Learning Spatio-Temporal Representation with Pseudo-3D Residual Networks

Zhaofan Qiu y, Ting Yao z, and Tao Mei z University of Science and Technology of China, Hefei, China Microsoft Research, Beijing, China

摘要

一些研究表明执行3D卷积来捕捉视频的空间和时间维度的信息是可行的，然而，从零开始开发非常深的3D CNN会导致昂贵的计算成本和内存需求。故本文作者提出把现有的2D网络应用到3D卷积神经网络中，本文在深度残差网络中设计了一个bottleneck结构的变体，通过在空间域上执行1×3×3卷积加3×1×1卷积来模拟3×3×3卷积，来在时间上相邻的特征图上构造时间联系。此外作者提出名叫Pseudo-3D Residual Net (P3D ResNet)的新架构，这个新架构遵循使用增加网络深度来增强网络结构多样性可以提升神经网络的能力的原则,探索不同变体的block在resnet中不同的放置位置对性能的影响。进一步在五个不同的benchmarks和三个不同的任务上检测了P3D ResNet的视频表现的泛化能力。

文章背景

近些年，卷积神经网络的崛起令人信服地证明了高度的学习视觉表达能力，特别是在图像领域。例如ResNet在ImageNet上得到了3.57%的错误率，比人类的成绩5.1%更低。

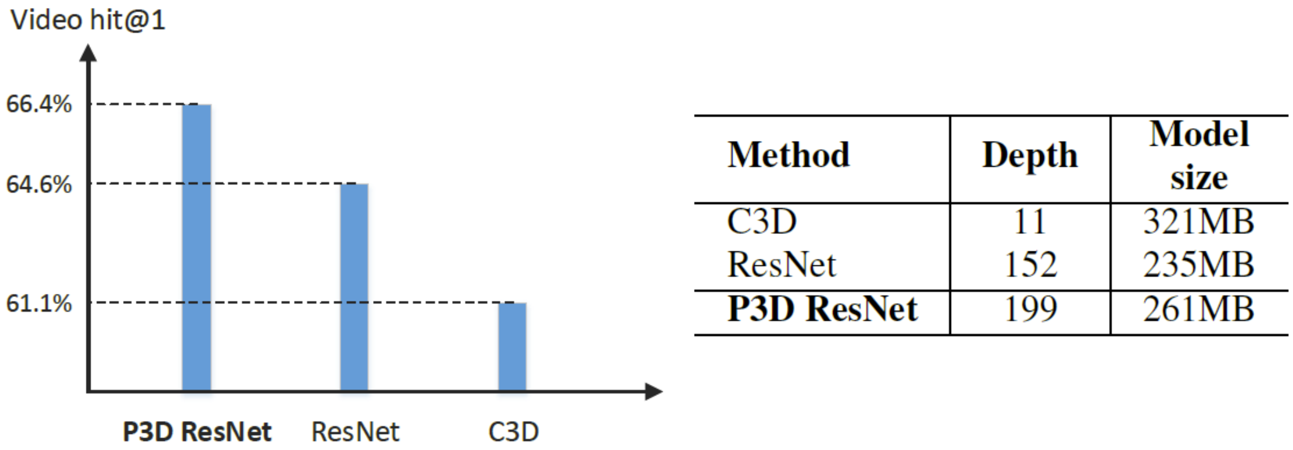
然而，视频是时间序列的帧，这些帧有着很大的变化性和复杂性，导致难以学习一个强大且通用的时空表达。

3D conv net

当前一个很自然方式来编码视频中的时空信息的方式是继承2D卷积神经网络，将其变成一个全新的3D 卷积神经网络。这样，网络不仅可以得到每个视频帧中存在的视觉外观表达，而且还可以获得连续帧中的时间变化。但是3D神经网络计算量巨大，而且模型尺寸也非常的大。

模型名	模型深度	模型大小
C3D network	11-layer 3D CNN	321MB
ResNet-152	152-layer 2D ResNet	235MB

对比可知，现有3D CNN架构的容量非常有限，计算成本和内存需求很高，因此很难训练出非常深的3D CNN，训练一个很深的3D CNN是很困难的。并且直接对ResNet-152使用sport-1M进行fine-tuning得到的准确度比在视频上直接训练的准确度更高。如下图所示



Method	Depth	Model size
C3D	11	321MB
ResNet	152	235MB
P3D ResNet	199	261MB

Figure 1. Comparisons of different models on Sports-1M dataset in terms of accuracy, model size and the number of layers.

