Project Proposals

- Next Wednesday (**09/15/2021**).
- A 5 minute presentation + 1-2 minutes for Q/A (time limit will be strictly enforced).
- Your presentation should cover: 1) your research problem, 2) the motivation, 3) basic methodology, 4) datasets that you plan to use, 5) experiments that you want to run.
- Send me the **PDF** slides by **September 14th, 11:59 PM**.
- Project report due **September 15th, 11:59 PM.**

SlowFast Networks for Video Recognition

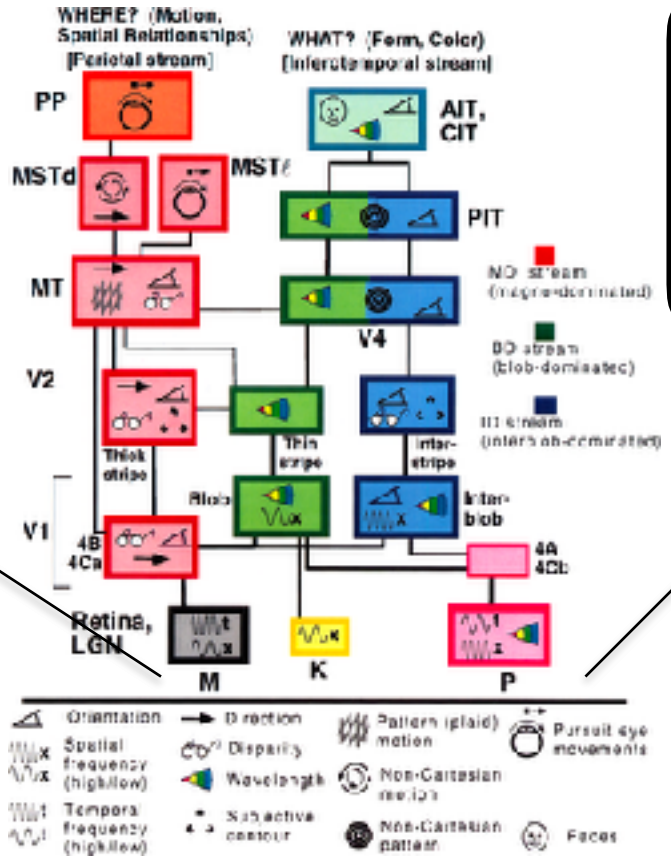
ICCV 2019

Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, Kaiming He

Motivation

- Processes information about motion & depth.
- Fast conduction rate.
- Minority of total cells (~20%)

Magno Cells

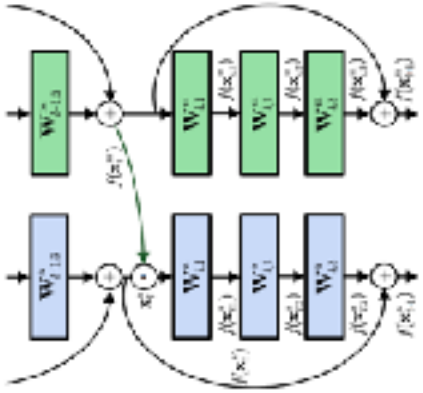


- Processes information about color.
- Slow conduction rate.
- Majority of total cells (~80%)

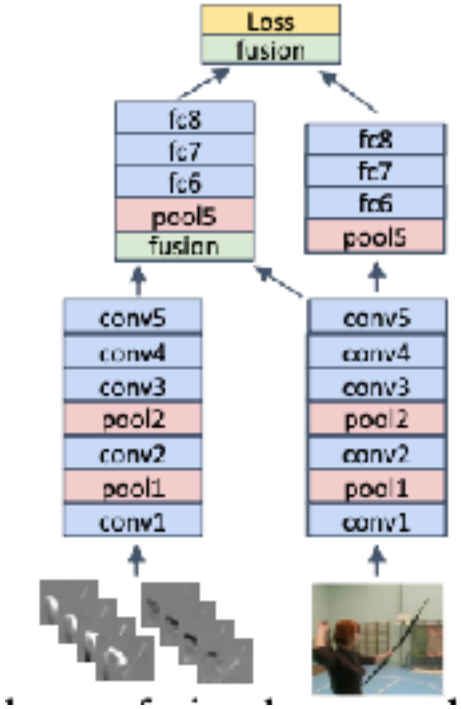
Parvo Cells

"Neural mechanisms of form and motion processing in the primate visual system", Essen et al., Neuron, 1994

