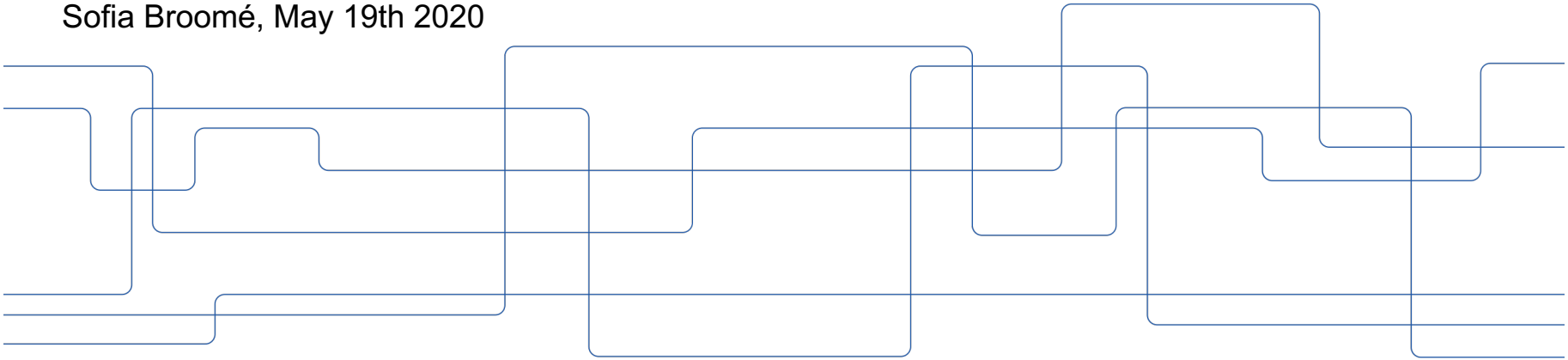
KTH ROYAL INSTITUTE OF TECHNOLOGY

SlowFast Networks for Video Recognition

Feichtenhofer, Fan, Malik & He, ICCV 2019

CV/DL Reading Group

Sofia Broomé, May 19th 2020



Outline

- SlowFast: Idea of the paper + why should we talk about it?
- Short about 3D convolutions
- In general: Recent years in state-of-the-art video architectures
- SlowFast: More details about architecture and method
- SlowFast: Experiments and results
- Discussion

Idea of the paper

- One **spatial** stream, one **temporal** stream:

  - Temporal stream uses input with high frame rate and less channels

  - Spatial stream uses input with low frame rate and normal amount of channels

- Biological analogy…

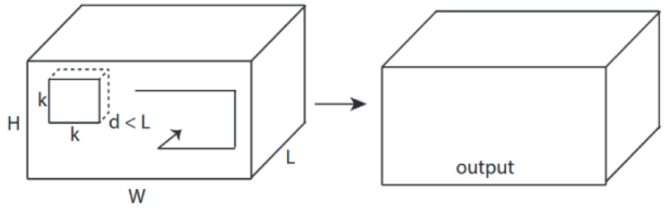
That sounds exactly like the 2-stream network, without optical flow. Is the presentation over?

Why talk about this paper?

- Has gotten attention so there is reason to read it critically
  - Oral presentation at ICCV 2019
- From Facebook AI Research (FAIR)
- See where the state-of-the-art for video is currently at

Hierarchical temporal modeling: 3D CNNs

- Kernels 3D tensors
  - Output another 3D volume



- Non-linearities between successive layers

- The same kernel is convolved across the entire sequence (linear blend of frames)
  - Time as a 3rd spatial axis

