论文分享

Spatiotemporal Residual Networks for Video Action Recognition

> Feichtenhofer C, Pinz A, Wildes R. Spatiotemporal residual networks for video action recognition[C]//Advances in neural information processing systems. 2016: 3468-3476.

摘要

- Two-stream Convolutional Networks ： 双流卷积网络在视频动作识别展现了强大的性能
- Residual Networks： 残差网络可以训练更深的网络结构

本文中设计了时空残差网络来结合这两项技术。并且在时空范围内通过两个方式引入了残差连接。

1. 在`appearance`和`motion`两个信息流中加入残差连接，使他们在空间上有交互。
2. 通过卷积核将预训练的图片卷积网络转变为时空网络。随着模型深度的增加，这种方法缓慢地增加了时空接收域。

背景介绍

常见方法

1.双流的卷积网络

> K. Simonyan and A. Zisserman. Two-stream convolutional networks for action recognition in videos. In NIPS, 2014.

2.循环网络

> Jeff Donahue, Lisa Anne Hendricks, Sergio Guadarrama, Marcus Rohrbach, Subhashini Venugopalan, Kate Saenko, and Trevor Darrell. Long-term recurrent convolutional networks for visual recognition and description. In Proc. CVPR, 2015.

网络的设计技巧

1. **避免信息瓶颈**，随着网络深度的增加，通道数逐渐增加，特征图输入输出的维度缓慢的减少。
2. 网络最后的**接受区域应当尽可能的大**，接受更多的输入信息。
3. 在卷积核(3*3)之前进行**降维操作**（1*1）。邻近卷积核输出的信息相关性大
4. **时空的不对称分解**可以有效的降低计算量和问题的难度。
5. 每一个通道上特征的**正则化**表示
6. 利用**残差连接**促进层数多的深度网络的训练