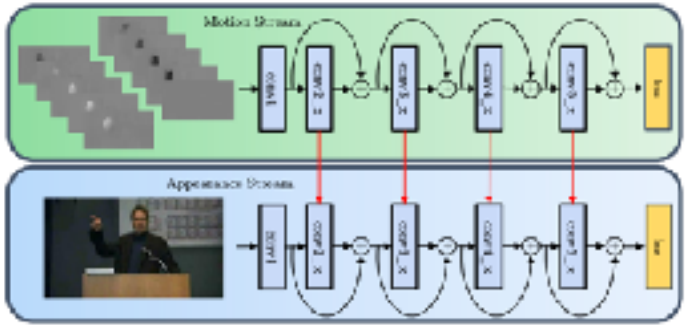
Motivation



"Spatiotemporal Multiplier Networks for Video Action Recognition", Feichtenhofer et al., CVPR 2017



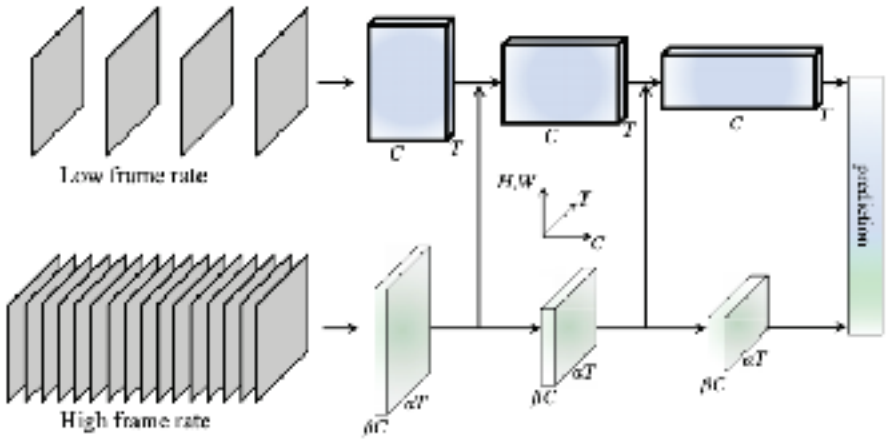
"Convolutional Two-Stream Network Fusion for Video Action Recognition", Feichtenhofer et al., CVPR 2016



"Spatiotemporal Residual Networks for Video Action Recognition", Feichtenhofer et al., NIPS 2016

SlowFast Networks

- The slow pathway (top) processes low frame rate, low temporal resolution video input.
- The fast pathway (bottom) processes high frame rate, high temporal resolution video input.
- The fast pathway contains much fewer channels than the slow pathway.
- Lateral connections fuse the two pathways.

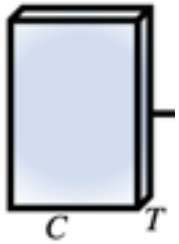


SlowFast Networks

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	5×7^2, 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	$\begin{bmatrix} 1 \times 1^2, 64 \\ 1 \times 3^2, 64 \\ 1 \times 1^2, 256 \end{bmatrix} \times 3$	$\begin{bmatrix} 3 \times 1^2, 8 \\ 1 \times 3^2, 8 \\ 1 \times 1^2, 32 \end{bmatrix} \times 3$	*Slow* : 4×56^2 *Fast* : 32×56^2
res_3	$\begin{bmatrix} 1 \times 1^2, 128 \\ 1 \times 3^2, 128 \\ 1 \times 1^2, 512 \end{bmatrix} \times 4$	$\begin{bmatrix} 3 \times 1^2, 16 \\ 1 \times 3^2, 16 \\ 1 \times 1^2, 64 \end{bmatrix} \times 4$	*Slow* : 4×28^2 *Fast* : 32×28^2
res_4	$\begin{bmatrix} 3 \times 1^2, 256 \\ 1 \times 3^2, 256 \\ 1 \times 1^2, 1024 \end{bmatrix} \times 6$	$\begin{bmatrix} 3 \times 1^2, 32 \\ 1 \times 3^2, 32 \\ 1 \times 1^2, 128 \end{bmatrix} \times 6$	*Slow* : 4×14^2 *Fast* : 32×14^2
res_5	$\begin{bmatrix} 3 \times 1^2, 512 \\ 1 \times 3^2, 512 \\ 1 \times 1^2, 2048 \end{bmatrix} \times 3$	$\begin{bmatrix} 3 \times 1^2, 64 \\ 1 \times 3^2, 64 \\ 1 \times 1^2, 256 \end{bmatrix} \times 3$	*Slow* : 4×7^2 *Fast* : 32×7^2
global average pool, concate, fc			# classes

Lateral Connections

Feature tensor from the slow pathway



T x S^2 x C

Feature tensor from the fast pathway

aT x S^2 x bC

- **Time-to-channel:** Feature tensor of shape (aT x S^2 x bC) is reshaped into (T, S^2, abC), i.e., all frames are packed into the channel dimension.
- **Time-strided sampling:** Only one frame out of every a frames is sampled.
- **Time-strided convolution:** 3D convolution with stride a is applied.