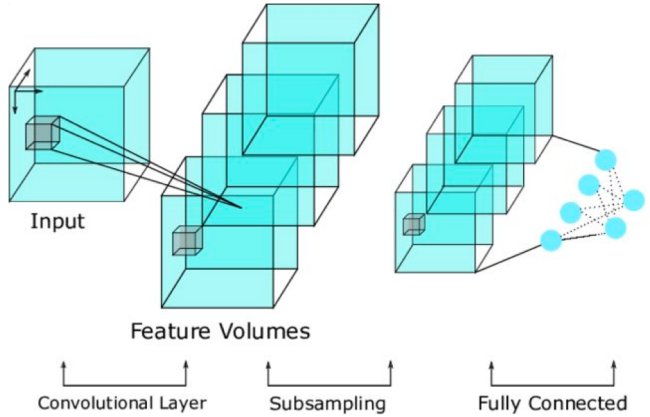


Figure from Roy & Mishra, ECCN: Activity Recognition Using Ensembled Convolutional Neural Networks, TENCON 2019

Motivation of the paper

Authors' motivation:

- Deep learning for video is still difficult
- Convolutions treat all dimensions symmetrically
- What about time? Not all spatiotemporal orientations are equally likely (slow more likely than fast)
- Hence: ”No reason for us to treat space and time symmetrically”

Motivation of the paper

From the introduction:

If all spatiotemporal orientations are not equally likely, then there is no reason for us to treat space and time symmetrically, as is implicit in approaches to video recognition based on spatiotemporal convolutions [49, 5]. We might instead "factor" the architecture to treat spatial structures and temporal events separately. For concreteness, let us

Motivation of the paper

But [my concerns]:

- Architecture treats space and time separately but still symmetrically
- Same 3D CNN backbone
- Local frames are modeled as bags
- Misses an important point: directionality

The resource-hungry (and for that reason typically private sector) lineage of SlowFast

Some early ideas, academia:

- 3D Convolutions (ECCV10), NYU
- The **two-stream model**, Simonyan & Zisserman (NeurIPS14), Oxford

Compute power parade:

- C3D, ICCV15 – FAIR
- **I3D** (Quo Vadis) (CVPR17) – DeepMind
- Pseudo-3D Residual Networks (ICCV17) – Microsoft Research
- **Non-Local Neural Networks** (CVPR18) – FAIR
- S3D (ECCV18) – Google Research
- R(2+1)D (CVPR18) – FAIR

