LONG BEACH
CALIFORNIA
June 16-20, 2019

Tutorial on Action Classification and Video Modeling

Revisiting Spatiotemporal Convolutions for Video Analysis

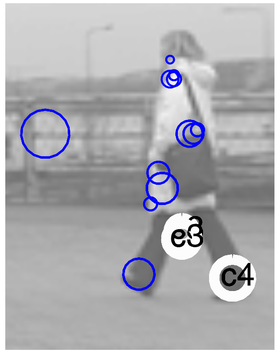
Lorenzo Torresani

facebook research

DARTMOUTH

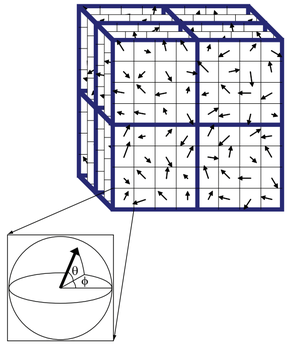
From handcrafted descriptors to spatiotemporal feature learning

STIPs



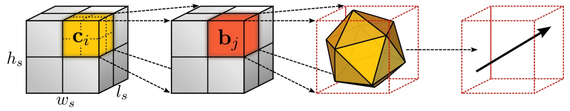
From [Laptev and Lindeberg, ICCV 2003]

SIFT-3D



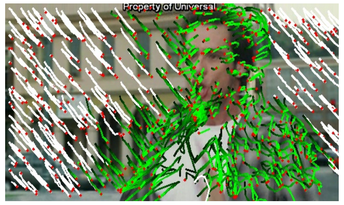
From [Scovanner et al., ACM MM 2007]

HOG-3D



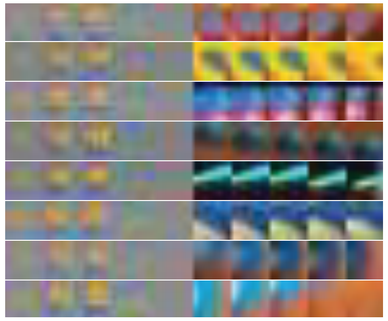
From [Kläser et al., BMVC 2008]

Improved Dense Trajectories



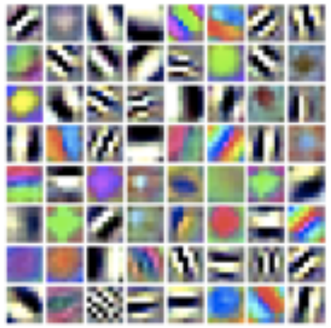
From [Wang and Schmid, ICCV 2013]

C3D



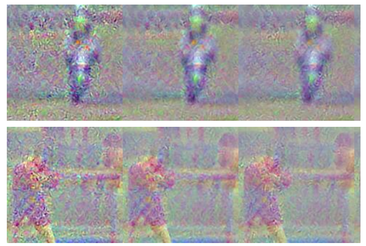
From [Tran et al., ICCV 2015]

I3D



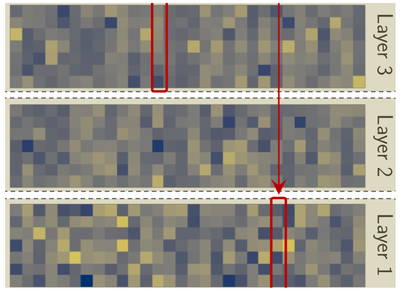
From [Carreira and Zisserman, CVPR 2017]

P3D



From [Qiu et al., ICCV 2017]

Timeception



From [Hussein et al., CVPR 2018]

3D convolution: related work

- 3D CNNs for recognizing human actions in video were arguably first proposed in [Baccouche et al., HBU 2011] and [Ji et al., TPAMI 2012]

- Studied in parallel for unsupervised spatiotemporal feature learning with Restricted Boltzmann Machines [Taylor et al., ECCV 2010] and stacked ISA [Le et al. CVPR 2011]

- Shown to lead to strong action recognition results when trained on large-scale datasets [Tran et al., ICCV 2015]

- Demonstrated to generalize well to other video tasks, e.g., action detection [Shou et al., CVPR 2016], video captioning [Pan et al., CVPR 2016], and hand gesture recognition [Molchanov et al., CVPR 2016]

Figure from [Baccouche et al., HBU 2011]

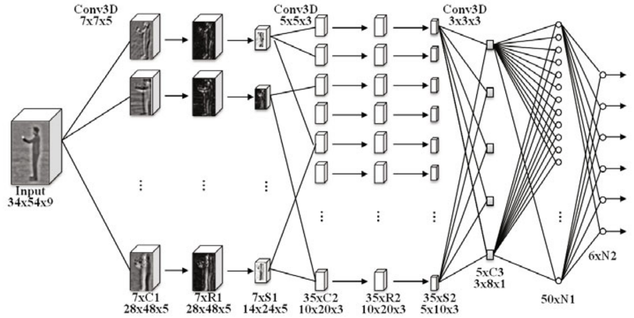


Figure from [Le et al., CVPR 2011]

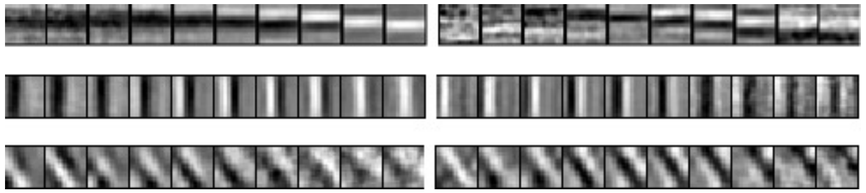
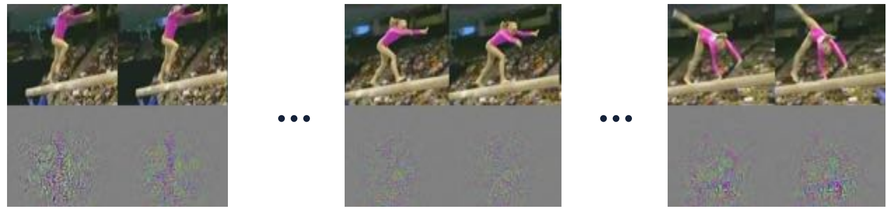
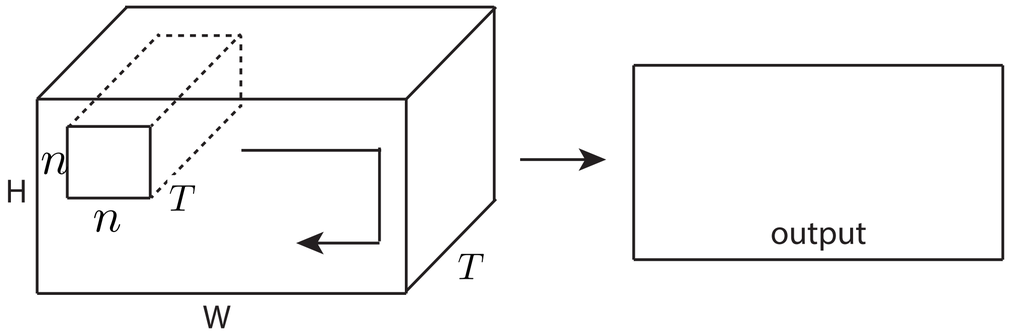


Figure from [Tran et al., ICCV 2015]



2D vs 3D convolution

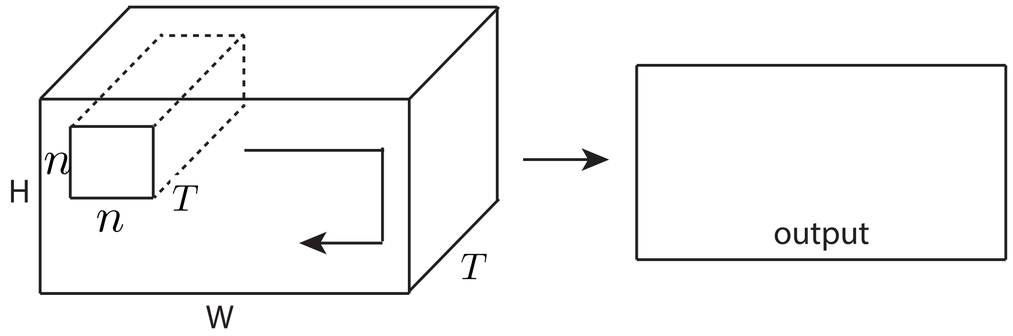
Simplified illustration based on single input channel:



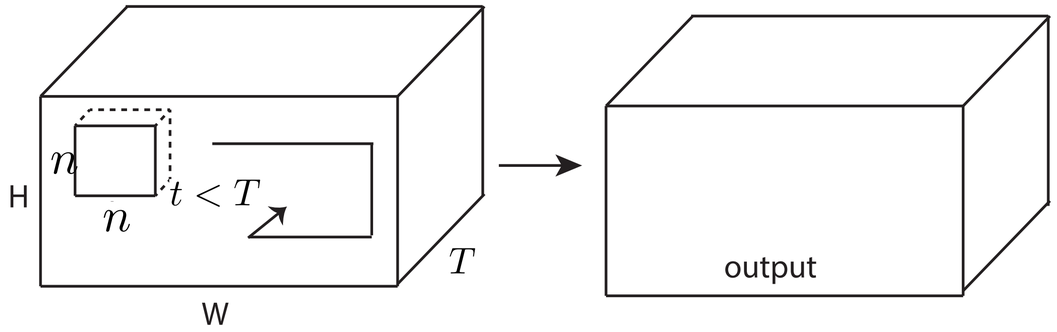
2D convolution on clip of T frames collapses temporal information

2D vs 3D convolution

Simplified illustration based on single input channel:



2D convolution
on clip of T frames
collapses temporal
information



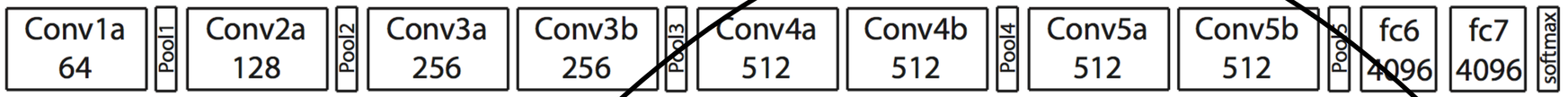
3D convolution
preserves temporal information

[Taylor et al., ECCV10; Le et al., CVPR11;
Ji et al. TPAMI13; Tran et al., ICCV15]

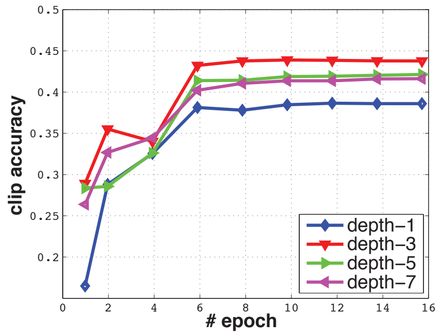
C3D [Tran, Bourdev, Fergus, Torresani, Paluri, ICCV 2015]

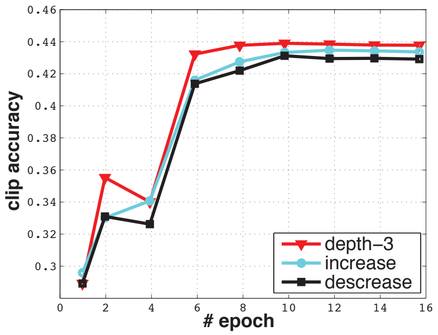
Large-scale training set: Sport1M [Karpathy et al., CVPR2014]
- 1.1M YouTube videos of 487 sport classes
- Train/test on provided split

