基于三维挤压激励模块的视频分类

李宁孝, 王国栋*, 王岩杰, 胡诗语, 王亮亮
青岛大学计算机科学技术学院, 山东 青岛 266071

摘要 针对视频分类中时序特征的融合问题,将二维卷积神经网络中的挤压激励(SE)网络与三维卷积残差网络相结合,提出了新的三维挤压激励网络结构模块,该模块比直接转化而来的三维挤压激励模块多了一个时间维度系数,时间维度系数记录了研究对象在时间轨迹上所进行的动作轨迹变化。新模块不仅可以记录某个时间点的特征,而且能够强化多个时间点的关联性。将具有时空纬度的挤压激励网络应用于人物的动作行为识别,检验了新模块的有效性。实验结果表明,新模块可加快损失收敛并有效提高视频分类精度。

关键词 图像处理; 信号处理; 视频分类; 挤压激励; 三维卷积; 残差网络; 深度学习
中图分类号 TP391.4 **文献标识码** A **doi**: 10.3788/LOP56.121004

Video Classification Based on Three-Dimensional Squeeze Excitation Module

Li Ningxiao, Wang Guodong*, Wang Yanjie, Hu Shiyu, Wang Liangliang
College of Computer Science & Technology, Qingdao University, Qingdao, Shandong 266071, *China*

Abstract To address the fusion problem of time sequence features in video classification, this paper proposes a new three-dimensional (3D) squeezing excitation (SE) network structure module that is constructed by combining the SE network in a two-dimensional convolutional neural network (CNN) with a 3D convolutional residual network. The new module adds an extra time-dimension coefficient to the coefficient set of a directly transformed 3D SE module, allowing it to record the changes in the motion trajectories of the research objects on time trajectories. The proposed module can not only record the characteristics of a specific time point, but also strengthen the relevance of multiple time points. To assess the effectiveness of the module, an SE network with a spatial and temporal latitude was used to perform character-action-behavior recognition. The experimental results indicate that the module can accelerate the loss convergence and effectively improve the accuracy of video classification.

Key words image processing; signal processing; video classification; squeeze excitation; three-dimensional convolution; residual network; deep learning
OCIS codes 100.2960; 100.3008; 100.4996; 100.6890

1 引 言

人物动作视频分类是计算机视觉中一个重要的研究领域,随着深度学习的普及,对人物动作视频分类的研究已取得了很大的进展,其中,使用三维(3D)卷积[1]解决分类任务是常用的方法,这一方法比二维(2D)卷积多了一个时间维度,因而拥有更多的参数。人物动作分类比多尺度感知行人检测[2]更侧重于时间维度上的空间轨迹变化,且两者需要学习的参数不同。当使用具有大量参数的深度卷积神经网络[3]时,大规模数据集是非常重要的。2D 卷积神经网络(CNNs),基于 ImageNet[4]大型数据集,在图像处理任务中具有很强的学习能力,这是因为在 ImageNet 上训练得很深的网络,如残差网络[5]等,有助于卷积神经网络获得通用的特征表示,使用这些特征可以有效地提高其他任务的性能。同样,当使用 3D 卷积解决视频分类任务时,也需要一个大型的数据集,如 Kinetics[6]。目前具有代表性的小

收稿日期: 2018-11-29; **修回日期**: 2018-12-20; **录用日期**: 2019-01-11
基金项目: 国家自然科学基金面上项目(61772294)、"十二五"国家科技支撑计划(2014BAG03B05)
* **E-mail**: doctorwgd@gmail.com

规模动作识别视频数据集有 UCF-101[7]、HMDB51[8]等，与图像识别数据集相比，当前可用于动作识别的数据集整体较小，其中 Kinetics 数据集是一个专注于人物动作识别的大型数据集。

2D CNNs 中有很多优秀的模型结构，例如残差连接网络、密集连接网络[9]、挤压激励[10](SE)结构等，在很多提出的新模型结构中都有可能使用到其中的残差连接或密集连接。

