基于特征蒸馏的实时人体动作识别*

Real-time Human Action Recognition Model Based on Feature Distillation

高 璇[1,2] 饶 鹏[2] 刘高睿[2] (1 上海大学通信与信息工程学院,上海 200444;
2 中国科学院上海技术物理研究所 中国科学院智能红外感知重点实验室,上海 200083)

摘要:针对当前人体动作识别算法不能实现准确快速识别的问题,提出了一种基于特征蒸馏的实时人体动作识别算法。首先,针对 C3D 算法准确率较低的问题,引入通道注意力机制,改善网络的性能,得到 SEC3D 后作为教师网络;然后,利用 SEC3D 指导学生网络的学习,将教师网络的知识通过蒸馏算法迁移到学生网络中。在动作视频数据集 UCF101 和 HMDB51 上进行实验,学生网络在比教师网络模型压缩了 23.7 倍、特征维度压缩了 25%的情况下,检测速度达到 358.7f/s,满足实时性的要求。实验结果表明,该算法在远高于原始 C3D 算法检测速度的同时,在尽量减少精度损失下,减少模型参数和计算量,构建了一种轻量型的精度和速度共存的实时人体动作识别模型。

关键词:蒸馏;实时性;通道注意力机制;动作识别

Abstract:Aiming at the problem that current human action recognition algorithms cannot achieve accurate and fast recognition,a real-time human motion recognition algorithm based on feature distillation is proposed.First,aiming at the problem of low accuracy of C3D algorithm,a channel attention mechanism is introduced to improve the performance of the network and get SEC3D as a teacher network.Then,SEC3D is used to guide the learning of the student network.The knowledge of the teacher network is transferred to the student network through a distillation algorithm.Experiments were performed on the action video datasets UCF101 and HMDB51.The student network was compressed by 23.7 times and the feature dimension was reduced by 25% compared with the teacher network model.The student network achieved a detection speed of 358.7f/s.

Keywords:distillation,real-time,channel attention mechanism,action recognition

基于视频的人体动作识别是计算机视觉领域中重要的研究方向,在公共视频监控、虚拟人机交互等都具有重要的价值。视频中包含丰富的时间和空间信息,为了有效地提取视频包含的时间和空间特征,研究人员进行了大量的实验研究。随着卷积神经网络(CNN)的迅速发展,也给人体动作识别带来了新的方向。为了提高网络的精度,研究人员往往会采用更深和更复杂的网络[1-5]。大型网络模型处理速度慢、特征维度高,需要大量的计算资源,使得网络只能在高性能的处理器上运行,因此无法移植到一些嵌入式平台中,如自动驾驶和机器人等平台,由于它们的硬件资源有限,所以设计一种轻量型、能实现实时同时准确率在可接受程度内的网络模型十分必要。

因此,为了解决基于深度卷积神经网络的人体动作识别模型大、计算速度慢的问题,本文提出了一种基于特征蒸馏的轻量型实时人体动作识别模型,通过知识蒸馏对人体动作识别模型进行压缩,对 C3D 算法的识别精度和速度进行了优化。

1 基于特征蒸馏的实时人体动作识别模型

本模型可以分为两部分,第一部分,对 C3D 网络进行优化改进,添加注意力机制进一步提升其准确率后作为教师网络;第二部分,对教师网络进行蒸馏,学生网络直接从教师网络的输出特征中进行学习,增强学生网络的特征表示能力。这样学生网络把教师网络的高维特征进行二次加工压缩,得到的低维特征在蕴含了教师网络和输入视频中知识的同时,又减小了网络模型,降低了特征维度,在保证精度的同时,提高了网络的处理速度。整体模型如图 1 所示。