C3D Architecture



- 16-frame clip as input
- Frames resized to 128x171
- 8 convolutional layers (3x3x3 kernels) + 2 fully connected layers
- 5 2x2x2 max-pooling





C3D: classification accuracy on Sports1M

[Tran, Bourdev, Fergus, Torresani, Paluri, ICCV 2015]

Method	Number of Nets	Clip hit@1
Deep Video's Single-Frame + Multires [19]	3 nets	42.4
Deep Video's Slow Fusion [19]	1 net	41.9
C3D (trained from scratch)	1 net	44.9
C3D (fine-tuned from I380K pre-trained model)	1 net	**46.1**

C3D: classification on Sports1M

[Tran, Bourdev, Fergus, Torresani, Paluri, ICCV 2015]

Facebook AI Research

DARTMOUTH

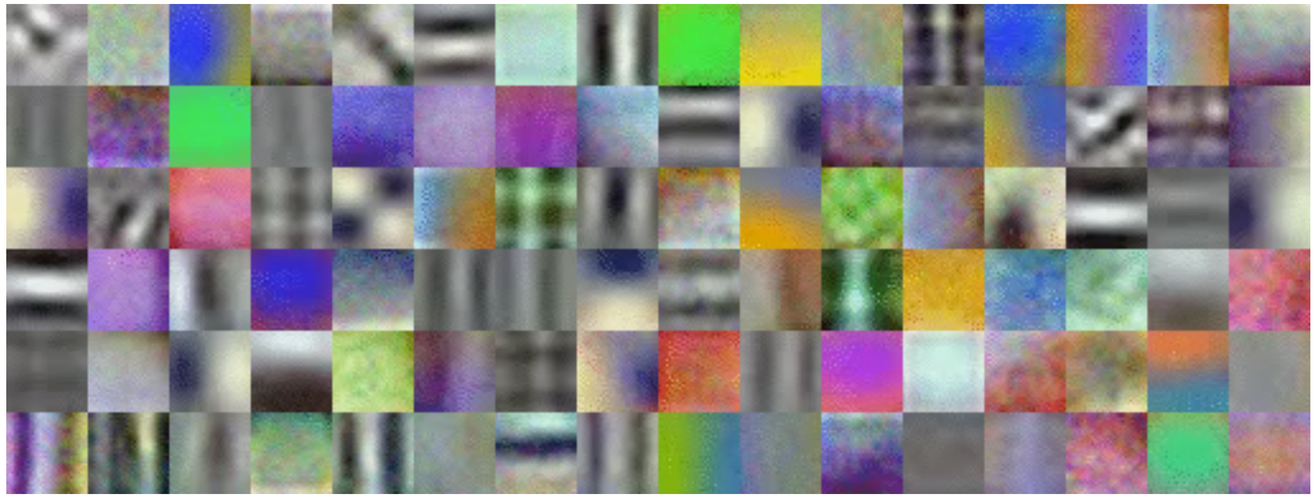
C3D: visualization of low-level features

[Tran, Bourdev, Fergus, Torresani, Paluri, ICCV 2015]

Facebook AI Research

DARTMOUTH

3D filters in 1st layer:



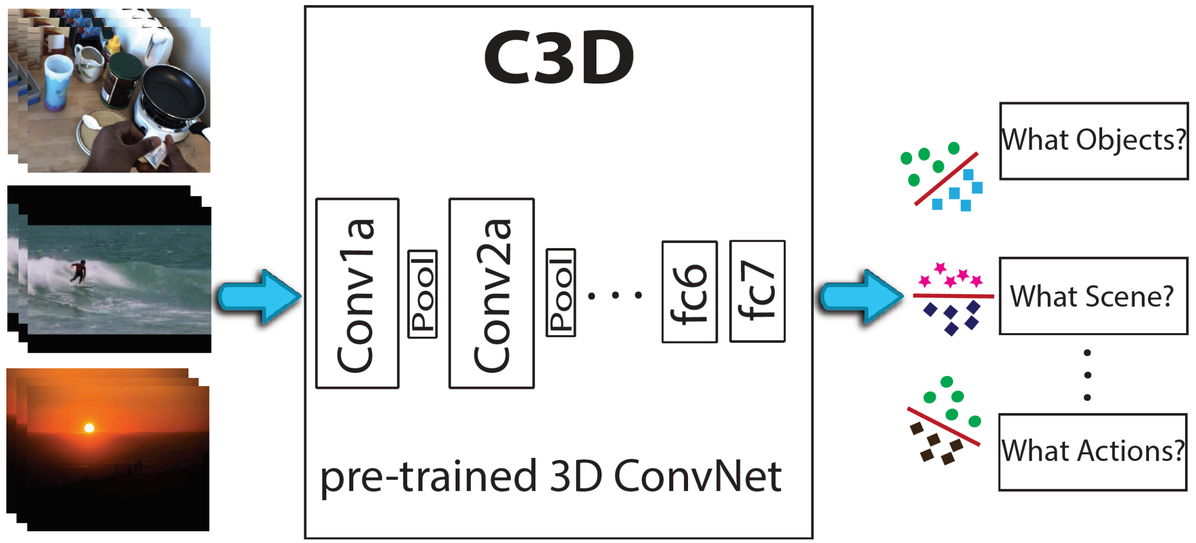
Transfer learning with C3D

Facebook AI Research

[Tran, Bourdev, Fergus, Torresani, Paluri, ICCV 2015]

DARTMOUTH

C3D as generic features:



Test on 4 video recognition tasks using simple linear classifiers trained on C3D features

Transfer learning with C3D

Facebook AI Research

[Tran, Bourdev, Fergus, Torresani, Paluri, ICCV 2015]

DARTMOUTH

Action categorization on UCF101:

linear SVM on iDT and frame-based CNN features

linear SVM on C3D and video-based CNNs from RGB

Method	Accuracy (%)
Imagenet + linear SVM	68.8
iDT w/ BoW + linear SVM	76.2
Deep networks [18]	65.4
Spatial stream network [36]	72.6
LRCN [6]	71.1
LSTM composite model [39]	75.8
C3D (1 net) + linear SVM	82.3
C3D (3 nets) + linear SVM	**85.2**

Transfer learning with C3D

Facebook AI Research

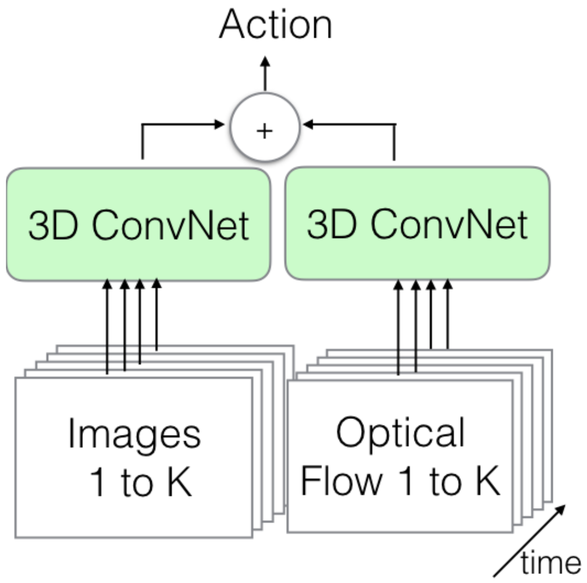
[Tran, Bourdev, Fergus, Torresani, Paluri, ICCV 2015]

DARTMOUTH

Generalization to other video analysis tasks:

ASLAN action similarity labeling	YUPENN scene classification	UMD scene classification	Object object recognition
[31]	[9]	[9]	[32]
68.7	96.2	77.7	12.0
78.3	**98.1**	**87.7**	**22.3**

I3D [Carreira, Zisserman, CVPR 2017]



Large-margin winner of the action recognition and temporal segmentation tracks @ CVPR 2017 Charade challenge

Key-features:

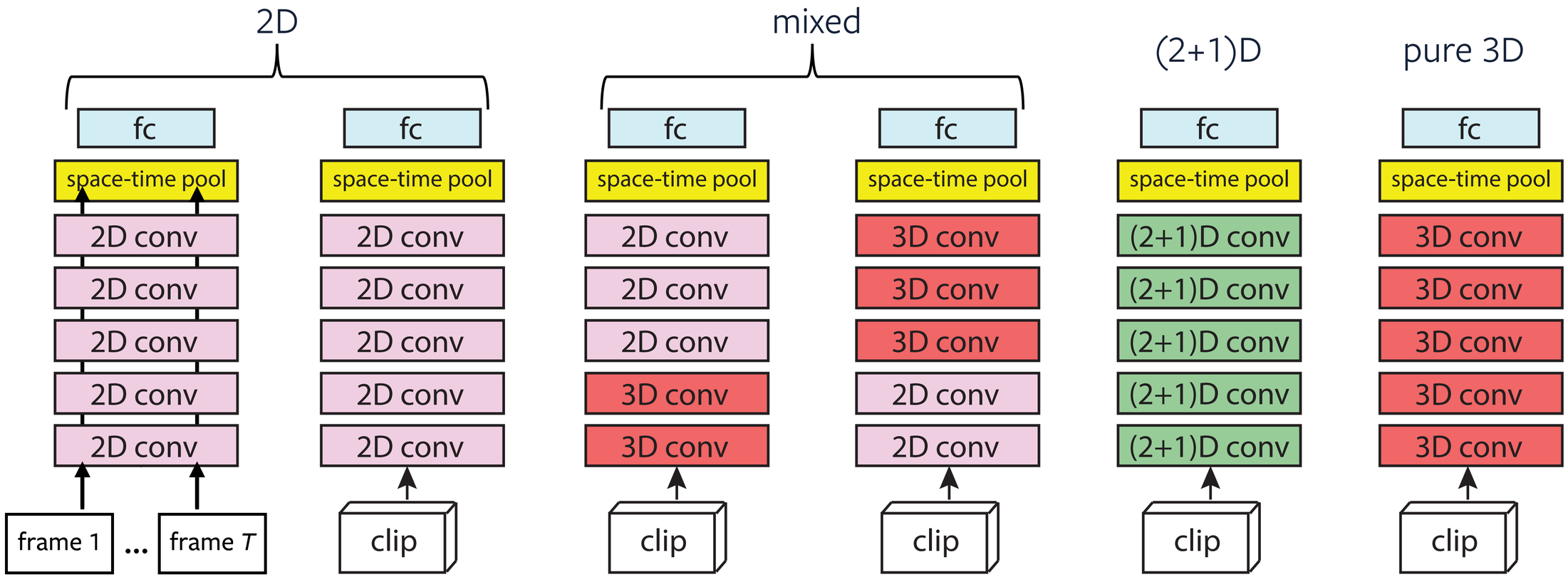
- 2D Inception architecture adopted into a 3D CNN
- 3D filters initialized by “temporally inflating” 2D filters learned from InageNet
- Large-scale training on new Kinetics dataset
  (240K training videos, 400 action classes)
- Two-stream architecture operating on RGB and optical flow