处理视频分类任务本质上是如何在一段具有时序特点的图片中提取深度特征，卷积神经网络在空间上的学习能力很好，但在时序上的特征关联性有待加强。为此，本文基于 Kinetics 数据集，使用由 2D 残差网络转化而来的 3D 残差神经网络(ResNet)[11]作为基准网络，引入优化的挤压激励模块来实现强化时间维度的特征学习，并用于视频分类。当前 3D CNNs 在视频分类中已有不错的表现，本文着重探讨 SE 模块是否会在时间轨迹上强化特征学习，不讨论网络深度对特征学习的影响。

2 使用优化后的 3D SE 模块提取深度特征

2.1 转化为 3D 卷积结构

3D 卷积是视频建模的一种方法，其比标准的卷积网络多一个时间纬度，能够直接创建时空数据的层次表示，因此，比 2D 卷积神经网络拥有更多的参数。模型参数在基于时空方向的人体行为识别[12]中至关重要，但也使得模型很难训练。深度 3D 卷积网络需要预训练像 Kinetics 这样的大型数据集[13]，这和 ImageNet 的预训练同等重要。

SE 模块可以很好地嵌入到 2D CNNs 中，在提升图像分类精度的同时增加的计算量很少。该模块通过学习并获取特征通道的重要程度，以抑制卷积层中不重要的特征，将这种思想应用到 3D CNNs 中能提升其图像的处理性能。经过改进的 3D CNNs，在空间维度和时间维度上均能学习特征通道的重要程度，并基于重要程度强化学习各维度上的重要特征，进而提高分类精度。

模型的单元结构如图 1 所示。将 2D 卷积核 K(高)×K(宽)转化为 3D 卷积核 K(时间维度)×K(高)×K(宽)，2D ResNet、3D ResNet 单元分别如图 1(a)、(b)所示；直接将 3D SE 模块嵌入到 3D ResNet 中，得到 3D SE_FC ResNet 单元如图 1(c)所示，为了反映 SE 模块在时间维度上对特征变化的影响，优化设计了 3D SE_FC ResNet 网络模型，在进行自适应平均池化[14]时，改变输出时间维度系数 T，把通道系数固定为 32，空间尺度不变，图 1(c)中，T(时间维度系数)×1(高)×1(宽)×4N(通道数)表示输出的尺度大小；通过实验对比它们在训练中的损失和测试精度的变化。

图 1 还给出了 3D ResNet50 与新模型卷积结构对比，两者均为由 2D 卷积转化而来的 3D 卷积结构，每个卷积层通过线性整流函数(RELU)[15]激活。其中，卷积层 conv2_x，conv3_x，conv4_x，conv5_x 的单元的个数为 x=3，4，6，3，每个单元都嵌入 3D SE_FC 模块。3D SE_FC 模块包括平均池化层、全连接层 FC1、RELU 激活、全连接层 FC2、归一化处理 Sigmoid，其中 FC1、FC2 的时间维度系数 T 分别为 1、4、8，它们图片的输入大小、卷积层 conv1 和最后的全连接层 FC 层是一致的。当新模型的系数 T 增大时，相应的全连接层参数也会成系数倍增加，如果 FC1、FC2 时间输入维度为 N，那么当 T 分别为 1，4，8 时，时间输出维度变为 N，4N，8N，这样可以拟合更多时间维度的特征。conv2_x 之前设置了卷积核为 1×3×3 的最大池化层，步长 stride 为 2；第 1 个卷积层 conv1 层在空间上用空间步长 stride 为 2 的下采样输入，最后的全连接层中 classes 为视频类别的数量。

2.2 视频分类特征提取流程

3D SE_FC 模块处理过程：1)进行 1×1×1，3×3×3，1×1×1 卷积；2)通过 3D 自适应平均池化输出 T×1×1 操作将空间维度和时间维度压缩，使时间维度特征融合进空间维度特征，通道数不变，这样不仅具有全局感受，而且具有时间上的联系，从而能够对每个通道的时间维度做出响应；3)融合后，通过 2 个 FC 全连接层更好地拟合空间上和时间上的相关性；4)通过 Sigmoid[16]获得 0～1 之间的归一化权重，最后将这个权重加权到卷积层中，从而加强重要特征的学习，抑制非重要特征的学习。