Recurrent Neural Networks

在每一帧的表达上使用池化策略或者RNN，通常是最后一层池化层的激活，或者2D CNN中的全连接。 然而，这类方法仅在顶层的高层特征上建立时间连接，而在低层形式（例如，底层的边缘）中的相关性未被充分利用。

本文通过发明一系列的bottleneck 结构的block来同时利用时间和空间信息，每个block中关键组件是是将一个1×3×3与一个3×1×1卷积核以串联或者并联的方式连接，以此来替代传统的3×3×3的卷积神经网络。

以此模型的尺寸能够得到显著的下降，（每一个block从3×3×3=27降到1×3×3+3×1×1=12），P3D ResNet中的时间连接是从下到上的每个级别构建的，并且学习的视频表示形式将与视频中的对象，场景和动作相关的信息封装起来，使其成为各种视频分析任务的通用对象。

P3D Blocks and P3D ResNet

3D convolutions

给定一段视频，尺寸为c×l×h×w，其中c是channels,视频片段的长度，h为高度，w为宽度。3D卷积像2D filters对空间信息进行建模，并且构造跨帧的时间连接。这里把3D convolution filter记为d×k×k，d是时间深度，k是卷积核尺寸。因此，假如我们有3D卷积核尺寸为3×3×3，可以被分解为1×3×3等于在空间上进行2D CNN，3×1×1为处理时空序列的1D CNN。这就是本文被叫做伪3D CNN的原因。用这种方法不仅降低了参数的数量，而且可以使用使用图片预训练的2D CNN 模型，赋予Pseudo 3D CNN更多利用从图像学习的场景和物体知识的能力。

pseudo-3D blocks

Residual Units

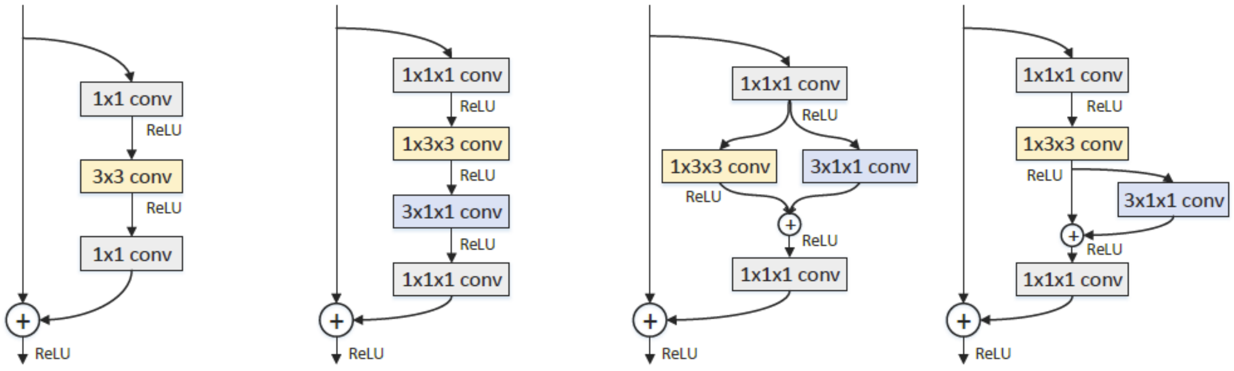
$$(\mathbf{I}+\mathbf{F})\cdot\mathbf{x}_t=\mathbf{x}_t+\mathbf{F}\cdot\mathbf{x}_t:=\mathbf{x}_t+\mathbf{F}\left(\mathbf{x}_t\right)=\mathbf{x}_{t+1},$$

