计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
Computer Systems & Applications,2021,30(3):272−275 [doi: 10.15888/j.cnki.csa.007840]

E-mail: csa@iscas.ac.cn
http://www.c-s-a.org.cn
Tel: +86-10-62661041

基于深度学习的多模态时空动作识别①

吴 敏, 王 敏

(河海大学 计算机与信息学院, 南京 211100)
通讯作者: 吴 敏, E-mail: wumwumwumin@163.com

摘 要: 针对视频理解中的时序难点以及传统方法计算量大的困难, 提出了一种带有时空模块的方法用于动作识别. 该方法采用残差网络作为框架, 加入时空模块提取图像以及时序信息, 并且加入 RGB 差值信息增强数据, 采用 NetVLAD 方法聚合所有的特征信息, 最后实现行为动作的分类. 实验结果表明, 基于时空模块的多模态方法具有较好的识别精度.

关键词: 时空模型; 多模态; 动作识别

引用格式: 吴敏,王敏.基于深度学习的多模态时空动作识别.计算机系统应用,2021,30(3):272–275. http://www.c-s-a.org.cn/1003-3254/7840.html

Multi-Modal Spatiotemporal Action Recognition Based on Deep Learning

WU Min, WANG Min

(College of Computer and Information, Hohai University, Nanjing 211100, China)

Abstract: In view of the time-series difficulty in video understanding and a large amount of calculation in traditional methods, we propose a method with spatio-temporal module for action recognition. With a residual network as the framework, this method adds spatio-temporal module to extract images and time series, adds RGB difference to enhance data, and finally uses the NetVLAD method to aggregate all feature information. In this way, actions are classified. The experimental results show that the multimodal method based on spatio-temporal module has better recognition accuracy.

Key words: spatio-temporal model; multi-modal; action recognition

1 引言

由于互联网的快速发展, 传播媒介的日渐丰富, 网络视频的数量以指数级的速度大量增长. 如何理解视频内容成为一个亟需解决的问题. 动作识别作为计算机视觉中的一个热门领域受到了广泛的关注, 在监控分析、人机交互、体育视频解读等领域内有着广阔的应用前景.

在动作识别中, 有两个关键的有效信息: 空间信息和运动信息. 一个识别系统的性能在很大程度上取决于它能否从中提取和利用相关信息. 然而, 由于许多复杂的因素, 例如比例变化、视角变化和相机运动等, 提取这些信息是非常困难的. 因此, 设计有效的特征表示和模型方法来处理这些挑战, 同时保留动作的有效分类信息就变的至关重要. 随着深度学习在图像、文本等领域内取得了成功后, 该方法在动作识别领域内也得到了广泛的应用, 由早期的手工特征的方法, 转变为基于深度学习的方法. 卷积神经网络有着强大的建模能力, 近年来, 机器设备计算能力的提升和大型数据集的出现, 使得基于深度学习的方法成为视频分析动作识别的参考标准.

动作识别作为动作预测领域和人体姿态分析的基础, 其主要目标就是对视频中的人物所做的行为动作进行理解分类, 那么如何有效利用视频中的各部分有效信息进行识别是首要问题. 视频识别和图像识别中最大的区分点就是时序信息的使用和建模. 早期采用时空描述符用于特征的提取和分类, Wang 等[1] 提出采