Revisiting the role of 3D convolution in video analysis

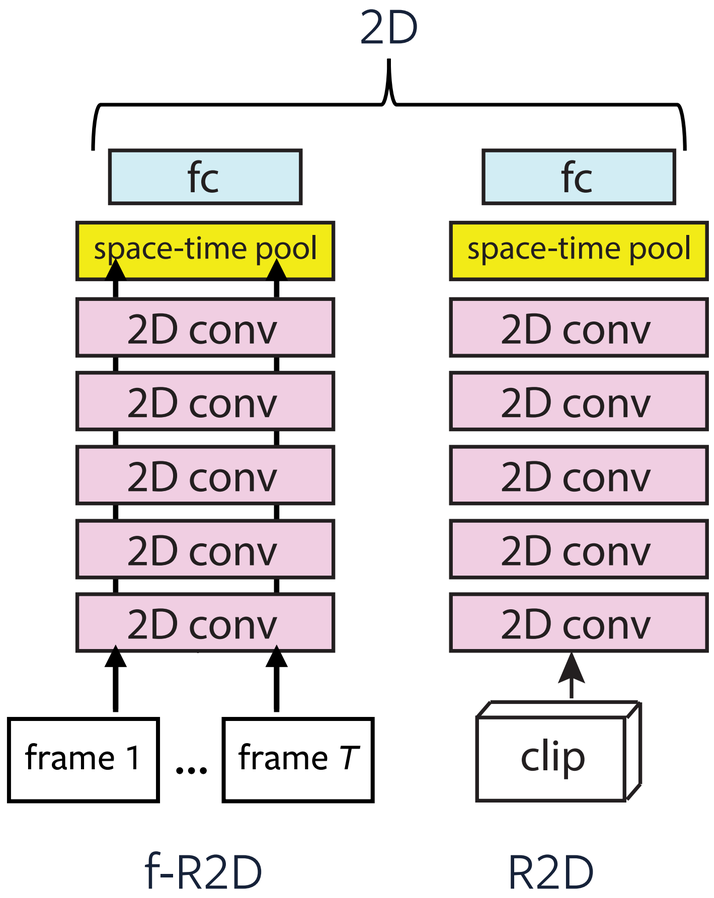
Several recent empirical studies [Qiu et al., CVPR 2017; Tran et al., CVPR 2018; Xie et al., ECCV 2018; Tran et al., arXiv 2019] aimed at addressing several fundamental questions:

- *Do we even need 3D convolution?*
- *If so, what layers should we make 3D, and what layers can be 2D?*
- *Is it beneficial to factorize spatiotemporal filters into disjoint space and time components?*
- *Is it useful to factorize spatiotemporal filters across channels?*

Empirical evaluation of different forms of spatiotemporal convolution [Tran et al., CVPR 2018]

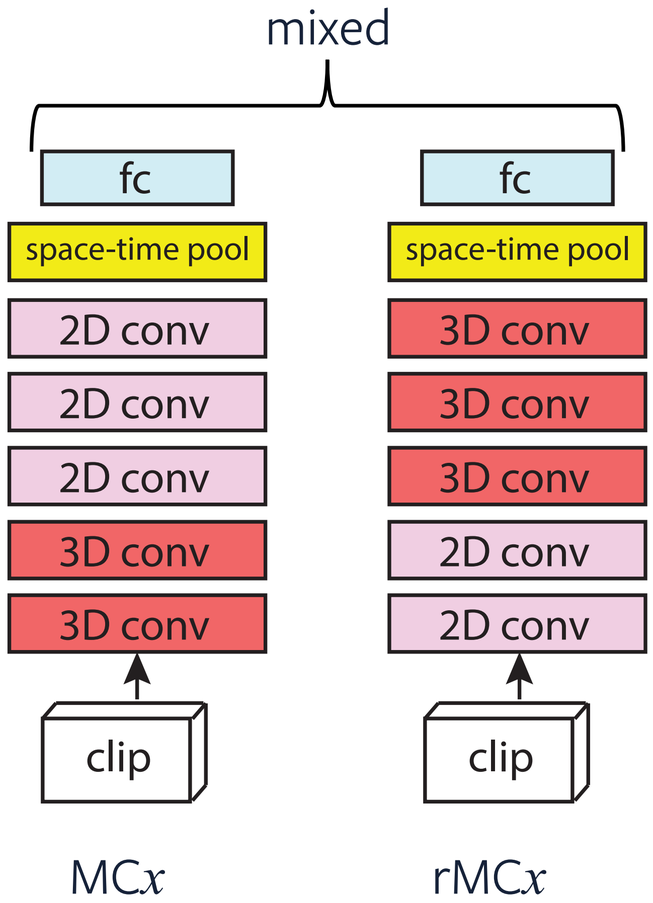


Empirical evaluation of different forms of spatiotemporal convolution [Tran et al., CVPR 2018]



- f-R2D processes the T frames independently
  - ✓ no temporal modeling whatsoever
- R2D treats the the T frames as channels
  - ✓ temporal information collapsed after the first layer

Empirical evaluation of different forms of spatiotemporal convolution [Tran et al., CVPR 2018]



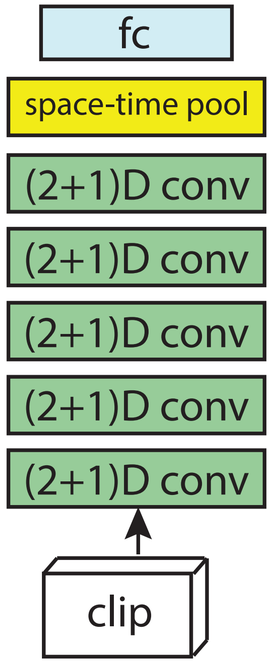
- MCx
  - ✓ 3D convs in first $(x\text{-}1)$ groups, 2D convs in top groups
- rMCx (reversed mixed convolutions)
  - ✓ 2D convs in first $(x\text{-}1)$ groups, 3D convs in top groups

Concurrently studied in [Xie et al., ECCV 2018] within I3D architecture:

- ✓ MCx are called "bottom-heavy" I3D
- ✓ rMCx are called "top-heavy" I3D

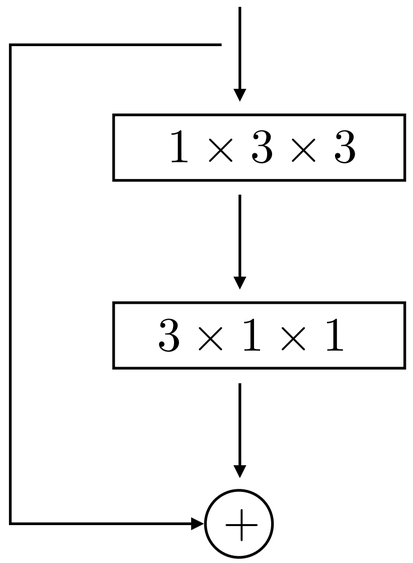
Empirical evaluation of different forms of spatiotemporal convolution [Tran et al., CVPR 2018]

(2+1)D



- (2+1)D: space-time factorization

R(2+1)D

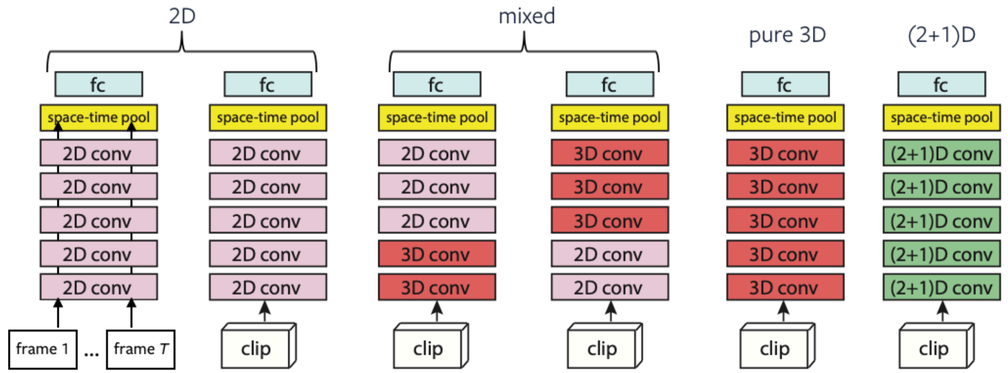


3D conv with 2D space filter

3D conv with 1D temporal filter

A similar space-time factorization was proposed in
[Qiu et al., CVPR 2017] within ResNet bottleneck blocks and in
[Xie et al., ECCV 2018] within I3D architecture

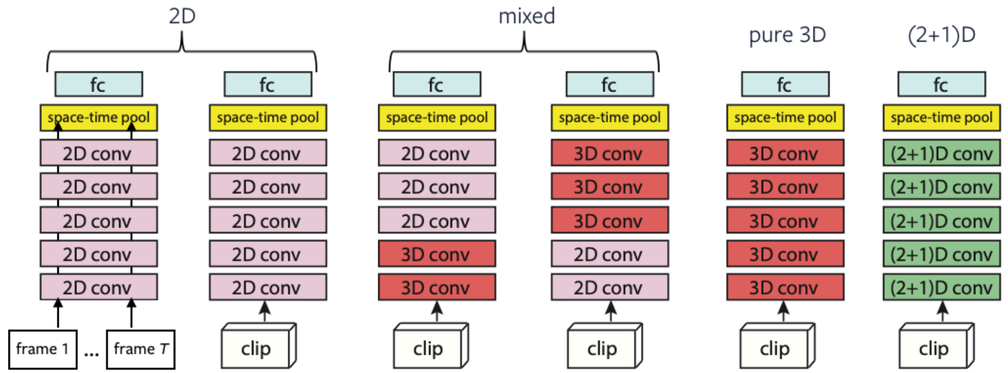
Empirical evaluation of different forms of spatiotemporal convolution [Tran et al., CVPR 2018]



Results on Kinetics-400 using ResNets of 18 layers:

Net	# params	Clip@1	Video@1
Input		16×112×112	
R2D	11.4M	47.0	58.9
f-R2D	11.4M	50.3	60.5
R3D	33.4M	52.5	64.2
MC2	11.4M	53.1	64.2
MC3	11.7M	53.7	64.7
MC4	12.7M	53.7	65.1
MC5	16.9M	53.7	65.1
rMC2	33.3M	53.1	64.9
rMC3	33.0M	53.2	65.0
rMC4	32.0M	53.4	65.1
rMC5	27.9M	52.1	63.1
R(2+1)D	33.3M	**56.8**	**68.0**

Empirical evaluation of different forms of spatiotemporal convolution [Tran et al., CVPR 2018]

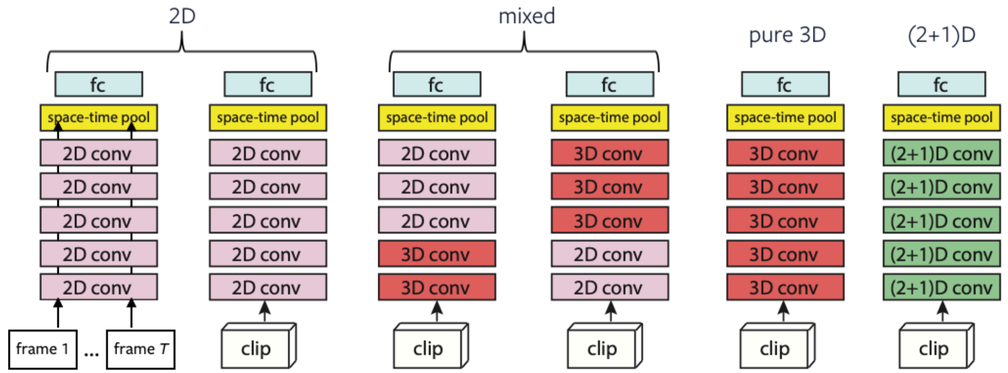


Results on Kinetics-400 using ResNets of 18 layers:

Net	# params	Clip@1	Video@1
Input		16×112×112	
R2D	11.4M	47.0	58.9
f-R2D	11.4M	50.3	60.5
R3D	33.4M	52.5	64.2
MC2	11.4M	53.1	64.2
MC3	11.7M	53.7	64.7
MC4	12.7M	53.7	65.1
MC5	16.9M	53.7	65.1
rMC2	33.3M	53.1	64.9
rMC3	33.0M	53.2	65.0
rMC4	32.0M	53.4	65.1
rMC5	27.9M	52.1	63.1
R(2+1)D	33.3M	**56.8**	**68.0**

Big accuracy gap between 2D and 3D CNNs

Empirical evaluation of different forms of spatiotemporal convolution [Tran et al., CVPR 2018]

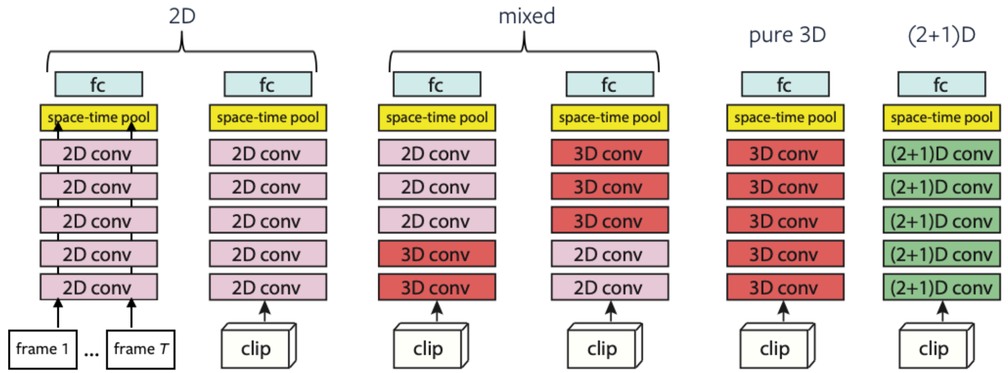


Results on Kinetics-400 using ResNets of 18 layers:

Net	# params	Clip@1	Video@1
Input		16×112×112	
R2D	11.4M	47.0	58.9
f-R2D	11.4M	50.3	60.5
R3D	33.4M	52.5	64.2
MC2	11.4M	53.1	64.2
MC3	11.7M	53.7	64.7
MC4	12.7M	53.7	65.1
MC5	16.9M	53.7	65.1
rMC2	33.3M	53.1	64.9
rMC3	33.0M	53.2	65.0
rMC4	32.0M	53.4	65.1
rMC5	27.9M	52.1	63.1
R(2+1)D	33.3M	**56.8**	**68.0**

R(2+1)D outperforms R3D and all other 3D CNNs

Empirical evaluation of different forms of spatiotemporal convolution [Tran et al., CVPR 2018]



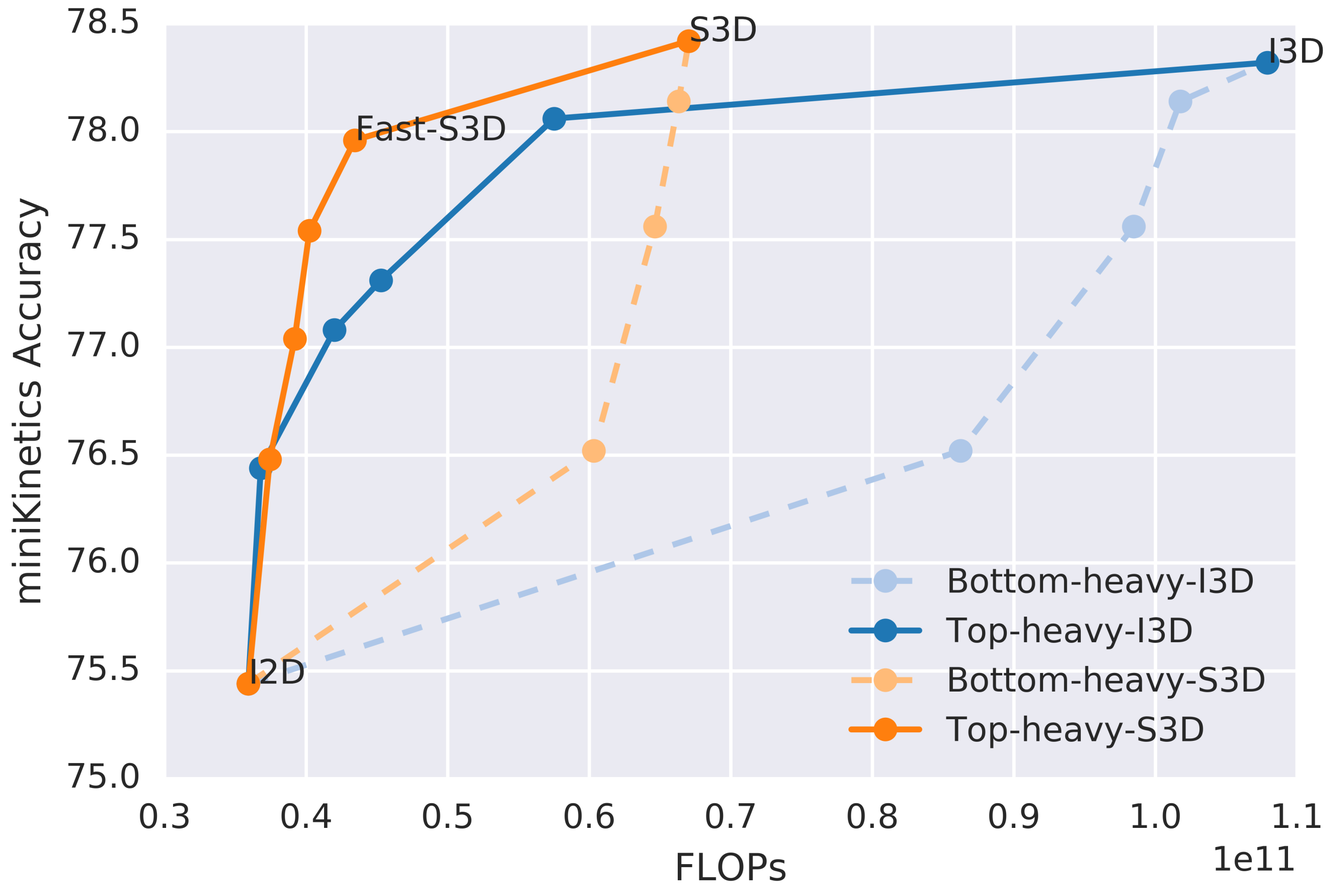
Results on Kinetics-400 using ResNets of 18 layers:

Net	# params	Clip@1	Video@1
Input		16×112×112	
R2D	11.4M	47.0	58.9
f-R2D	11.4M	50.3	60.5
R3D	33.4M	52.5	64.2
MC2	11.4M	53.1	64.2
MC3	11.7M	53.7	64.7
MC4	12.7M	53.7	65.1
MC5	16.9M	53.7	65.1
rMC2	33.3M	53.1	64.9
rMC3	33.0M	53.2	65.0
rMC4	32.0M	53.4	65.1
rMC5	27.9M	52.1	63.1
R(2+1)D	33.3M	**56.8**	**68.0**

Mixed 2D/3D CNNs do better than pure 3D

Empirical evaluation of different forms of spatiotemporal convolution

Independent results in [Xie et al., ECCV 2018] confirm these findings:

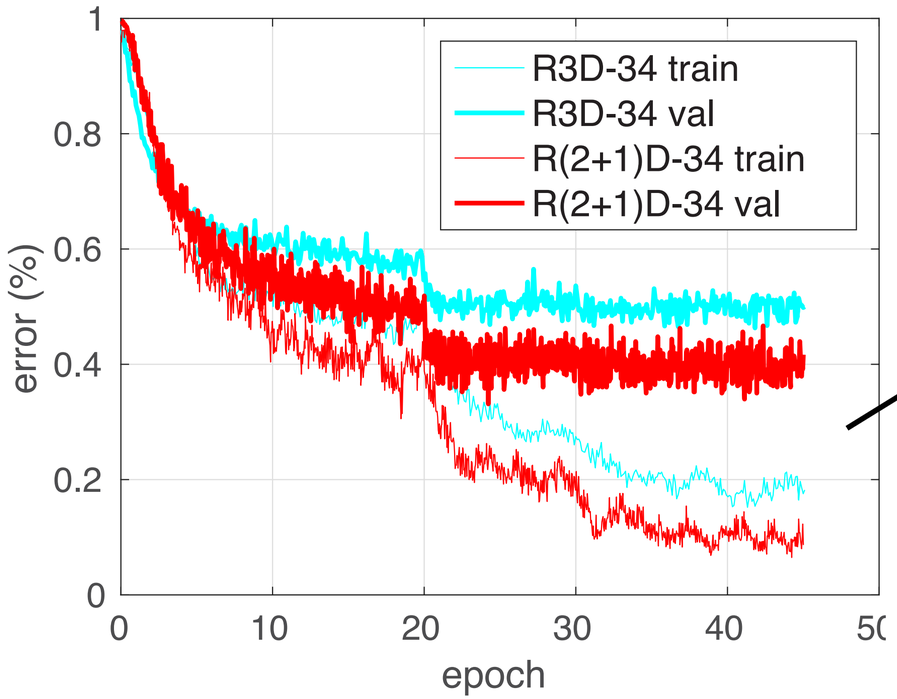


- Orange vs blue:
  S3D (space-time factorized I3D) outperforms I3D

- Solid vs dotted:
  top-heavy mixed convolutions do better than bottom-heavy mixed convolutions for same # FLOPs

Why is space-time factorization beneficial? [Tran et al., CVPR 2018]

- For the same number of parameters, (2+1)D factorization doubles number of nonlinearities (additional ReLU between spatial and temporal filtering)
- Space-time factorization renders optimization easier:

