



JOURNAL OF IMAGE AND GRAPHICS

主办: 中国科学院空天信息创新研究院
中国图象图形学学会
北京应用物理与计算数学研究所

中国图象学报 中国图形学报

2022
04
VOL.27

ISSN1006-8961
CN11-3758/TB



口罩人脸姿态分类 P1125



第27卷第4期（总第312期）
2022年4月16日

中国精品科技期刊
中国国际影响力优秀学术期刊
中国科技核心期刊
中文核心期刊

版权声明

凡向《中国图象图形学报》投稿，均视为同意在本刊网站及CNKI等全文数据库出版，所刊载论文已获得著作权人的授权。本刊所有图片均为非商业目的使用，所有内容，未经许可，不得转载或以其他方式使用。

Copyright

All rights reserved by Journal of Image and Graphics, Institute of Remote Sensing and Digital Earth, CAS. The content (including but not limited text, photo, etc) published in this journal is for non-commercial use.

主管单位 中国科学院
主办单位 中国科学院空天信息创新研究院
中国图象图形学学会
北京应用物理与计算数学研究所

主 编 吴一戎
编辑出版 《中国图象图形学报》编辑出版委员会
通信地址 北京市海淀区北四环西路19号
邮 编 100190
电子信箱 jig@aircas.ac.cn
电 话 010-58887035
网 址 www.cjig.cn

广告发布登记号 京朝工商广登字20170218号
总 发 行 北京报刊发行局
订 购 全国各地邮局
海外发行 中国国际图书贸易集团有限公司
(邮政信箱: 北京399信箱 邮编: 100048)
印刷装订 北京科信印刷有限公司

Journal of Image and Graphics

Title inscription: Song Jian | Monthly, Started in 1996

Superintended by Chinese Academy of Sciences
Sponsored by Aerospace Information Research Institute, CAS
China Society of Image and Graphics
Institute of Applied Physics and Computational Mathematics

Editor-in-Chief Wu Yirong
Editor, Publisher Editorial and Publishing Board of Journal of Image and Graphics
Address No. 19, North 4th Ring Road West, Haidian District, Beijing, P. R. China
Zip code 100190
E-mail jig@aircas.ac.cn
Telephone 010-58887035
Website www.cjig.cn

Distributed by Beijing Bureau for Distribution of Newspapers and Journals
Domestic All Local Post Offices in China
Overseas China International Book Trading Corporation
(P.O.Box 399, Beijing 100048, P.R.China)
Printed by Beijing Kexin Printing Co., Ltd.

CN 11-3758/TB
ISSN 1006-8961
CODEN ZTTXFZ

国外发行代号 M1406
国内邮发代号 82-831
国内定价 60.00元



残差密集结构的东巴画渐进式重建(第1084页)



结合双模板融合与孪生网络的鲁棒视觉目标跟踪(第1191页)



注意力机制的曲面沉浸式投影系统补偿(第1238页)

综述

中国图像工程: 2021

章毓晋 1009

人脸伪造及检测技术综述

曹申豪, 刘晓辉, 毛秀青, 邹勤 1023

数据集论文

MTMS300: 面向显著物体检测的多目标多尺度基准数据集

李楚为, 张志龙, 李树新 1039

图像处理和编码

利用条件生成对抗网络的光场图像重聚焦

谢柠宇, 丁宇阳, 李明悦, 刘渊, 律睿悠, 晏涛 1056

二阶变分图像恢复模型的重启动快速ADMM方法

宋田田, 潘振宽, 魏伟波, 李青 1066

残差密集结构的东巴画渐进式重建

蒋梦洁, 钱文华, 徐丹, 吴昊, 柳春宇 1084

图像分析和识别

融合自注意力机制的生成对抗网络跨视角步态识别

张红颖, 包雯静 1097

着装场景下双分支网络的人体姿态估计

吕中正, 刘骊, 付晓东, 刘利军, 黄青松 1110

嵌入双尺度分离式卷积块注意力模块的口罩人脸姿态分类

陈森徽, 刘文波, 张弓 1125

联合损失优化下的高相似度奶山羊身份识别

尚诚, 王美丽, 宁纪锋, 李群辉, 姜雨, 王小龙 1137

对抗一致性约束的无监督域自适应绝缘子检测

李梅玉, 李仕林, 赵明, 方正云, 张亚飞, 余正涛 1148

注意力机制改进轻量SSD模型的海面小目标检测

贾可心, 马正华, 朱蓉, 李永刚 1161

注意力引导网络的显著性目标检测

何伟, 潘晨 1176

结合双模板融合与孪生网络的鲁棒视觉目标跟踪

陈志良, 石繁槐 1191

流形正则化约束的图像语义分割

肖振久, 宗佳旭, 兰海, 魏宪, 唐晓亮 1204

空洞可分离卷积和注意力机制的实时语义分割

王因, 侯志强, 蒲磊, 马素刚, 程环环 1216

自适应多任务学习的自动艺术分析

杨冰, 向学勤, 孔万增, 施妍, 姚金良 1226

虚拟现实与增强现实

注意力机制的曲面沉浸式投影系统补偿

雷清桦, 杨婷, 程鹏 1238

遥感图像处理

遥感影像空间分治快速匹配

卫春阳, 乔彦友 1251

Chinagraph 2020

结合门循环单元和生成对抗网络的图像文字去除

王超群, 全卫泽, 侯诗玉, 张晓鹏, 严冬明 1264

面向人体骨骼运动数据优化的双自编码器网络

李书杰, 朱海生, 王磊, 刘晓平 1277

采用蒸馏训练的时空图卷积动作识别融合模型

杨清山, 穆太江 1290

面向时变体数据的特征可视化方法

刘力 1302

ChinaVR 2020

带法向约束的隐式T样条曲线重构

任浩杰, 寿华好, 莫佳慧, 张航 1314

观看经度联合加权全景图显著性检测算法

孙耀, 陈纯毅, 胡小娟, 李凌, 邢琦玮 1322

CONTENTS

JOURNAL OF IMAGE AND GRAPHICS



Gradual model reconstruction of Dongba painting based on residual dense structure(P1084)



Double template fusion based siamese network for robust visual object tracking(P1191)



An attention mechanism based inter-reflection compensation network for immersive projection system(P1238)

Review

- Image engineering in China: 2021
Zhang Yujin 1009
- A review of human face forgery and forgery-detection technologies
Cao Shenhao, Liu Xiaohui, Mao Xiuqing, Zou Qin 1023

Dataset

- MTMS300: a multiple-targets and multiple-scales benchmark dataset for salient object detection
Li Chuwei, Zhang Zhilong, Li Shuxin 1039

Image Processing and Coding

- Light field image re-focusing based on conditional generative adversarial networks leverage
Xie Ningyu, Ding Yuyang, Li Mingyue, Liu Yuan, Lyu Ruimin, Yan Tao 1056
- Restart fast ADMM methods for second-order variational models of image restoration
Song Tiantian, Pan Zhenkuan, Wei Weibo, Li Qing 1066
- Gradual model reconstruction of Dongba painting based on residual dense structure
Jiang Mengjie, Qian Wenhua, Xu Dan, Wu Hao, Liu Chunyu 1084

Image Analysis and Recognition

- The cross-view gait recognition analysis based on generative adversarial networks derived of self-attention mechanism
Zhang Hongying, Bao Wenjing 1097
- Dual branch network for human pose estimation in dressing scene
Lyu Zhongzheng, Liu Li, Fu Xiaodong, Liu Lijun, Huang Qingsong 1110
- A embedded dual-scale separable CBAM model for masked human face poses classification
Chen Senqiu, Liu Wenbo, Zhang Gong 1125
- Joint loss optimization based high similarity identification for milch goats
Shang Cheng, Wang Meili, Ning Jifeng, Li Qunhui, Jiang Yu, Wang Xiaolong 1137
- Unsupervised domain adaptation insulator detection based on adversarial consistency constraints
Li Meiyu, Li Shilin, Zhao Ming, Fang Zhengyun, Zhang Yafei, Yu Zhengtao 1148
- Attention-mechanism-based light single shot multiBox detector modelling improvement for small object detection on the sea surface
Jia Kexin, Ma Zhenghua, Zhu Rong, Li Yonggang 1161
- The salient object detection based on attention-guided network
He Wei, Pan Chen 1176
- Double template fusion based siamese network for robust visual object tracking
Chen Zhiliang, Shi Fanhuai 1191
- Image semantic segmentation based on manifold regularization constraint
Xiao Zhenjiu, Zong Jiaxu, Lan Hai, Wei Xian, Tang Xiaoliang 1204
- Real-time semantic segmentation analysis based on cavity separable convolution and attention mechanism
Wang Nan, Hou Zhiqiang, Pu Lei, Ma Sugang, Cheng Huanhuan 1216
- Automatic art analysis based on adaptive multi-task learning
Yang Bing, Xiang Xueqin, Kong Wanzeng, Shi Yan, Yao Jinliang 1226

