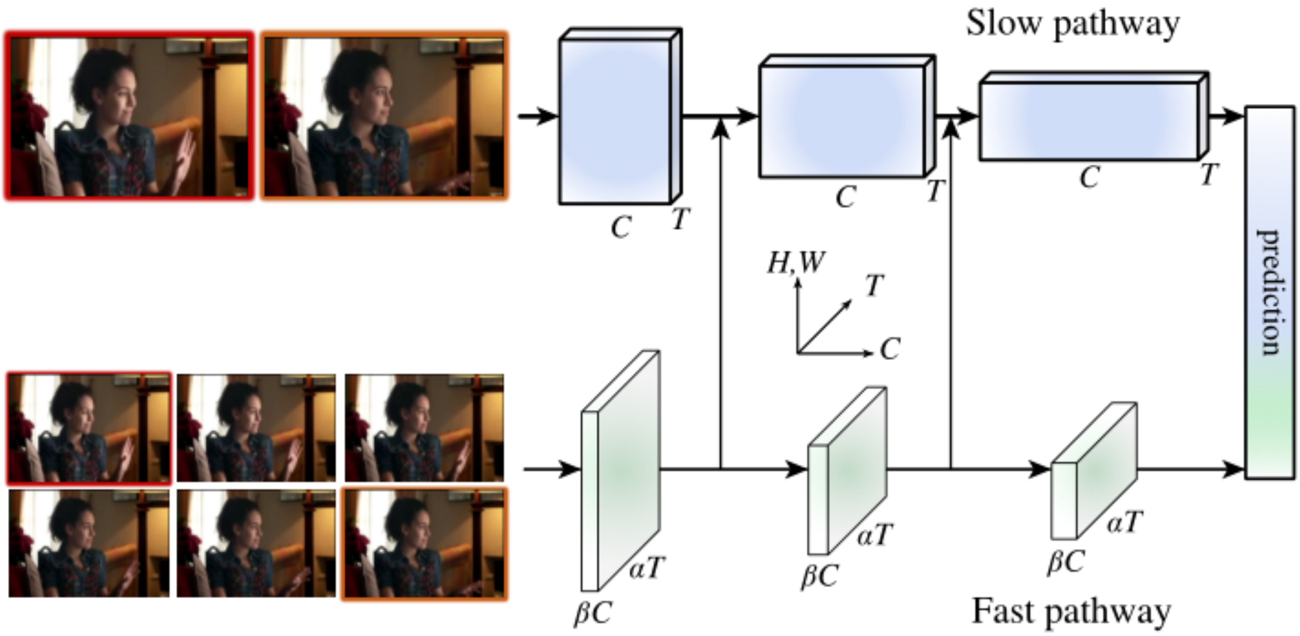
SlowFast Networks for Video Recognition



最近一直在看视频分类，时序行为检测的文章。斗胆讲讲最近看到的一篇文章，FAIR出品的"SlowFast Networks for Video Recognition"。看了一下第一作者的介绍，也是在这领域研究了多年的大牛。文章整体给人感觉就是大厂的感觉，对比消融实验都做得非常详细。尤其是128块GPU 想想就很刺激了。

摘要：提出了一种快慢结合的网络来用于视频分类。其中一路为Slow网络，输入为低帧率，用来捕获空间语义信息。另一路为Fast网络，输入为高帧率，用来捕获运动信息。而且Fast网络是一个轻量级的网络，其channel比较小。当然了在Kinetics达到了79%的精度。。。在AVA上也达到了28.3mAP

的 state-of-the-art的水平。

1. **介绍**

作者一开始提出了一个很有趣的问题，对于图像I(x,y)，我们很自然的将其分为x,y两个维度。但是对于视频I(x,y,t)呢? 时间维度并不能和空间维度等同来看待，这也是当前c3d等工作的效果难以达到最优水平的原因之一。作者从生物学方面获得启发，认为慢运动更符合人类的运动感受刺激。所以才提出对运动维度（时间维度）和空间维度分而治之的思想。

对于空间维度，空间语义信息是变化缓慢的。比如，挥手的动作中，手的语义信息是不发生变化的。一个人无论走还是跑，他仍然是一个人。但对于运动维度，运动相比于发生运动的实体来说，变化是非常快的。基于这些，作者提出来一个双路的SlowFast网络。正如**摘要** 所说，一路为Slow网络，输入为低帧率，用来捕获空间语义信息。另一路为Fast网络，输入为高帧率，用来捕获运动信息，Fast网络是一个轻量级的网络。

