]_{qd(k)}\cdot\log([\sigma_{\mathrm{cls}}(S_n^{d(k)})]_{qd(k)}) \tag{9}$$

其中:$z_n\in\{0,1\}^{D\times 2}$ 为对应真实目标的标签,当 d 序列中的预测框 R_i 对应的类为 q 时 $[z_n]_{qd}=1$,否则为0;N 为处理单个序列时循环迭代的总次数。实例判别损失定义如下:

$$L_{\mathrm{ins}}=-\frac{1}{N}\sum_{n=1}^{N}\sum_{d=1}^{D}[z_n]_{+d}\cdot\log([\sigma_{\mathrm{ins}}(S_n^d)]_{+d}) \tag{10}$$

和 L_{cls} 不同的是,实例判别损失函数 L_{ins} 仅计算的是 d 序列中预测框 R_i 对应类为目标的样本,用符号+表示。该损失函数会使得目标物体在当前序列中的分数变得较大,而在其

他序列中变得较小，最终使得网络模型在多个物体含有相似语义信息但是属于不同类别的情况下，能够有效保持其判别能力的鲁棒性。

2.3 SAMDNet的离线训练

SAMDNet离线训练时，其离线训练模型的多分支全连接层对应不同的视频序列，卷积层是共享的。离线网络模型中卷积层参数的初始化是直接加载已在ImageNet上预训练好的VGG-M网络中对应卷积层的参数，全连接层中的参数由标准正态分布初始化，整个网络的所有参数均是可学习的。在离线训练的每次迭代中，训练数据来自于每个视频序列的正负样本集，样本集是以目标真实框为中心，以高斯随机的方式在其周围选取固定数目的图像块而来的，离线训练的每次迭代只处理单个视频序列，单次迭代中随机抽选当前序列中的4帧图片，每帧图片提取4个正样本和12个负样本，因此网络的输入为16个正本和48个负样本组成的样本集。图片网络样本集中每个样本和目标真实框的重叠率大于0.7的被标记为正样本，重叠率小于0.5的样本被标记为负样本。

训练数据集为ImageNet-Vid，该数据集包括3 862个短时视频序列作为训练集，555个作为验证集，937个用于测试集。该数据集基于目标检测任务建立，其中的每个短时视频序列都是经过精心挑选的，全方位地考虑了各种因素，如运动类型、视频背景干扰、遮挡等。

2.4 SAMDNet的测试追踪

SAMDNet中为了预估目标的位置，预设的候选目标集合是由前一帧确定的目标框为中心在位移和尺度两个方向按照高斯分布生成的N个候选框$X^i=(x^i,y^i,s^i)(i=1,2,\cdots,N)$，其中$x^i$、$y^i$为前一帧目标框中心坐标，$s^i$为缩放比例，$X^i$的协方差是对角矩阵，对角线值满足$(0.09r^2,0.09r^2,0.25)$，其中$r$是前一帧的目标框的宽和高的均值，每个候选框的大小为前一帧目标框大小的1.05^{s^i}倍。当前帧目标框的确定方式如下，记当前为第t帧，$\{x_t^i\}_{i=1,2,\cdots,N}$为$t$帧中按高斯分布生成的$N$个候选框集合，则当前帧的目标框$x_t^*$定义方式如式(11)所示：

$$x_t^*=\arg\max_{x_t^i} f^+(x_t^i) \tag{11}$$

其中$f^+(x_t^i)$为候选框图像经过网络模型后得到的正分数，分数最高的候选框作为当前帧的预测目标框。当选定的目标框分数$f^+(x_t^*)\geqslant 0.5$时，在下一帧处理前使用边界框回归算法[11]来修正当前得到的目标框，使得当前帧更加贴合真实框。同时在负样本生成过程中使用了负样本挖掘技术[16]选取分数最接近正样本阈值的负样本作为在线训练的负样本，以此来提高模型区分正样本和负样本的能力。

网络模型的在线更新方式分为长时更新和短时更新两种：长时更新是指网络以固定处理帧数为间隔更新网络，目的是为了保证模型的鲁棒性；短时更新发生在获得的目标框分数$f^+(x_t^*)<0.5$时，使用保存的历史正负样本特征立即进行网络的更新，目的是为了提高网络模型的自适应性。