本方法的结构

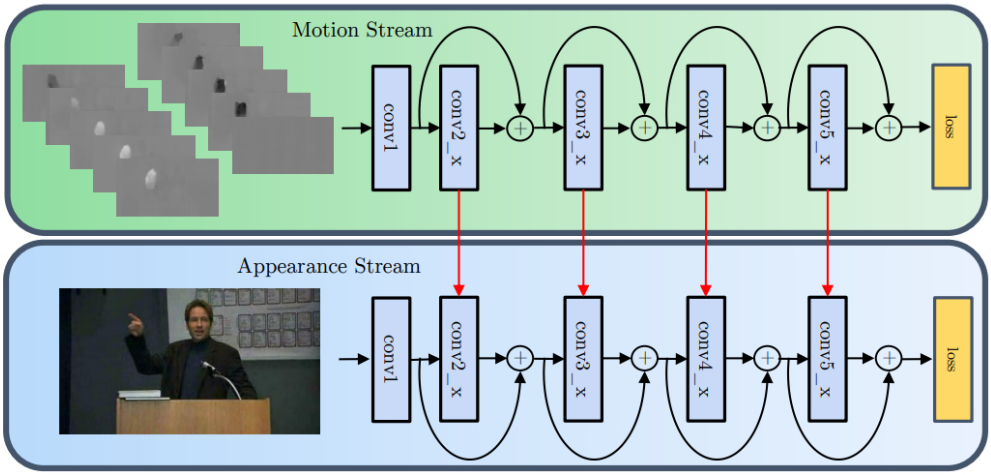
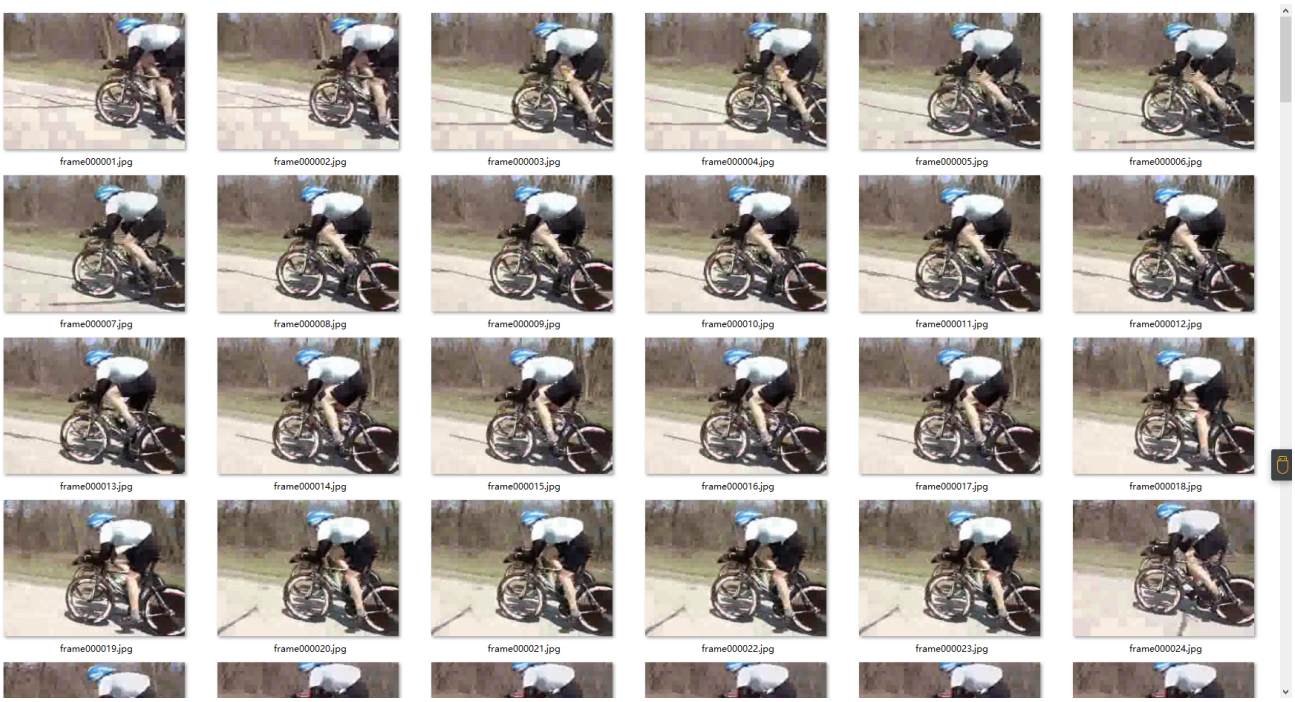


Figure 1: Our method introduces residual connections in a two-stream ConvNet model [20]. The two networks separately capture spatial (appearance) and temporal (motion) information to recognize the input sequences. We do not use residuals from the spatial into the temporal stream as this would bias both losses towards appearance information.