① 收稿时间: 2020-07-19; 修改时间: 2020-08-28; 采用时间: 2020-09-01; csa 在线出版时间: 2021-03-03

用 Fisher 向量对密集运动轨迹 DT 进行编码表示. 基于此, Wang 等[2] 提出改进后的 IDT 算法, 改进特征正则化方式和特征编码方式, 在动作识别上取得了显著的成果. 深度学习的出现使得能够更好地进行特征的提取和学习. 2D 卷积建模用于视频理解主要是对单帧视频进行特征的提取, 不能够很好地对时序信息进行建模. Simonyan 等[3] 提出了将基于外观的信息与运动相关的信息分离出来, 使用两个并行的卷积网络处理 RGB 和光流输入, 基于空间流和光流图的双流卷积网络方法用于动作识别, 识别率高. Wang 等[4]基于双流网络提出时间片段网络, 将整段视频分割成连续的视频片段, 将每段视频分别输入到网络中, 它将稀疏时间采样策略和基于视频的监督相结合, 使用整个视频有效的学习. 3D 卷积[5] 能够对时空信息进行更好的捕捉, 但是所需的计算成本太大. Ji 等[6] 首先提出了扩展时间信息后的 3D 卷积网络用于动作识别, 使用三维核从空间和时间维度中提取特征. Tran 等[7]使用三维卷积和三维池化进一步改进 3D 卷积网络并命名为 C3D. 近年来对于识别的实时性要求不断提高, 网络架构转向采用轻量级的模块来替代传统的光流方法来减少计算量. Lee 等[8] 提出了包含运动模块的运动特征网络 MFNet, 该运动块可以在端到端训练的统一网络中的相邻帧之间编码时空信息. Jiang 等[9] 将 2D 网络作为主干架构, 提出一个简单高效的 STM 模块用于编码空间和运动信息. Feichtenhofer 等[10] 提出了快慢结合的模型, 使用了一个慢速高分辨率 CNN(Slow 通道) 来分析视频中的静态内容, 同时使用一个快速低分辨率 CNN(Fast 通道) 来分析视频中的动态内容, 对同一个视频片段应用两个平行的卷积神经网络, 取得了显著的效果. Zhao 等[11] 将 RGB 和光流嵌入到一个具有新层的二合一流网络中, 在运动条件层从流图像中提取运动信息, 在运动调制层利用这些信息生成用于调制低层 RGB 特征的变换参数, 进行端到端的训练, 利用运动条件对 RGB 特征进行调制可以提高检测精度.

本文采用时空模块提取图像以及时序信息, 使用平移部分通道的方法来实现时空信息的融合, 减少计算量, 同时加入 RGB 差值信息增强数据, 最后采用 NetVLAD 聚合所有的特征信息实现行为动作的分类.

2 基于深度学习的多模态时空动作识别

2.1 移位时空模块

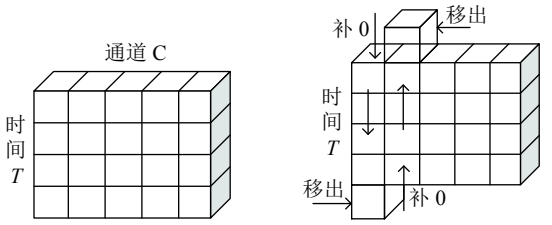
该模块的主要功能是对不同时间点提取的视频帧特征图进行信息交换, 从而实现时序特征的提取. 地址的移位用于图像识别取得了较好的效果. 根据文献 [12] 中的模型, 不同于卷积操作, 移位操作本身不需要参数或浮点运算, 相反移位操作包含一系列的记忆性操作, 可以通过移位操作融合 1×1 卷积来提取聚合特征信息, 从而减少计算量. 以普通一维卷积举例来说, 预测值表示为对不同输入进行加权求和的结果值, 如式 (1) 所示. 换一个角度如果将输入值看成是当前时间点和相邻时间点的输入, 也就是输入值看成移位后的−1, 0, 1 三个时间点的输入值后, 如式 (2) 所示, 再进行乘性相加, 如式 (3) 所示. 由此移位卷积可以概括为移位和乘性相加两个过程的结果.

$$Y_i = \omega_1 X_{i-1} + \omega_2 X_i + \omega_3 X_{i+1} \tag{1}$$

$$X_i^{-1} = X_{i-1}, X_i^0 = X_i, X_i^{+1} = X_{i+1} \tag{2}$$

$$Y = \omega_1 X^{-1} + \omega_2 X^0 + \omega_3 X^{+1} \tag{3}$$

将 T 帧图片 C 通道的输入进行排列后的张量, 如图 1 所示.



(a) 移位前原始特征图 (b) 双向移位后的特征图

图 1 移位模块的特征示意图

每行的各图片通道都表示的是不同时间点获取的图片帧特征值. 对于不同时间点下的同一通道的特征值沿着时间维度进行平移, 部分通道值向下平移一格, 部分通道值向上平移一格, 移位后空缺的部分补 0, 多出的特征图通道值移出, 从而实现双向平移, 相邻帧的特征信息在移动后与当前帧混合. 但并不是平移越多, 交换的信息也越多. 如果移位的比例太小, 时间建模的能力可能不足以处理复杂的时间关系; 如果移位的比例太大, 空间特征学习能力可能会降低过多. 为了进行有效的时空信息融合, 避免移动过多的通道而导致空间建模能力下降, 只移动部分通道, 从而达到平衡空间特征学习和时间特征学习的模型能力.

本文将该移位模块加入到残差网络的每个分支残差块中, 在卷积操作前进行移位操作, 不增加 3D 计算

量的情况下实现时空信息的融合, 对于每个插入的移位时空模块, 时间感受野被放大 2 倍, 由此进行复杂的时间建模.