The two-stream model

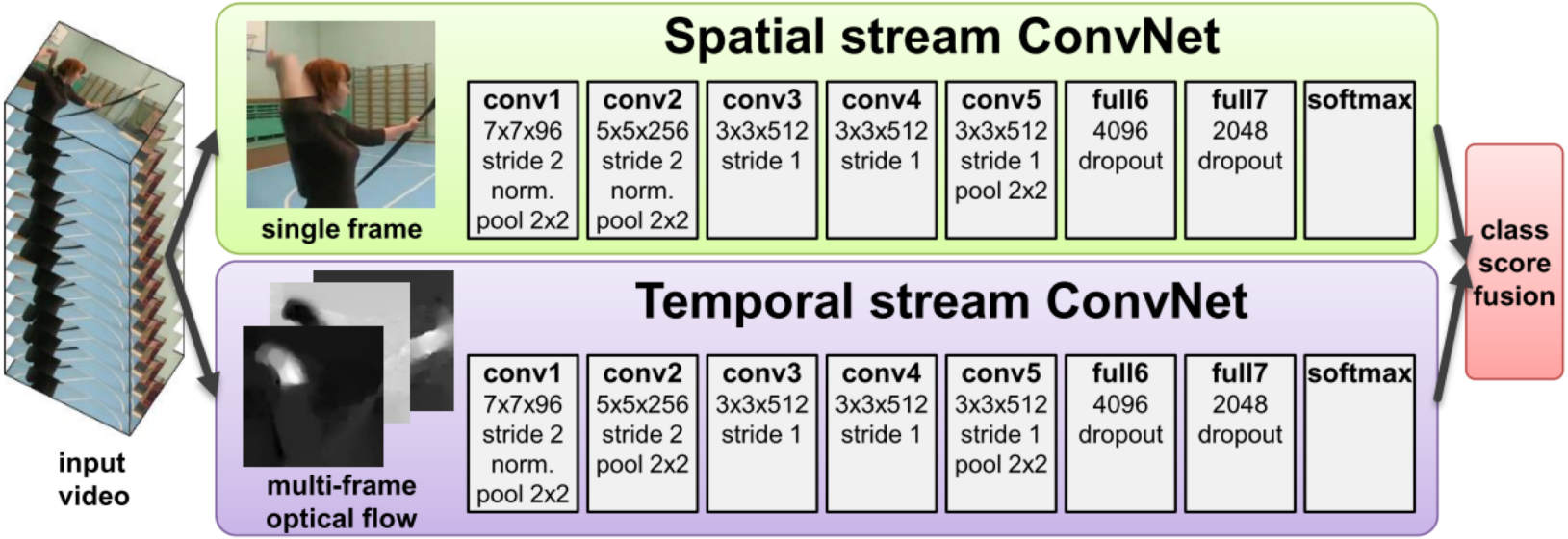
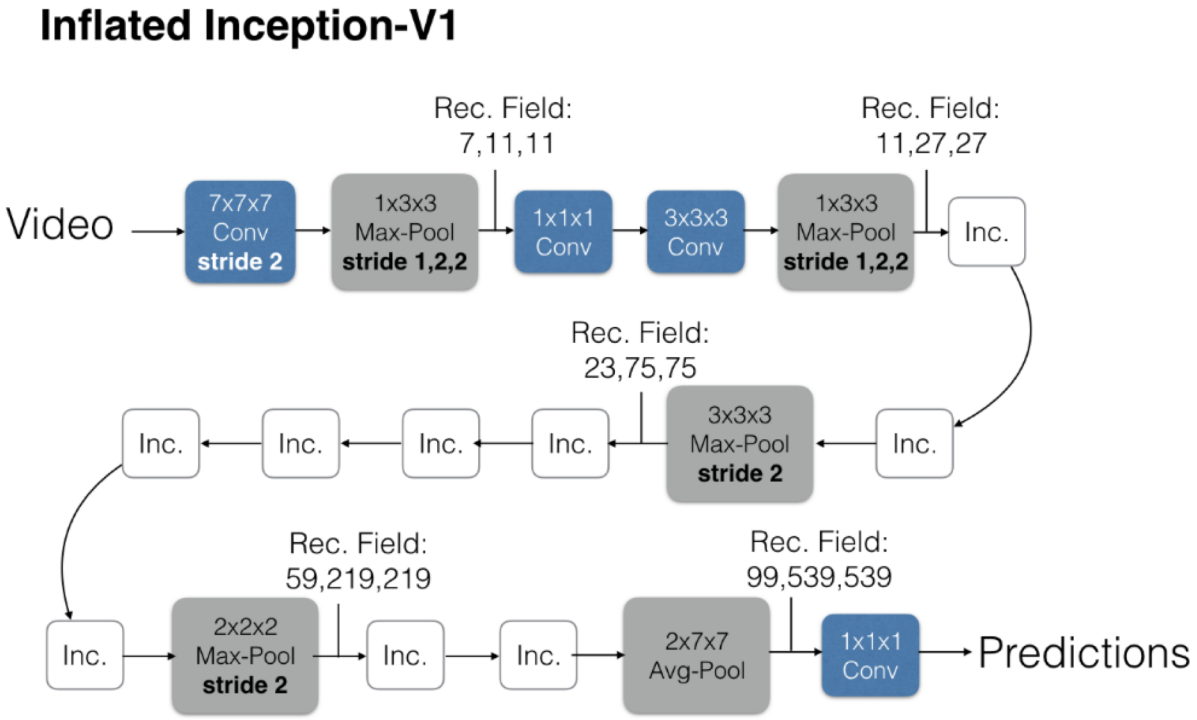
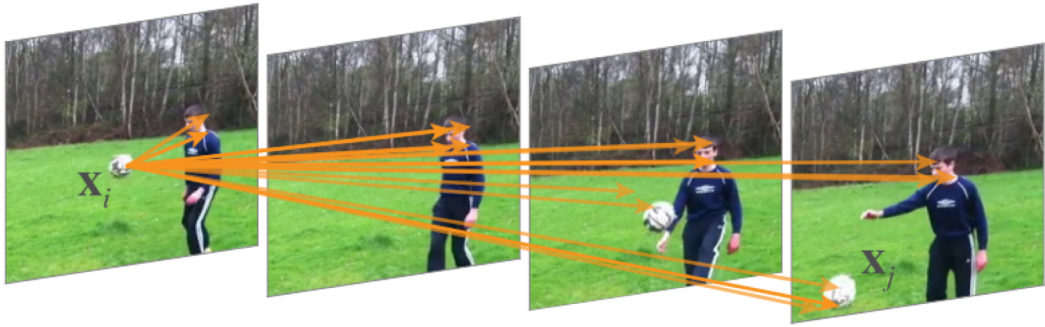