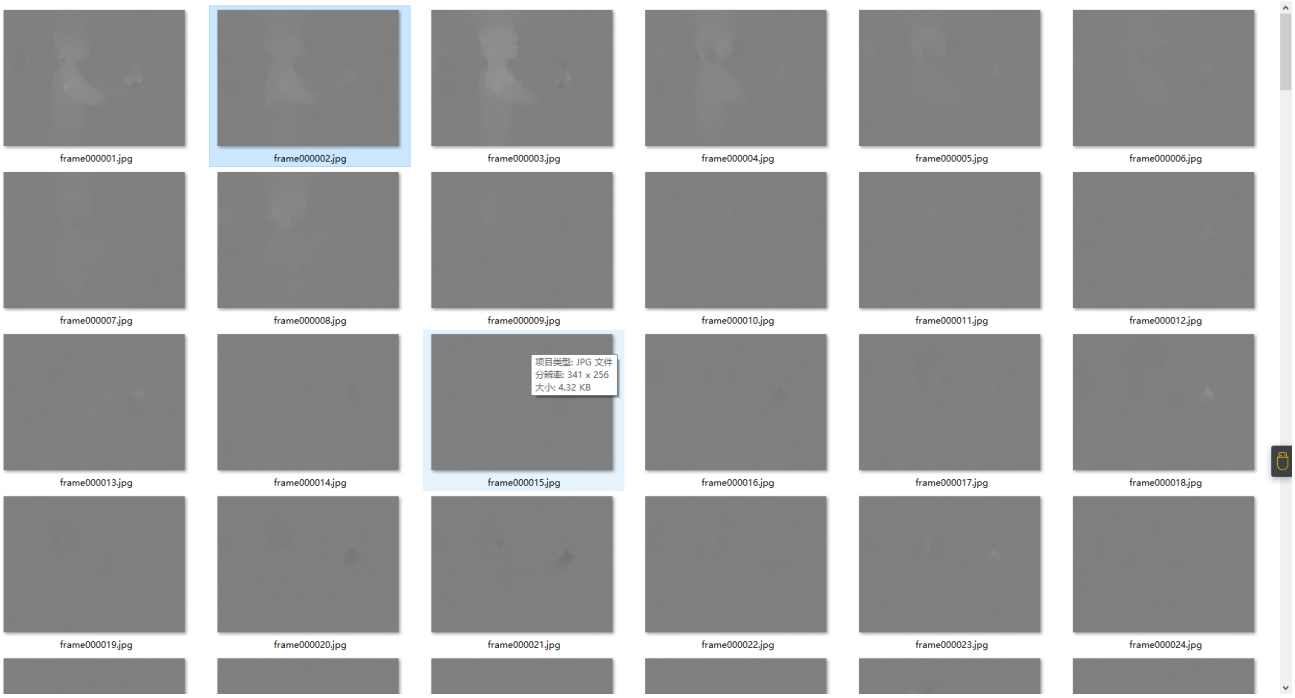
双流结构：

- spatial appearance stream： RGB images
- a temporal motion stream： optical flow information

RGB images



optical flow

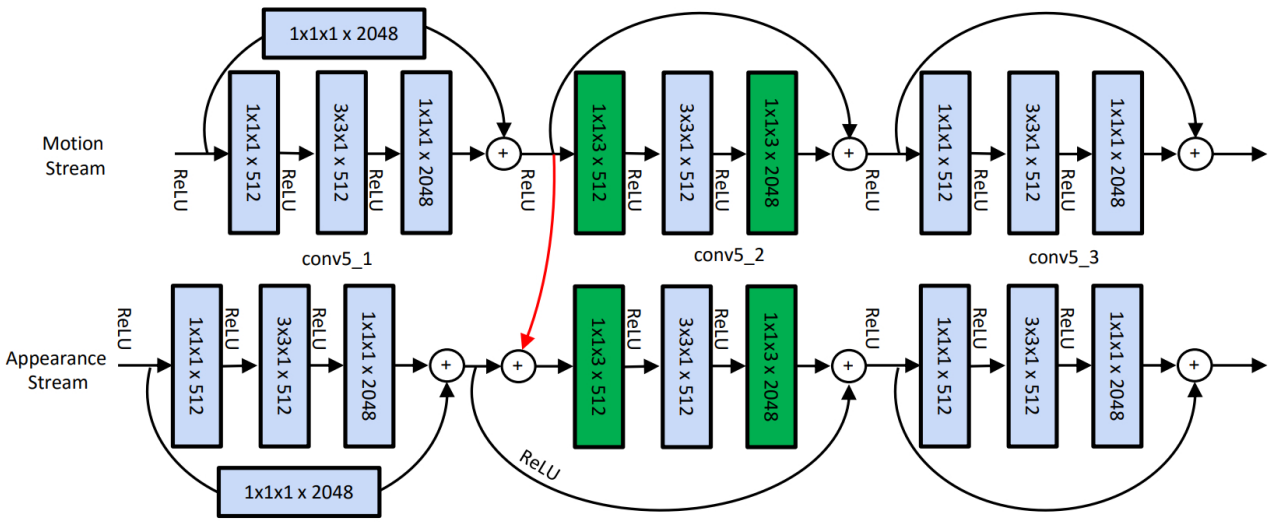


特点：

1. 将残差网络引入到视频动作识别中
2. 在两个信息流中加入残差连接
3. 克服了原方法中时间接受域大小限制的问题。

具体实现

双流残差网络



> The network sees an input of size 224×224 that is reduced five times in the network by stride 2 convolutions followed by a global average pooling layer of the final 7×7 feature map and a fullyconnected classification layer with softmax.