3 实验结果与分析

为了验证本文提出算法的有效性，选择了近年来数个先进算法来进行对比实验。本文所提出的模型是在Pytorch 1.1.0框架上训练的，实验平台为一台配置了Intel Xeon Bronze 3106 CPU和一块NVIDIA GeForce TITAN XP显卡的计算机。

3.1 在OTB50和OTB2015上的评估

本文算法在两个公开且被广泛使用的测试基准集OTB50[17]和OTB2015[18]上进行测试。OTB2015包括100个视频序列，该数据集中充分包含11种追踪过程中可能遇到的挑战性问题，如平面内旋转、遮挡、关照变化、运动模糊等；OTB50为OTB2015中前50个视频序列，而且这50个序列几乎涵盖了整个测试集中最复杂的视频序列。两个测试基准集有两个相同的度量标准：成功率和精确率。成功率是指预测框和标记框交集区域像素个数和并集区域像素个数之比；精确率是指预测框和标记框的中心误差在一个特定的阈值内的视频帧数占总帧数的百分比，本文算法评估采用的阈值为20。

本实验中采用的评估方法为单次通过方式(One-pass Evaluation，OPE)，除了和MDNet算法对比外，另选择了多个性能顶尖的追踪算法，分别是：高效卷积算法ECO[19]，ECO算法不使用CNN特征的版本ECO-HC[19]，2016年视觉目标跟踪挑战(VOT2016)的冠军——连续域卷积相关滤波算法C-COT[20]，空间约束相关滤波器(Spatially Regularized Correlation Filter，SRDCF)添加了深度CNN特征的版本DeepSRDCF[21]，作为深度视觉追踪的基准对比算法CNN-SVM[22]，全卷积孪生网络算法(Fully-Convolutional Siamese network，SiamFC)的多尺度版本SiamFC-3s[23]，以及判别式相关滤波器网络(Discriminant Correlation Filter Network，DCFNet)[24]。

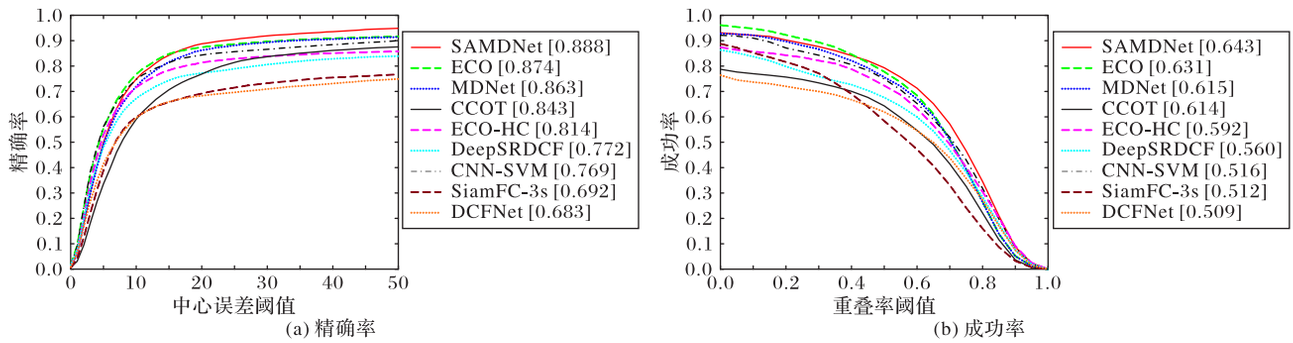
图4为OTB50数据集下的测试结果，可以看出本文的SAMDNet算法相较于原算法MDNet在精确率率指标上提高了2.5个百分点，在成功率指标上提高了1.6个百分点。本文算法在精确率和成功率两个指标上全面超过了CCOT和ECO-HC算法，在精确率指标上与ECO算法相比提高了1.4个百分点。图5为OTB2015数据集下的测试结果，可以看出本文的SAMDNet算法在精确率指标上取得了0.907的优异表现，超过MDNet算法1.8个百分点，也超过了CCOT算法接近1个百分点，此外和CCOT的改进版本ECO算法仅仅相差0.3个百分点，SAMDNet在成功率指标上也超过了CCOT和MDNet等其他算法。