实验中，输入是 3(通道数)×16(时间维度)×112(高)×112(宽)。整个特征提取过程如图 2 所示。1) 在空间上进行步长为 2 的下采样作为输入，其中时间步长为 1，卷积核为 7×7×7；2)进行 1×3×3 的最大池化，为保留更多的时间维度信息，将卷积核中的时间维度设置为 1；3) 经过 3D SE_FC ResNet 单元的处理，得到具有固定维度的 FC 层，维度数和类别数相等；4) 进行分类预测，其中最大得分表示已识别的类标签。

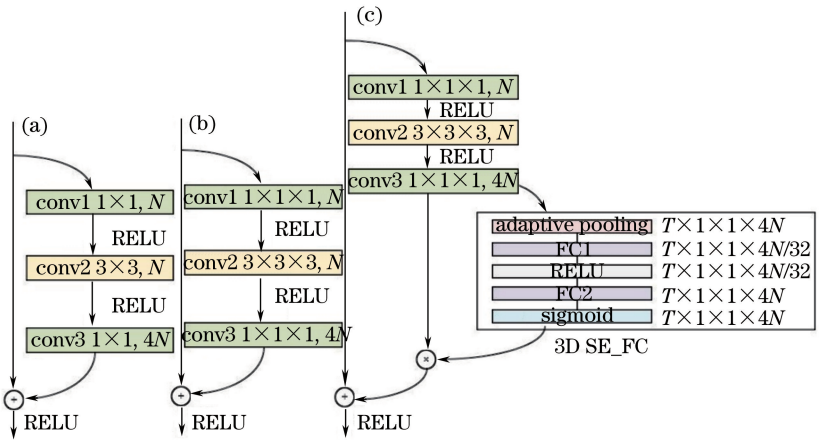

图 1 模型的单元结构。(a) 2D ResNet 单元;(b) 3D ResNet 单元;(c) 3D SE_FC ResNet 单元

Fig. 1 Unit structure of module. (a) 2D ResNet unit; (b) 3D ResNet unit; (c) 3D SE_FC ResNet unit

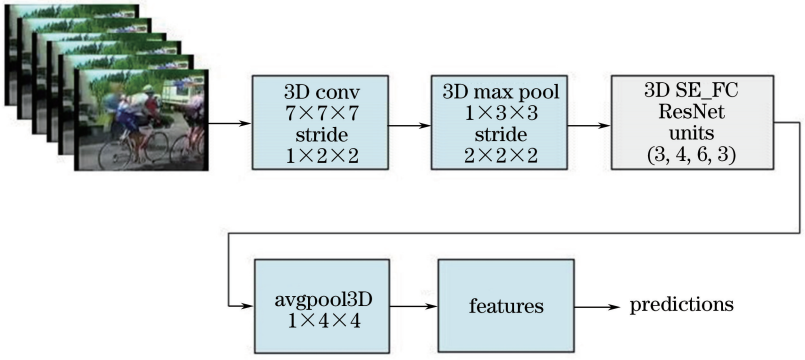

图 2 特征提取过程

Fig. 2 Flow chart of feature extraction

3 实验配置

3.1 实验设计

实验以 3D ResNet50 作为基准网络结构,分别从新模块嵌入之前、嵌入之后、嵌入后时间维度系数 T 的变化以及不同视频数据集的使用 4 个方面进行分析。实验用到 3D ResNet50、3D SE_FC ResNet50(T=1)、3D SE_FC ResNet50(T=4)、3D SE_FC ResNet50(T=8)4 个网络模型,通过对比在不同数据集上训练过程中的衰减趋势和平均精度,衡量新模块在不同时间维度系数下的增益效果。实验 1 是在没有使用 Kinetics 数据集进行预训练的情况下训练模型,实验 2 使用 3D ResNet50 在 Kinetics 数据集上进行预训练,再将其部分模型参数迁移到 3D SE_FC ResNet50 中进行微调。