The residual units are defined as:

$$\mathbf{x}_{l+1} = f\left(\mathbf{x}_l + \mathcal{F}(\mathbf{x}_l; \mathcal{W}_l)\right), \tag{1}$$

双流框架的缺点:

无法在空间上连接 `appearance` 和 `motion` 的信息。为了解决这样的问题，加入了**Motion Residuals**.

Motion Residuals

$$\hat{\mathbf{x}}^{\mathrm{a}}_{l+1} = f(\mathbf{x}^{\mathrm{a}}_l) + \mathcal{F}\Big(\mathbf{x}^{\mathrm{a}}_l + f(\mathbf{x}^{\mathrm{m}}_l), \mathcal{W}^{\mathrm{a}}_l\Big), \tag{2}$$

$$\frac{\partial \mathcal{L}}{\partial \mathbf{x}^{\mathrm{a}}_l} = \frac{\partial \mathcal{L}}{\partial \hat{\mathbf{x}}^{\mathrm{a}}_{l+1}} \frac{\partial \hat{\mathbf{x}}^{\mathrm{a}}_{l+1}}{\partial \mathbf{x}^{\mathrm{a}}_l} = \frac{\partial \mathcal{L}}{\partial \hat{\mathbf{x}}^{\mathrm{a}}_{l+1}} \left(\frac{\partial f(\mathbf{x}^{\mathrm{a}}_l)}{\partial \mathbf{x}^{\mathrm{a}}_l} + \frac{\partial}{\partial \mathbf{x}^{\mathrm{a}}_l} \mathcal{F}\Big(\mathbf{x}^{\mathrm{a}}_l + f(\mathbf{x}^{\mathrm{m}}_l), \mathcal{W}^{\mathrm{a}}_l\Big) \right) \tag{3}$$

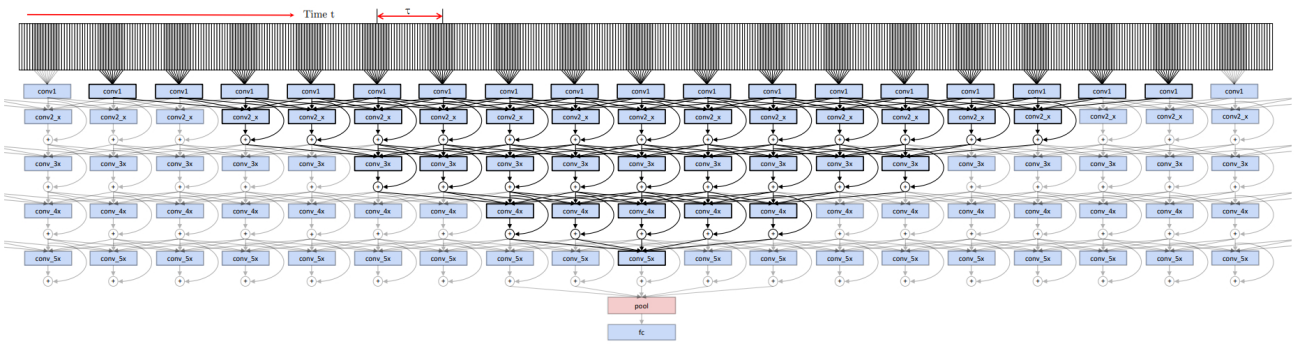
$$\frac{\partial \mathcal{L}}{\partial \mathbf{x}^{\mathrm{m}}_l} = \frac{\partial \mathcal{L}}{\partial \mathbf{x}^{\mathrm{m}}_{l+1}} \frac{\partial \mathbf{x}^{\mathrm{m}}_{l+1}}{\partial \mathbf{x}^{\mathrm{m}}_l} + \frac{\partial \mathcal{L}}{\partial \hat{\mathbf{x}}^{\mathrm{a}}_{l+1}} \frac{\partial}{\partial \mathbf{x}^{\mathrm{a}}_l} \mathcal{F}\Big(\mathbf{x}^{\mathrm{a}}_l + f(\mathbf{x}^{\mathrm{m}}_l), \mathcal{W}^{\mathrm{a}}_l\Big), \tag{4}$$

