基于时空异构双流卷积网络的行为识别

丁雪琴 朱轶昇 朱浩华 刘光灿*
(南京信息工程大学自动化学院 江苏 南京 210000)

摘 要 目前大多数基于双流卷积网络的行为识别方法采用同样的时空网络结构,双流合并时会产生大量的冗余信息,从而降低识别的精确度。对此提出一种基于双流网络的时空异构网络结构。该网络采用两种不同的时空网络结构对行为进行分类。此外,对视频序列的长时间结构采用分段形式进行建模,使整个行为视频的学习变得高效。在 UCF101 和 HMDB51 数据集上进行实验,结果证明该时空异构双流网络优于时空同构双流网络。

关键词 深度学习 行为识别 ResNet 双流网络

中图分类号 TP3 **文献标志码** A DOI:10.3969/j.issn.1000-386x.2022.03.025

ACTION RECOGNITION BASED ON SPATIOTEMPORAL HETEROGENEOUS TWO-STREAM CNN

Ding Xueqin Zhu Yisheng Zhu Haohua Liu Guangcan*
(*College of Automation*, *Nanjing University of Information Science & Technology*, *Nanjing 210000*, *Jiangsu*, *China*)

Abstract Most current action recognition methods based on the two-stream convolutional network use the same structure for spatio-temporal networks, and a large amount of redundant information is generated when two streams are merged, thereby reducing the accuracy of recognition. Based on the above problems, this paper proposes a spatiotemporal heterogeneous network structure based on a two-stream network. This two-stream network used two different spatiotemporal network structures to classify actions. In addition, the long-term structure of the video sequence was modeled in segments to make the learning of the entire action video efficient. The experiments were performed on UCF101 and HMDB51 datasets. The results prove that the proposed spatiotemporal heterogeneous two-stream network is superior to the spatiotemporal homogeneous two-stream network.

Keywords Deep learning Action recognition ResNet Two-stream network

0 引 言

行为识别是计算机视觉研究的一个热点,目标是从一个未知的视频或图像序列中自动分析其中正在进行的行为。它在视频监控、行为分析、智能家居、视频检索和人机智能交互等领域发挥着重要的作用,但由于视点变化、背景杂乱和光照条件等限制,行为识别仍然面临着重大挑战。近年来,深度卷积网络(ConvNets)[1]在图像和语音识别方面取得了巨大的突破。此后,计算机视觉的研究人员一直试图将卷积网络转移到行为识别上来应用。

与图像领域的成功相比,深度学习在基于视频的行为识别领域发展相对缓慢。主要有两个原因:(1) 与图像数据集相比,视频数据的规模和多样性是不可比拟的,因此需要建立一个用于深度网络训练的大规模标记视频数据库;(2) 与二维图像相比,视频包含更多的时序信息,引入了比图像更复杂的分析工作。

为了解决上述问题,近年来人们针对基于深度卷积网络的视频行为识别进行了许多尝试,也获得快速

收稿日期:2020-02-24。国家自然科学基金优秀青年基金项目(61622305);国家自然科学基金青年基金项目(61502238);江苏省自然科学基金杰出青年基金项目(BK20160040)。**丁雪琴**,硕士生,主研领域:图像处理。**朱轶昇**,硕士生。**朱浩华**,硕士生。**刘光灿**,教授。