实验 1 中,分别使用 3D ResNet50 和嵌入 SE_FC 模块的 3D ResNet50 在 UCF101、HMDB51 数据集上进行训练,保存每个阶段的最优模型,并在相同实验条件下对比各阶段结果。在训练 UCF101 数据集时,使用 kaiming 初始化[17]参数,对比前 100 个 epoch 的衰减下降趋势和训练精度,再基于此模型,改变学习率为 0.001 并训练 70 个 epoch,检查衰减收敛情况和训练精度变化,本训练中使用随机梯度下降法[18](SGD)优化参数函数。在 HMDB51 数据集进行训练时,使用在 UCF101 数据集上训练获得的模型作为预训练模型,本次训练使用自适应估计[19](Adam)梯度优化函数,初始学习率为 0.001 时训练 300 个 epoch,并对比衰减的变化和训练精度的变化。

实验 2 中,使用 Kinetics 数据集预训练 3D ResNet,接着在 UCF101 和 HMDB51 上进行微调,只在 conv5_x 和最后的 FC 层进行该步骤,SE_FC 模块只嵌入 conv5_x 中。最后,将获得的结果与最新的视频分类方法进行对比。

3.2 视频数据集和数据预处理

UCF-101 和 HMDB51 以及 Kinetics 视频数据集是被业界公认的、在动作识别领域比较成功的数据集,现在仍然被用来作为基准。因此,选用这些数

据集验证挤压激励模块的有效性。

UCF-101 数据集是人类行为识别数据集中的典型代表，一个视频只包含一类人类行为，数据集包括 13320 个动作实例，包含运动、瑜伽、乐器等 101 个人类行为类别，共 27 h。其中非动作帧被删除，平均持续时间大约 7 s，数据集提供了 3 个训练、测试集的分布集合，其中 70% 为训练集，30% 为测试集。

HMDB51 数据集包括 51 个人类动作，总共有 6849 个视频剪辑，每个类别至少 101 个剪辑，动作分为 5 种类型：面部动作、面部动作与对象、身体运动、身体运动与对象、人体互动。每个视频的平均长度为 3 s，数据集提供了 3 个训练、测试的分布集合，其中 70%为训练集，30%为测试集。

Kinetics 数据集关注于人类行为，而不是活动或事件。动作类包括：例如画画、喝酒、大笑、拳击等单个人物动作；例如拥抱、亲吻、握手等多人物动作；例如打开礼物、修剪草坪、洗碗等人物对象动作。有些动作是细粒度的，需要时间推理进行区分，例如不同类型的游泳。有些动作需要通过强调物体依赖性加以区分，例如演奏不同类型的管乐器。数据集有 400 个人类动作类，每个类有 400 个或更多的剪辑，每个剪辑来自一个独特的视频，共有 24 万个训练视频。剪辑持续 10 s 左右，没有未剪辑的视频。测试集由每个类的 100 个剪辑组成。

数据预处理过程中，将视频的大小调整为 320 pixel×240 pixel。所有的视频根据帧数切割成相应的图片并保存，这些图片用于训练和测试。训练时选用 16 张图片进行 3D 卷积。

3.3 模型训练

为了在不同的参数更新方法下对比衰减的收敛趋势，实验中使用了 SGD 法和 Adam，其中 Adam 梯度更新速度快于 SGD 法。训练中参数包括动量衰减 0.001 和 0.9，SGD 的初始学习率为 0.01，衰减权重为 1×10^{-5}，Adam 的初始学习率为 0.001，计算梯度和平方梯度的系数为 0.9 和 0.999，eps 为 1×10^{-8}，权重衰减为 0。

实验 1 中把在 UCF101 上训练得到的模型用于训练 HMDB51 数据，因为其数据规模太小，所以很容易过拟合。UCF101 数据集是从头开始训练。视频样本使用均匀采样的方式，在同一个时间段获取的样本是等量的，即在时空上等量滑动，生成不重叠的片段，然后分别进行 3D 卷积，获得每个片段类的得分，最后获得最大值的类即为样本所标的标签。

实验 2 中把 3D ResNet50 在 Kinetics 数据集上预训练获得的模型作为预训练模型，使用 SGD 法分别在 UCF101 和 HMDB51 数据集上进行微调。