Figure 1: **Two-stream architecture for video classification.**

I3D


Inflated Inception-V1
Video
7x7x7 Conv stride 2
1x3x3 Max-Pool stride 1,2,2
Rec. Field: 7,11,11
1x1x1 Conv
3x3x3 Conv
1x3x3 Max-Pool stride 1,2,2
Rec. Field: 11,27,27
Inc.
Rec. Field: 23,75,75
3x3x3 Max-Pool stride 2
Inc.
Inc.
Inc.
Inc.
Inc.
Inc.
Rec. Field: 59,219,219
2x2x2 Max-Pool stride 2
Inc.
Inc.
Inc.
Rec. Field: 99,539,539
2x7x7 Avg-Pool
1x1x1 Conv
Predictions

Non-local Neural Networks



- Builds on Non-local means method for denoising
- g a linear embedding
- f some pairwise function, e.g. Gaussian

$$\mathbf{y}_i = \frac{1}{\mathcal{C}(\mathbf{x})} \sum_{\forall j} f(\mathbf{x}_i, \mathbf{x}_j) g(\mathbf{x}_j). \qquad (1)$$

- Insert a non-local block for example at residual connection

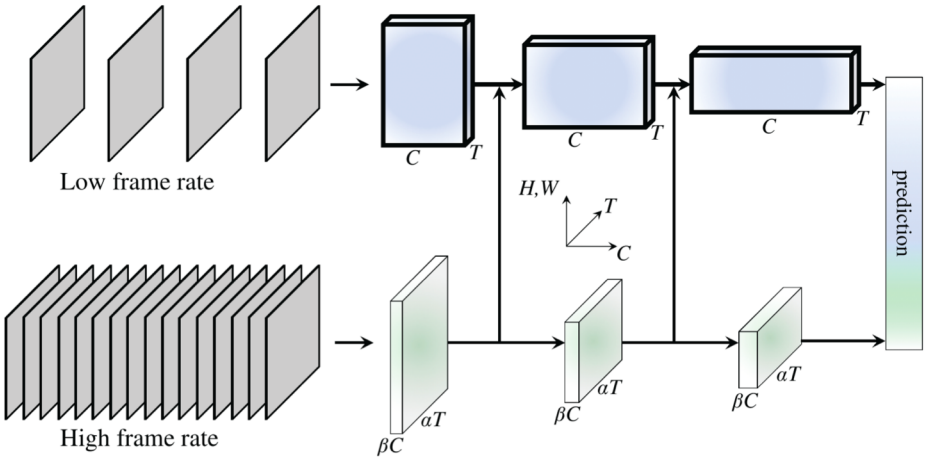
$$\mathbf{z}_i = W_z \mathbf{y}_i + \mathbf{x}_i, \qquad (6)$$

model		top-1	top-5
R50	baseline	71.8	89.7
	space-only	72.9	90.8
	time-only	73.1	90.5
	spacetime	**73.8**	**91.0**
R101	baseline	73.1	91.0
	space-only	74.4	91.3
	time-only	74.4	90.5
	spacetime	**75.1**	**91.7**

The resource-hungry (and for that reason typically private sector) lineage of SlowFast

- Why does it matter where this research comes from?
  - Hyper-parameter search space difficult to inspect with less resources
  - The large datasets often used as benchmarks collected by Google, Fb
  - How can we know if this is progress?
  - Viable to treat video recognition the same way as object recognition?