Virtual Reality and Augmented Reality

- An attention mechanism based inter-reflection compensation network for immersive projection system
Lei Qinghua, Yang Ting, Cheng Peng 1238

Remote Sensing Image Processing

- Spatial divide and conquer based remote sensing image quick matching
Wei Chunyang, Qiao Yanyou 1251

Chinagraph 2020

- Gate recurrent unit and generative adversarial networks for scene text removal
Wang Chaoqun, Quan Weize, Hou Shiyu, Zhang Xiaopeng, Yan Dongming 1264
- Dual auto-encoder network for human skeleton motion data optimization
Li Shujie, Zhu Haisheng, Wang Lei, Liu Xiaoping 1277
- Action recognition using ensembling of different distillation-trained spatial-temporal graph convolution models
Yang Qingshan, Mu Taijiang 1290
- A feature visualization method for time-varying volume data
Liu Li 1302

ChinaVR 2020

- Implicit T-spline curve reconstruction with normal constraint
Ren Haojie, Shou Huahao, Mo Jiahui, Zhang Hang 1314
- Saliency detection algorithm of panoramic images using joint weighting with observer's attention longitude
Sun Yao, Chen Chunyi, Hu Xiaojuan, Li Ling, Xing Qiwei 1322

中图法分类号: TP399 文献标识码: A 文章编号: 1006-8961(2022)04-1290-12

论文引用格式: Yang Q S and Mu T J. 2022. Action recognition using ensembling of different distillation-trained spatial-temporal graph convolution models. Journal of Image and Graphics, 27(04): 1290-1301 (杨清山, 穆太江. 2022. 采用蒸馏训练的时空图卷积动作识别融合模型. 中国图象图形学报, 27(04): 1290-1301) [DOI: 10.11834/jig.200791]

采用蒸馏训练的时空图卷积动作识别融合模型

杨清山, 穆太江*

清华大学计算机科学与技术系, 北京 100084

摘要: **目的** 基于深度学习的动作识别方法识别准确率显著提升,但仍然存在很多挑战和困难。现行方法在一些训练数据大、分类类别多的数据集以及实际应用中鲁棒性较差,而且许多方法使用的模型参数量较大、计算复杂,提高模型准确度和鲁棒性的同时对模型进行轻量化仍然是一个重要的研究方向。为此,提出了一种基于知识蒸馏的轻量化时空图卷积动作识别融合模型。**方法** 改进最新的时空卷积网络,利用分组卷积等设计参数量较少的时空卷积子模型;为了训练该模型,选取两个现有的基于全卷积的模型作为教师模型在数据集上训练,在得到训练好的教师模型后,再利用知识蒸馏的方法结合数据增强技术训练参数量较少的时空卷积子模型;利用线性融合的方法将知识蒸馏训练得到的子模型融合得到最终的融合模型。**结果** 在广泛使用的 NTU RGB + D 数据集上与前沿的多种方法进行了比较,在 CS(cross-subject)和 CV(cross-view)两种评估标准下,本文模型的准确率分别为 90.9% 和 96.5%,与教师模型 2s-AGCN(two-stream adaptive graph convolutional networks for skeleton-based action)相比,分别提高了 2.4% 和 1.4%;与教师模型 DGNN(directed graph neural network)相比,分别提高了 1.0% 和 0.4%;与 MS-AAGCN(multi-stream attention-enhanced adaptive graph convolutional neural network)模型相比,分别提高了 0.9% 和 0.3%。**结论** 本文提出的融合模型,综合了知识蒸馏、数据增强技术和模型融合的优点,使动作识别更加准确和鲁棒。**关键词:** 动作识别;知识蒸馏;深度学习;融合模型;图卷积

Action recognition using ensembling of different distillation-trained spatial-temporal graph convolution models

Yang Qingshan, Mu Taijiang*

Department of Computer Science and Technology, Tsinghua University, Beijing 100084, China

Abstract: **Objective** Skeleton-based action recognition, which is an intensively studied field, aims to classify human actions represented by a sequence of selected key points of the human body into action categories. Skeleton-based action recognition has a wide variety of applications, including human-computer interaction, elderly monitoring, and video understanding. Recognition accuracy has improved significantly in recent years due to the development of deep learning. However, few studies have focused on the number of parameters and the robustness of the model. In previous skeleton-based action recognition methods, convolution with a big kernel size has been used to extract spatial and temporal features for the broad receptive field, leading to an increase in model parameters and more complicated calculations. Many previous studies have confirmed that graph convolution has better performance in skeletal data. However, graph convolution operators are

收稿日期: 2020-12-30; 修回日期: 2021-02-24; 预印本日期: 2021-03-03

* 通信作者: 穆太江 taijiang@tsinghua.edu.cn

基金项目: 国家自然科学基金项目(61902210, 61521002)

Supported by: National Natural Science Foundation of China (61902210, 61521002)

designed manually, and their versatility and robustness are insufficient. Therefore, we hope to design a lightweight temporal convolution module that preserves the large receptive field for temporal feature learning. In the spatial dimension, we aimed for better robustness of the spatial convolution module constructed using two kinds of graph convolutions. We will improve the performance of the model with the help of data enhancement technology to increase the diversity of input data and improve the generalization ability of the model for different perspectives. To this end, a distillation training method that can improve the accuracy of lightweight models is used for model training and a multi-stream spatiotemporal graph convolutional ensemble model is constructed to improve the current methods and increase the accuracy of the skeleton-based action recognition. **Method** In this study, we propose a skeleton-based multi-stream ensemble model composed of six sub-networks for action recognition. These sub-networks are divided into two types: directed graph convolutional sub neural network (DGCNNet) and adaptive graph convolutional sub neural network (AGCNNet). Each sub-network is constructed with temporal convolution modules, spatial convolution modules, and attention modules. The temporal convolution module in the sub-network is designed with a 2D depth-wise group convolution layer with a convolution kernel size of 9×1 and a normal convolution layer with a convolution kernel size of 1×1 . Two types of graph convolution-directed graph convolution and adaptive graph convolution-are used in the spatial convolution module to extract spatial features and enhance the robustness of the model. Three self-attention modules between the spatial convolution and the temporal convolution modules are applied over the channel dimension, spatial dimension, and temporal dimension of the features to underscore the informative features. We also introduce a cross-modal distillation training method that can be used to train lighter and more accurate student models with trained teacher models and ground truth to train the sub-networks. The distillation training consists of two steps: teacher model training and student model training. The teacher model is trained using training data and its weights are fixed after the training is completed. The student model is trained with the feature vector encoded by the frozen teacher model and the ground-truth labels of training data. Two previous methods, 2s-AGCN (two-stream adaptive graph convolutional networks for skeleton-based action) and directed graph neural networks (DGNN), are used as teacher models to train the DGCNNet sub-network and the AGCNNet sub-network, respectively, according to the type of graph convolution in the spatial convolution module. Cross-entropy loss is used for teacher model training and a combination of mean squared error loss and cross-entropy loss is used as the final loss for student model training. The student model trained using the distillation training method is not only lighter but also more accurate than the corresponding teacher model. In addition to joints, we also take bones and affine-transformation-augmented data as input to train the student model. Finally, a multi-stream spatiotemporal graph convolutional ensemble model is constructed with six lightweight sub-networks, and it has better robustness and higher accuracy. The accuracies of our model for cross-subject benchmark and cross-view benchmark of NTU RGB + D dataset are 90.9% and 96.5%, respectively, higher than many other currently best approaches. **Result** We compared our model with the other 14 best models thus far on the widely used NTU RGB + D dataset. Our model achieved a 90.9% cross-subject accuracy and a 96.5% cross-view accuracy in terms of the benchmark. A comparison of our model with the teacher model, 2s-AGCN, indicated that the accuracy increased by 2.4% and 1.4%. When compared with another teacher model, DGNN, the accuracy of our model increased by 1.0% and 0.4%; and when compared with the base-line method, spatial temporal graph convolution networks (ST-GCN), the accuracies of our model are 9.4% and 8.2% higher, respectively. In addition, extensive experiments indicated the effectiveness of knowledge distillation on this task and we also explored the effects of the different combinations of input modalities on the final accuracy of the model. **Conclusion** In this article, we propose a new multi-stream ensemble model that contains six sub-models trained using the distillation training method, and each sub-model is constructed with spatial convolution modules, temporal convolution modules, and attention modules. The results of the experiment indicate that our model outperforms several state-of-the-art skeleton-based action recognition approaches and that the ensemble algorithm can improve the performance.