ResNet的主要思想是参考输入Xt,学习添加的残差函数F,这里的Xt输入是通过shortcut connetction实现的，以此代替直接学习无参考的非线性函数。

P3D Blocks design

为了将2D残差模块更改为3D结构，以用来编码视频的时空信息。在改造的过程中，并不直接涉及两个问题，首先是：这里空间维度的2D就卷积核与1D时间序列的卷积核是否需要直接影响对方，两个维度直接影响意味着空间维度空间2D空间输出作为时间序列1D卷积核的输入，例如，用级联模式。两种卷积核之间不直接的连接，解耦了联系，使得两种卷积核在网络中走不通的路径。第二个问题是，两种卷经济是否需要直接影响最终的输出。同样的，在此上下文中的直接影响表示每种类型的滤波器的输出应该直接连接到最终输出。

在这两种设计议题上，文章作者提出三种不同的P3D blocks，分别叫做P3D-A，P3D-B,P3D-C。



(a) Residual Unit [7]　(b) P3D-A　(c) P3D-B　(d) P3D-C

Figure 3. Bottleneck building blocks of Residual Unit and our Pseudo-3D.

P3D-A

第一种结构把时序1D卷积核跟在空间2D卷积核后边，使用级联的方式，因此这两种卷积核，可以在同一路径上直接影响对方，并且只有时间1D卷积核被直接连接到最终的输出，公式如下：

$$(\mathbf{I}+\mathbf{T}\cdot\mathbf{S})\cdot\mathbf{x}_t := \mathbf{x}_t+\mathbf{T}\left(\mathbf{S}\left(\mathbf{x}_t\right)\right)=\mathbf{x}_{t+1}.$$

P3D-B

第二种结构与第一种结构相似，除了两种卷积核之间互不直接影响，尽管两个S和T之间没有直接影响，两者被直接累积作为最后的输出，公式如下：

$$(\mathbf{I}+\mathbf{S}+\mathbf{T})\cdot\mathbf{x}_t := \mathbf{x}_t+\mathbf{S}\left(\mathbf{x}_t\right)+\mathbf{T}\left(\mathbf{x}_t\right)=\mathbf{x}_{t+1}.$$

P3D-C

第三种结构把P3D-A与P3D-B结合起来了，通过直接构建S,T和输出之间的之间影响。特别的，为了使得基于P3D-A的S和最终输出之间的直接联系，作者在S和最终输出之间建立了一个shortcut connection，公式如下

$$(\mathbf{I}+\mathbf{S}+\mathbf{T}\cdot\mathbf{S})\cdot\mathbf{x}_t := \mathbf{x}_t+\mathbf{S}\left(\mathbf{x}_t\right)+\mathbf{T}\left(\mathbf{S}\left(\mathbf{x}_t\right)\right)=\mathbf{x}_{t+1}.$$