为了进一步验证本文算法的有效性，选择了7个性能先进的算法进行了属性对比实验，结果如表1所示，选择的测试基准集为OTB2015。本文选取的对比属性有背景杂斑(Background Clutter，BC)、尺度变化(Scale Variation，SV)、遮挡(OCclusion，OC)、形变(DEformation，DE)、平面内旋转(In-Plane Rotation，IPR)、平面外旋转(Out-of-Plane Rotation，OPR)、超出视野(Out of View，OV)、光照变换(Illumination Variation，IV)、运动模糊(Motion Blur，MB)、快速移动(Fast Motion，FM)、低分辨率(Low Resolution，LR)，选取的评价指标为成功率。如表1所示，加粗的为对比实验中该列的最大值，可以看出，在11个属性中，本文算法SAMDNet相较于MDNet算法在8个属性上均有不同程度的提升，相较于全部选择的对比算法SAMDNet在属性SV、DE、IPR、OPR上达到了最佳表现，其他属性上也与各自最优值极其接近。

3.2 消融实验

基于卷积神经网络的视觉追踪中，传统网络模型中池化层的存在是为了压缩网络模型的参数量和降低过拟合，但同时也导致了大量数据损失，而且这种损失会随着网络模型层数的增加而逐渐叠加，虽然深度特征抽象度高、鲁棒性好，但深度特征的空间信息损失也越多，这不利于需要精确空间定位的追踪任务。基于此，本文所提算法中将空间注意力模块设置在CONV1和CONV2之间，在出现空间信息损失之前，最大限度地使相似的特征彼此相关，突出相似特征的空间信息。

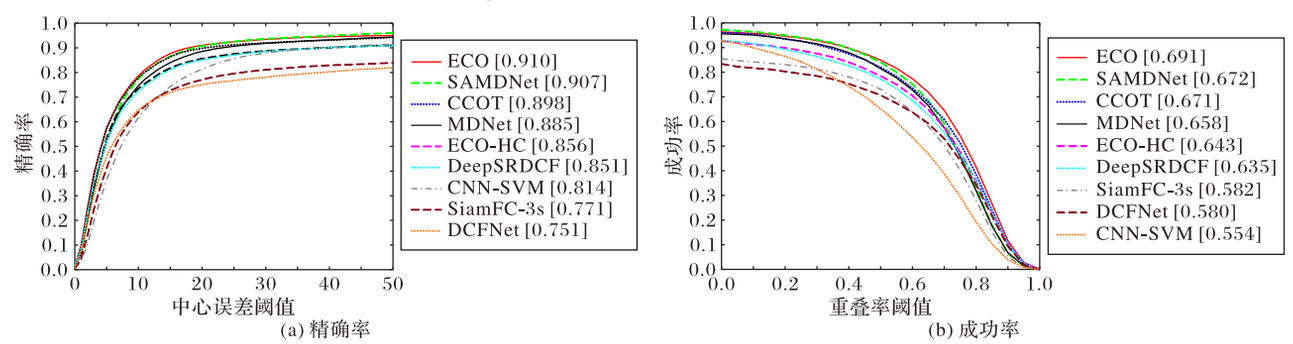
对于通道注意力模块而言,其注意力矩阵的计算方式和空间注意力类似,但是以每个通道的特征图之间的互相映射的角度来衡量彼此间的依赖关系,为了有选择地突出互相关联的通道的重要性的同时最大化通道注意力矩阵信息的丰富性,因此本文算法中通道注意力模块加载到卷积层通道数最多的CONV3之后。