发展。Karpathy 等[2]比较了几种用于行为识别的卷积网络体系结构,并在一个非常大的 Sports-1M 数据集上进行了相应的训练过程。Tran 等[3]介绍了一种基于三维卷积网络的动作识别方法。Simonyan 等[4]提出了一种基于双流网络的性能优化方法。虽然这些方法在一定限度上利用了视频中的时间信息,但它们只关注短期的运动变化,没有捕获视频中的长时间信息。为了解决这个问题,Wang 等[5]提出了一种从视频数据中提取长时间信息的时域网络(TSN)。对于时间跨度较长的视频行为识别而言,单帧或者是单个短片段中单帧堆栈的数据量是不够的,需要采用密集时间采样的方式来获取长范围时间结构,但是这样会存在视频连续帧之间的冗余,因此要用稀疏的时间采样来代替密集的时间采样,也就是对视频做抽帧的时候采取较为稀疏的抽帧方式,这样可以去除一些冗余信息,同时降低计算量。Cho 等[6]提出了一个新的时空融合网络(STFN),它集成了整个视频的外观和运动信息的时间动态,然后将捕获的时间动态信息进行融合,以获得更好的视频级表示,并通过端到端训练进行学习。Martinez 等[7]利用细粒度识别方面的进展来改进行为识别的模型,将重点放在如何提高网络的表示能力,也就是改进网络的最后一层,在这一层中变化对计算成本的影响很小。Torpey 等[8]使用三维卷积神经网络从视频采样片段中分别提取局部外观和运动特征,将局部特征连接起来形成全局表示,然后用全局表示训练一个线性支持向量机来执行行为分类。

基于以上方法,本文提出一种基于行为识别的双流卷积网络结构。在原双流网络结构中,时间网络和空间网络具有相同的结构,但人们对表观和运动的理解是两个截然不同的过程,因此空间和时间网络应该是不一样的。为了解决这一难题,本文提出了一种基于时空异双流网络的行为识别方法。此外,为了从视频序列中提取长时间信息,将视频分段[5]的思想引入到提出的时空异构网络中。实验结果表明,本文时空异构双流网络的性能优于时空同构网络。

1　时空异构双流卷积网络模型

本文基于双流卷积网络,提出了时空异构的双流网络结构,在此基础上,将 BN-Inception 和 ResNet 引入作为时空异构双流网络的基本网络,最后引入视频分段的思想,建立了视频分段的时空异构双流卷积网络模型,整体框架如图 1 所示。

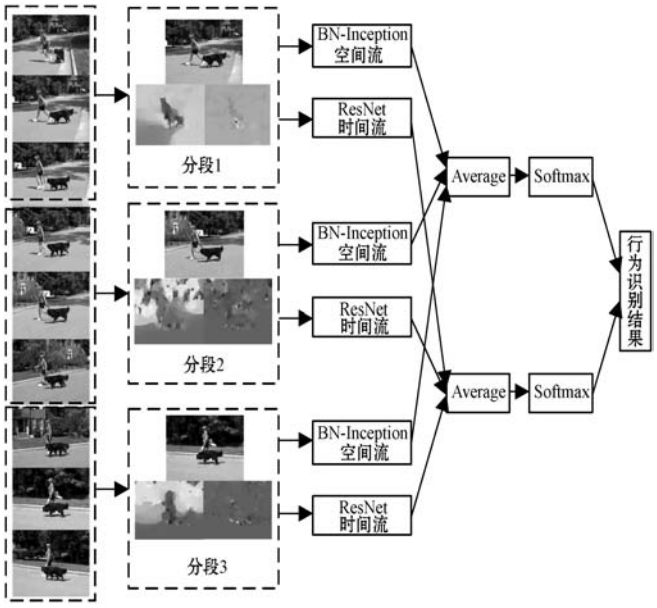

图 1　整体框架

1.1　时空异构双流网络

时空异构双流网络结构如图 2 所示,其采用了不同的网络结构。可以看出,设计时空异构双流网络有两个动机:(1) 当双流网络中的时空网络具有相同的结构即时空同构时,双流合并时会产生大量的冗余信息;(2) 由于人对表观和运动的理解是两个截然不同的过程,所以时空的网络结构应该是不一样的。

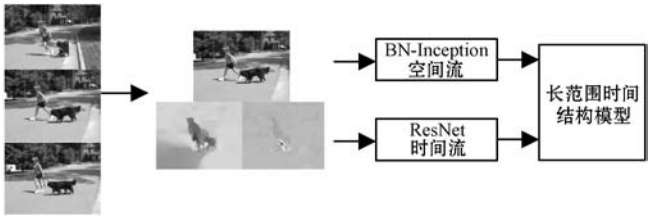

图 2　时空异构双流结构

输入数据的形式是 RGB 图像和光流场,如图 3 所示。单个 RGB 图像是对视频的中的某一帧的静态外观进行编码,光流场是视频的光流信息用来获取运动信息。与原始的双流卷积神经网络[1]一样,空间卷积神经网络对单个 RGB 图像进行操作,而时间卷积神经网络以一组连续的光流场作为输入。