4 实验结果和讨论

UCF101 第 1、2 阶段训练损失以及 HMDB51 训练损失如图 3～5 所示。由图可见，在模型训练的过程中，由 2D 转化而来的新模块可以提升模型的训练效率，通过调整时间维度 $T(1,4)$ 可以在一定程度上增强模型的特征学习能力。训练结束后选择验证精度最高的模型进行对比，数据集 UCF101 和 HMDB51 的平均验证精度对比如表 1～2 所示。当新模块的时间维度系数 $T=4,8$ 时，精度均超过了基准网络 3D ResNet50 的精度，且当 $T=4$ 时精度最高，由此说明调整时间维度系数可以提高视频分类精度。为了进一步说明新模块的优势，将微调后得到的模型与其他方法对比，如表 2 所示，在 UCF101 数据集上新模块嵌入 50 层的 3D ResNet 得到的精度超过了嵌入 200 层的 3D ResNet，同时效果也好于表中其他方法。平均每个视频分类时间对比如表 3 所示，由表可见，在实际视频分类应用中，新模块在提高分类精度的同时对运行效率的影响很小。后期研究将考虑使用双流卷积神经网络[20]，把时间和空间分开并进行联合检测。

UCF101 第 1 阶段训练损失如图 3 所示，给出了在数据集 UCF101(split3)上从头开始训练的前 100 个 epoch 的衰减下降趋势图，学习率为 0.01，采用 SGD 梯度下降法。由图 3 可以看到，3D ResNet50 嵌入 3D SE_FC($T=1,4$)模块后，前期下降趋势明显快于 3D ResNet，虽然最后都收敛于同一饱和状态，但是 3D SE 模块可以更快地进入饱和状态。

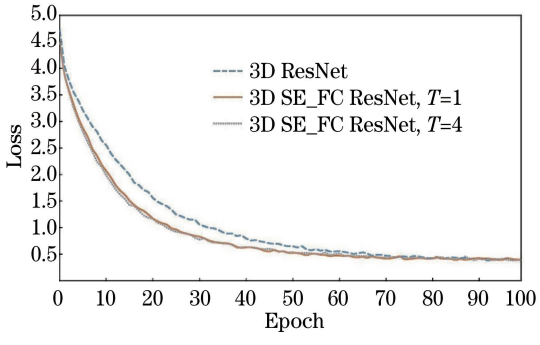

图 3 UCF101 第 1 阶段训练损失

Fig. 3 Loss at first stage of UCF101 training

UCF101 第 2 阶段训练损失如图 4 所示。这是

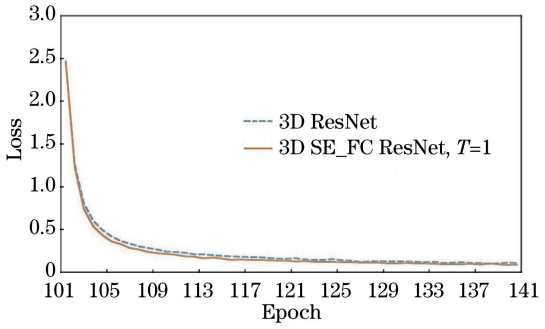

图 4 UCF101 第 2 阶段训练损失

Fig. 4 Loss at second stage of UCF101 training

在数据集 UCF101(split3)上训练，加载前 100 个 epoch 作为预训练模型继续训练 50 个 epoch，学习率为 0.001，依然采用 SGD 法，由图可见，嵌入 3D SE_FC(T=1)模块的 3D ResNet 的衰减略低于 3D ResNet。

HMDB51 训练损失如图 5 所示，这是在数据集 HMDB51(split3)上训练 300 个 epoch。使用在 UCF101 上所得到的训练模型作为预训练模型，学习率为 0.001，计算梯度和平方梯度的系数为 0.9 和 0.999，eps 为 1×10^{-8}，权重衰减为 0。梯度下降法使用了 Adam 方法，这样学习率可以实现动态自适应，3D SE_FC 模块中的时间维度 T=1，4。对比衰减变化趋势，3D SE_FC ResNet 稍快于 3D ResNet，并且整体上衰减略低于 3D ResNet。