1.1 特征蒸馏

基于特征蒸馏的算法,首先训练一个精度较高的卷积神经网络作为教师网络,利用教师网络得到训练集的软标签,然后将

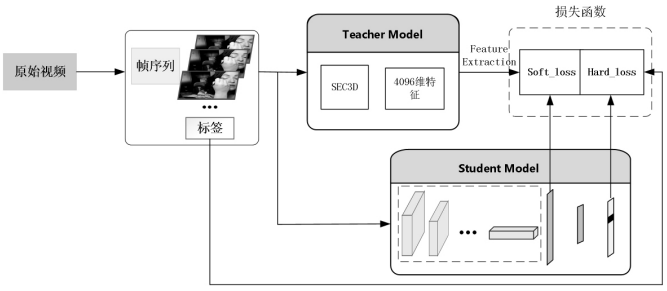

图 1 基于特征蒸馏的动作识别网络结构示意图

真实硬标签和软标签共同训练学生模型,并通过参数 α 作为两部分损失函数的调节因子。

在特征蒸馏模型训练中,通过在原 softmax 函数中引入温度系数 t,得到改进后的 softmax 函数,让输出层产生一个经过软化后的概率向量,如下式:

$$S_i^{'}=\frac{e^{z_i}/t}{\sum_j e^{z_j}/t} \tag{1}$$

其中,t 为蒸馏温度,调节 t 可以改变教师网络和学生网络输出结果的软化程度。

本模型中设计的特征蒸馏训练的损失函数 $L_{KD}(W_s)$ 为:

$$L_{KD}(W_s)=\alpha t^2 * soft\_loss+(1-\alpha) * hard\_loss$$
$$=\alpha t^2 L_{CE}(Q_s^{\tau},Q_t^{\tau})+(1-\alpha)L_{CE}(Q_s,I) \tag{2}$$

其中,α 为两部分损失函数的权重参数,用于调节两部分损失值的权重;t 为温度系数;Q_s^{τ}、Q_t^{τ} 分别为学生网络和教师网络的软化结果,即经过改进 softmax 函数后的输出结果;L_{CE} 为交

* 国家自然科学基金资助项目(105100202)

叉熵损失函数；Q_s 为学生网络经过 softmax 函数后的输出结果；I 为输入视频的真实标签值。

损失函数由两部分组成，第一部分是学生网络与真实标签间的交叉熵损失值，让学生网络向真实标签值进行优化；第二部分是教师网络和学生网络的交叉熵损失，目的是让学生网络向一个软化后的分布进行优化，然后将这两部分的加权和作为特征蒸馏模型训练的损失函数。

1.2 网络结构

2015 年，文献[6]提出了 C3D 网络，利用若干个连续视频帧作为输入，通过 5 个 3D 卷积层和 5 个 3D 池化层构建 CNN，C3D 网络精度虽然不如其他深度三维卷积网络，但是其速率远远高于其他方法。文献[6]在 UCF101 数据集上进行训练，模型识别率为 82.3%，速度为 314f/s。因此基于上述考虑，本模型的教师网络基于 C3D 网络进行改进，引入通道注意力操作[7]，提出了 SEC3D 网络，模型结构如图 2 所示。

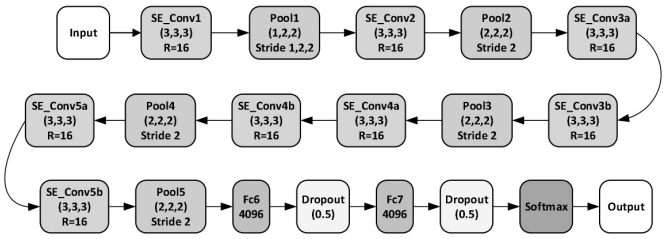

图 2 SEC3D 网络结构图

SE 模块通过对卷积层输出特征图各通道间的依赖关系进行建模以提高网络的特征表达能力和识别准确率。SEC3D 网络由 8 个添加通道注意力的卷积层、5 个池化层、2 个全连接层和 2 个比率为 0.5 的 Dropout 层以及 softmax 层组成。引入注意力机制的 SEC3D 结构在参数较少的情况下，可以实现良好的特征提取效果。