作者专门强调了SlowFast网络受到生物学中灵长类视觉系统中视网膜节细胞的启发。在视网膜节细胞中，80%是P-cell, 20%是M-cell，其中M-cell，接受高帧率信息，负责响应运动变化，对空间和颜色信息不敏感。P-cell处理低帧率信息，负责精细的空间和颜色信息。而这正对应于SlowFast网络的两路。

2. 方法

- **Slow pathway**

对于一个video clip, Slow 网络的每 τ帧采样一帧作为输入。假定该网络的输入为*T 帧，则该视频clip的长度为$\tau \times$T 。*

- **Fast pathway**

*高帧率：*Fast网络相比于Slow网络，处理高帧率的信息，则每τ/α帧采样一帧作为输入，也就是输入为 αT帧。（α=8 默认）

*高分辨率的空间特征：*不使用空间降采样层。

*轻量级：*相比于Slow网络，channel为其 β倍（β<1）。一般计算复杂度（FLOPs）于channel为二次关系，所以在SlowFast中，Fast网络占到20%左右的计算量。

- **侧连接**

侧连接连接Fast和Slow网络，达到信息融合的目的。在每个阶段，将Fast输出链接到Slow中。作者也尝试了双向连接，但是没有效果的提升。

最后是全局平局池化，双路信息串联，后接一个全连接层用来分类。

3. 实例化

SlowFast网络是generic的，backbone可以为各种state-of-the-art的网络。本文作者也尝试了3D-Resnet和non-local模块。

一个基于3D-ResNet-50的网络结构如下表所示。

stage	*Slow* pathway	*Fast* pathway	output sizes $T\times S^2$
raw clip	-	-	64×224^2
data layer	stride 16, 1^2	stride **2**, 1^2	*Slow* : 4×224^2 *Fast* : 32×224^2
conv_1	1×7^2, 64 stride 1, 2^2	$\underline{5\times7^2}$, 8 stride 1, 2^2	*Slow* : 4×112^2 *Fast* : 32×112^2
pool_1	1×3^2 max stride 1, 2^2	1×3^2 max stride 1, 2^2	*Slow* : 4×56^2 *Fast* : 32×56^2
res_2	$\begin{bmatrix} 1\times1^2, 64 \\ 1\times3^2, 64 \\ 1\times1^2, 256 \end{bmatrix}\times3$	$\begin{bmatrix} \underline{3\times1^2}, 8 \\ 1\times3^2, 8 \\ 1\times1^2, 32 \end{bmatrix}\times3$	*Slow* : 4×56^2 *Fast* : 32×56^2
res_3	$\begin{bmatrix} 1\times1^2, 128 \\ 1\times3^2, 128 \\ 1\times1^2, 512 \end{bmatrix}\times4$	$\begin{bmatrix} \underline{3\times1^2}, 16 \\ 1\times3^2, 16 \\ 1\times1^2, 64 \end{bmatrix}\times4$	*Slow* : 4×28^2 *Fast* : 32×28^2
res_4	$\begin{bmatrix} \underline{3\times1^2}, 256 \\ 1\times3^2, 256 \\ 1\times1^2, 1024 \end{bmatrix}\times6$	$\begin{bmatrix} \underline{3\times1^2}, 32 \\ 1\times3^2, 32 \\ 1\times1^2, 128 \end{bmatrix}\times6$	*Slow* : 4×14^2 *Fast* : 32×14^2
res_5	$\begin{bmatrix} \underline{3\times1^2}, 512 \\ 1\times3^2, 512 \\ 1\times1^2, 2048 \end{bmatrix}\times3$	$\begin{bmatrix} \underline{3\times1^2}, 64 \\ 1\times3^2, 64 \\ 1\times1^2, 256 \end{bmatrix}\times3$	*Slow* : 4×7^2 *Fast* : 32×7^2
	global average pool, concate, fc		# classes

Table 1. **An example instantiation of the SlowFast network**. The dimensions of kernels are denoted by $\{T\times S^2, C\}$ for temporal, spatial, and channel sizes. Strides are denoted as {temporal stride, spatial stride^2}. Here the speed ratio is $\alpha = 8$ and the channel ratio is $\beta = 1/8$. τ is 16. The green colors mark *higher* temporal resolution, and orange colors mark *fewer* channels, for the Fast pathway. Non-degenerate temporal filters are underlined. Residual blocks are shown by brackets. The backbone is ResNet-50.