P3D-D

出于结构多样性的考虑，混合了三种结构。

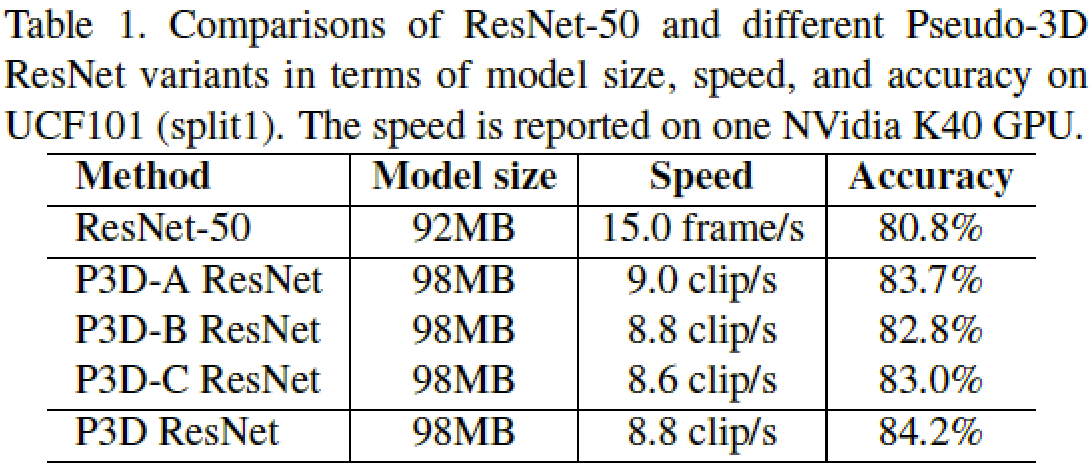


Table 1. Comparisons of ResNet-50 and different Pseudo-3D ResNet variants in terms of model size, speed, and accuracy on UCF101 (split1). The speed is reported on one NVidia K40 GPU.

Method	Model size	Speed	Accuracy
ResNet-50	92MB	15.0 frame/s	80.8%
P3D-A ResNet	98MB	9.0 clip/s	83.7%
P3D-B ResNet	98MB	8.8 clip/s	82.8%
P3D-C ResNet	98MB	8.6 clip/s	83.0%
P3D ResNet	98MB	8.8 clip/s	84.2%

Bottleneck architectures

为了降低计算复杂度，本文使用Bottleneck architectures来降低计算复杂度，将传统的3×3变为3层，第一层为1×1，第二层为3×3，第三层为1×1，第一层和第三层的1×1卷积被用来降低和恢复输入样例的维度，这样的Bottleneck architectures使得中间的3×3卷积有着更小的输入和输出维度，在本文中，作者将两个1×1×1的卷积放在路径的开始和结束，详细见图3，这使得时空序列的输入和输出的维度都得以降低。

Pseudo3D ResNet

Comparisons between P3D ResNet variants

table1展示的在UCF101上ResNet-50 和Pseudo-3D ResNet 变体的性能和时间效率，所有的P3D ResNet变体都有着比ResNet-50更好的性能，而且模型的size并没有增加很多。同时P3D ResNet变体的速度达到8.6-9.0每秒。

Mixing different P3D Blocks

为了增强结构多样性， 本文提出如图4的结构，将各种不同的block混合起来。本文将残差块换为本文设计的三种block的链，以下图的顺序。实验结果表明，这种结构分别只用A,B,C提高了0.5%，1.4%，1.2%，这表明了增加结构多样性，然后把网络构建地更深，能够提升神经网络的结果。

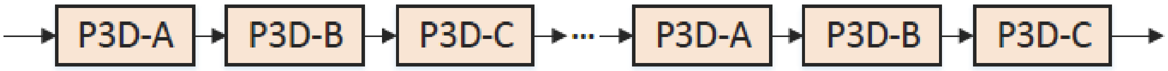


Figure 4. P3D ResNet by interleaving P3D-A, P3D-B and P3D-C.

Spatio-Temporal Representation Learning

P3D ResNet的是在Sports-1M上训练的，这个数据集包含1.13million视频，487个运动标记，每个label包含1k-3k的视频，超过5%的视频包含多于一个label。本文使用70%训练，10%验证，20%测试。

在Sports-1M上对比的准确度如下图，这个结果基本上展示了通过把3D卷积核分解为2D空间卷积核和1D时间卷积的优点。

Table 2. Comparisons in terms of pre-train data, clip length, Top-1 clip-level accuracy and Top-1&5 video-level accuracy on Sports-1M.

Method	**Pre-train Data**	**Clip Length**	**Clip hit@1**	**Video hit@1**	**Video hit@5**
Deep Video (Single Frame) [10]	ImageNet1K	1	41.1%	59.3%	77.7%
Deep Video (Slow Fusion) [10]	ImageNet1K	10	41.9%	60.9%	80.2%
Convolutional Pooling [37]	ImageNet1K	120	70.8%	72.3%	90.8%
C3D [31]	–	16	44.9%	60.0%	84.4%
C3D [31]	I380K	16	46.1%	61.1%	85.2%
ResNet-152 [7]	ImageNet1K	1	46.5%	64.6%	86.4%
P3D ResNet (ours)	ImageNet1K	16	47.9%	66.4%	87.4%

Table 3. Performance comparisons with the state-of-the-art methods on UCF101 (3 splits). TSN: Temporal Segment Networks [36]; TDD: Trajectory-pooled Deep-convolutional Descriptor [35]; IDT: Improved Dense Trajectory [34]. We group the approaches into three categories, i.e., end-to-end CNN architecture which are fine-tuned on UCF101 at the top, CNN-based video representation extractors with linear SVM classifier in the middle an approaches fused with IDT at the bottom. For the methods in th first direction, we report the performance of only taking frame and frames plus optical flow (in brackets) as inputs, respectively.