图 3　输入数据形式

1.2　网络架构

一个好的视频网络结构应该提取更多不同的时空信息。为了最大限度地挖掘时空异构双流网络的潜力,本文在时空异构双流网络中引入 ResNet 和 BN-Inception 网络作为提取时空特征的网络结构。

1.2.1 残差网络

为了提取更多的判别信息,采用具有较深层次的 ResNet[9] 作为基本网络。通常网络层数增加,网络就会出现退化现象。为了解决这个问题,He 等[9] 提出了残差网络,该方法没有直接拟合期望的底层映射 $H(x)$,而是通过拟合残差映射 $F(x):=H(x)-x$ 来训练深度网络,其结构设如图 4 所示。

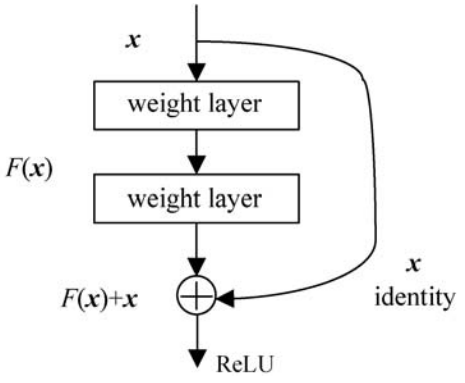

图 4 残差单元结构

残差单元被定义为[9]:

$$\boldsymbol{x}_{l+1}=\sigma(\boldsymbol{x}_l+F(\boldsymbol{x}_l;\boldsymbol{w}_l)) \tag{1}$$

式中:$\boldsymbol{x}_l$ 和 $\boldsymbol{x}_{l+1}$ 分别为第 l 层的输入和输出;$F(\boldsymbol{x}_l;\boldsymbol{w}_l)$ 是非线性残差映射;$\sigma(\cdot)$ 表示 ReLU 函数[10]。残差单元的主要优势是跨层连接的方式可以从第一层直接传播到网络中的任何层,避免了梯度爆炸和消失的问题。同时,跨层连接不会引入额外的参数和计算复杂度,而且可以加快网络的收敛速度。

1.2.2 BN-Inception

BN-Inception[11] 用一个非常有效的正则化方法,使大型卷积网络的训练速度加快,同时收敛后的分类准确率也得到大幅提高。它不再依赖于具有技巧性的参数初始化点,可以使用更大的学习率加快训练过程,另外其正则化手段可以有效缓解 Sigmoid 或 tanh 等激活函数的梯度消失问题,同时在一定程度上也降低了对 Dropout 等手段的依赖。

由于 ResNet 能够通过增加相当的深度来提高准确率,BN-Inception 网络用一个非常有效的正则化方法,让大型卷积网络的训练速度加快,同时收敛后的分类准确率也得到大幅提高。因此本文将 ResNet 和 BN-Inception 网络作为基本网络,构建了一个更深层次的时空异构双流网络。与双流网络使用的 VGG 网络相比,ResNet 具有更少的滤波器和更低的计算复杂度。虽然增加了 ResNet 的深度,但 ResNet-50(38 亿次)和 ResNet-101(76 亿次)的计算复杂度仍然低于 VGG-16(153 亿次)和 VGG-19(196 亿次)。

1.3 建模长范围时间结构

视频中的长时间信息对行为识别也起着非常重要的作用。从 TSN[5] 中得到引导,通过视频分段来提取视频序列中长时间的时间信息来提高时空异构双流网络的性能。根据时间的长短,将视频分成 K 个等长片段 $\{S_1,S_2,\cdots,S_K\}$,基于分段的空时异构双流卷积网络 Y 对行为的识别可以表示为:

$$Y(T_1,T_2,\cdots,T_K)=H(g(F(T_1;W),F(T_2;W),\cdots,F(T_K;W))) \tag{2}$$

式中:$(T_1,T_2,\cdots,T_K)$ 是一个片段序列,每个代码片段 T_K 从其对应的片段 S_K 中随机采样,在空间网络对应的是 RGB 帧图像,时间网络是光流;$F(T_K;W)$ 是一个带有参数 W 的卷积神经网络函数,该函数对代码片段 T_K 进行操作,生成所有类的类分数;分段融合函数 $g(\cdot)$ 将多个短片段的输出融合,得到空间网络或时间网络的特征。利用输出函数 $H(\cdot)$ 对识别结果进行分类,利用 Softmax 函数得到各行为类别的概率值。