SlowFast: Model Details



- Same backbone in both pathways
- $\alpha > 1$ ratio between fast and slow pathways' numbers of sampled frames. Typical value: 8
- $\beta < 1$ is the ratio between fast and slow pathways' number of channels Typical value: 1/8
- No temporal downsampling in temporal pathway
- Lateral fusion

SlowFast: Model Details
Lateral fusion

- Need to match feature dimensions between the pathways
- 4 fusions for ResNets

- **Three variants:**

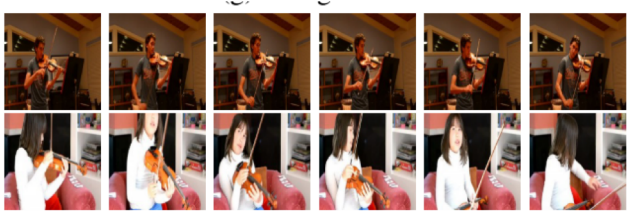
  - Time-to-channel

  - Time-strided sampling

  - Time-strided convolution

Datasets used in the article

- Kinetics-400, DeepMind (400 classes, ~650 hours) classification
- Kinetics-600, DeepMind (600 classes, ~1350 hours) classification



(i) playing violin

(k) braiding hair

(m) dribbling basketball

Fine-tuning after Kinetics:

- Charades, CMU (157 classes, ~80 hours) classification
- AVA, Google (60 classes, ~110 hours, 400+ hours of tracklets) detection

Insensitivity to temporal direction of Kinetics

	Kinetics-Full		Something-something	
Model	Normal (%)	Reversed (%)	Normal (%)	Reversed (%)
I3D	71.1	71.1	45.8	15.2
I2D	67.0	67.2	34.4	35.2

SlowFast: Model Details

- **Training (Kinetics):**
  - 250 or 500 epochs
  - Large minibatch training on 128 GPUs

- **Inference:**
  - Sample 10 clips from a video along temporal axis
  - Take 3 spatial crops from each clip
  - Refer to one crop as one **view**
  - Average the 30 views' softmax scores for prediction

SlowFast Experiments: **Kinetics-400**

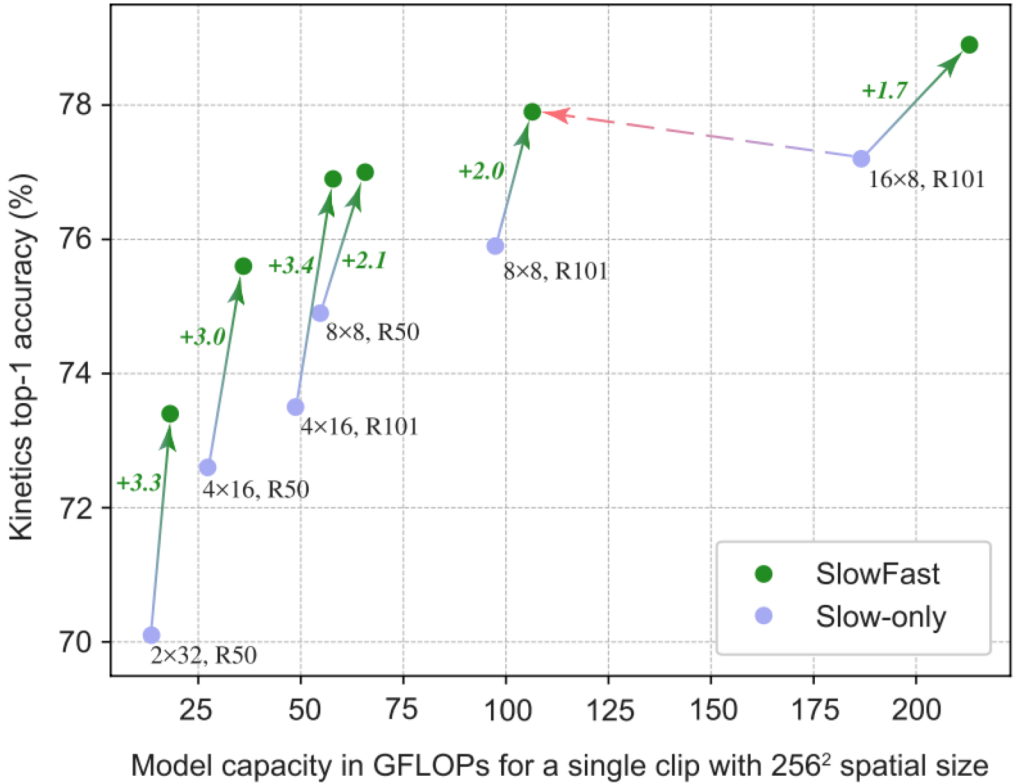
- Table 2: low inference cost and SotA
- Found ±0.3% for Imagenet pre-training