(a) 精确率　　(b) 成功率

图4　SAMDNet在OTB50的测试结果

Fig. 4　Test results of SAMDNet on OTB50



(a) 精确率　　(b) 成功率

图5　SAMDNet在OTB2015的测试结果

Fig. 5　Test results of SAMDNet on OTB2015

表1　不同追踪算法的11个属性的成功率对比

Tab. 1　Success rate comparison of 11 attributes of different tracking algorithms

算法	BC	SV	OC	DE	IPR	OPR	OV	IV	MB	FM	LR
SAMDNet	0.677	**0.654**	0.627	**0.637**	**0.653**	**0.657**	0.635	0.676	0.663	0.647	0.583
CCOT	**0.679**	0.652	**0.673**	0.616	0.623	0.649	**0.645**	**0.679**	**0.703**	**0.675**	0.597
MDNet	0.638	0.642	0.626	0.609	0.647	0.646	0.633	0.658	0.664	0.648	**0.625**
ECO-HC	0.638	0.608	0.627	0.592	0.582	0.608	0.592	0.638	0.627	0.634	0.590
DeepSRDCF	0.624	0.607	0.603	0.567	0.589	0.607	0.553	0.624	0.642	0.628	0.475
CNN-SVM	0.537	0.489	0.514	0.547	0.548	0.548	0.488	0.537	0.578	0.546	0.403
SiamFC-3s	0.523	0.556	0.547	0.510	0.557	0.558	0.506	0.574	0.550	0.568	0.592
DCFNet	0.588	0.569	0.578	0.501	0.557	0.575	0.557	0.588	0.544	0.541	0.533

为了进一步验证本文算法SAMDNet中各模块的有效性,设计了如表2、3所示的消融实验:C1p含义为在卷积层CONV1之后设置了空间注意力模块,C3c含义是在CONV3之后设置了通道注意力模块;I含义为预训练模型是采用了包含实例判别函数的复合损失函数;15指的是测试基准集为OTB2015。实验结果表明,各个模块对算法的性能提升都起着积极作用,最佳的表现来自于三个模块的共同作用。

表2　SAMDNet的子模块组合实验的精确率和成功率对比

Tab. 2　Comparison of SAMDNet' sub-module combinations on precision and success rate

组合	精确率	成功率
C1p_C3c_I_15	**0.907**	**0.672**
C3c_I_15	0.889	0.661
C1p_I_15	0.887	0.658
MDNet	0.885	0.656

表3为注意力模块的组合实验,实验中通道注意力模块的位置是固定的,空间注意力模块的位置是变量。实验结果表明当空间注意模块处于CONV1后可以获得最佳性能,同时也验证了算法模型设计的合理性。

表3　不同注意力模块组合的精确率和成功率对比

Tab. 3　Comparison of different attention module combinations on precision and success rate

组合	精确率	成功率
C1p_C3c_I_15	**0.907**	**0.672**
C2p_C3c_I_15	0.898	0.664
C3p_C3c_I_15	0.885	0.658
MDNet	0.885	0.656

4　结语

本文是基于多域卷积神经网络(MDNet)的算法改进,通过引入自注意力机制来解决原算法中模型漂移问题,此外通过实例判别函数提升了模型对包含相似语义信息目标的判别

能力。由于基于卷积神经网络的深度追踪中目标位置信息容易随着网络深度的加深而逐渐损失,最后影响追踪目标的精确定位,而空间注意力和通道注意力的应用使模型获得了更鲁棒的位置表示,目标定位更加准确。本文算法在广泛使用的测试基准集OTB50和OTB2015上取得了优秀的表现,显著超越MDNet算法的同时,也在两个测试基准集上全面超过了VOT2016的冠军算法CCOT,同时在OTB50精确率指标上也超过了2017年的ECO算法。但是本文算法也存在一些不足,如实验过程中发现该算法在目标出现运动模糊、快速移动和低分辨率等情况时,追踪效果并不理想,所以未来将进一步研究克服该算法存在的不足。

参考文献(References)

[1] HENRIQUES J F, CASEIRO R, MARTINS P, et al. High-speed tracking with kernelized correlation filters[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2015, 37(3): 583-596.

[2] 樊佳庆,宋慧慧,张开华. 通道稳定性加权补充学习的实时视觉跟踪算法[J]. 计算机应用, 2018, 38(6):1751-1754. (FAN J Q, SONG H H, ZHANG K H. Real-time visual tracking via channel stability weighted complementary learning[J]. Journal of Computer Applications, 2018, 38(6): 1751-1754.)

[3] 熊昌镇,车满强,王润玲. 基于稀疏卷积特征和相关滤波的实时视觉跟踪算法[J]. 计算机应用, 2018, 38(8): 2175-2179. (XIONG C Z, CHE M Q, WANG R L. Real-time visual tracking algorithm based on correlation filters and sparse convolutional features [J]. Journal of Computer Applications, 2018, 38(8): 2175-2179.)