分段融合的最终损失函数定义为:

$$L(y,G)=-\sum_{i=1}^{C}y_i\left(G_i-\log\sum_{j=1}^{C}\exp G_j\right) \tag{3}$$

式中:C 表示动作类别的数量;y_i 表示关于类别 i 的基准标签;$G_i=g(F(T_1;W),F(T_2;W),\cdots,F(T_K;W))$ 是类 i 的类得分,通过对 K 个片段的同一类别的得分进行平均得到。本文利用多个片段,用标准的反向传播算法联合优化模型参数 W。反向传播过程中,W 的梯度对时空异构双流网络行为识别损失值 L 可以推导出如下公式:

$$\frac{\partial L(y,G)}{\partial W}=\frac{\partial L}{\partial G_k}\sum_{k=1}^{K}\frac{\partial g}{\partial F(T_K)}\frac{\partial F(T_K)}{\partial W} \tag{4}$$

然后,通过小批量随机梯度下降法得到相关的模型参数。从式(4)可以看出,使用 K 个小片段的类别融合 G 来更新参数。使用此类优化方式,能学习到视频级的模型参数,进而获得长期的时间信息。

2 实 验

2.1 数据集

本文在 UCF101[12] 和 HMDB51[13] 两大数据集上验证方法的有效性。UCF101 数据集包含 101 个动作类和 13 320 个视频剪辑。HMDB51 由 51 个动作类别的 6 766 个视频剪辑组成。对于这两个数据集,本文遵循 THUMOS13 挑战机制[14] 的评估方案,在训练和测试过程中,将每个数据集分为三组,以三组数据的平均准确性作为评价模型效果的指标。

2.2 基本参数设置

本次实验是基于 PyTorch 0.3.0 深度学习框架。

采用 MBGD 来学习网络参数,批量参数为 256,动量参数为 0.9,使用来自 ImageNet 的预训练模型初来始化网络权重。在实验中设置了一个较小的学习率。对于空间网络,初始化学习率为 0.001,每 2 000 次迭代减少到它的 1/10 次。整个训练过程在 4 500 次迭代停止。对于时间网络,设置初始化学习率为 0.005,经过 12 000 和 18 000 次迭代后,学习率降低到它的 1/10,最大迭代设置为 20 000。

在测试过程中按照双流网络结构[4]的测试方法。在相同的时间间隔内,从动作视频中采样 25 帧 RGB 帧或光流堆栈。对于每个采样帧,通过裁剪 4 个角、1 个中心和其水平翻转来获得网络的 10 个输入。本文融合时空网络采用的是加权平均,设置空间网络和时间网络的权值比为 1∶1.5。以下所有的实验都是在 UCF101 第一组数据集上进行。

2.3　不同分段数目性能分析

将视频分为 K 个等长的片段来对长范围时间视频进行建模。当视频段数较少时,会导致行为信息提取不足,训练模型过于简单;当视频段数较多时,将导致数据冗余,增加计算量。表 1 显示了使用 ResNet50/101 网络时,不同视频段下时间网络的识别性能。结果表明,将视频分成三段时有较好的识别性能。因此在以下实验中,视频片段的数目都设置为 3。

表 1　时间网络中不同视频段数的行为识别准确率对比(%)

网络结构	$K=1$	$K=2$	$K=3$	$K=4$	$K=5$
ResNet50	77.23	77.89	78.63	78.46	77.84
ResNet101	78.54	79.75	80.90	79.88	78.65