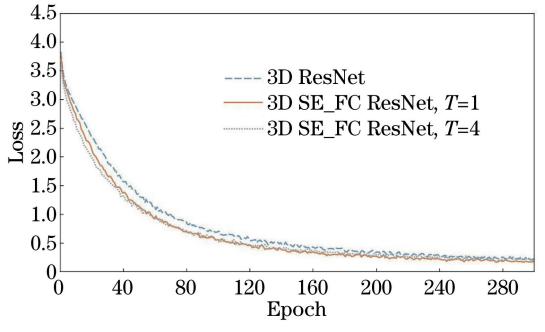

图 5 HMDB51 训练损失

Fig. 5 Loss of HMDB51 training

数据集 UCF101 平均验证精度对比如表 1 所示。数据集 HMDB51 平均验证精度对比如表 2 所示。表中 3D ResNet50、3D ResNet101 为 50 层和 101 层的原始 3D 卷积神经网络，3D SE_FC ResNet50 为新提出的 3D 挤压激励卷积神经网络，时间维度系数 T 分别为 1，4，8。通过调整学习率继续在 UCF101(split3)上进行训练，将学习率调整为0.0001，继续训练，其衰减已经处于饱和状态，对比它们预测概率最大(Top1)和预测概率最大前 5 名(Top5)的验证平均精度，并对比数据集 HMDB51 前 300 个 epoch 模型的验证平均精度。由于没有使用 Kinetics 预训练模型，所得到的结果精度低于使用 Kinetics 预训练模型所得到的模型，具体请查看文献[13]，其中 T 为 3D SE_FC 模块中的时间维度。

表 1 数据集 UCF101 的平均验证精度对比

Table 1 Average validation accuracy comparison on UCF101 dataset

Method	Pretraining dataset	Average validation accuracy /%	
		Top1	Top5
3D ResNet50	No	43.8	67.3
3D ResNet101		43.4	68.8
3D SE_FC ResNet50, T=1		42.5	66.1
3D SE_FC ResNet50, T=4		44.9	69.6
3D SE_FC ResNet50, T=8		45.5	68.2

表 2 数据集 HMDB51 的平均验证精度对比

Table 2 Average validation accuracy comparison on HMDB51 dataset

Method	Pretraining dataset	Average validation accuracy /%	
		Top1	Top5
3D ResNet50	UCF101	15.6	43.0
3D ResNet101		15.2	41.0
3D SE_FC ResNet50, T=1		16.1	43.6
3D SE_FC ResNet50, T=4		18.7	46.2
3D SE_FC ResNet50, T=8		17.7	44.6

平均每个视频分类时间对比如表 3 所示，时间包括整个视频预处理时间。由表 3 可见，嵌入新模块后所增加的时间很少，因此，新模块对视频分类效率的影响很小。

表 3 平均每个视频分类时间对比

Table 3 Average time comparison for each video classification

Method	Time /ms
3D ResNet50	104
3D ResNet101	111
3D SE_FC ResNet50, T=1	112
3D SE_FC ResNet50, T=4	110
3D SE_FC ResNet50, T=8	114

Kinetics 数据集预训练后，数据集 HMDB51 和 UCF101 的测试精度对比如表 4 所示。3D ResNet18，3D ResNet34，3D ResNet50，3D ResNet101 和 3D ResNet200 分别为 18，34，50，101，200 层的原始 3D 卷积神经网络，3D DenseNet-121 为 121 层的 3D 密集链接网络。其中 SE_FC 只

嵌入到 conv5_x 中，其他层保持不变，训练只微调 conv5_x 和 FC 层。将使用 Kinetics 预训练的模型所得到 Top1 平均测试精度与其他 3D 模型以及其他最新方法对比，结果表明，50 层的 SE_FC ResNet 在 UCF101 数据集上可以达到 200 层的 3D ResNet 所得到的精度，并且超过了 50 层的 3D ResNet 所得到的精度。

表 4 数据集 HMDB51 和 UCF101 的测试精度对比

Table 4 Test accuracy comparison on HMDB51 and UCF101 datasets

Method	Pretraining dataset	Test accuracy /%	
		HMDB51	UCF101
3D ResNet18	Kinetics	56.4	84.4
3D ResNet34		59.1	87.7
3D ResNet50		61.0	89.3
3D ResNet101		61.7	88.9
3D ResNet200		63.5	89.6
3D DenseNet-121		59.6	87.6
Method in Ref. [21]	—	59.4	88.0
Method in Ref.[22]	—	—	88.6
Method in Ref. [23]	—	—	85.9
Method in Ref. [24]	—	—	88.6
Method in Ref.[25]	sports 1M	—	82.3
3D SE_FC ResNet50, T=1(ours)	Kinetics	61.3	89.0
3D SE_FC ResNet50, T=4(ours)		59.6	90.1
3D SE_FC ResNet50, T=8(ours)		59.0	89.5