Results on Kinetics

Fusing Slow and Fast pathways with lateral connections is better than the Slow and Fast only baselines.

	lateral	top-1	top-5	GFLOPs
Slow-only	-	72.6	90.3	27.3
Fast-only	-	51.7	78.5	**6.4**
SlowFast	-	73.5	90.3	34.2
SlowFast	TtoC, sum	74.5	91.3	34.2
SlowFast	TtoC, concat	74.3	91.0	39.8
SlowFast	T-sample	75.4	91.8	34.9
SlowFast	T-conv	**75.6**	**92.1**	36.1

Results on Kinetics

Varying values of β, the channel capacity ratio of the Fast pathway to make SlowFast lightweight.

	top-1	top-5	GFLOPs
Slow-only	72.6	90.3	27.3
$\beta = 1/4$	75.6	91.7	54.5
$1/6$	**75.8**	92.0	41.8
$1/8$	75.6	**92.1**	36.1
$1/12$	75.2	91.8	32.8
$1/16$	75.1	91.7	30.6
$1/32$	74.2	91.3	28.6

Results on Kinetics

The proposed training recipe achieves comparable results without ImageNet pre-training.

model	pre-train	top-1	top-5	GFLOPs
3D R-50 [56]	ImageNet	73.4	90.9	36.7
3D R-50, recipe in [56]	-	69.4	88.6	36.7
3D R-50, our recipe	-	73.5	90.8	36.7

Results on Kinetics

Comparison to the state-of-the-art

model	flow	pretrain	top-1	top-5	GFLOPs×views
I3D [5]		ImageNet	72.1	90.3	108 × N/A
Two-Stream I3D [5]	✓	ImageNet	75.7	92.0	216 × N/A
S3D-G [61]	✓	ImageNet	77.2	93.0	143 × N/A
Nonlocal R50 [56]		ImageNet	76.5	92.6	282 × 30
Nonlocal R101 [56]		ImageNet	77.7	93.3	359 × 30
R(2+1)D Flow [50]	✓	-	67.5	87.2	152 × 115
STC [9]		-	68.7	88.5	N/A × N/A
ARTNet [54]		-	69.2	88.3	23.5 × 250
S3D [61]		-	69.4	89.1	66.4 × N/A
ECO [63]		-	70.0	89.4	N/A × N/A
I3D [5]	✓	-	71.6	90.0	216 × N/A
R(2+1)D [50]		-	72.0	90.0	152 × 115
R(2+1)D [50]	✓	-	73.9	90.9	304 × 115
SlowFast 4×16, R50		-	75.6	92.1	36.1 × 30
SlowFast 8×8, R50		-	77.0	92.6	65.7 × 30
SlowFast 8×8, R101		-	77.9	93.2	106 × 30
SlowFast 16×8, R101		-	78.9	93.5	213 × 30
SlowFast 16×8, R101+NL		-	**79.8**	**93.9**	234 × 30

Results on AVA

Comparison to the state-of-the-art

model	flow	video pretrain	val mAP	test mAP
I3D [20]		Kinetics-400	14.5	-
I3D [20]	✓	Kinetics-400	15.6	-
ACRN, S3D [46]	✓	Kinetics-400	17.4	-
ATR, R50+NL [29]		Kinetics-400	20.0	-
ATR, R50+NL [29]	✓	Kinetics-400	21.7	-
9-model ensemble [29]	✓	Kinetics-400	25.6	21.1
I3D [16]		Kinetics-600	21.9	21.0
SlowFast		Kinetics-400	26.3	-
SlowFast		Kinetics-600	26.8	-
SlowFast, +NL		Kinetics-600	27.3	**27.1**
SlowFast*, +NL		Kinetics-600	**28.2**	-

Contributions

- A framework that achieves great results on a variety of action recognition datasets.
- Very effective optimization protocol for training video models from scratch.
- An extension to spatiotemporal localization task.

Weaknesses

- Questionable motivation.
- Cumbersome two-stream architecture, which could have been avoided with simple dilation mechanisms.
- Model architecture seems less important than the optimization procedure.
- Extremely large computational cost during training (e.g., ~400 epochs on Kinetics).

Discussion Questions

- What do you think about neuroscience inspired motivation?

Discussion Questions

- What do you think about neuroscience inspired motivation?
- What do you think about the proposed two-stream architecture?

Discussion Questions

- What do you think about neuroscience inspired motivation?
- What do you think about the proposed two-stream architecture?
- What are your thoughts on the state-of-the-art comparisons?