- **侧连接**

每层的输出，Slow为{ T, S^2, C}，而Fast为{ $\alpha T, S^2, \beta C$}，需要将两者尺寸匹配。为此作者尝试了多种方式

*Time-to-channel: *将{ $\alpha T, S^2, \beta C$}reshape为{ $T, S^2, \alpha\beta C$}再融合。

Time-strided sampling：将{ $\alpha T, S^2, \beta C$}进行采样成{ $T, S^2, \beta C$}再融合。

*Time-strided convolution: *用3D卷积，其中卷积核为 5×1^2，个数为 $2\beta C$，步长2。

4. 实验

作者多次强调自己的网络是trained from scratch，感觉也是在强调恺明大神的新作“Rethinking ImageNet Pre-training”（个人理解，求轻拍）。

- **Ablations 实验**

做的非常详细(毕竟128块GPU)。直接上图：

model	pre-train	$T \times \tau$	t-reduce	top-1	top-5	GFLOPs
3D R-50 [50]	ImageNet	32×2	2^3	73.3	90.7	33.1
3D R-50 (our recipe)	-	32×2	2^3	73.0	90.4	33.1
3D R-50 [50]	ImageNet	8×8	2^1	73.4	90.9	28.1
3D R-50, our recipe	-	8×8	2^1	73.5	90.8	28.1

(a) **Baselines trained from scratch**: Using the same structure as [50], our training recipe achieves comparable results *without* ImageNet pre-training. "t-reduce" is the temporal downsampling factor in the network.

model	$T \times \tau$	t-reduce	top-1	top-5	GFLOPs
3D R-50	8×8	2^1	73.5	90.8	28.1
3D R-50	8×8	1	**74.6**	**91.5**	44.9
our Slow-only, R-50	4×16	1	72.6	90.3	**20.9**
our Fast-only, R-50	32×2	1	51.7	78.5	**4.9**

(b) **Individual pathways**: Training our Slow-only or Fast-only pathway alone, using the structure specified in Table 1. "t-reduce" is the total temporal downsampling factor within the network.

	lateral	top-1	top-5	GFLOPs
Slow-only	-	72.6	90.3	20.9
SlowFast	-	73.5	90.3	26.2
SlowFast	TtoC, concat	74.3	91.0	30.5
SlowFast	TtoC, sum	74.5	91.3	26.2
SlowFast	T-sample	75.4	91.8	26.7
SlowFast	T-conv	**75.6**	**92.1**	27.6

(c) **SlowFast fusion**: Fusing Slow and Fast pathways with various lateral connections is consistently better than the Slow-only baseline. Backbone: R-50.

	top-1	top-5	GFLOPs
Slow-only	72.6	90.3	20.9
$\beta = 1/4$	75.6	91.7	41.7
$1/6$	**75.8**	92.0	32.0
$1/8$	75.6	**92.1**	27.6
$1/12$	75.2	91.8	25.1
$1/16$	75.1	91.7	23.4
$1/32$	74.2	91.3	21.9

(d) **Channel capacity ratio**: Varying values of β, the channel capacity ratio of the Fast pathway. Backbone: R-50.

Fast pathway	spatial	top-1	top-5	GFLOPs
RGB	-	**75.6**	**92.1**	27.6
RGB, β=1/4	*half*	74.7	91.8	26.3
gray-scale	-	**75.5**	**91.9**	**26.1**
time diff	-	74.5	91.6	26.2
optical flow	-	73.8	91.3	26.9

(e) **Weaker spatial input to Fast pathway**: Various ways of weakening spatial inputs to the Fast pathway in SlowFast models. β=1/8 unless specified otherwise. Backbone: R-50.

	top-1	top-5	GFLOPs
Slow-only	72.6	90.3	20.9
SlowFast	**75.6**	**92.1**	**27.6**
2-Slow ens.	73.2	90.8	41.8
"SlowSlow"	70.5	88.6	75.6

(f) ***vs.* Slow+Slow**: Ensembling 2 Slow-only models (ens.), or replacing the Fast pathway with a Slow pathway ("SlowSlow") . Backbone: R-50.

	$T \times \tau$	α	top-1	top-5	GFLOPs
Slow-only	4×16	-	72.6	90.3	20.9
SlowFast	4×16	8	75.6	92.1	27.6
Slow-only	8×8	-	74.9	91.5	41.9
SlowFast	8×8	4	**77.0**	**92.6**	50.3
SlowFast	2×32	8	73.4	90.8	**13.9**
SlowFast	4×16	4	75.3	91.7	25.2
SlowFast	6×16	8	76.8	92.2	41.1
SlowFast	8×12	4	76.8	92.5	50.3

(g) **Various SlowFast instantiations**, compared to Slow-only counterparts. Here all SlowFast models use β=1/8 for the Fast pathway. Backbone: R-50.