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

Table 2. **Comparison with the state-of-the-art on Kinetics-400.**

SlowFast Experiments: **Kinetics-400**

- Instantiations
  T x τ (input sampling, temporal stride)
- All cases, higher accuracy than Slow-only
- Higher accuracy && lower cost than a temporally heavier Slow-only



SlowFast Experiments: **Kinetics-600**

- New dataset, limited other results
- Table 3: low inference cost and SotA

model	pretrain	top-1	top-5	GFLOPs × views
I3D [3]	-	71.9	90.1	108 × N/A
StNet-IRv2 RGB [21]	ImgNet+Kin400	79.0	N/A	N/A
SlowFast 4×16, R50	-	78.8	94.0	36.1 × 30
SlowFast 8×8, R50	-	79.9	94.5	65.7 ×30
SlowFast 8×8, R101	-	80.4	94.8	106 × 30
SlowFast 16×8, R101	-	81.1	95.1	213 × 30
SlowFast 16×8, R101+NL	-	**81.8**	**95.1**	234 × 30

Table 3. **Comparison with the state-of-the-art on Kinetics-600.** SlowFast models the same as in Table 2.

SlowFast Experiments: **Charades**

- Table 4: low inference cost and SotA

model	pretrain	mAP	GFLOPs × views
CoViAR, R-50 [59]	ImageNet	21.9	N/A
Asyn-TF, VGG16 [42]	ImageNet	22.4	N/A
MultiScale TRN [62]	ImageNet	25.2	N/A
Nonlocal, R101 [56]	ImageNet+Kinetics400	37.5	544 × 30
STRG, R101+NL [57]	ImageNet+Kinetics400	39.7	630 × 30
our baseline (Slow-only)	Kinetics-400	39.0	187 × 30
SlowFast	Kinetics-400	42.1	213 × 30
SlowFast, +NL	Kinetics-400	42.5	234 × 30
SlowFast, +NL	Kinetics-600	**45.2**	234 × 30

Table 4. **Comparison with the state-of-the-art on Charades**. All our variants are based on $T \times \tau = 16 \times 8$, R-101.

SlowFast Experiments: Kinetics-400
Ablations on fast pathway, fusion

	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

(a) **SlowFast fusion**: Fusing Slow and Fast pathways with various types of lateral connections throughout the network hierarchy is consistently better than the Slow and Fast only baselines.

SlowFast Experiments: Kinetics-400
Ablations on fast pathway, channel capacity

	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

(b) **Channel capacity ratio**: Varying values of β, the channel capacity ratio of the Fast pathway to make SlowFast lightweight.

SlowFast Experiments: Kinetics-400
Ablations on fast pathway, weaker spatial input

Fast pathway	spatial	top-1	top-5	GFLOPs
RGB	-	**75.6**	**92.1**	36.1
RGB, β=1/4	*half*	74.7	91.8	34.4
gray-scale	-	**75.5**	**91.9**	**34.1**
time diff	-	74.5	91.6	34.2
optical flow	-	73.8	91.3	35.1

(c) **Weaker spatial input to Fast pathway**: Alternative ways of weakening spatial inputs to the Fast pathway in SlowFast models. β=1/8 unless specified otherwise.