2.2 多模态

除了充分利用时空信息之外, 本文还加入叠加的 RGB 差值进行多模态的输入, 实现信息增强的效果. 常用的提取光流图来表征运动信息的方法计算量大, 在光流图计算过程中的关键步骤是将像素值沿时间方向求偏导, 所以本文将其简化成 RGB 差值来作为输入, 来表示外观变化和显著运动的区域, 从中训练学习运动信息, 从而大大节省了光流提取的时间. 得出的预估分数与时空特征得出的分数进行相加平均用于识别结果.

2.3 NetVLAD 方法

VLAD 方法[13] 在图像检索领域中作为局部聚合描述符向量, 对提取的图像特征进行后处理编码用于图像的表示, 近年来开始应用到端到端的卷积神经网络中用来表示图像特征. 本文采用 NetVLAD 方法[14] 来作为池化层加入到卷积层的最后, 作为池化层来聚合特征信息.

对于一张特征图 x, 需要从空间位置 $i \in \{1 \cdots N\}$ 获取 D 维的特征向量 $x_i \in R^D$ 来表示该特征图. 首先给定 K 个聚类中心 c_k 将特征空间 R^D 划分成 K 个单元. 每一个特征向量 x_i 都对应着一个单元, 并用残差向量 $x_i - c_k$ 表示特征向量和聚类中心的差值, 由此得到的差分向量表示为:

$$V(j,k) = \sum_{i=1}^{N} \frac{e^{-\alpha \|x_i - c_k\|^2}}{\sum_{k'} e^{-\alpha \|x_i - c_{k'}\|^2}} (x_i(j) - c_k(j)) \tag{4}$$

其中, $x_i(j)$ 和 $c_k(j)$ 表示特征向量 x_i 和聚类中心 c_k 的第 j 个分量, α 是可训练的超参数. 输出矩阵 V 中的第 K 列表示的是第 K 个单元中聚合的特征向量, 接着将矩阵按列进行归一化, 以及 L_2-归一化后化为一维向量 $v \in R^{KD}$ 表征特征图. 最后将输出值送入到全连接层用于分类.

3 实验分析

3.1 实验环境与实验数据

实验硬件配置为 GTX 1080 Ti, 编程语言为 Python, 基于 PyTorch 框架. 数据集是 UCF101[15] 和 HMDB51[16]. UCF101 数据集包含 101 个动作类别, 共 13 320 个视频片段. HMDB51 数据集是一个包含电影、网络视频等多个来源的真实动作视频的集合, 共 51 个类别, 6766 个视频片段. 数据集都提供了相应的训练集和测试集的划分. 调整视频帧为 224×224 作为网络的输入.

训练参数为: 50 个 epoch, 初始学习率为 0. 01, 权值衰减率为 1e–4, 批处理大小为 16, dropout 值为 0. 5. 本文使用从 Kinetics 数据集[17] 预先训练的权重进行微调,并冻结批处理规范化层. 对于残差移位模块, 根据文献 [18] 中的研究结果, 当部分移位信道 1/4 (双向移位每个方向 1/8) 时, 性能达到峰值.

时空模块的部分采用的是 ResNet50 框架, 将时空模块加入到残差网络的分支残差块中, 获得更好的空间特征学习能力.

3.2 实验结果分析

从表 1 可以看出, 本文中加入时空模块以及多模态的方法确实能够对识别精度有一定程度的提升, 对比 C3D、ArtNet 方法, 在预训练数据集相同, 浮点运算的数量级相同的条件下, 对两个数据集的识别精度分别达到了不同程度的提升. 对比 TSN 方法, 识别精度得到了很大的提升, 在两个数据集上分别提升了 8.8 和 19.4 个百分点, 也能够看出使用大型动作数据集进行预训练得出的参数优化能够使实验结果精度得到更大的提升. 在计算资源充足的条件下, 预训练能够对识别的精度起到较大的提升影响. 同时对比 I3D 方法, 本文方法在基于 2D 模型下的浮点计算量, 能够达到与之相匹敌的识别精度, 实现了计算量和识别精度两方面的平衡.

表 1 实验结果

方法	预训练数据集	基础架构	FLOPs (GB)	UCF101	HMDB51
TSN[19]	ImageNet+Kinetics	Inception V2	16	86. 4	53. 7
C3D[7]	Kinetics	ResNet-18	20	89. 8	62. 1
ArtNet[20]	Kinetics	ResNet-18	24	94. 3	70. 9
I3D[21]	ImageNet+Kinetics	3D ResNet-50	153	95. 6	74. 8
本文	Kinetics	ResNet-50	33	95. 2	73. 1

4 结论与展望

本文提出了一种带时空模块的多模态方法. 该方法将时空模块引入到 2D 卷积网络中, 实现时空信息的提取融合, 不增加浮点运算, 同时加入 RGB 差值进行信息增强, 并采用 NetVLAD 方法聚合所有的特征信息, 最后实现行为动作的分类, 在数据集 UCF101 和 HMDB51

上达到了比较理想的识别精度, 且与 3D 方法的计算量相比较, 较好地实现了计算量和识别精度的平衡.