5 结束语

嵌入 3D SE_FC 模块可以有效提高衰减收敛的速度，同时分类精度也有一定提升，因此，3D SE_FC 模块通过强化时间维度的学习可以提高训练的效率和精度，同时能在空间维度增强学习。这表明 3D SE_FC 模块对时间维度特征具有一定增益。但是，模型的复杂度也会增加，对于部分数据集，单纯地将其嵌入并不会提高精度，而需要更多的数据拟合模型参数。

目前，2D CNNs 在图像处理的各项任务中都有很大的进展，证明了卷积神经网络在空间上具有较强的学习能力。但是，当使用 3D CNNs 解决视频分类任务时，如何更好地处理空间和时间上的特征，并获得更好的效果仍是需要重点考虑的。实际情况中，3D CNNs 的表现不佳，相比于光流法以及其他双流相结合的方法并没有优势。因此，在时间上的深度特征学习很重要，很多方法的思路是将空间和时间分开，并结合 3D CNNs 分别提取空间特征和时间特征，取得了较好的效果。无论是全部分开还是通过不同的卷积核的形式，都是为了更好地整合空间维度特征和时间维度特征。由于 3D CNNs 在空间维度卷积网络已经有了很强的学习能力，因此在原有的 3D CNNs 基础上嵌入优化的 3D 卷积挤压激励模块，就是为了增强时间维度的学习。为了更好地整合时间维度特征，将此结构作为时间维度的放大器(并非完全的放大，其中也包括空间因素)，如何进一步发挥此结构的优势是下一步研究的重点。

当前的卷积网络可以很好地学习空间维度的特征，而如何更好地学习时间维度特征仍是一个关键的问题。此外，3D 卷积的学习参数远大于 2D 卷积，如果数据量太小很容易造成过拟合现象，因此，需要综合考虑这些问题，设计出用于视频分类的更好的 3D 卷积模型。

参 考 文 献

[1] Wang K Z, Wang X L, Lin L, *et al*. 3D human activity recognition with reconfigurable convolutional neural networks[C]// Proceedings of the 22nd ACM International Conference on Multimedia, November 3-7, 2017, Orlando, Florida, USA. New York: ACM Press, 2014: 97-106.

[2] Liu H, Peng L, Wen J W. Multi-scale aware pedestrian detection algorithm based on improved full convolutional network[J]. Laser & Optoelectronics Progress, 2018, 55(9): 091504.
刘辉, 彭力, 闻继伟. 基于改进全卷积网络的多尺度感知行人检测算法[J]. 激光与光电子学进展, 2018, 55(9): 091504.

[3] 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, June 23-28, 2014, Columbus, OH, USA. New York: IEEE, 2014: 1725-1732.

[4] Deng J, Dong W, Socher R, *et al*. ImageNet: a large-scale hierarchical image database[C]// 2009 IEEE Conference on Computer Vision and Pattern Recognition, June 20-25, 2009, Miami, FL, USA. New York: IEEE, 2009: 248-255.

[5] 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), June 27-30, 2016, Las Vegas, NV, USA. New York: IEEE, 2016: 770-778.

[6] Kay W, Carreira J, Simonyan K, *et al*. The kinetics human action video dataset[EB/OL]. (2017-05-19)[2018-11-15]. https://arxiv.org/abs/1705.06950.

[7] Soomro K, Zamir A R, Shah M. UCF101: a dataset of 101 human actions classes from videos in the wild [EB/OL]. (2012-12-03)[2018-11-15]. https://arxiv.org/abs/1212.0402.

[8] Kuehne H, Jhuang H, Stiefelhagen R, *et al*. HMDB51: a large video database for human motion recognition[M]// Nagel W, Kröner D, Resch M. High Performance Computing in Science and Engineering '12. Berlin, Heidelberg: Springer, 2012: 571-582.