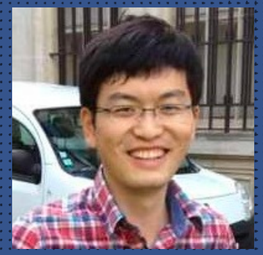


(2+1)D factorization
lowers the **training error**
in addition to the test error

facebook
Artificial Intelligence Research

Channel-Separated 3D Networks


Du Tran


Heng Wang

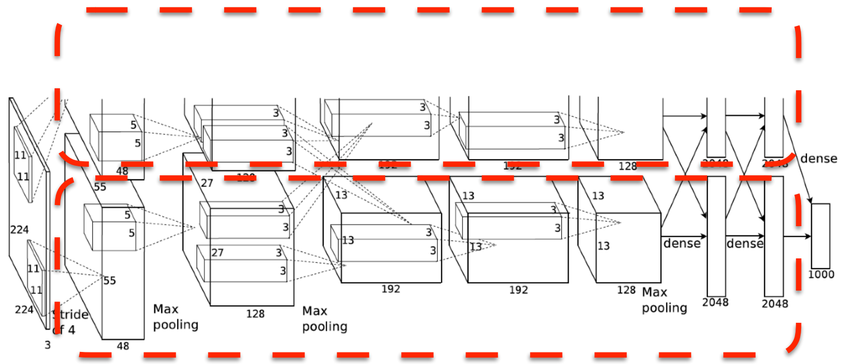

Lorenzo Torresani


Matt Feiszli

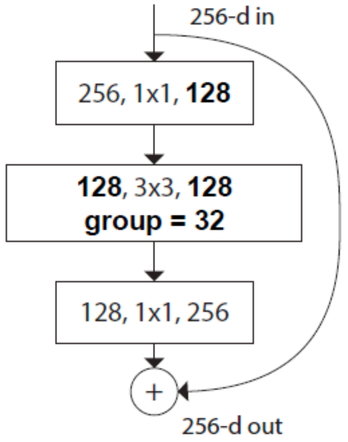
Preprint available at https://arxiv.org/abs/1904.02811

group convolution: prior work

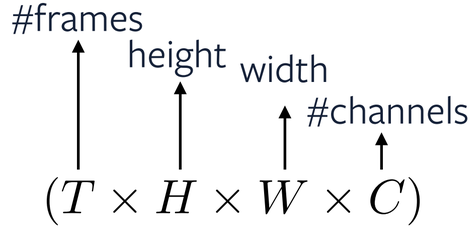
- To reduce #parameters and #FLOPS of 2D CNNs:
  - ✓ Adopted in AlexNet [Krizhevsky et al., NIPS12] to overcome GPU memory limits
  - ✓ Frequently used for mobile application, e.g., in MobileNets [Howard et al., arXiv2017], Xception [Chollet et al., CVPR17], ShuffleNet [Zhang et al. CVPR18]



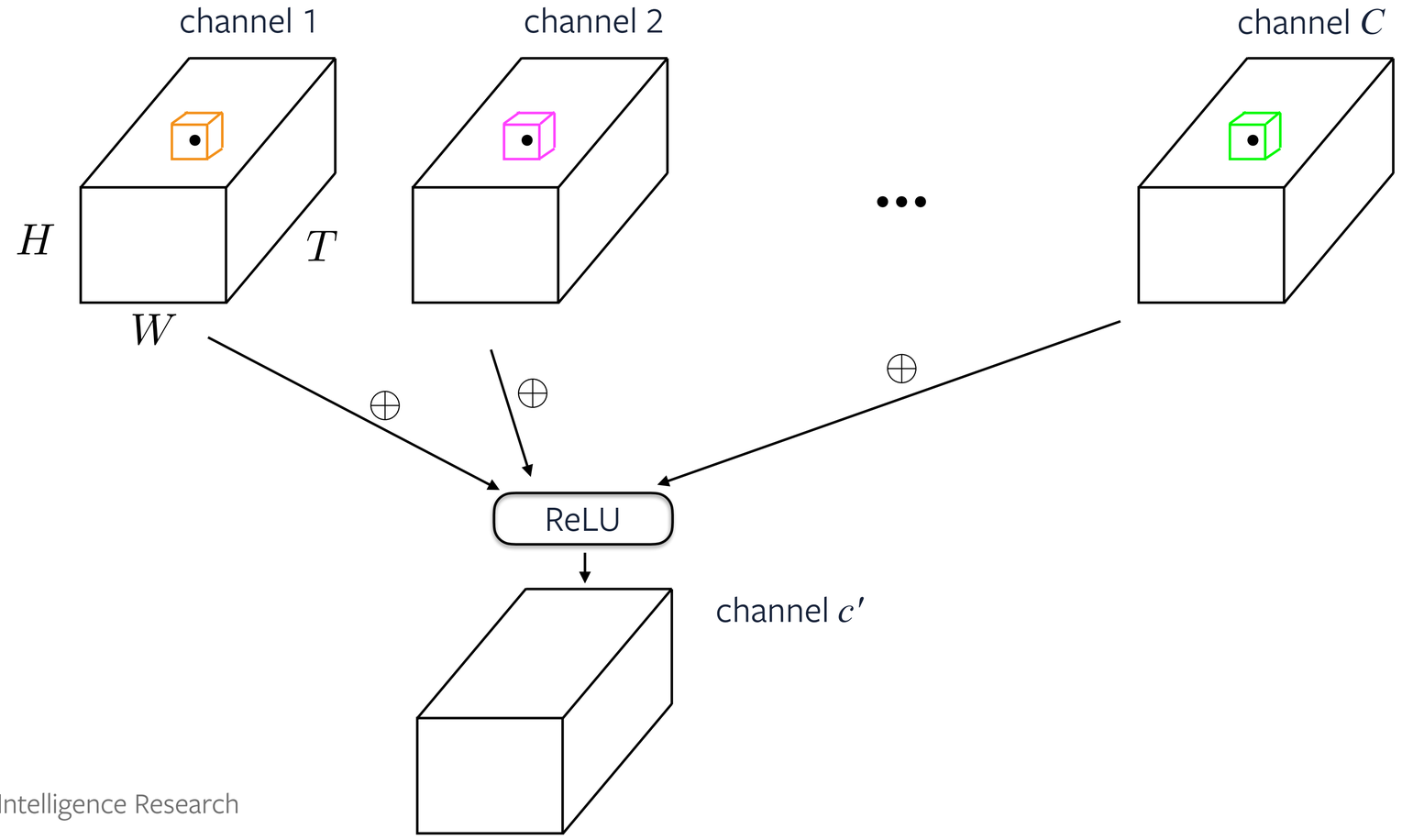
- To improve accuracy of 2D CNNs:
  - ✓ ResNeXt [Xie et al., CVPR17] is a ResNet [He at al., CVPR16] with group convolutions, yielding better accuracy for the same number of parameters

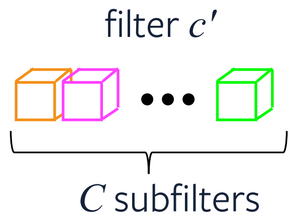


3D convolution



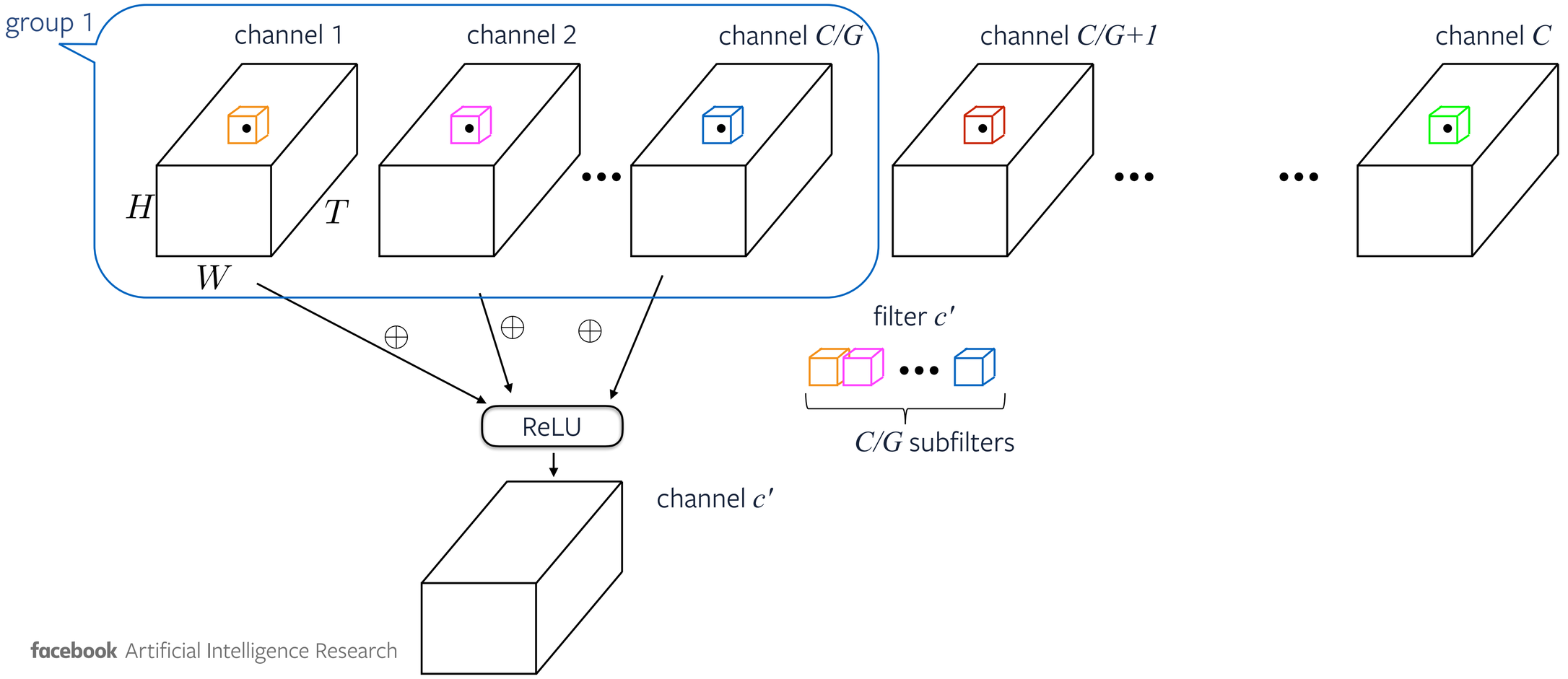
- Each filter operates on all C channels of the 4D spatiotemporal tensor
- Each filter consists of C 3D subfilters $(n \times n \times n)$, each applied to a 3D spatiotemporal channel $(T \times H \times W)$