为了降低模型复杂度，学生网络将模型压缩为 2 个卷积层、2 个池化层、1 个全连接层及 softmax 层。卷积核的大小为 3×3×3，池化核的大小为 2×2×2，步长均为 2，滤波器数量分别为 32 和 64。这里并没有使用传统的卷积核来进行特征提取，而采用深度可分离卷积，可在减少模型参数的情况下提高模型训练速度。

深度可分离卷积使用一个深度卷积和一个 1＊1 的卷积来代替传统卷积，分别表示为 Depthwise 过程和 Pointwise 过程，前者使用 C 个不同的卷积核分别对 C 个输入通道进行卷积，收集每个通道的空间特征，即 Depthwise 特征。后者对所有的输入做 1*1 的普通卷积，收集每个点的特征，即 Pointwise 特征，这样不仅仅在理论上更高效，而且由于大量的 1＊1 卷积被使用，可以直接使用矩阵相乘来完成操作，并且不需要额外的预处理操作，在提升运算效率的同时减少学生网络的检测框架的参数量。如图 3 所示。

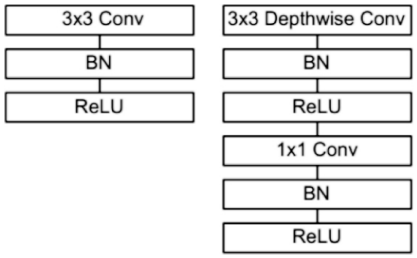

图 3 深度可分离卷积(右)与传统卷积(左)的不同

1.3 模型分析

原 C3D 的模型大小为 299.11MB，添加通道注意力的教师网络 SEC3D 的模型大小为 299.71MB，而进行特征蒸馏后的学生网络模型大小为 12.67MB，即学生网络比教师网络模型大小压缩了 23.7 倍，比 C3D 网络压缩了 23.6 倍。教师网络输出 4096 维特征，学生网络输出 1024 维特征，特征维度压缩了 25%，减小了模型计算时间和存储空间，模型分析如表 1 所示。

表 1 模型分析表

网络	参数量	特征维度	模型大小
C3D	78,409,573	4096	299.11MB
SEC3D	78,567,889	4096	299.71MB
学生网络	3,322,547	1024	12.67MB

2 实验

2.1 实验数据集与设置

本文实验数据集为国际公开的视频动作识别数据集 UCF101[8]和 HDBM51[9]。按照划分标准，将两个数据集划分为 70%的训练集和 30%的测试集。本文分段提取 16 帧图片序列作为输入，对输入的图片序列，本文先将其 resize 成 171×128 大小，然后通过随机裁剪得到 112×112 大小的图像，最后再对图片进行归一化。

本文实验在 Linux 操作系统(Ubuntu16.04)GPU 环境下搭建的 Pytorch 平台上进行，实验的具体细节及超参数设置如下：使用 Adam 优化算法，在训练教师网络时，通过在 SEC3D 中加入 dropout 层来尽量避免模型过拟合，dropout 率设置为 0.5。在 SE 模块中，压缩比例设置为 16。学习率设置为 10^{-5}，批量大小设置为 20，损失函数为交叉熵损失函数，训练迭代 40 次。在训练学生网络时，学习率设置为 10^{-5}，批量大小设置为 20，训练迭代 40 次。蒸馏温度 t 设置为 20，损失函数的权重参数 α 设置为 0.8。

2.2 实验结果与分析

图 4 和图 5 分别展示了教师网络和学生网络在 UCF101 数据集上的网络训练和测试迭代收敛图。从图中可以看出，在 UCF101 测试集中，教师网络在第 17 个 epoch 时，loss 接近平稳，准确率接近 90%；学生网络在第 20 个 epoch 时，loss 接近平稳，准确率接近 85%。