SlowFast Experiments: AVA detection

- Spatiotemporal localization of human actions
- mAP, IoU threshold 0.5
- Faster R-CNN but with SlowFast backbone
- Off-the-shelf person detector

SlowFast Experiments: AVA detection

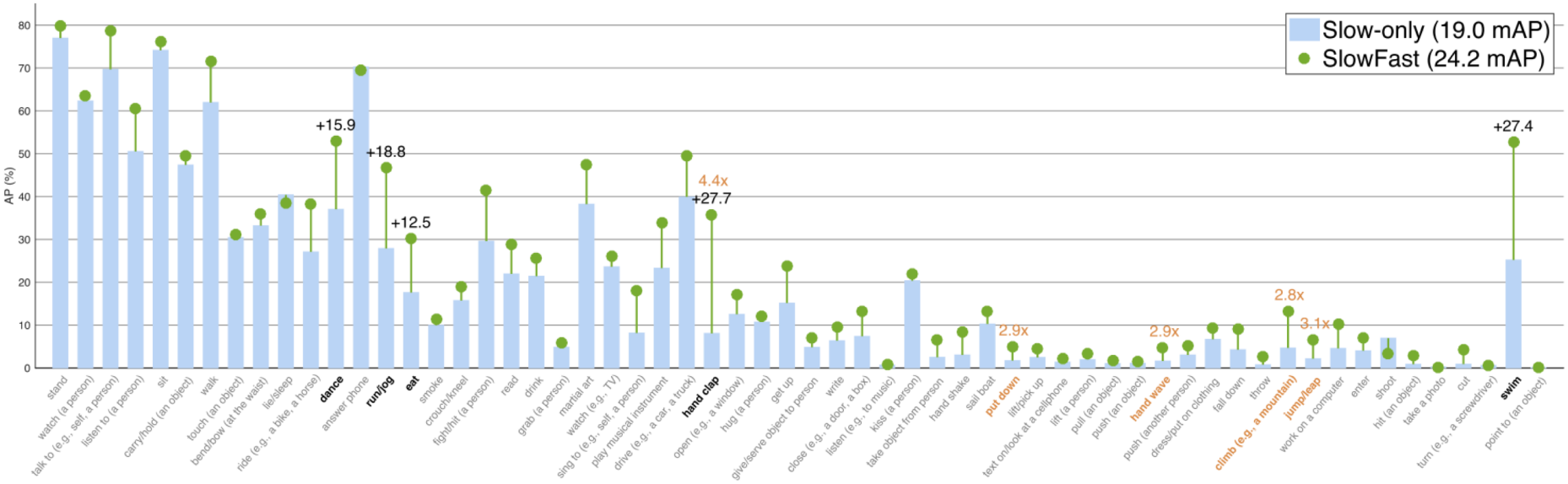
- Improvement from baseline
- Discuss: This is *solely* contributed by our SlowFast idea.

model	$T \times \tau$	α	mAP
Slow-only, R-50	4×16	-	19.0
SlowFast, R-50	4×16	8	**24.2**

Table 9. **AVA action detection baselines**: Slow-only *vs*. SlowFast.

SlowFast Experiments: AVA detection

- Improvement from baseline



SlowFast Experiments: AVA detection

- Better relative improvement, compared to **optical flow** for others (e.g. +1.1 mAP vs +5.2 mAP)
- +5.6 mAP higher than previous best model (21.7 mAP) under similar setting
- Different improvements, obtain 30.7 mAP in best setting, and 34.3 for an ensemble of 7 models

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**	-

SlowFast Conclusion

6. Conclusion

The time axis is a special dimension. This paper has investigated an architecture design that contrasts the speed along this axis. It achieves state-of-the-art accuracy for video action classification and detection. We hope that this SlowFast concept will foster further research in video recognition.

Discussion points

- No principal difference in how space and time are being modeled
- Is more than smart bag-modeling of these datasets needed? (Spatial information)
- What do you think of reporting one result in this way?
- What is a better way of measuring video capabilities?
- Something else you thought of?