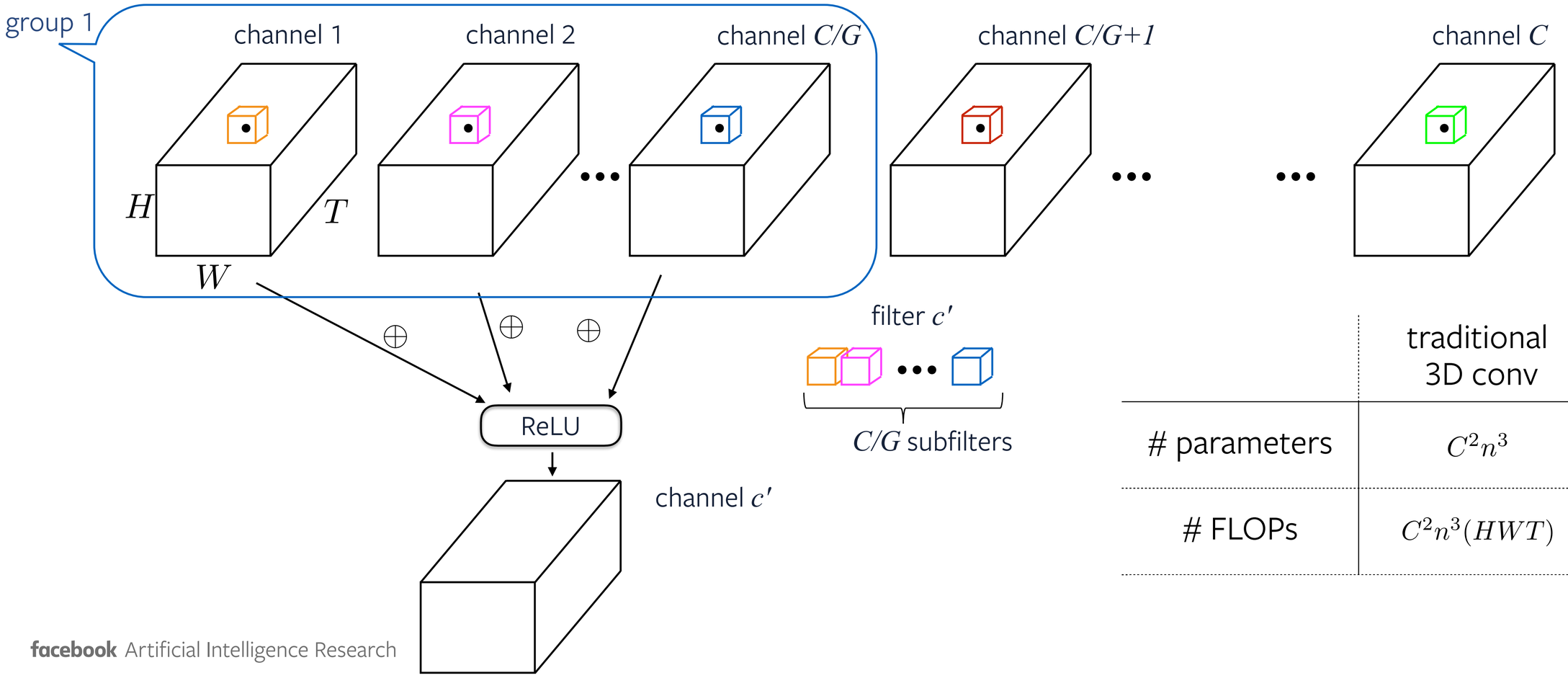
group 3D convolution

- Each filter operates on a subset of $\frac{C}{G}$ channels (G denotes # groups):
  each filter consists of $\frac{C}{G}$ 3D subfilters $(n \times n \times n)$, each applied to a 3D spatiotemporal channel $(T \times H \times W)$



group 3D convolution

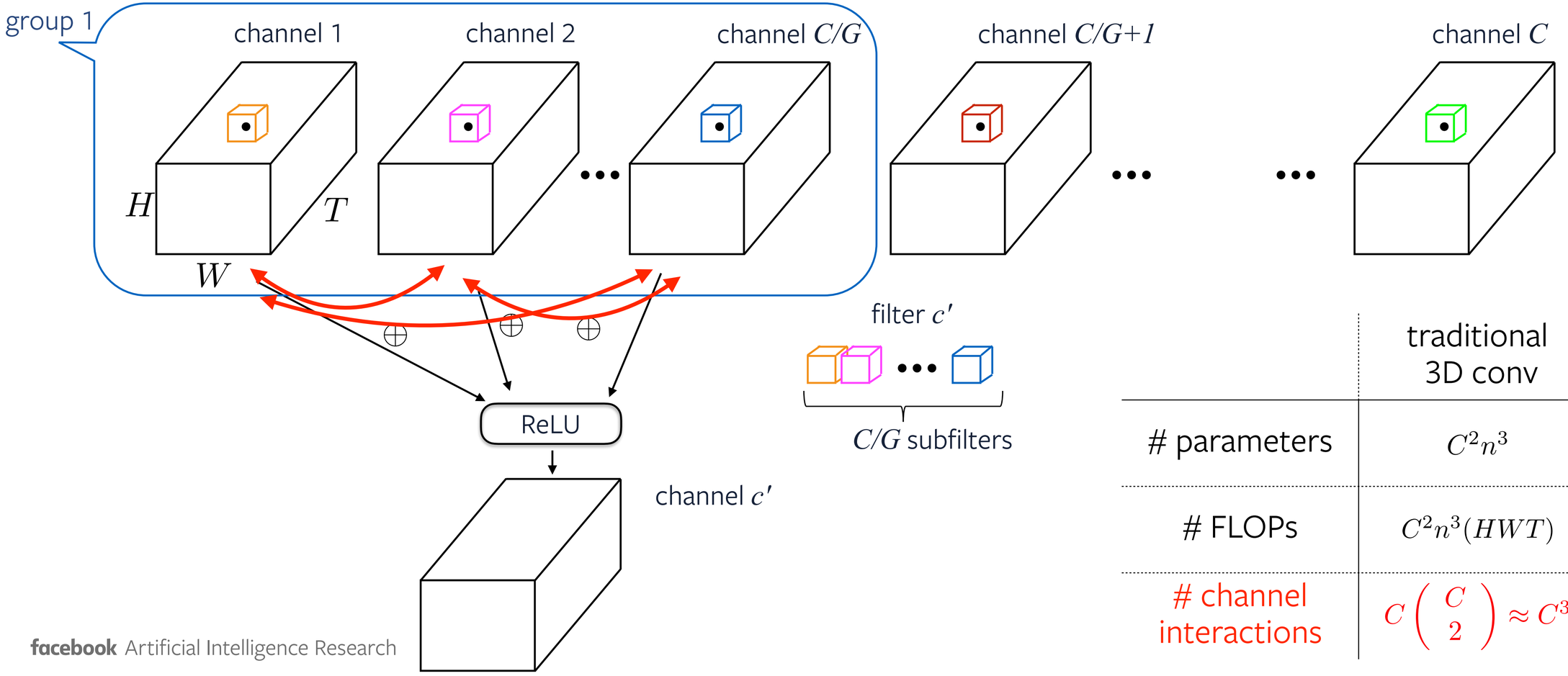
- Each filter operates on a subset of $\frac{C}{G}$ channels (G denotes # groups):
  each filter consists of $\frac{C}{G}$ 3D subfilters $(n \times n \times n)$, each applied to a 3D spatiotemporal channel $(T \times H \times W)$



	traditional 3D conv	group 3D conv
# parameters	C^2n^3	$\frac{C^2n^3}{G}$
# FLOPs	$C^2n^3(HWT)$	$\frac{C^2n^3(HWT)}{G}$

group 3D convolution

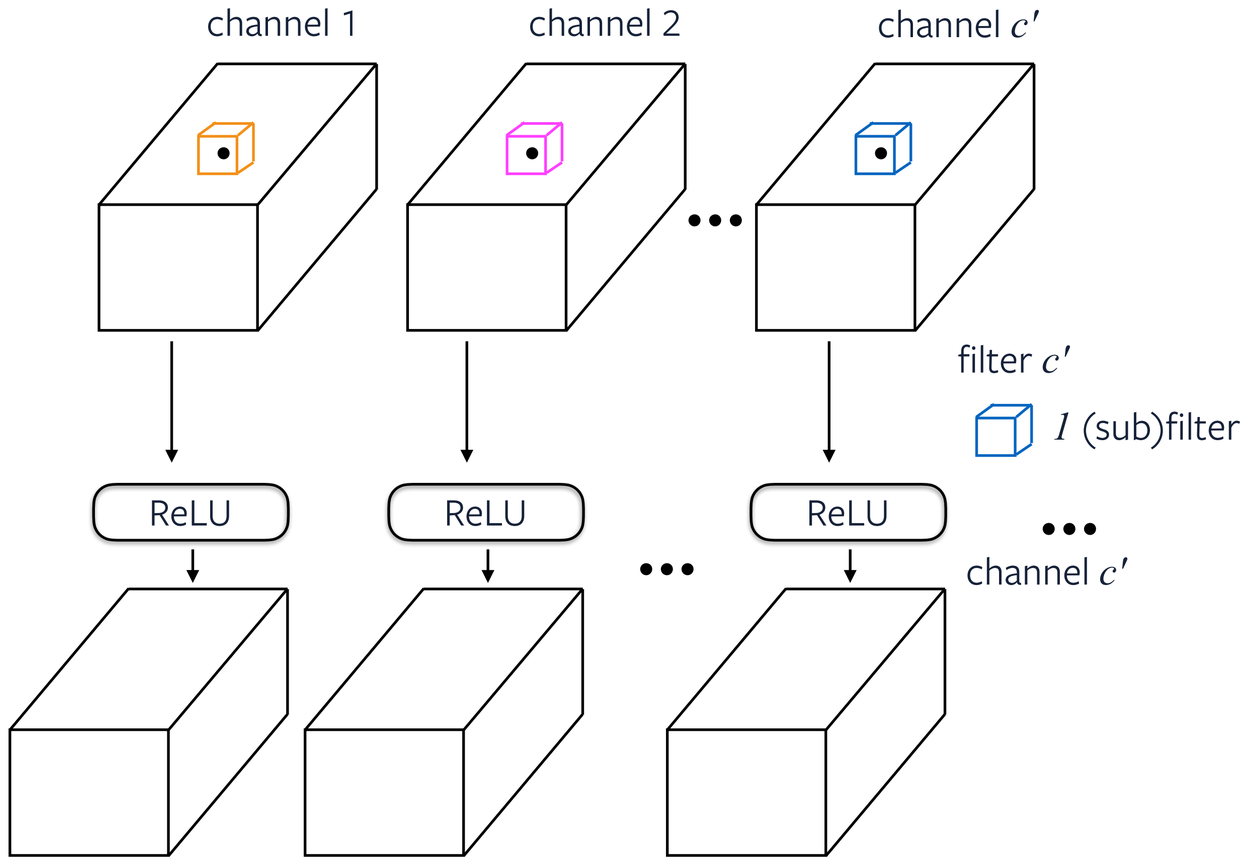
- Each filter operates on a subset of $\frac{C}{G}$ channels (G denotes # groups):
  each filter consists of $\frac{C}{G}$ 3D subfilters $(n \times n \times n)$, each applied to a 3D spatiotemporal channel $(T \times H \times W)$