Key words: action recognition; knowledge distillation; deep learning; ensemble model; graph convolution

0 引言

动作识别有着广泛的应用,包括人机交互、视频理解与检索、场景理解和老人监护等,是计算机视觉一个活跃的研究领域。早期人们主要研究基于 RGB 数据、深度数据和光流数据等的动作识别,随着深度学习技术对人体姿势和关键点估计准确度的提升,催生了很多工具和设备(Cao 等,2017,2021; Fankhauser 等,2015),可以方便地估计人体骨架数据,从而吸引了众多的研究者研究基于骨架的动作识别。

骨架(skeleton)是人体结构和姿势的一种简单表示,每一帧骨架数据包含多个关键点(joints),不同时刻的骨架数据组合在一起构成一个骨架序列(skeleton series)即可表示一个动作。骨架数据由于简单、冗余信息少和计算快捷等特点,广泛用于人体动作识别。大多数方法采用与循环神经网络(recurrent neural network, RNN)、卷积神经网络(convolutional neural network, CNN)或者图卷积网络(graph convolutional network, GCN)相关的经典深度学习网络结构构建模型。GCN 类的方法可以用邻接矩阵等显式地表示关键点的空间位置关系,并基于空间拓扑关系和时域信息设计模型中数据的更新方式。总体来说,与其他类型的方法相比,GCN 类的方法在多个数据集上取得了更好的结果,更适用于基于骨架关键点的动作识别任务。然而,动作识别既要处理关键点空间上的位置信息,还要处理时序上的变化信息,目前很多前沿的 GCN 类方法的模型结构比较复杂,层数比较深,参数量较大,因此需要研究更加简单轻量化和更加鲁棒的模型。另外,关键点的坐标、角度和相机视野等都是重要的信息,不同形式的输入数据对模型准确率影响较大。在 GCN 类的方法中,Shi 等人(2019a, 2020)和 Yan 等人(2018)使用了用于模拟光流的 motion 数据以及 joints 和 bone 数据等,整体提高了识别准确率,但是单独使用某一种数据时模型的准确率较低。

基于当前方法存在的问题,本文研究如何将数据增强、知识蒸馏、模型融合和分组卷积等与深度学习相关的技术和方法更好地应用于该任务,在降低模型参数量的同时进一步提高模型的识别准确率。具体而言,用分组卷积替换全卷积设计时空卷积模

型,降低模型的参数量;为了保证简单模型的准确性,模型训练时使用知识蒸馏的技术,用训练好的教师模型指导学生模型训练;受 Shi 等人(2019a, b)的启发,采用有向图卷积和自适应图卷积两种图卷积形式构建模型的空间卷积(spatial convolution)模块,使两种结构的子模型得以优势互补,结合仿射变换的数据增强方式,最后训练得到一个多股并联的融合模型。

本文有 3 个主要贡献:1)使用分组卷积等技术设计了轻量的时序卷积模块,降低了模型参数量。首次将知识蒸馏应用于基于骨架数据的动作识别问题,保证参数量减少后模型的准确性。2)利用仿射变换等数据增强技术增加了新的输入数据形式,为模型增加了额外观察动作的角度,增加了模型的鲁棒性。3)提出了一个多股并联的融合模型,在 NTU RGB + D 数据集上有更高的识别准确率。

1 相关工作

1.1 基于骨架的动作识别方法

基于骨架的动作识别方法可以分为传统机器学习类的方法和深度学习类的方法。对于传统机器学习类的方法,通常需要手动设计用于动作分类的特征,然后用其他常规的机器学习的方法对动作进行识别和分类。Climent-Pérez 等人(2012)使用遗传算法(genetic algorithm)选择对不同动作有贡献的点,然后使用 K 均值聚类(K-means)方法分类动作;Wang 等人(2013)根据身体部位将关键点分为不同的组,然后使用数据挖掘技术在时间和空间域上挖掘位姿信息,并且使用支持向量机对动作进行分类。此类传统的机器学习相关算法需要手动设计分类特征,表达能力较弱,基本无法胜任分类类别多、数据集庞大的分类任务。深度学习类的方法又可以分为 RNN 类的方法、GCN 类的方法和其他 CNN 类的方法。RNN 类的方法中,研究最多的是使用长短期记忆(long short-term memory, LSTM)结构解决该问题,前人设计了包含不同 LSTM 变体结构的模型,如 ST-LSTM(spatio-temporal long short-term network)(Liu 等,2016)、Part-aware LSTM(Shahroudy 等,2016)等。LSTM 结构的模型在处理语音和文本等时序数据类的任务,例如语音识别、文本翻译等任务中表现突出,尽管此处的骨架数据一定程度上也可以看做时

间序列的数据,所不同的是,动作识别中,非常依赖骨架数据在空间和时间各个维度的变化,使此类方法难以发挥其处理时间序列数据的优势。其他CNN类的方法,有使用3维卷积的如Two-Stream 3DCNN(Liu等,2017a),或者将骨架数据通过不同方式排列以及设计定制化的卷积核大小,从而满足2维卷积操作的方法,如TCN(temporal convolutional networks)(Kim和Reiter,2017)、Synthesized CNN(Liu等,2017b)等。但因为骨架关键点在人体中是有自然连接关系的,这类方法一个共同的不足就是没有很好地利用骨架数据这种“内在信息”;另外,其他使用3D卷积的方法一个比较明显的劣势就是3D卷积导致模型参数多,计算代价大。而GCN是最近两年在这一问题上使用最多的一类方法。

1.2 图卷积动作识别网络

Yan等人(2018)提出的ST-GCN(spatial temporal graph convolution networks)模型,首次将基于图的卷积网络应用于基于骨架的动作识别这一任务,与其他类型的方法相比,显著地提高了在NTU RGB+D数据集上的准确率,证明了图卷积网络在这一任务上强大的表达和生成能力。自此以后,图卷积成为研究该问题的热门方法,不同结构的图卷积模型相继出现。Shi等人(2019b)提出使用端到端的方式学习一个数据驱动的矩阵,作为表示骨架空间连接关系拓扑结构的邻接矩阵的补充,从而使这种拓扑结构更具弹性。Shi等人(2019a)使用一个无环的有向图表示骨架的拓扑结构,设计了一个有向图卷积网络用于提取和更新关键点(joints)和连接关键点之间的骨(bone)的特征,并根据这些特征进行动作识别。Shi等人(2020)在图卷积网络层中增加了注意力(attention)模块,帮助网络更加注意输入数据中重要的点、帧和特征。Wu等人(2019)在图卷积模型中设计了一个新的残值层和稠密链接块,帮助模型更好地学习时域和空域上的有效信息。这些方法都证明了图卷积在基于骨架的动作识别任务中的有效性。

1.3 知识蒸馏

模型蒸馏训练是“机器教机器”学习范式的一种深度学习模型训练方式。蒸馏知识(Hinton等,2015)的目的是让复杂度低的模型学习复杂模型处理任务的方式,从而将知识由复杂模型转移到简单模型用于解决问题,降低模型的计算量。另一个相