SlowFast	$T \times \tau$	α	top-1	top-5	GFLOPs
R-50	4×16	8	75.6	92.1	27.6
R-50 + NL	4×16	8	76.3	92.2	33.8
R-50	8×8	4	77.0	92.6	50.3
R-50 + NL	8×8	4	**77.7**	**93.1**	65.5
R-101	4×16	8	76.9	92.7	44.5
R-101 + NL	4×16	8	77.4	92.7	47.4
R-101	8×8	4	77.9	93.2	81.5
R-101 + NL	8×8	4	**79.0**	**93.6**	88.0

(h) **Advanced backbones for SlowFast models**, with ResNet-101 [21] and/or non-local (NL) blocks [50]. NL blocks are added to $res_{3,4}$ for R-50 and to res_4 for R-101.

- **Comparison with state-of-the-art results**

直接放图：

model	flow	pretrain	top-1	top-5	inference GFLOPs×views
I3D [3]		ImageNet	72.1	90.3	108 × N/A
Two-Stream I3D [3]	✓	ImageNet	75.7	92.0	216 × N/A
S3D-G [53]	✓	ImageNet	77.2	93.0	143 × N/A
Nonlocal R-50 [50]		ImageNet	76.5	92.6	282 × 30
Nonlocal R-101 [50]		ImageNet	77.7	93.3	359 × 30
R(2+1)D Flow [45]	✓	-	67.5	87.2	152 × 115
STC [7]		-	68.7	88.5	N/A × N/A
ARTNet [48]		-	69.2	88.3	23.5 × 250
S3D [53]		-	69.4	89.1	66.4 × N/A
ECO [54]		-	70.0	89.4	N/A × N/A
I3D [3]	✓	-	71.6	90.0	216 × N/A
R(2+1)D [45]		-	72.0	90.0	152 × 115
R(2+1)D [45]	✓	-	73.9	90.9	304 × 115
SlowFast, R50 (4×16)		-	75.6	92.1	36.1 × 30
SlowFast, R50		-	77.0	92.6	65.7 × 30
SlowFast, R50 + NL		-	77.7	93.1	80.8 × 30
SlowFast, R101		-	77.9	93.2	106 × 30
SlowFast, R101 + NL		-	**79.0**	**93.6**	115 × 30

Table 3. **Comparison with the state-of-the-art on Kinetics-400**. In the column of computational cost, we report the cost of a single "view" (temporal clip with spatial crop) and the numbers of such views used. Details of the SlowFast models in this table are in Table 2h. "N/A" indicates the numbers are not available for us. The SlowFast models are the $T\times\tau = 8\times8$ versions, unless specified.

model	pretrain	top-1	top-5	inference GFLOPs×views
I3D [2]	-	71.9	90.1	108 × N/A
StNet-IRv2 RGB [18]	ImgNet+Kinetics400†	79.0	N/A	N/A
SlowFast, R50	-	79.9	94.5	65.7 ×30
SlowFast, R101	-	80.4	94.8	106 × 30
SlowFast, R101 + NL	-	**81.1**	**94.9**	115 × 30

Table 4. **Kinetics-600 results**. SlowFast models are with $T\times\tau = 8\times8$. †: The Kinetics-400 training set partially overlaps with the Kinetics-600 validation set, and "*it is therefore not ideal to evaluate models on Kinetics-600 that were pre-trained on Kinetics-400*" [2].

跑完kinetics-400，再跑-600 （跪了。）

*‘在最新的2018年ActivityNet比赛，冠军的最佳单模模型，精度为79.0%。我们的方法达到了81.1%。’*很皮。

- **AVA action detection 结果**

model	flow	video pretrain	val mAP	test mAP
I3D [17]		Kinetics-400	14.5	-
I3D [17]	✓	Kinetics-400	15.6	-
ACRN, S3D [41]	✓	Kinetics-400	17.4	-
ATR, R50 + NL [26]		Kinetics-400	20.0	-
ATR, R50 + NL [26]	✓	Kinetics-400	21.7	-
9-model ensemble [26]	✓	Kinetics-400	25.6	21.1
I3D [13]		Kinetics-**600**	21.9	21.0
SlowFast, R101		Kinetics-400	26.1	-
SlowFast, R101		Kinetics-**600**	26.8	**26.6**
SlowFast, R101 + NL		Kinetics-**600**	27.3	-
SlowFast++, R101 + NL		Kinetics-**600**	**28.3**	-

Table 7. **Comparison with the state-of-the-art on AVA**. Here "++" indicates a version of our method that is tested with multi-scale and horizontal flipping augmentation (testing augmentation strategies for existing methods are not always reported).

5. 总结

直接放上大佬的原话吧。

We hope that this SlowFast concept will foster further research in video recognition.

个人感悟：

感觉从TSN之后，大家开始更多关注在如何稀疏采样上。

同时如何在时间维度上更好的处理运动信息，也是大家重点研究的问题。

最后题外话，如何去国内大厂实习？求带。