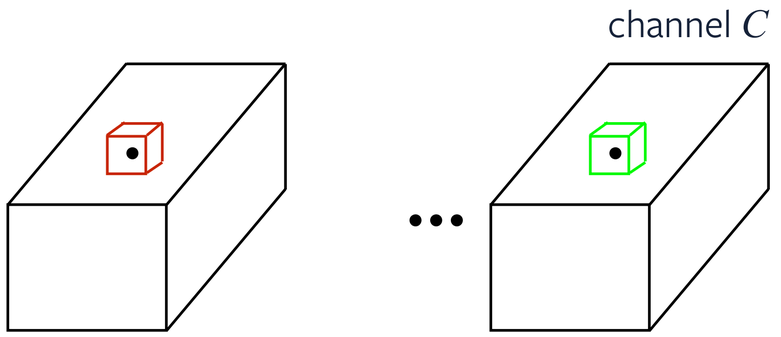


	traditional 3D conv	group 3D conv
# parameters	C^2n^3	$\frac{C^2n^3}{G}$
# FLOPs	$C^2n^3(HWT)$	$\frac{C^2n^3(HWT)}{G}$
# channel interactions	$C\binom{C}{2} \approx C^3$	$C\binom{C/G}{2} \approx \frac{C^3}{G^2}$

channel-separated 3D convolution

- Set # groups $G = C$ → each filter operates on *one* channel only

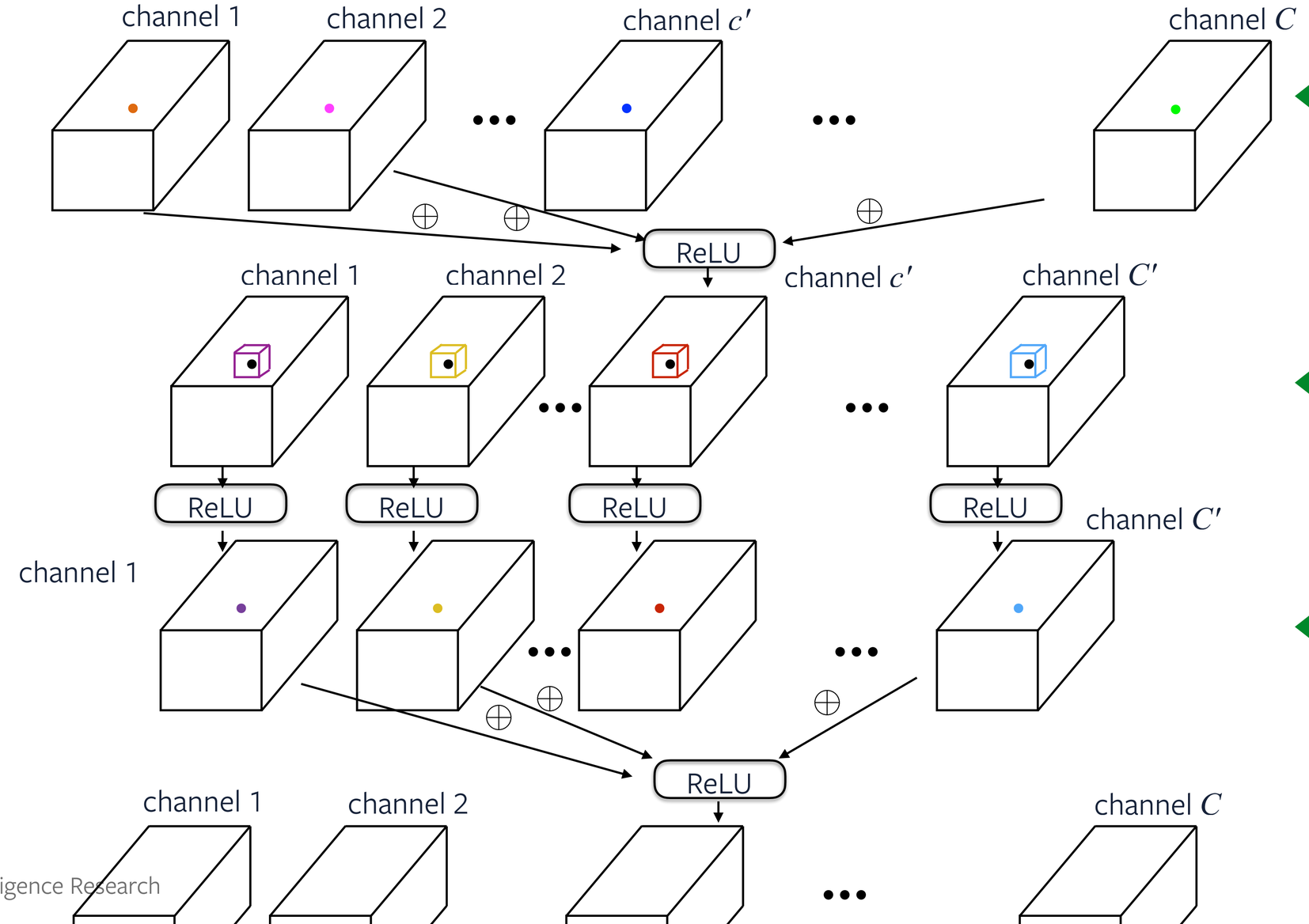




	traditional 3D conv	group 3D conv	channel-separated 3D conv
# parameters	C^2n^3	$\frac{C^2n^3}{G}$	Cn^3
# FLOPs	$C^2n^3(HWT)$	$\frac{C^2n^3(HWT)}{G}$	$Cn^3(HWT)$
# channel interactions	$C\binom{C}{2} \approx C^3$	$C\binom{C/G}{2} \approx \frac{C^3}{G^2}$	0

channel-separated 3D network

- Use point-wise $1 \times 1 \times 1 \times C$ convs to restore some channel interactions before & after channel separation:



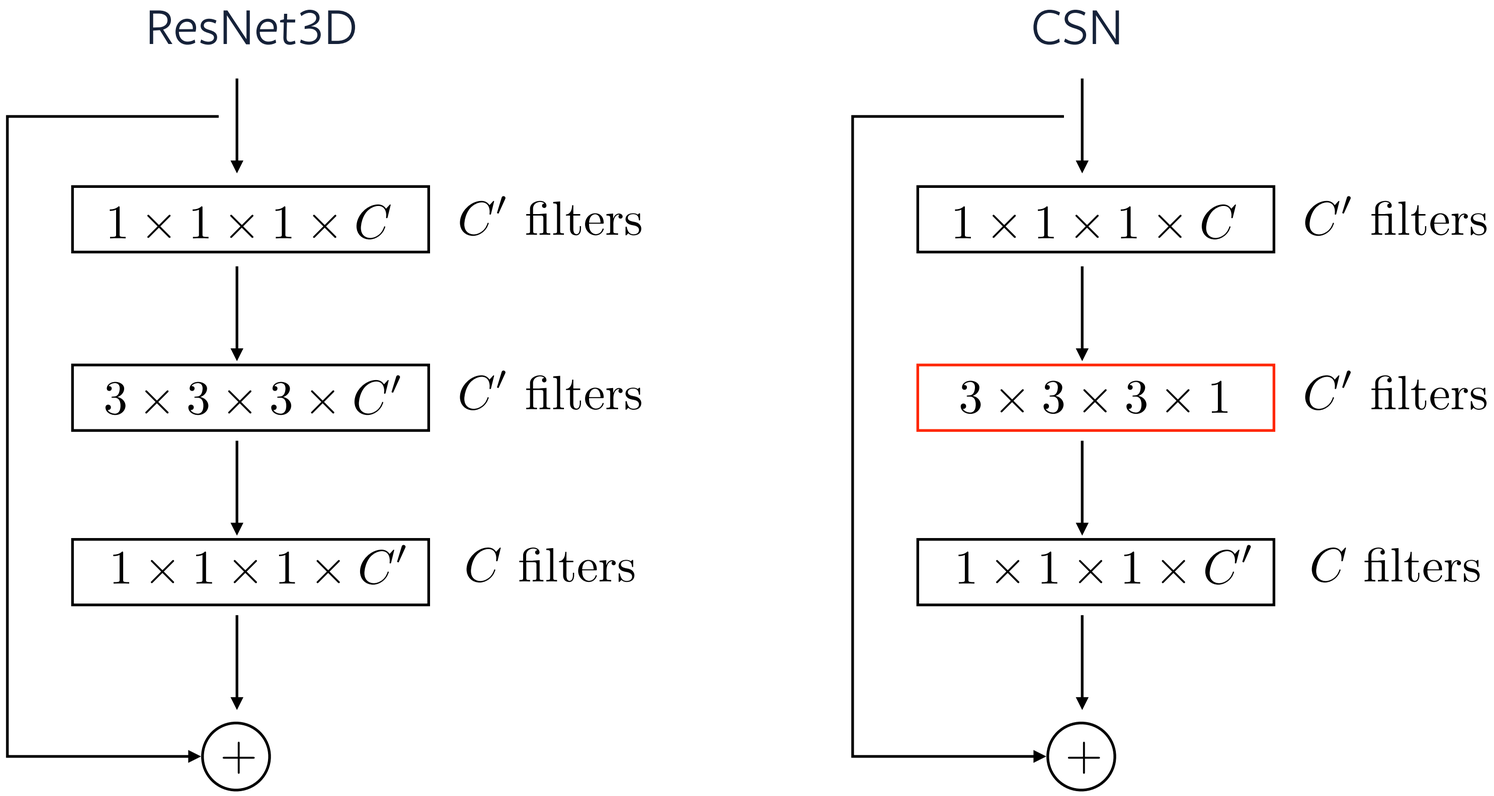
point-wise $1 \times 1 \times 1 \times C$ convs across *all* channels

channel-separated $n \times n \times n \times 1$ 3D convs

point-wise $1 \times 1 \times 1 \times C'$ convs across *all* channels

CSN using ResNet blocks

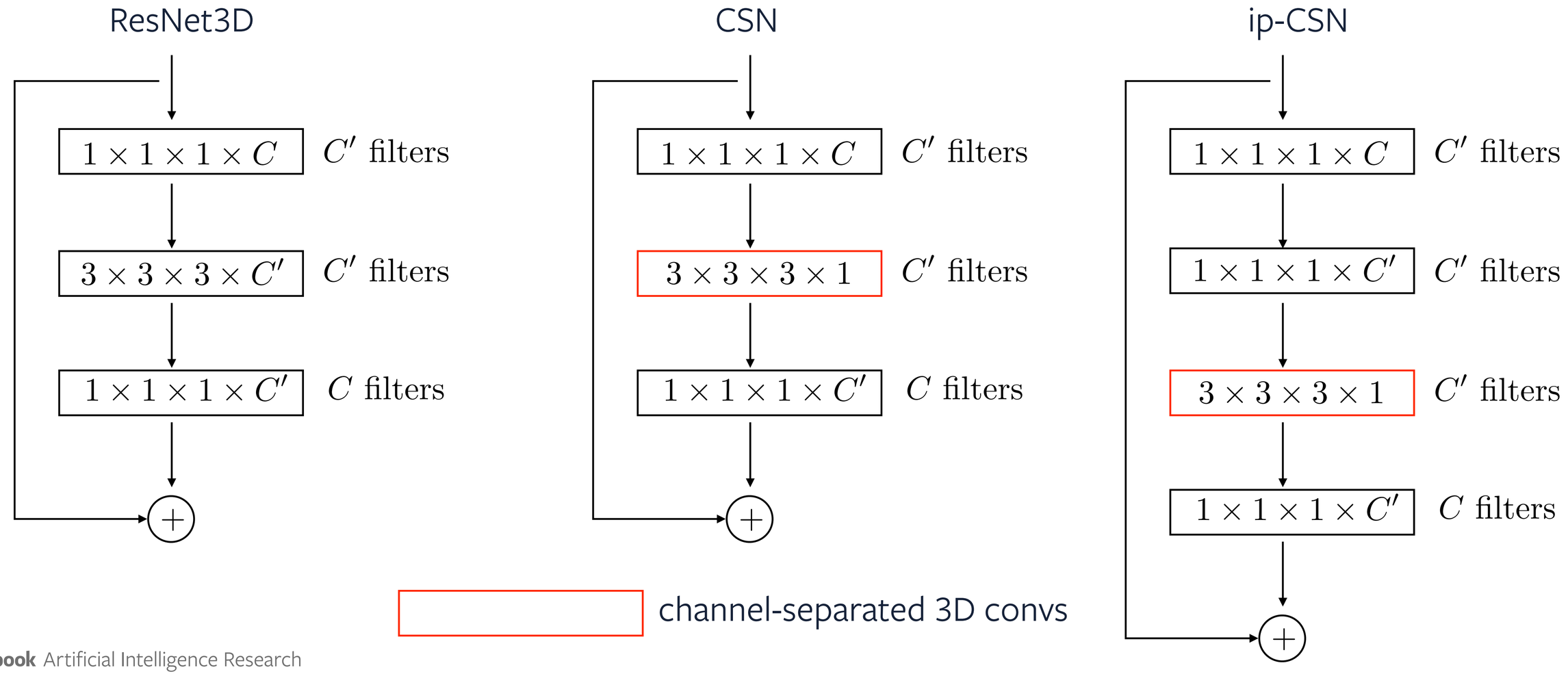
- Comparison between Channel-Separated Network (CSN) and ResNet3D
- Both using ResNet bottleneck block but CSN performs channel-separated 3D convs in all blocks