关的概念是特权信息(privileged information),Vapnik和Izmailov(2015)介绍了使用特权信息训练的方法,对于“特征—标签”数据,在训练时加上特权信息成为“特征—特权信息—标签”数据用于训练,而用于测试的“特征—标签”数据则并不包含这种特权信息。通过这种方式训练的模型在测试集上能有更高的准确率。综合这些思想,用模型A蒸馏训练模型B的过程为:使用模型A在“特征—标签”数据上训练得到教师模型,然后用教师模型全连接层之前的特征作为特权信息添加到训练数据。最后,“特征—特权信息—标签”数据用于训练模型B,模型B为最终的测试模型。

2 模型构建

模型融合是在深度学习各个领域使用较多的一种策略,其目的是融合一些表达力较弱的分支网络,以构建一个全局优化的整体模型。在基于骨架的动作识别这个问题中,图卷积是有效的一种方法,但不同结构的图卷积模型提取的特征,以及模型做出识别判断时依据的特征信息不一样,融合模型能起到优势互补的作用。

模型的整体示意图如图1所示,模型由有向图卷积子模型(directed graph convolutional sub neural network, DGCNNet)和自适应图卷积子模型(adaptive graph convolutional sub neural network, AGC-NNNet)两种结构的图卷积子模型组成,其中,DGCNNet和AGC-NNNet中的空间卷积模块分别采用与有向图神经网络(directed graph neural networks, DGNN)(Shi等,2019a)和自适应图卷积神经网络(adaptive graph convolutional networks for skeleton-based action, AGCN)(Shi等,2019b,2020)相同的图卷积模块,时序卷积模块使用分组卷积设计,子模型的参数量较少。

2.1 DGCNNet子模型

DGCNNet是本文提出的融合模型中的一个子模型,包含10层light-DGC(light directed graph convolutional)基础层,结构如图2所示,虚线框表示当前后张量通道(channel)数不匹配的时候使用卷积,其中的空间卷积模块DGC模块采用与DGNN(Shi等,2019a)相同的图卷积结构,用于在空间上更新骨架信息。除此之外,在light-DGC基础层中进行了如下设计:

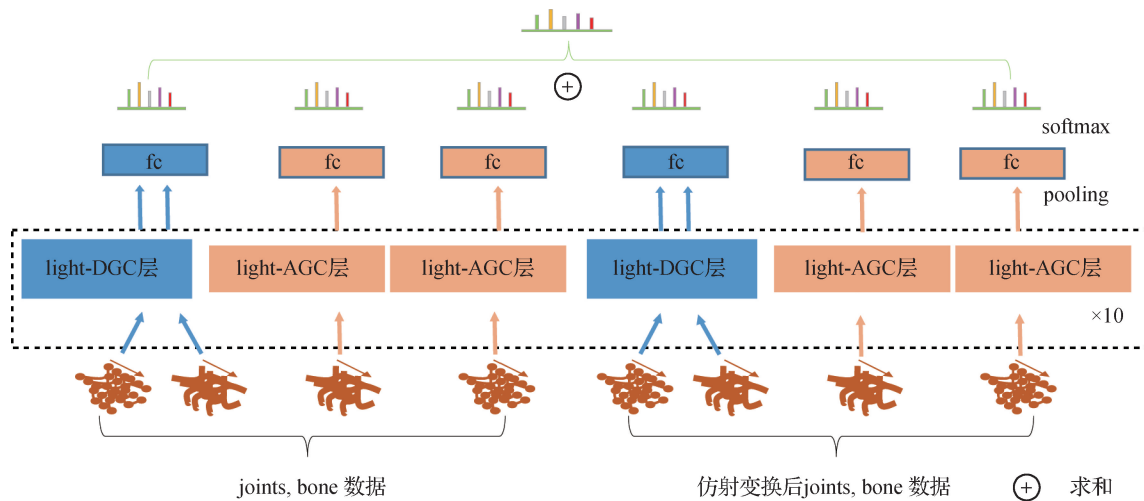


图1 融合模型示意图

Fig. 1 The proposed ensemble model

1) 在空间卷积模块之后串联自注意 (self-attention) 模块。在深度学习神经网络中, 自注意机制 (Parikh 等, 2016) 是一种用来计算输入数据不同位置特征重要性的机制, 每个自注意模块学习一个权重张量, 用于表示各位置特征的“重要程度”, 已经成功应用于语音识别、文本翻译以及计算机视觉的多种任务中。具体而言, 在空间卷积模块之后串联了 3 个自注意模块, 其作用是特征图在时间、空间和特征 3 个维度分别学习一个系数向量, 用于增强特征图中重要的点、重要的帧和重要的特征通道对模型分类中的影响。

2) 为了更好地学习时间上的变化信息, 时域维度的卷积核通常较大, 使用普通卷积导致模型的参

数量较大, 因此, 用于更新数据时域信息的 light-TCN 模块中使用参数量较少的分组卷积替换普通的全卷积。light-TCN 模块的结构如图 3 所示, 使用一个卷积核为 (9, 1) 的逐通道的 2 维卷积和一个核为 (1, 1) 的普通 2 维卷积处理特征图的时域信息, 替换卷积核为 (9, 1) 的普通 2 维卷积。使用该分组卷积策略后, 用 C_i 和 C_o 分别表示输入和输出特征图通道数, 在只考虑卷积的情况下, 该模块的参数量由 $C_o \times C_i \times 9 \times 1$ 减少为 $C_o \times 1 \times 1 \times 9 + C_o \times C_i \times 1 \times 1$ 。