2.4　不同分段融合函数性能分析

在式(2)中,分段融合函数由函数 $g(\cdot)$ 定义。本文评估了最大池化、平均池化和加权平均池化三个融合方案来作为融合函数的形式。实验结果见表 2。可以看出,平均池化函数可以获得最佳性能,最大池化的方式整体性能较差,可能是由于视频分段中内容不同会导致判别误差比较大。因此在以下实验中,本文选择平均池化作为默认的分段融合函数。

表 2　基于 BN-Inception 结构下不同融合方式准确率对比(%)

融合函数	空间网络	时间网络	双流融合
最大池化	84.86	86.70	91.44
平均池化	85.56	87.23	93.39
加权平均池化	85.64	87.16	92.45

2.5　时空异构和时空同构网络分析

本节中的所有实验都是在 UCF101 的第一组数据上进行的。本文将时空异构网络分为同一类型的不同深度的网络和不同类型的网络。测试使用了 ResNet-50、ResNet-101 和 BN-Inception[11]。比较了三种不同网络结构的性能,分别为:(1) 具有相同结构的时空网络;(2) 深度不同但结构相同的时空网络;(3) 具有不同网络结构的时空网络。在实验中可以发现结构相同但深度不同的时空网络的性能要优于时空同构网络,实验结果见表 3。从双流融合的结果来看,ResNet-101 是时间网络的最佳选择。选择 ResNet-101 作为时间网络,选择不同结构的 BN-Inception 作为空间网络时,其对 UCF101 的第一组数据的准确率为 92.24%。实验表明,时空异构网络的性能优于时空同构网络。

表 3　时空异构和时空同构网络的准确率比较(%)

网络结构	空间网络	时间网络	双流融合
(1) spatio_ResNet50 + temp_ResNet50	83.46	82.65	90.74
(2) spatio_ResNet50 + temp_ResNet101	83.46	87.39	91.19
(3) spatio_BN-Inception + temp_ResNet101	86.21	87.39	92.24

2.6　与现有方法对比

表 4 将本文方法与现有方法进行比较,如基于稠密轨迹编码方式的 DT[15] 和 iDT[16] 表示方法、基于深度学习方法的 3D 卷积网络(C3D)[17]、双流卷积网络(Two Stream)[4]、空间时间分解卷积网络($F_{ST}CN$)[18] 和长期卷积网络(LTC)[21]。从表 4 中 UCF 101 和 HMDB51 数据集可以看出,本文方法优于其他方法。与双流方法(Two Stream)[4] 相比,其准确率分别提高了 4.3 百分点和 3.1 百分点。验证了时空异构双流网络在基于长时间结构上的建模是效果显著的,相比于时空同构双流网络,时空异构双流网络的性能有一定的提高。

表 4　本文方法与其他方法的准确率比较(%)

序号	方法	UCF101	HMDB51
1	DT + MVSV[15]	83.5	55.9
2	iDT + FV[16]	85.9	57.2
3	C3D(3 nets)[17]	85.2	/
4	Two Stream[4]	88.0	59.4
5	$F_{ST}CN$(SCI fusion)[18]	88.1	59.1
6	Two stream + LSTM[19]	88.6	/
7	TDD + FV[20]	90.3	63.2
8	LTC[21]	91.7	64.8
9	本文方法	92.3	62.5

3 结 语

本文提出了一种用于人体行为识别的时空异构双流网络。由于人类对表象和运动的认识和理解是两个完全不同的过程,本文改进了现有的方法,设计了不同的网络结构来提取时空信息。通过实验研究在性能上对时空异构双流网络和时空同构双流网络进行比较,从结果可见时空异构双流网络的性能更好。同时为了发掘时空异构网络的最大潜力,以 ResNets 和 BN-Inception 作为基本网络来提取更多的表观和运动特征。在此基础上,建立了视频的长时间时间信息提取结构。通过端到端培训,该网络在 HMDB51 和 UCF101 数据集上的性能显著提高。

参 考 文 献

[1] LeCun Y, Boser B, Denker J S, et al. Backpropagation applied to handwritten zip code recognition[J]. Neural Computation,1989,1(4):541 - 551.