跨时间卷积残差网络

利用时间残差的原因：

- 时空的相干性在处理随时间变化的数据十分的重要
- 时间平滑很重要，要求数据随时间变换缓慢
- 卷积网络可以捕获随时间重复的运动模式，。例如 `锤击` 动作。
- 求和操作是一个低通滤波器，会平滑掉高频的运动特征。并且 反向传播也无法弥补这样的问题。

  操作的办法就是建立一个 `时间卷积核` 。



$$\hat{\mathbf{w}}_l(i,j,t,c) = \frac{\mathbf{w}_l(i,j,c)}{T'}, \forall t \in [1, T'], \tag{5}$$

在时间维度上平均特征。

总框架

Layers	conv1	pool1	conv2_x	conv3_x	conv4_x	conv5_x	pool5
Blocks	7×7×1, 64	3×3×1 max stride 2	[1×1×1, 64; 3×3×1, 64; 1×1×1, 256] skip-stream [1×1×3, 64; 3×3×1, 64; 1×1×3, 256] [1×1×1, 64; 3×3×1, 64; 1×1×1, 256]	[1×1×1, 128; 3×3×1, 128; 1×1×1, 512] skip-stream [1×1×3, 128; 3×3×1, 128; 1×1×3, 512] [1×1×1, 128; 3×3×1, 128; 1×1×1, 512] ×2	[1×1×1, 256; 3×3×1, 256; 1×1×1, 1024] skip-stream [1×1×3, 256; 3×3×1, 256; 1×1×3, 1024] [1×1×1, 256; 3×3×1, 256; 1×1×1, 1024] ×4	[1×1×1, 512; 3×3×1, 512; 1×1×1, 2048] skip-stream [1×1×3, 512; 3×3×1, 512; 1×1×3, 2048] [1×1×1, 512; 3×3×1, 512; 1×1×1, 2048]	7×7×1 avg 1×1×5 max stride 2
Output size	112×112×11	56×56×11	56×56×11	28×28×11	14×14×11	7×7×11	1×1× 4
Recept. Field	7×7×1	11×11×1	35×35×5τ	99×99×9τ	291×291×13τ	483×483×17τ	675 × 675× 47τ

实验

三步训练

1. Motion and appearance streams
2. ST-ResNet
3. ST-ResNet* ： with a temporal max-pooling layer after pool5

Dataset	Appearance stream	Motion stream	Two-Streams	ST-ResNet	ST-ResNet*
UCF101	82.29%	79.05%	89.47%	92.76%	93.46%
HMDB51	43.42%	55.47%	60.59%	65.57%	66.41%

与其他算法比较

Method	UCF101	HMDB51	Method	UCF101	HMDB51
Two-Stream ConvNet [20]	88.0%	59.4%	IDT [29]	86.4%	61.7%
Two-Stream+LSTM[18]	88.6%	-	C3D + IDT [26]	90.4%	-
Two-Stream (VGG16) [1, 31]	91.4%	58.5%	TDD + IDT [30]	91.5%	65.9%
Transformations[31]	92.4%	62.0%	Dynamic Image Networks + IDT [2]	89.1%	65.2%
Two-Stream Fusion[5]	92.5%	65.4%	Two-Stream Fusion[5]	93.5%	69.2%
ST-ResNet*	**93.4%**	**66.4%**	ST-ResNet* + IDT	**94.6%**	**70.3%**

Table 4: Mean classification accuracy of the state-of-the-art on HMDB51 and UCF101 for the best ConvNet approaches (left) and methods that additionally use IDT features (right). Our ST-ResNet obtains best performance on both datasets.

结论

提出了一个时空残差网络用于视频动作识别

1. 双流网络和残差网络结合
2. 通过时间卷积核将空间特征转换到时空上。