2.2 AGCNet 子模型

用邻接矩阵表示人体骨架关键点之间的连接关系, 并基于连接关系设置卷积方式, 各点的数据由与其相连的点更新, 是另外一类使用广泛的图卷积模

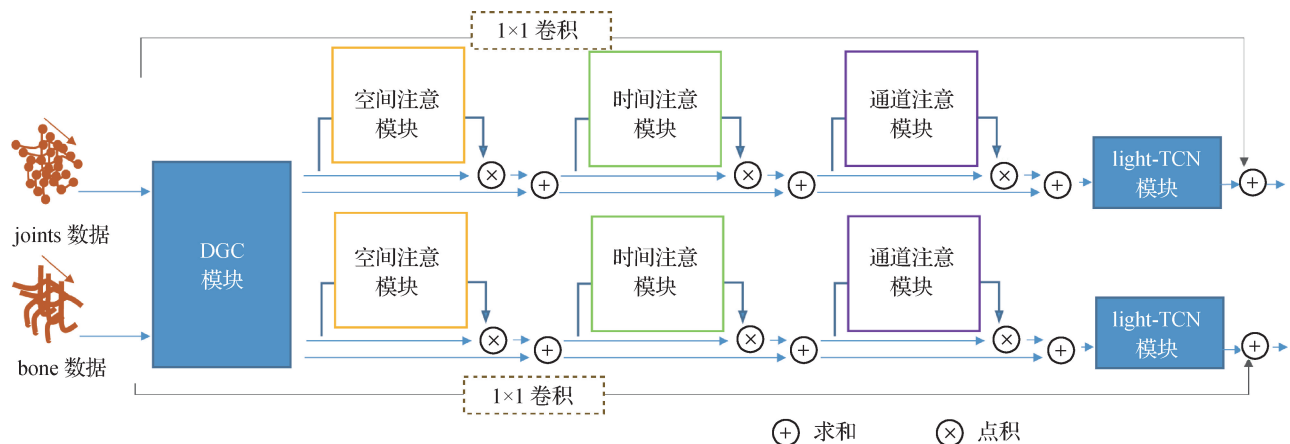


图2 light-DGC 层

Fig. 2 light-DGC layer

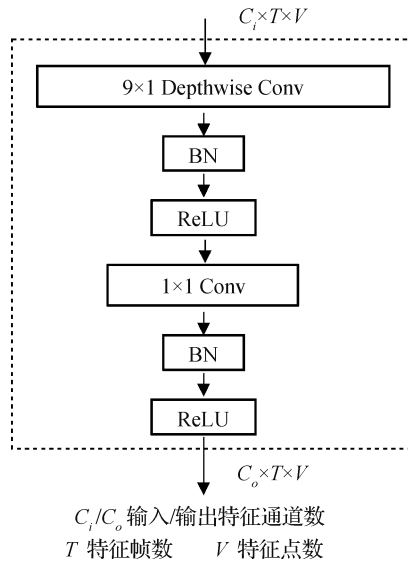


图 3 light-TCN 模块

Fig. 3 light-TCN block

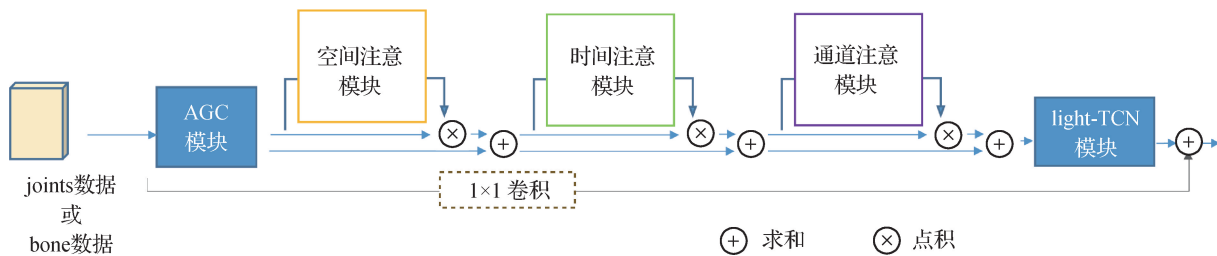


图 4 light-AGC 层

Fig. 4 light-AGC layer

2015) 可以将同一任务中复杂模型学到的知识迁移到简单模型, 提高简单模型的表达力, 从而达到用简单模型处理复杂问题的目的。本文使用的蒸馏训练的方法如图 5 所示, 训练过程分为两步: 1) 使用训练数据训练一个教师模型, 保存模型的参数; 2) 将教师模型全连接层之前的特征作为“特权信息”, 结合训练数据共同训练学生模型, 损失函数与 Crasto 等人 (2019) 使用的相同, 用 fc_{tea} 表示教师模型“特权信息”特征层, 将其与学生模型对应特征层 fc_{stu} 的均方误差 (mean squared error, MSE) 作为损失函数的其中一项, 与模型预测值 $\hat{Y}$ 和数据标签 Y 之间的交叉熵 (cross entropy) 一起, 共同构成最终的损失函数。为了调整均方误差项的比重, 在损失函数中增加了一个均方误差项的权重系数 α 。用于训练的最终的损失函数为

$$L = crossEntropy(\hat{Y}, Y) + \alpha \times MSE(fc_{tea}, fc_{stu}) \quad (1)$$

式中, $crossEntropy()$ 表示计算交叉熵损失, $MSE()$

型。为了能够在模型训练过程中更加关注重要的帧、重要的点和重要的特征, Shi 等人 (2020) 在模型中引入了注意力增强模块。受 Shi 等人 (2019b, 2020) 工作的启发, 用自适应图卷积作为模型的空间卷积模块, 构建融合模型 AGCNet。AGCNet 由 10 层 light-AGC 基础层构成, light-AGC 基础层如图 4 所示, 虚线框表示当前后张量通道数不匹配的时候使用卷积。与 Shi 等人 (2020) 的方法相比, light-AGC 基础层使用参数更少的 light-TCN 模块更新特征图的时域信息, 因为在 Shi 等人 (2020) 的方法中, 用来处理时域的卷积模块的卷积核较大, 导致模型的参数量大, 此处将其替换为 light-TCN 模块, 显著降低了模型的参数量。

2.3 知识蒸馏模型训练

知识蒸馏 (Hinton 等, 2015; Vapnik 和 Izmailov,

表示计算均方差损失。

蒸馏训练时, 用 DGCNet 和 AGCNet 模型作为学生模型, 用 AGCN 和 DGNN 分别作为教师模型蒸馏训练子模型。通过比较学生模型与教师模型的准确率证明使用的蒸馏训练方法的有效性。除了这种教师模型与学生模型结构不同的情况之外, 本文也设置了大量的用同一种模型结构, 在不同的数据上分别训练教师模型和学生模型的跨数据的蒸馏训练实验, 用于证明蒸馏训练的有效性。对 AGCN 进行跨数据蒸馏训练时, 首先使用在 joints 数据上训练得到的模型作为教师模型, 然后用 bone 数据蒸馏训练学生模型; 对 DGNN 模型做跨数据蒸馏训练时, 首先使用骨架数据训练教师网络, 然后用仿射变换后的骨架数据蒸馏训练学生网络。

2.4 数据增强

为了提高模型的准确率, 前人的研究中使用了

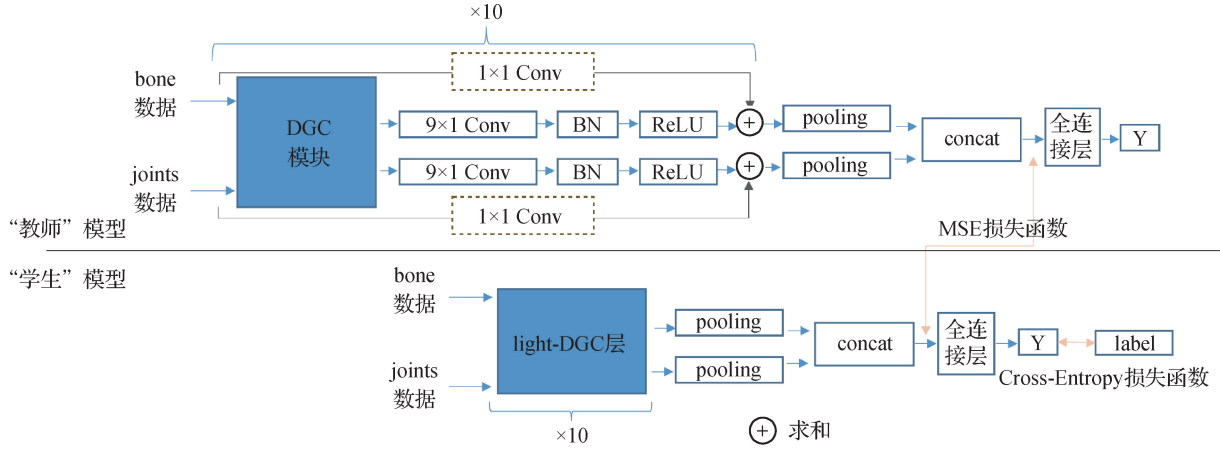


图5 模型蒸馏训练

Fig. 5 The framework of the distillation training

多种增广数据,其中,motion 数据使用较广泛。然而通常情况下,使用 motion 数据单独训练模型时得到的准确率较低。本文选择使用仿射变换作为数据增强方法,以仿射变换后的骨架数据作为增广数据,以提高模型对不同视角的鲁棒性。在 NTU RGB + D 数据集中,人体关键点 joints 是 3 维空间中的点,在数据变换时保持与地面垂直坐标轴的坐标值不变,对其他维坐标值进行仿射变换。对于关键点 X ,经过仿射变换,得到变换后的数据 $X_a = X \cdot A + b$,其

$$\text{中, } A = \begin{pmatrix} \cos\theta & -\sin\theta & 0 \\ \sin\theta & \cos\theta & 0 \\ 0 & 0 & 1 \end{pmatrix}。$$

考虑到在 NTU RGB + D 数据集采集数据时,相机设置的偏移角度为 $\pm 45^\circ$,为了不失一般性以及简化运算,此处设置 $\theta = 60^\circ$, $b = 0$ 。

2.5 模型融合

由于使用仿射变换增加了数据的输入形式,DGCNN 的输入包括骨架数据(joints + bone)和仿射变换后的骨架数据两种形式,而 AGCNN 的输入则包括 joints、bone、仿射变换后的 joints 和仿射变换后的 bone 等 4 种形式,使用蒸馏训练方法,训练得到 6 个参数量较小的子模型。最后测试时,将子模型的 softmax 函数的输出相加,作为模型最后的输出。用 $\hat{Y}_i (i = 1, \dots, 6)$ 表示第 i 个子模型的 softmax 函数的输出,最终的预测值为

$$\hat{Y} = \sum_{i=1}^6 \hat{Y}_i \quad (2)$$

最终的融合模型如图 1 所示。

3 实验分析

实验在 NTU RGB + D 数据集上进行,该数据集有 cross-subject 和 cross-view 两种评判标准。通过在两种标准下设置实验,训练和评估本文提出的模型,并与前沿的方法对比,以说明方法的有效性。