[4] SONG Y, MA C, WU X, et al. VITAL: visual tracking via adversarial learning [C]// Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2018: 8990-8999.

[5] GOODFELLOW I J, POUGET-ABADIE J, MIRZA M, et al. Generative adversarial nets [C]// Proceeding of the 27th International Conference on Neural Information Processing Systems. Cambridge: MIT Press, 2014: 2672-2680.

[6] FAN H, LING H. SANet: structure-aware network for visual tracking [C]// Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops. Piscataway: IEEE, 2017: 2217-2224.

[7] PINEDA F J. Generalization of back propagation to recurrent and higher order neural networks [C]// Proceedings of the 1987 International Conference on Neural Information Processing Systems. Cambridge: MIT Press, 1987: 602-611.

[8] NAM H, BAEK M, HAN B. Modeling and propagating CNNs in a tree structure for visual tracking[EB/OL]. [2019-12-20]. https://arxiv.org/pdf/1608.07242.pdf.

[9] NAM H, HAN B. Learning multi-domain convolutional neural networks for visual tracking[C]// Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2016: 4293-4302.

[10] KRISTAN M, MATAS J, LEONARDIS A, et al. The visual object tracking VOT2015 challenge results[C]// Proceedings of the 2015 IEEE International Conference on Computer Vision Workshop. Piscataway: IEEE, 2015: 564-586.

[11] GIRSHICK R, DONAHUE J, DARRELL T, et al. Rich feature hierarchies for accurate object detection and semantic segmentation [C]// Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2014: 580-587.

[12] RUSSAKOVSKY O, DENG J, SU H, et al. ImageNet large scale visual recognition challenge[J]. International Journal of Computer Vision, 2015, 115(3): 211-252.

[13] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[C]// Proceedings of the 31st International Conference on Neural Information Processing Systems. Red Hook, NY: Curran Associates Inc., 2017: 6000-6010.

[14] ZHANG H, GOODFELLOW I J, METAXAS D N, et al. Self-attention generative adversarial networks[C]// Proceedings of the 36th International Conference on Machine Learning. New York: JMLR.org, 2019: 7354-7363.

[15] WANG X, GIRSHICK R, GUPTA A, et al. Non-local neural networks[C]// Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2018: 7794-7803.

[16] SUNG K K, POGGIO T. Example-based learning for view based human face detection[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 1998, 20(1): 39-51.

[17] WU Y, LIM J, YANG M H. Online object tracking: a benchmark [C]// Proceedings of the 2013 IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2013: 2411-2418.

[18] WU Y, LIM J, YANG M H. Object tracking benchmark [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2015, 37(9): 1834-1848.

[19] DANELLJAN M, BHAT G, SHAHBAZ KHAN F, et al. ECO: efficient convolution operators for tracking[C]// Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2017: 6931-6939.

[20] DANELLJAN M, ROBINSON A, SHAHBAZ KHAN F, et al. Beyond correlation filters: learning continuous convolution operators for visual tracking [C]// Proceedings of the 14th European Conference on Computer Vision. Cham: Springer, 2016: 472-488.

[21] DANELLJAN M, HAGER G, SHAHBAZ KHAN F, et al. Adaptive decontamination of the training set: a unified formulation for discriminative visual tracking[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2016: 1430-1438.

[22] HONG S, YOU T, KWAK S, et al. Online tracking by learning discriminative saliency map with convolutional neural network [C]// Proceedings of the 32nd International Conference on Machine Learning. New York: JMLR.org, 2015: 597-606.

[23] BERTINETTO L, VALMADRE J, HENRIQUES J F, et al. Fully-convolutional Siamese networks for object tracking [C]// Proceedings of the 2016 European Conference on Computer Vision, LNCS 9914. Cham: Springer, 2016: 850-865.

[24] WANG Q, GAO J, XING J, et, al. DCFNet: discriminant correlation filters network for visual tracking[EB/OL]. [2019-12-20]. https://arxiv.org/pdf/1704.04057v1.pdf.

This work is partially supported by the National Natural Science Foundation of China (61871260).

LI Shengwu, born in 1994, M. S. candidate. His research interests include visual tracking, deep learning.

ZHANG Xuande, born in 1979, Ph. D., professor. His research interests include image quality evaluation, image restoration.