Method	**Accuracy**
End-to-end CNN architecture with fine-tuning	
Two-stream ConvNet [25]	73.0% (88.0%)
Factorized ST-ConvNet [29]	71.3% (88.1%)
Two-stream + LSTM [37]	82.6% (88.6%)
Two-stream fusion [6]	82.6% (92.5%)
Long-term temporal ConvNet [33]	82.4% (91.7%)
Key-volume mining CNN [39]	84.5% (93.1%)
ST-ResNet [4]	82.2% (93.4%)
TSN [36]	85.7% (94.0%)
CNN-based representation extractor + linear SVM	
C3D [31]	82.3%
ResNet-152	83.5%
P3D ResNet	**88.6%**
Method fusion with IDT	
IDT [34]	85.9%
C3D + IDT [31]	90.4%
TDD + IDT [35]	91.5%
ResNet-152 + IDT	92.0%
P3D ResNet + IDT	**93.7%**

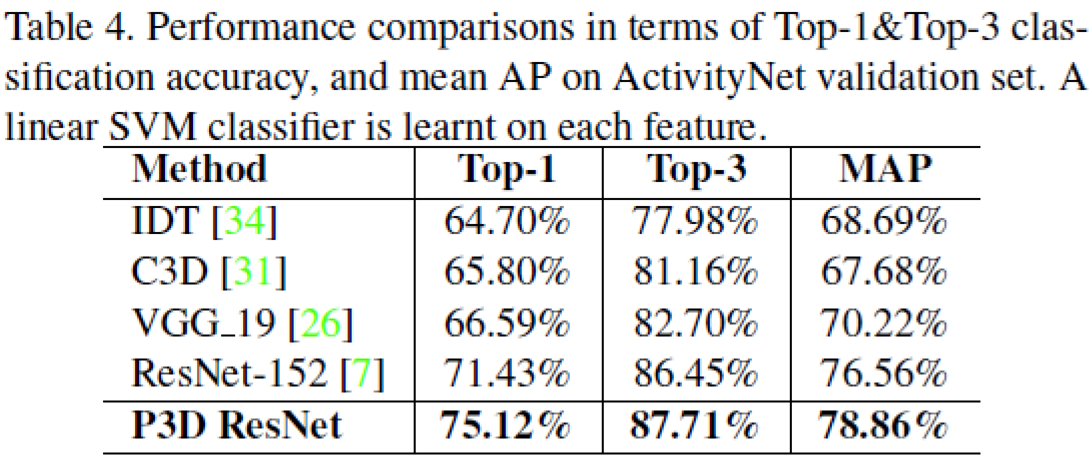


Table 4. Performance comparisons in terms of Top-1&Top-3 classification accuracy, and mean AP on ActivityNet validation set. A linear SVM classifier is learnt on each feature.

Method	Top-1	Top-3	MAP
IDT [34]	64.70%	77.98%	68.69%
C3D [31]	65.80%	81.16%	67.68%
VGG_19 [26]	66.59%	82.70%	70.22%
ResNet-152 [7]	71.43%	86.45%	76.56%
P3D ResNet	**75.12%**	**87.71%**	**78.86%**