3.1 数据集介绍

NTU RGB + D 数据集是使用 3 个微软的 Kinect v2 相机拍摄的动作识别相关的数据集。整个数据集包含 60 类动作,超过 56 000 个动作实例和 400 万帧数据。每个动作实例包含 RGB 视频、深度数据、人体关键点数据和红外数据。其中,本文使用的人体关键点数据为:每个人体有 25 个关键点,每个关键点表示为相机坐标系的 3 维空间点。NTU RGB + D 提供两种训练和评判的标准:CS(cross-subject) 和 CV(cross-view),前者使用 20 个志愿者的数据作训练集训练,剩下 20 个志愿者的数据作测试集评估模型;后者则使用其中两个视角拍摄的数据作训练集训练,剩余一个视角拍摄的数据作测试集评估模型(Shahrourdy 等,2016)。对于输入数据需要做一些预处理,包括坐标变换使动作主体位于视野中央,将所有动作序列扩展至 300 帧等,处理过程与 Shahrourdy 等人(2016)和 Shi 等人(2019a)的方法相同,得到形状为 $C \times T \times V$ 的多维 joints 数据,其中 C 为坐标维数、 T 为帧数、 V 为点数。除此之外,模型还用到了连接关键点的 bone 数据。bone 数据的产生方式与 Shi 等人(2019a,2020)的方法相同。

3.2 参数设置

模型采用 Pytorch 深度学习框架实现。对于 DGCNNNet 和 AGCNet 子网络,分别使用 10 层 light-DGC 基础层和 10 层 light-AGC 基础层搭建模型,两支子网络的输出通道数相同,10 层的输出通道数分别为 (64、64、64、64、128、128、128、256、256、256)。训练数据批大小为 32,优化器使用随机梯度下降法 (stochastic gradient descent, SGD),学习率初始化为 0.1。训练教师网络时,在训练到 40 和 90 代后将学习率缩小为 1/10,模型总共训练 120 代。蒸馏训练学生网络时,先后在训练到 30 和 40 代后将学习率缩为 1/10,模型总共训练 60 代。训练前 10 代,反向传播不修改 DGCNNNet 和 AGCNet 网络中表示连接关系的矩阵的值,10 代以后正常训练。对于蒸馏训练时损失函数中 α 的取值,通过实验确定,用预训练好并固定权重的 AGCN 模型作教师模型,以 AGCNet 作为学生模型在 bone 数据上进行训练和测试,不同的 α 取值及对应的测试准确率如图 6 所示,根据实验结果,将所有子模型的蒸馏训练的 α 值都设置为 50。

3.3 学生模型与教师模型对比

学生模型使用轻量的时序卷积模块,显著降低了模型参数,同时利用模型蒸馏训练的方法使模型在降低参数量的情况下拥有较高的准确率。对 DGCNNNet 子模型的蒸馏训练,首先在骨架数据 joints + bone 上训练 DGNN,由于原文 (Shi 等, 2019a) 未公开源码,此处使用的 DGNN 网络单层结构如图 5

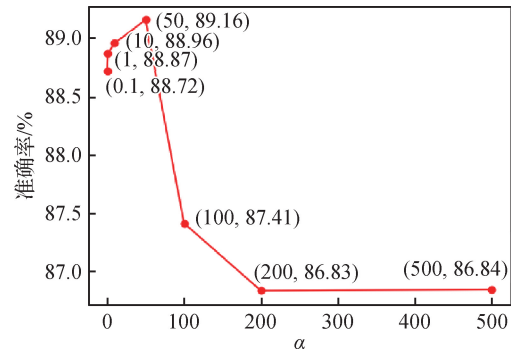


图6 模型准确率与损失函数系数 α 的关系

Fig. 6 The relationship between model accuracy and loss function coefficient α

中教师网络所示,可能与原文代码有微小差别。然后用其蒸馏训练学生模型 DGCNNNet。对 AGCNet 子模型的蒸馏训练,首先在 bone 数据上训练教师模型 AGCN,然后用教师模型蒸馏训练学生模型 AGCNet。对蒸馏得到的 AGCNet 和 DGCNNNet 学生模型与对应的教师模型在参数量、收敛速度和 top1 准确率 3 项指标上进行对比,结果如表 1 所示。从表 1 可以发现,两种结构的学生模型的参数量都有显著降低(减少约 50%),通过蒸馏训练的模型准确率比没有蒸馏训练的高,甚至比参数量大的教师模型的准确率高。这是因为学生模型中使用了轻量的时域卷积模块,降低了模型的参数量,而且训练时通过损失函数增加了来自教师模型的约束,利用模型蒸馏训练的方法保证了学生模型的识别准确率。

表1 学生模型与教师模型对比

Table 1 Comparison of student model and teacher model

模型	top1 准确率/%		参数量/MB	训练代数
	CS 标准	CV 标准		
DGNN (joints + bone)	89.2	95.2	6.17	120
DGCNNNet (joints + bone)	88.7	94.5	3.99	60
DGCNNNet (joints + bone 蒸馏)	89.3	95.5	3.99	60
AGCN (bone)	88.4	94.7	3.78	120
AGCNet (bone)	88.5	94.6	1.73	60
AGCNet (bone 蒸馏)	89.2	95.1	1.73	60

注:加粗字体表示各列最优结果。

模型时域卷积模块量化对比如表 2 所示, g 表示分组卷积的分组数。由于 DGNN 未公开源码,用做教师网络的 DGNN 模型在具体实现时,时域卷积

使用两个核大小为 9×1 的 2 维卷积分别处理 joints 和 bone 数据。表中展示的是模型最后一层的参数,其他层同理。

表 2 模型最后一层时序卷积层参数对比

Table 2 Comparison of last temporal convolution layer in student model and teacher model

模型	参数量	输入通道数	时序卷积层
DGNN	1 179 648	256	$(9 \times 1, 256) (9 \times 1, 256)$
light-DGC	135 680	256	$(9 \times 1, 256, g = 256) (1 \times 1, 256) (9 \times 1, 256, g = 256) (1 \times 1, 256)$
2s-AGCN	589 824	256	$(9 \times 1, 256)$
light-AGC	67 840	256	$(9 \times 1, 256, g = 256) (1 \times 1, 256)$

3.4 数据增强的作用

为了提高融合模型的准确率,本文使用仿射变换数据增强技术,其效果相当于为模型增加了观察动作的视角,增加模型的鲁棒性。实验设置时,分别用仿射变换前后的骨架数据 joints + bone 作为输入,采用蒸馏训练方法,分别对 DGCNNet 和 AGC-NNet 进行蒸馏训练。使用不同输入进行训练得到的子模型的结果,以及将各子模型融合得到多股并联的融合模型的结果如表 3 所示。实验数据表明,利用仿射变换做数据增强能够帮助模型提高准确

率,同时进一步证明了本文提出模型的有效性。

模型对 NTU RGB + D 数据集中各个类别的识别准确率如图 7 所示,有些动作例如 falling、standing up、jump up 在动作发生过程中,骨架信息在空间和时间维度变化很大,这类动作通过融合不同结构的图卷积网络(AGCNNet 和 DGCNNet)能够准确识别;而有的动作例如 writing、playing with phone 等,骨架信息变化不大,对于这类动作,尽管模型融合了不同结构图卷积网络,并使用了数据增强,识别的准确率依然不高,需要更深入的研究。

表 3 本文模型在不同输入数据上的识别准确率

Table 3 The accuracy of our model on different input modalities

/%			
模型	数据类型	CS 标准	CV 标准
DGCNNet	joints + bone	89.3	95.5
DGCNNet	仿射变换后 joints + bone	89.0	95.4
AGCNNet	bone	89.2	95.1
AGCNNet	仿射变换后 bone	89.2	95.4
AGCNNet	joints	88.2	94.9
AGCNNet	仿射变换后 joints	88.5	94.7
融合模型(ensemble model)	-	90.9	96.5