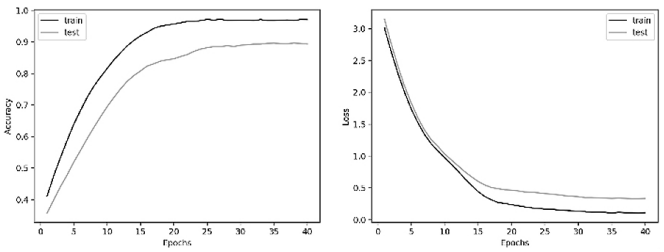

图 4 教师网络在 UCF101 的准确率和损失值

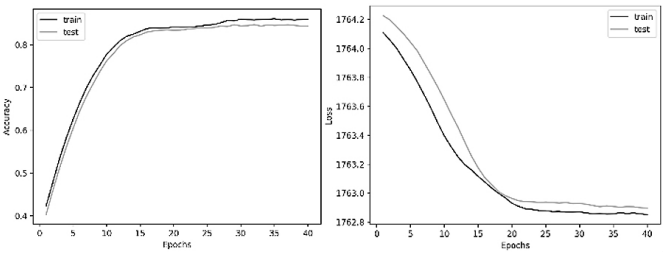

图 5 学生网络在 UCF101 的准确率和损失值

将本文方法和 C3D 进行比较和分析，结果如表 2 所示。

由表 2 可以看出，在 UCF101 数据集上，在模型压缩了 23.7 倍、特征维度压缩了 25%的情况下，学生网络识别精度可以达到 86%，相比教师网络的 90%准确率只降低了 4%；而与 C3D 网络相比，在模型压缩了 23.6 倍、特征维度压缩了 25%的

表 2 本文方法和 C3D 对比表

方法	数据集	精度（%）	识别速度(s)
C3D(1 net) (re-implement) (CPU)	UCF101	81.6	0.6
	HMDB51	-	
SEC3D （CPU）	UCF101	90	0.61
	HMDB51	61	
学生网络 （CPU）	**UCF101**	**86**	**0.0446**
	HMDB51	**56**	

情况下，学生网络识别精度比 C3D 复现网络的精度提高了 4.4%。同时为了测试本文方法的有效性，本文在 CPU 上对模型的识别速度进行了测试，教师模型处理一个样本的数据用时约 0.61s，而经过蒸馏压缩后的学生模型处理一个样本用时约 44.6ms，满足实时的要求，充分证明了本文提出的基于特征蒸馏的模型是有效的，在保证精度的同时，能够实现对人体动作的实时识别，增强了动作识别在移动端的应用性。

本文的主要思想并不是仅为追求较高的准确率，而是在尽量减少精度损失下，减少模型参数和计算量，降低模型对高性能硬件设备的依赖，构建一个轻量型的精度和速度共存的实时人体动作识别模型。本文与其他经典动作识别方法在 UCF101 进行了识别速度的比较，结果见表 3。

表 3 本文方法和其他方法对比表

方法	精度	FPS	模型大小（MB）
C3D(1 net)[6] (GPU)	82.3%	313.9	299.11
Two-stream CNNs (GPU)[1]	88%	14.3	427
EMV+RGB-CNN[10] （GPU）	86.4%	390.7	427
教师网络(CPU)	90%	26.2	299.71
学生网络(CPU)	**86%**	**358.7**	**12.67**

由表 3 可见，本文方法在 CPU 上，在 UCF101 数据集上达到了 358.7fps，满足实时性的要求，远远快于教师网络的检测速度。C3D 文献中在 GPU 上，只使用一个网络，特征维度为 4096 维的情况下，识别准确率为 82.3%，FPS 为 313.9。EMV+RGB-CNN 模型使用了双流的基本结构，用运动向量(MV)代替光流，实现了较快的速度和较高的精度。本文学生网络因为使用单流网络，虽然比 EMV+RGB-CNN 模型略逊一筹，但是本文模型具有结构简单，计算量少的优点，比 EMV+RGB-CNN 模型压缩了 97%，在 CPU 上，在 UCF101 数据集上的准确率为 86%，FPS 为 358.7，实现了降低模型对高性能硬件设备的要求的同时，构建一个轻量型的精度和速度共存的实时人体动作识别模型。