[2] Karpathy A, Toderici G, Shetty S, et al. Large-scale video classification with convolutional neural networks[C]//2014 IEEE Conference on Computer Vision and Pattern Recognition. IEEE,2014:1725 - 1732.

[3] Tran D, Bourdev L, Fergus R, et al. Learning spatiotemporal features with 3D convolutional networks[C]//2015 IEEE International Conference on Computer Vision(ICCV). IEEE, 2015:4489 - 4497.

[4] Simonyan K, Zisserman A. Two-stream convolutional networks for action recognition in videos[C]//2014 Annual Conference on Neural Information Processing Systems, 2014: 568 - 576.

[5] Wang L M, Xiong Y J, Wang Z, et al. Temporal segment networks: Towards good practices for deep action recognition [C]//2016 14th European Conference on Computer Vision. Springer,2016:20 - 36.

[6] Cho S, Foroosh H. Spatio-temporal fusion networks for action recognition[C]//2018 Asian Conference on Computer Vision. Springer,2018:347 - 364.

[7] Martinez B, Modolo D, Xiong Y J, et al. Action recognition with spatial-temporal discriminative filter banks[C]//2019 IEEE/CVF International Conference on Computer Vision(ICCV). IEEE,2019:5482 - 5491.

[8] Torpey D, Celik T. Human action recognition using local two-stream convolution neural network features and support vector machines[EB]. arXiv:2002.09423,2002.

[9] He K M, Zhang X Y, Ren S Q, et al. Deep residual learning for image recognition[C]//2016 IEEE Conference on Computer Vision and Pattern Recognition(CVPR). IEEE, 2016:770 - 778.

[10] Krizhevsky A, Sutskever I, Hinton G E. Imagenet classification with deep convolutional neural networks[C]//2012 26th Annual Conference on Neural Information Processing Systems,2012:1106 - 1114.

[11] Ioffe S, Szegedy C. Batch normalization: Accelerating deep network training by reducing internal covariate shift[EB]. arXiv:1502.03167,2015.

[12] Soomro K, Zamir A R, Shah M. UCF101: A dataset of 101 human actions classes from videos in the wild[EB]. arXiv: 1212.0402,2012.

[13] Kuehne H, Jhuang H, Garrote E, et al. HMDB: A large video database for human motion recognition[C]//2011 International Conference on Computer Vision. IEEE, 2011: 2556 - 2563.

[14] Idrees H, Zamir A R, Jiang Y G, et al. The THUMOS challenge on action recognition for videos "in the wild"[J]. Computer Vision and Image Understanding, 2017, 155: 1 - 23.

[15] Cai Z W, Wang L M, Peng X J, et al. Multi-view super vector for action recognition[C]//2014 IEEE Conference on Computer Vision and Pattern Recognition. IEEE,2014:596 - 603.

[16] Wang H, Schmid C. Action recognition with improved trajectories[C]//2013 IEEE International Conference on Computer Vision. IEEE,2013:3551 - 3558.

[17] Tran D, Bourdev L, Fergus R, et al. Learning spatiotemporal features with 3D convolutional networks[C]//2015 IEEE International Conference on Computer Vision(ICCV). IEEE, 2015:4489 - 4497.

[18] Sun L, Jia K, Yeung D Y, et al. Human action recognition using factorized spatio-temporal convolutional networks [C]//2015 IEEE International Conference on Computer Vision(ICCV). IEEE,2015:4597 - 4605.

[19] Ng J Y H, Hausknecht M, Vijayanarasimhan S, et al. Beyond short snippets: Deep networks for video classification [C]//2015 IEEE Conference on Computer Vision and Pattern Recognition(CVPR). IEEE,2015:4694 - 4702.

[20] Wang L M, Qiao Y, Tang X O. Action recognition with trajectory-pooled deep-convolutional descriptors [C]//2015 IEEE Conference on Computer Vision and Pattern Recognition(CVPR). IEEE,2015:4305 - 4314.

[21] Varol G, Laptev I, Schmid C. Long-term temporal convolutions for action recognition[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017, 40(6): 1510 - 1517.