注:加粗字体表示各列最优结果,“-”表示融合模型需要输入本列中的所有数据。

3.5 蒸馏训练的作用

模型蒸馏训练是本文提高模型准确率的一个重要策略,蒸馏训练的损失函数包含均方误差和交叉熵两项,均方误差项的系数通过实验确定,准确率与系数的关系如图 6 所示,根据实验结果,损失函数中均方误差项的系数取值为 50。通过大量的实验,证明了蒸馏训练的有效性,详细数据见表 1 和表 4。表 1 展示了使用蒸馏训练的学生网络有更少的参数量和更高的准确率,而且蒸馏训练的收敛速度快,训练代数少。表 4 展示了使用相同结构的模型在不同

数据上蒸馏训练的效果,训练结果显示通过蒸馏训练能够提高模型的准确率。另外,蒸馏训练也有一定的局限性,因为在模型训练阶段需要首先训练教师网络,训练时增加了一定的计算代价。

3.6 与其他方法对比

为了验证本文提出的融合模型的性能,与其他前沿方法进行对比。由于训练使用的 NTU RGB + D 动作识别数据集较大,只有基于深度学习的相关算法表现出了优异性能,因此选择基于深度学习的算法进行对比,这些算法大致可以分为基于 RNN 的深

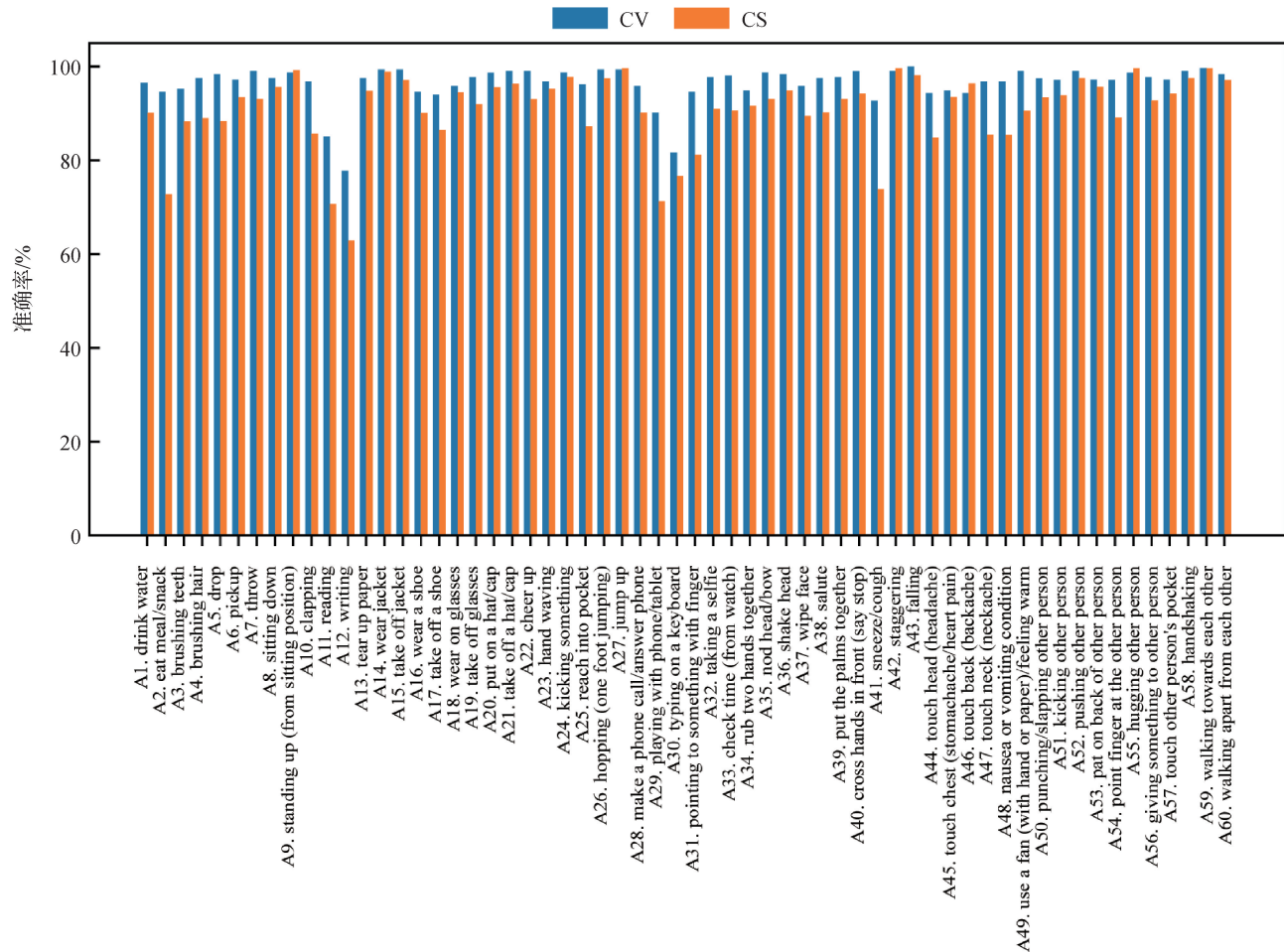


图 7 融合模型对 NTU RGB + D 数据集各类别动作的识别准确率

Fig. 7 Each class accuracy of the proposed ensemble model on NTU RGB + D dataset

表 4 知识蒸馏对教师模型准确率的影响

Table 4 Effect of knowledge distillation on the accuracy of teacher model

模型	数据类型	CV 标准
AGCN	joints	94.9
AGCN	bone	94.7
AGCN(蒸馏)	bone	95.1
DGNN	joints + bone	95.2
DGNN	仿射变换后 joints + bone	95.4
DGNN(蒸馏)	仿射变换后 joints + bone	95.6

注:加粗字体表示最优结果。

度学习算法、基于 CNN 的深度学习算法以及 GCN 相关的深度学习算法 3 类,对比结果如表 5 所示。本文模型在数据集的 CS 和 CV 两种评判标准下准确率分别达到了 90.9% 和 96.5%,明显优于基于 GCN 的基准方法 (Yan 等,2018)的 81.5% 和

表 5 不同方法的准确率对比

Table 5 Comparison of accuracy among different methods

方法	CS 标准	CV 标准
H-RNN (Du 等,2015)	59.1	64.0
Deep LSTM (Shahroudy 等,2016)	60.7	67.3
ST-LSTM (Liu 等,2016)	69.2	77.7
VA-LSTM (Zhang 等,2017)	79.4	87.6
Ind-RNN (Li 等,2018)	81.8	88.0
Two-Stream 3DCNN (Liu 等,2017a)	66.8	72.6
TCN (Kim 和 Reiter,2017)	74.3	83.1
ST-GCN (Yan 等,2018)	81.5	88.3
2s-AGCN (Shi 等,2019b)	88.5	95.1
GCN-NAS (Peng 等,2020a)	89.4	95.7
2s-SDGCN (Wu 等,2019)	89.6	95.7
Mix dimension (Peng 等,2020b)	89.7	96.0
DGNN (Shi 等,2019a)	89.9	96.1
MS-AAGCN (Shi 等,2020)	90.0	96.2
本文	90.9	96.5

注:加粗字体表示各列最优结果。

88.3%,与其他优秀的前沿方法相比,同样有比较强的竞争力。另外,与其他方法一样,模型在CS评判标准下的准确度比在CV评判标准下的低,这与两种标准的设置是密切相关的,在CV标准下,测试集和训练集分别是不同角度同步拍摄的相同动作,而在CS标准下,同一个动作在训练集和测试集由不同的人完成,难以保证一致性,再加上骨架估计模块的误差导致其比CV标准更具挑战性。尽管如此,本文提出的融合模型在CS评判标准下测试集的识别准确率依然达到90.9%,是所有实验方法中的最优结果。