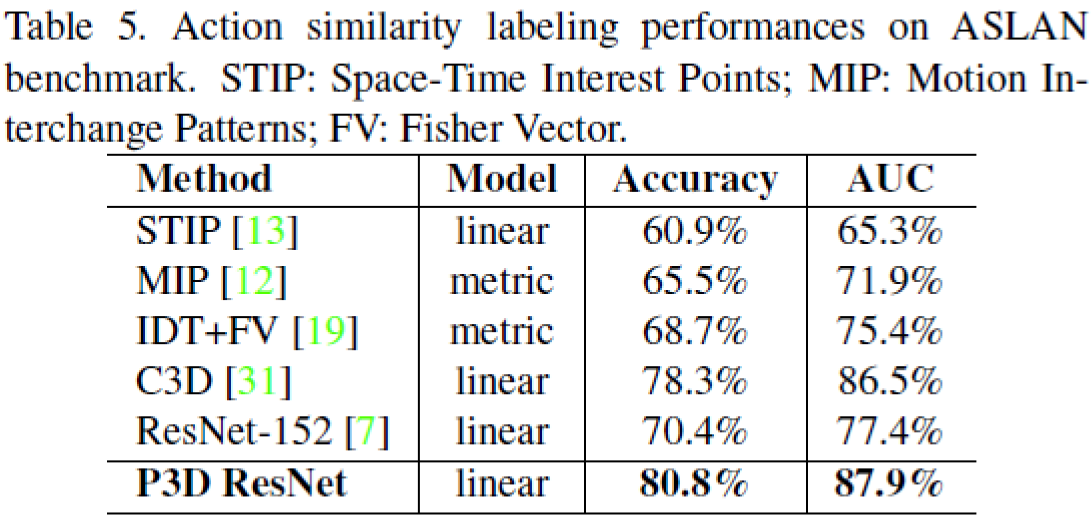


Table 5. Action similarity labeling performances on ASLAN benchmark. STIP: Space-Time Interest Points; MIP: Motion Interchange Patterns; FV: Fisher Vector.

Method	Model	Accuracy	AUC
STIP [13]	linear	60.9%	65.3%
MIP [12]	metric	65.5%	71.9%
IDT+FV [19]	metric	68.7%	75.4%
C3D [31]	linear	78.3%	86.5%
ResNet-152 [7]	linear	70.4%	77.4%
P3D ResNet	linear	**80.8%**	**87.9%**

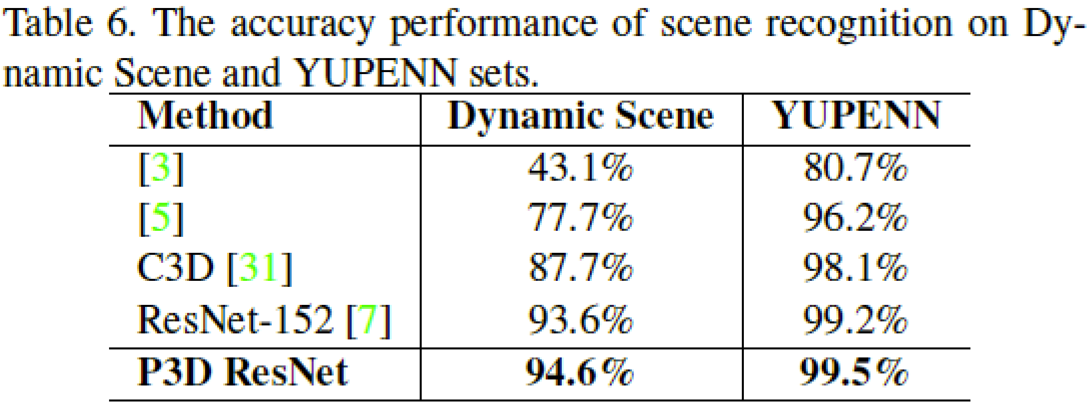


Table 6. The accuracy performance of scene recognition on Dynamic Scene and YUPENN sets.

Method	Dynamic Scene	YUPENN
[3]	43.1%	80.7%
[5]	77.7%	96.2%
C3D [31]	87.7%	98.1%
ResNet-152 [7]	93.6%	99.2%
P3D ResNet	**94.6%**	**99.5%**

Conclusion

以后的工作：

1. 注意力机制会被应用。
2. 会进一步阐述在训练时增加每个视频clip中的帧会怎样影响结果.
3. 会扩充P3D ResNet学习别的类型的输入，例如光学流，或者声音。