3 结束语

本文提出了基于特征蒸馏的人体动作识别模型，利用添加通道注意力的 C3D 网络结构作为教师网络，指导轻量级的学生网络的训练学习。学生网络在比教师网络模型压缩了 23.7 倍、特征维度压缩了 25%的情况下，速度提高了 92.7%，满足实时性的要求，在 UCF101 和 HMDB51 数据集上分别能达到 86%和 56%的准确率。实验证明，通过本文提出的特征蒸馏算法最终训练生成的动作识别模型，应用于嵌入式设备或移动端设备中时可满足视频人体动作实时识别的需求，但是学生模型在满足实时识别要求时，难免牺牲了一些准确率，所以下一步工作考虑进一步提高学生模型的准确率。

参考文献

[1]Simonyan K,Zisserman A.Two-stream convolutional networks for action recognition in videos[J]. Advances in Neural Information Processing Systems, 2014,1(4):568-576

[2]C Feichtenhofer, A Pinz, R P Wildes. Spatiotemporal residual networks for video action recognition [J].arXiv preprint arXiv: 1611.02155, 2016

[3]Simonyan K, Zisserman A. Very Deep Convolutional Networks for Large-Scale Image Recognition [J].arXiv preprint arXiv: 1409.1556, 2014

[4]João Carreira, Andrew Zisserman. Quo Vadis, Action recognition? A new model and the kinetics dataset [C]//Proceedings of IEEE Conference on Computer Vision and Pattern Recognition. Washington DC: IEEE Computer Society, 2017: 4724-4733

[5]Limin Wang, Wei Li, Wen Li, et al. Appearance-and-relation networks for video classification [C]//IEEE Conference on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE Press, 2018:1430-1439

[6]Du Tran, Lubomir Bourdev, Rob Fergus, et al. Learning spatiotemporal features with 3D convolutional networks[C]//Proceedings of IEEE International Conference on Computer Vision.Washington DC: IEEE Computer Society,2015:4489-4497

[7]Jie Hu, Li Shen,Samuel Albanie,et al.Squeeze-and-excitation networks [C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.Piscataway,NJ:IEEE Press,2018: 7132-7141

[8]Soomro K, Zamir A R, Shah M. UCF101: a dataset of 101 human action classes from videos in the wild [J].arXiv. org, arXiv: 1212. 0402,2012

[9]Hilde K, Hueihan J, Rainer S, et al. HMDB51: a large video database for human motion recognition[C]//High Performance Computing in Science and Engineering. Berlin: Springer, 2013:571-582

[10]Zhang B,Wang L, Wang Z, et al. Real-Time Action Recognition with Enhanced Motion Vector CNNs [C]//2016 IEEE International Conference on Computer Vision and Pattern Recognition (CVPR'16). IEEE, 2016

[收稿日期:2020.5.27]

（上接第 96 页）

cess [J].Chinese Journal of Chemical Engineering,2013,21(2): 127-134

[3]Huang Z,He F,Zheng A Q,et al.Synthesis gas production from biomass gasification using steam coupling with natural hematite as oxygen carrier [J].Journal of Chinese People´s Armed Police Force Academy,2013,53(1):244-251

[4]赵坤,何方,黄振,李海滨.生物质化学链气化制取合成气模拟研究[J].煤炭转化,2011,34(4):87-92

[5]Abad A,Adanez J,Garcia-Labiano F,et al.Mapping of the range of operational conditions for Cu-,Fe-, and Ni-based oxygen carriers in chemical-looping combustion[J].Chemical Engineering Science,2007,62(1):533-549

[6]Idziak K,Czakiert T,Krzywanski J,et al.Safety and environmental reasons for the use of Ni-,Co-,Cu-,Mn- and Fe-based oxygen carriers in CLC/CLOU applications:An overview [C]// 32nd International Conference on Efficiency,Cost,Optimization, Simulation and Environmental Impact of Energy Systems,2020

[7]Wang S,Song T,Yin S Y,et al.Syngas,tar and char behavior in chemical looping gasification of sawdust pellet in fluidized bed[J].Fuel,2020,270

[收稿日期:2020.5.20]