4 结 论

本文首先采用两种不同结构的图卷积构建的空间卷积模块,结合通道、空间、时间3个维度上的自注意模块,并使用逐通道的分组卷积设计的时序卷积模块构建了多支轻量化的子模型。然后用蒸馏训练的方法从参数量大的教师模型中蒸馏知识,训练这些学生子模型。为了增加模型的鲁棒性,训练和测试时使用了仿射变换等数据增强技术,用于增广输入数据。最终,训练得到了6支更轻量更准确的子模型,并进一步用其融合构建了一支参数量较少且准确性和鲁棒性更好的多股并联融合模型。

在NTU RGB+D数据集上采用两种评判标准对本文提出的融合模型进行了大量实验,实验结果表明,模型的准确率比同样是基于图卷积的基准方法有了很大提高,优于众多现行的基于骨架的动作识别前沿算法。在未来的工作中,拟研究解决骨架信息变化小的动作识别问题。这类动作识别与环境及周边物体有密切关系,需要使用其他数据例如RGB数据结合物体检测和识别等技术构建人物与环境及周边物体的关系图。

致 谢 研究过程中模型训练使用的服务器等研究设备得到了“北京市高等学校工程研究中心”和“清华—腾讯互联网创新技术联合实验室”的支持,在此表示感谢。

参考文献 (References)

Cao Z, Hidalgo G, Simon T, Wei S E and Sheikh Y. 2021. OpenPose: realtime multi-person 2D pose estimation using part affinity fields.

IEEE Transactions on Pattern Analysis and Machine Intelligence, 43(1): 172-186 [DOI: 10.1109/TPAMI.2019.2929257]

Cao Z, Simon T, Wei S E and Sheikh Y. 2017. Realtime multi-person 2D pose estimation using part affinity fields//Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Honolulu, USA; IEEE: 1302-1310 [DOI: 10.1109/CVPR.2017.143]

Climent-Pérez P, Chaaraoui A A, Padilla-López J R and Flórez-Revuelta F. 2012. Optimal joint selection for skeletal data from RGB-D devices using a genetic algorithm//Proceedings of the 11th Mexican International Conference on Artificial Intelligence. San Luis Potosí, Mexico; Springer: 163-174 [DOI: 10.1007/978-3-642-37798-3_15]

Crasto N, Weinzaepfel P, Alahari K and Schmid C. 2019. MARS: motion-augmented RGB stream for action recognition//Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Long Beach, USA; IEEE: 7874-7883 [DOI: 10.1109/CVPR.2019.00807]

Du Y, Wang W and Wang L. 2015. Hierarchical recurrent neural network for skeleton based action recognition//Proceedings of 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Boston, USA; IEEE: 1110-1118 [DOI: 10.1109/CVPR.2015.7298714]

Fankhauser P, Bloesch M, Rodriguez D, Kaestner R, Hutter M and Siegwart R. 2015. Kinect v2 for mobile robot navigation: evaluation and modeling//Proceedings of 2015 International Conference on Advanced Robotics (ICAR). Istanbul, Turkey; IEEE: 388-394 [DOI: 10.1109/ICAR.2015.7251485]

Hinton G, Vinyals O and Dean J. 2015. Distilling the knowledge in a neural network [EB/OL]. [2020-12-30]. <http://export.arxiv.org/pdf/1503.02531.pdf>

Kim T S and Reiter A. 2017. Interpretable 3D human action analysis with temporal convolutional networks//Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). Honolulu, USA; IEEE: 1623-1631 [DOI: 10.1109/CVPRW.2017.207]

Li S, Li W Q, Cook C, Zhu C and Gao Y B. 2018. Independently recurrent neural network (IndRNN): building a longer and deeper RNN//Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA; IEEE: 5457-5466 [DOI: 10.1109/CVPR.2018.00572]

Liu H, Tu J and Liu M. 2017a. Two-stream 3D convolutional neural network for skeleton-based action recognition [EB/OL]. [2020-12-30]. <http://export.arxiv.org/pdf/1705.08106.pdf>

Liu J, Shahroudy A, Xu D and Wang G. 2016. Spatio-temporal LSTM with trust gates for 3D human action recognition//Proceedings of the 14th European Conference on Computer Vision. Amsterdam, the Netherlands; Springer: 816-833 [DOI: 10.1007/978-3-319-46487-9_50]

- Liu M Y, Liu H and Chen C. 2017b. Enhanced skeleton visualization for view invariant human action recognition. *Pattern Recognition*, 68: 346-362 [DOI: 10.1016/j.patcog.2017.02.030]
- Parikh A P, Täckström O, Das D and Uszkoreit J. 2016. A decomposable attention model for natural language inference//*Proceedings of 2016 Conference on Empirical Methods in Natural Language Processing*. Austin, USA: Association for Computational Linguistics: 2249-2255 [DOI: 10.18653/v1/D16-1244]
- Peng W, Hong X P, Chen H Y and Zhao G Y. 2020a. Learning graph convolutional network for skeleton-based human action recognition by neural searching//*Proceedings of the 34th AAAI Conference on Artificial Intelligence*. New York, USA: AAAI: 2669-2676 [DOI: 10.1609/aaai.v34i03.5652]
- Peng W, Shi J G, Xia Z Q and Zhao G Y. 2020b. Mix dimension in poincaré geometry for 3D skeleton-based action recognition//*Proceedings of the 28th ACM International Conference on Multimedia*. Lisboa, Portugal: ACM: 1432-1440 [DOI: 10.1145/3394171.3413910]
- Shahroudy A, Liu J, Ng T T and Wang G. 2016. NTU RGB + D: A large scale dataset for 3D human activity analysis//*Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. Las Vegas, USA: IEEE: 1010-1019 [DOI: 10.1109/CVPR.2016.115]
- Shi L, Zhang Y F, Cheng J and Lu H Q. 2019a. Skeleton-based action recognition with directed graph neural networks//*Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. Long Beach, USA: IEEE: 7904-7913 [DOI: 10.1109/CVPR.2019.00810]
- Shi L, Zhang Y F, Cheng J and Lu H Q. 2019b. Two-stream adaptive graph convolutional networks for skeleton-based action recognition//*Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. Long Beach, USA: IEEE: 12018-12027 [DOI: 10.1109/CVPR.2019.01230]
- Shi L, Zhang Y F, Cheng J and Lu H Q. 2020. Skeleton-based action recognition with multi-stream adaptive graph convolutional networks. *IEEE Transactions on Image Processing*, 29: 9532-9545 [DOI: 10.1109/TIP.2020.3028207]
- Vapnik V and Izmailov R. 2015. Learning using privileged information: similarity control and knowledge transfer. *The Journal of Machine Learning Research*, 16(1): 2023-2049
- Wang C Y, Wang Y Z and Yuille A L. 2013. An approach to pose-based action recognition//*Proceedings of 2013 IEEE Conference on Computer Vision and Pattern Recognition*. Portland, USA: IEEE: 915-922 [DOI: 10.1109/CVPR.2013.123]
- Wu C, Wu X J and Kittler J. 2019. Spatial residual layer and dense connection block enhanced spatial temporal graph convolutional network for skeleton-based action recognition//*Proceedings of 2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW)*. Seoul, Korea (South): IEEE: 1740-1748 [DOI: 10.1109/ICCVW.2019.00216]
- Yan S J, Xiong Y J and Lin D H. 2018. Spatial temporal graph convolutional networks for skeleton-based action recognition//*Proceedings of the 32nd AAAI Conference on Artificial Intelligence*. New Orleans, USA: AAAI: 7444-7452
- Zhang P F, Lan C L, Xing J L, Zeng W J, Xue J R and Zheng N N. 2017. View adaptive recurrent neural networks for high performance human action recognition from skeleton data//*Proceedings of 2017 IEEE International Conference on Computer Vision (ICCV)*. Venice, Italy: IEEE: 2136-2145 [DOI: 10.1109/ICCV.2017.233]

作者简介



杨清山, 1993 年生, 男, 硕士研究生, 主要研究方向为计算机视觉、深度学习、行为理解。
E-mail: 17710660314@163.com



穆太江, 通信作者, 男, 助理研究员, 主要研究方向为视觉媒体学习、SLAM 和 3 维重建、人机交互。
E-mail: taijiang@tsinghua.edu.cn