[9] Huang G, Liu Z, Maaten L V D, *et al*. Densely connected convolutional networks[C]// 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), July 21-26, 2017, Honolulu, HI, USA. New York: IEEE, 2017: 2261-2269.

[10] Hu J, Shen L, Sun G. Squeeze-and-excitation networks[C]// 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, June 18-23, 2018, Salt Lake City, UT, USA. New York: IEEE, 2018: 7132-7141.

[11] Hara K, Kataoka H, Satoh Y. Learning spatio-temporal features with 3D residual networks for action recognition[C]// 2017 IEEE International Conference on Computer Vision Workshops (ICCVW), October 22-29, 2017, Venice, Italy. New York: IEEE, 2017: 3154-3160.

[12] Xu H Y, Kong J, Jiang M, *et al*. Action recognition based on histogram of spatio-temporal oriented principal components[J]. Laser & Optoelectronics Progress, 2018, 55(6): 061009.
徐海洋, 孔军, 蒋敏, 等. 基于时空方向主成分直方图的人体行为识别[J]. 激光与光电子学进展, 2018, 55(6): 061009.

[13] Hara K, Kataoka H, Satoh Y. Can spatiotemporal 3D CNNs retrace the history of 2D CNNs and ImageNet?[C]// 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, June 18-23, 2018, Salt Lake City, UT, USA. New York: IEEE, 2018: 6546-6555.

[14] Liu D, Zhou Y Z, Sun X Y, *et al*. Adaptive pooling in multi-instance learning for web video annotation [C]// 2017 IEEE International Conference on Computer Vision Workshops (ICCVW), October 22-29, 2017, Venice, Italy. New York: IEEE, 2017: 318-327.

[15] Nair V, Hinton G E. Rectified linear units improve restricted Boltzmann machines [C]// 27th International Conference on Machine Learning, 2010, Haifa, Israel. Omnipress, 2010: 807-814.

[16] Hochreiter S, Schmidhuber J. Long short-term memory[J]. Neural Computation, 1997, 9(8): 1735-1780.

[17] He K M, Zhang X Y, Ren S Q, *et al*. Delving deep into rectifiers: surpassing human-level performance on ImageNet classification [C]// 2015 IEEE International Conference on Computer Vision (ICCV), December 7-13, 2015, Santiago, Chile. New York: IEEE, 2015: 1026-1034.

[18] Bordes A, Bottou L, Gallinari P. SGD-QN: careful quasi-Newton stochastic gradient descent[J]. Journal of Machine Learning Research, 2009, 10(Jul):1737-1754.

[19] Kingma D P, Ba J. Adam: a method for stochastic optimization[EB/OL]. (2017-01-30)[2018-11-15]. https://arxiv.org/abs/1412.6980.

[20] Liu F, Liu P Y, Zhang J N, *et al*. Joint detection of RGB-D images based on double flow convolutional neural network [J]. Laser & Optoelectronics Progress, 2018, 55(2): 021503.
刘帆, 刘鹏远, 张峻宁, 等. 基于双流卷积神经网络的RGB-D图像联合检测[J]. 激光与光电子学进展, 2018, 55(2): 021503.

[21] Szegedy C, Liu W, Jia Y Q, *et al*. Going deeper with convolutions[C]// 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 7-12, 2015, Boston, MA, USA. New York: IEEE, 2015: 7298594.

[22] 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), June 7-12, 2015, Boston, MA, USA. New York: IEEE, 2015: 4694-4702.

[23] Wang H, Schmid C. Action recognition with improved trajectories[C]// 2013 IEEE International Conference on Computer Vision, December 1-8, 2013, Sydney, NSW, Australia. New York: IEEE, 2013: 3551-3558.

[24] Qiu Z F, Yao T, Mei T. Learning spatio-temporal representation with pseudo-3D residual networks[C]

// 2017 IEEE International Conference on Computer Vision (ICCV), October 22-29, 2017, Venice, Italy. New York: IEEE, 2017: 5534-5542.

[25] Tran D, Bourdev L, Fergus R, *et al*. Learning spatiotemporal features with 3D convolutional networks[C] // 2015 IEEE International Conference on Computer Vision (ICCV), December 7-13, 2015, Santiago, Chile. New York: IEEE, 2015: 4489-4497.