参考文献

1 Wang H, Klaser A, Schmid C, *et al*. Action recognition by dense trajectories. Proceedings of CVPR 2011. Providence, RI, USA. 2011. 3169–3176.

2 Wang H, Schmid C. Action recognition with improved trajectories. Proceedings of 2013 IEEE International Conference on Computer Vision. Sydney, Australia. 2013. 3551–3558.

3 Simonyan K, Zisserman A. Two-stream convolutional networks for action recognition in videos. Proceedings of the 27th International Conference on Neural Information Processing Systems. Montreal, QC, Canada. 2014. 568–576.

4 Wang LM, Xiong YJ, Wang Z, *et al*. Temporal segment networks: Towards good practices for deep action recognition. Proceedings of the 14th European Conference on Computer Vision. Amsterdam, the Netherlands. 2016. 20–36.

5 Carreira J, Zisserman A. Quo vadis, action recognition? A new model and the kinetics dataset. Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, HI, USA. 2017. 4724–4733.

6 Ji SW, Xu W, Yang M, *et al*. 3D convolutional neural networks for human action recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2013, 35(1): 221–231. [doi: 10.1109/TPAMI.2012.59]

7 Tran D, Bourdev L, Fergus R, *et al*. Learning spatiotemporal features with 3D convolutional networks. Proceedings of 2015 IEEE International Conference on Computer Vision. Santiago, Chile. 2015. 4489–4497.

8 Lee M, Lee S, Son S, *et al*. Motion feature network: Fixed motion filter for action recognition. Proceedings of the 15th European Conference on Computer Vision. Munich, Germany. 2018. 392–408.

9 Jiang BY, Wang MM, Gan WH, *et al*. STM: Spatiotemporal and motion encoding for action recognition. Proceedings of 2019 IEEE/CVF International Conference on Computer Vision. Seoul, Republic of Korea. 2019. 2000–2009.

10 Feichtenhofer C, Fan HQ, Malik J, *et al*. SlowFast networks for video recognition. Proceedings of 2019 IEEE/CVF International Conference on Computer Vision. Seoul, Republic of Korea. 2019. 6201–6210.

11 Zhao JJ, Snoek CGM. Dance with flow: Two-in-one stream action detection. Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, CA, USA. 2019. 9927–9936.

12 Wu BC, Wan A, Yue XY, *et al*. Shift: A zero FLOP, zero parameter alternative to spatial convolutions. Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, UT, USA. 2018. 9127–9135.

13 Jégou H, Douze M, Schmid C, *et al*. Aggregating local descriptors into a compact image representation. Proceedings of 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. San Francisco, CA, USA. 2010. 3304–3311.

14 Arandjelovic R, Gronat P, Torii A, *et al*. NetVLAD: CNN architecture for weakly supervised place recognition. Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, NV, USA. 2016. 5297–5307.

15 Soomro K, Zamir AR, Shah M. Ucf101: A dataset of 101 human actions classes from videos in the wild. arXiv: 1212.0402, 2012.

16 Kuehne H, Jhuang H, Stiefelhagen R, *et al*. HMDB51: A large video database for human motion recognition. Nagel WE, Kröner DH, Resch MM. High Performance Computing in Science and Engineering' 12. Berlin, Heidelberg: Springer, 2013. 571–582.

17 Kay W, Carreira J, Simonyan K, *et al*. The kinetics human action video dataset. arXiv: 1705.06950, 2017.

18 Lin J, Gan C, Han S. TSM: Temporal shift module for efficient video understanding. Proceedings of 2019 IEEE/CVF International Conference on Computer Vision. Seoul, Republic of Korea. 2019. 7082–7092.

19 Wang LM, Xiong YJ, Wang Z, *et al*. Temporal segment networks for action recognition in videos. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2019, 41(11): 2740–2755. [doi: 10.1109/TPAMI.2018.2868668]

20 Wang LM, Li W, Li W, *et al*. Appearance-and-relation networks for video classification. Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, UT, USA. 2018. 1430–1439.

21 Wang XL, Gupta A. Videos as space-time region graphs. Proceedings of the 15th European Conference on Computer Vision. Munich, Germany. 2018. 413–431.