channel-separated 3D convs

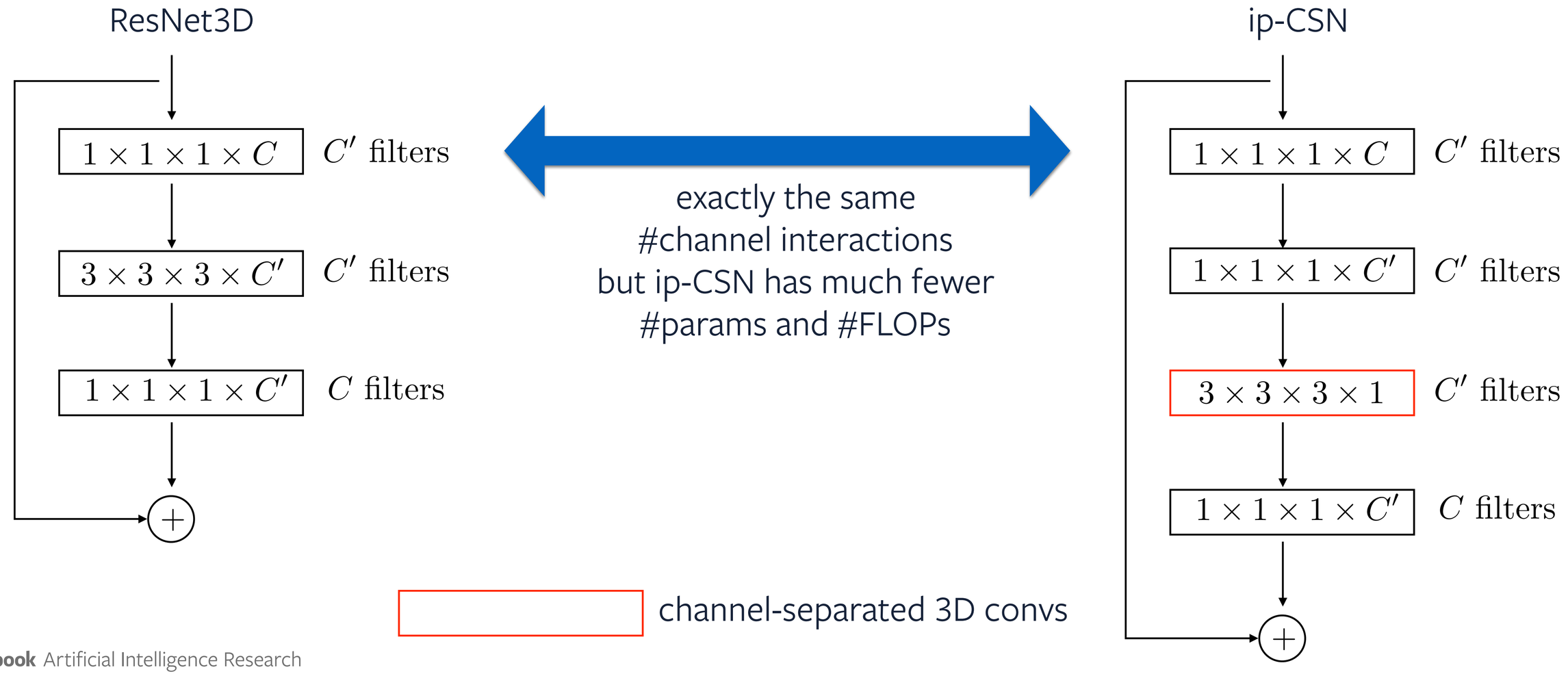
CSN using ResNet blocks

- Comparison between Channel-Separated Network (CSN) and ResNet3D
- Both using ResNet bottleneck block but CSN performs channel-separated 3D convs in all blocks



CSN using ResNet blocks

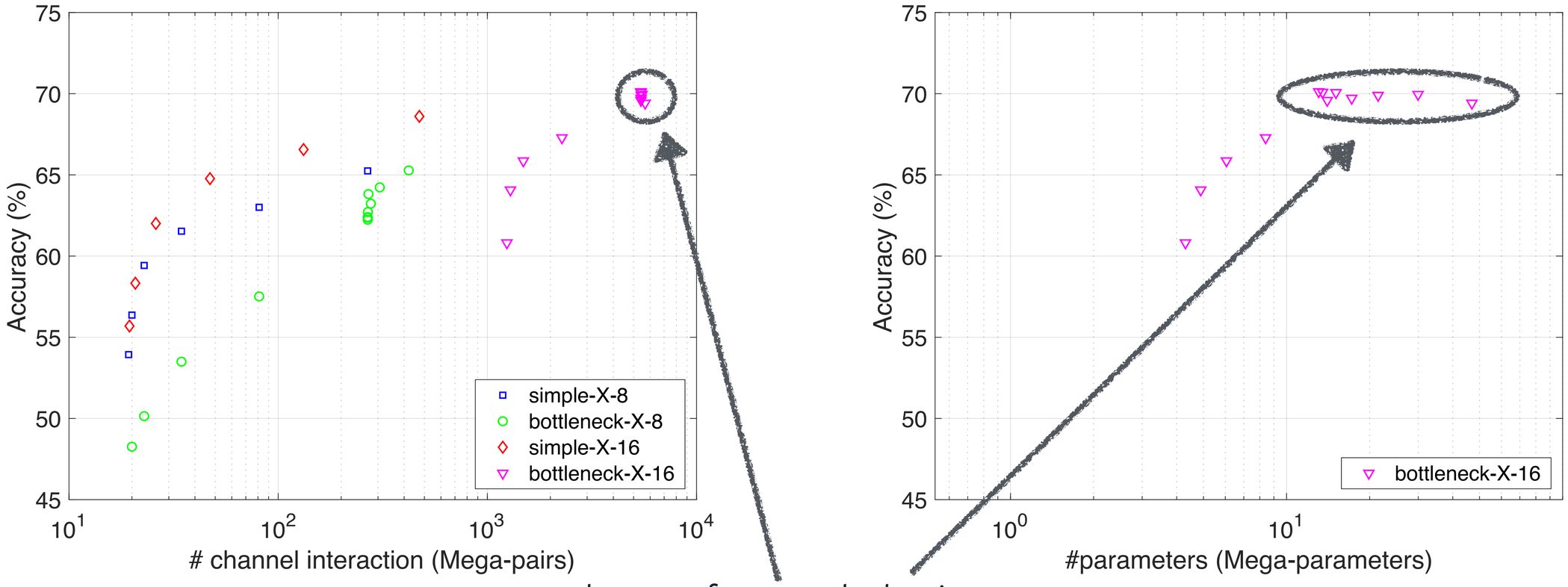
- Comparison between Channel-Separated Network (CSN) and ResNet3D
- Both using ResNet bottleneck block but CSN performs channel-separated 3D convs in all blocks



Experimental Comparison

- Results on Kinetics-400:

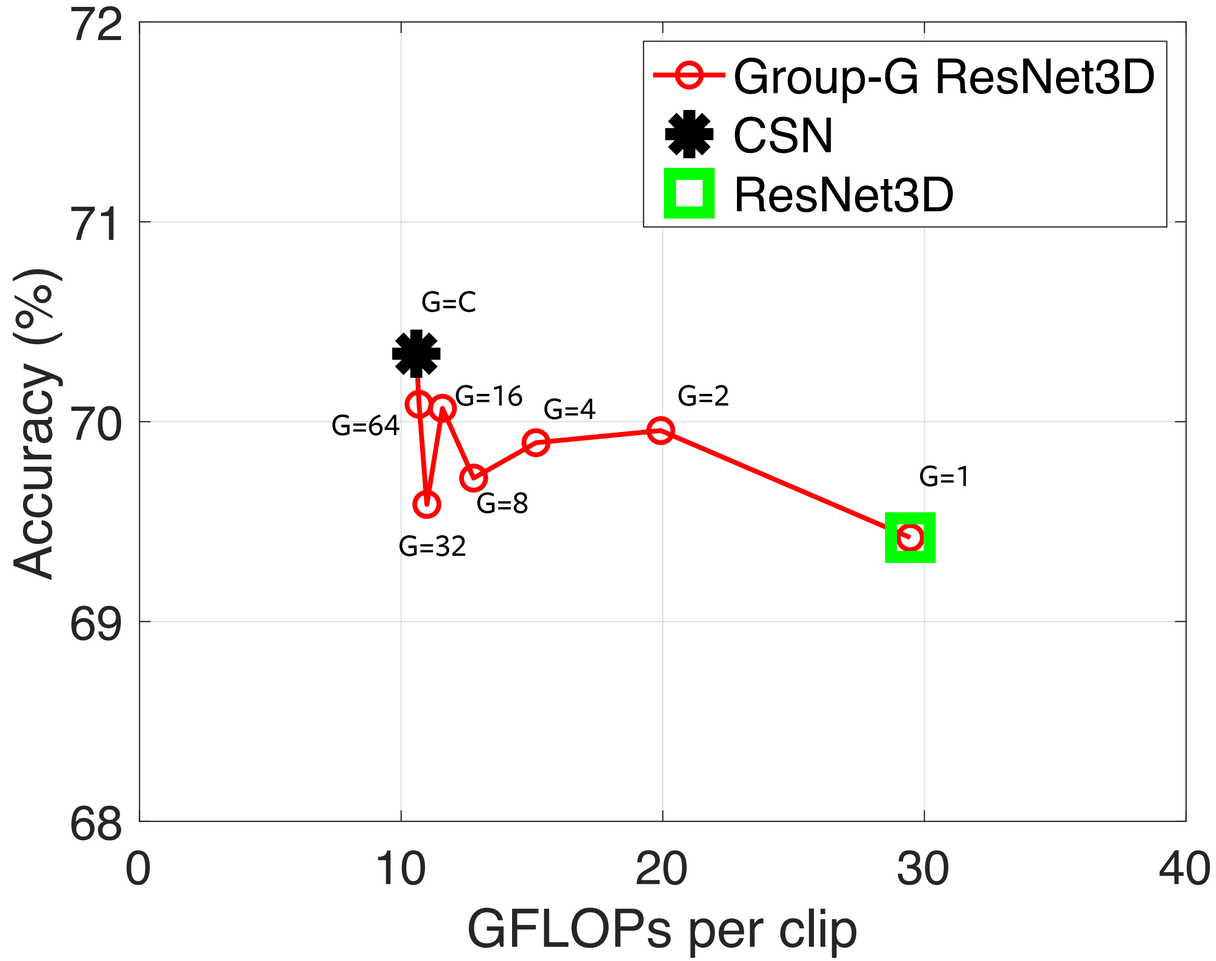
  "# channel interaction is a better predictor of accuracy than #parameters"



cluster of networks having
similar # channel interactions
but largely different #parameters

Experimental Comparison

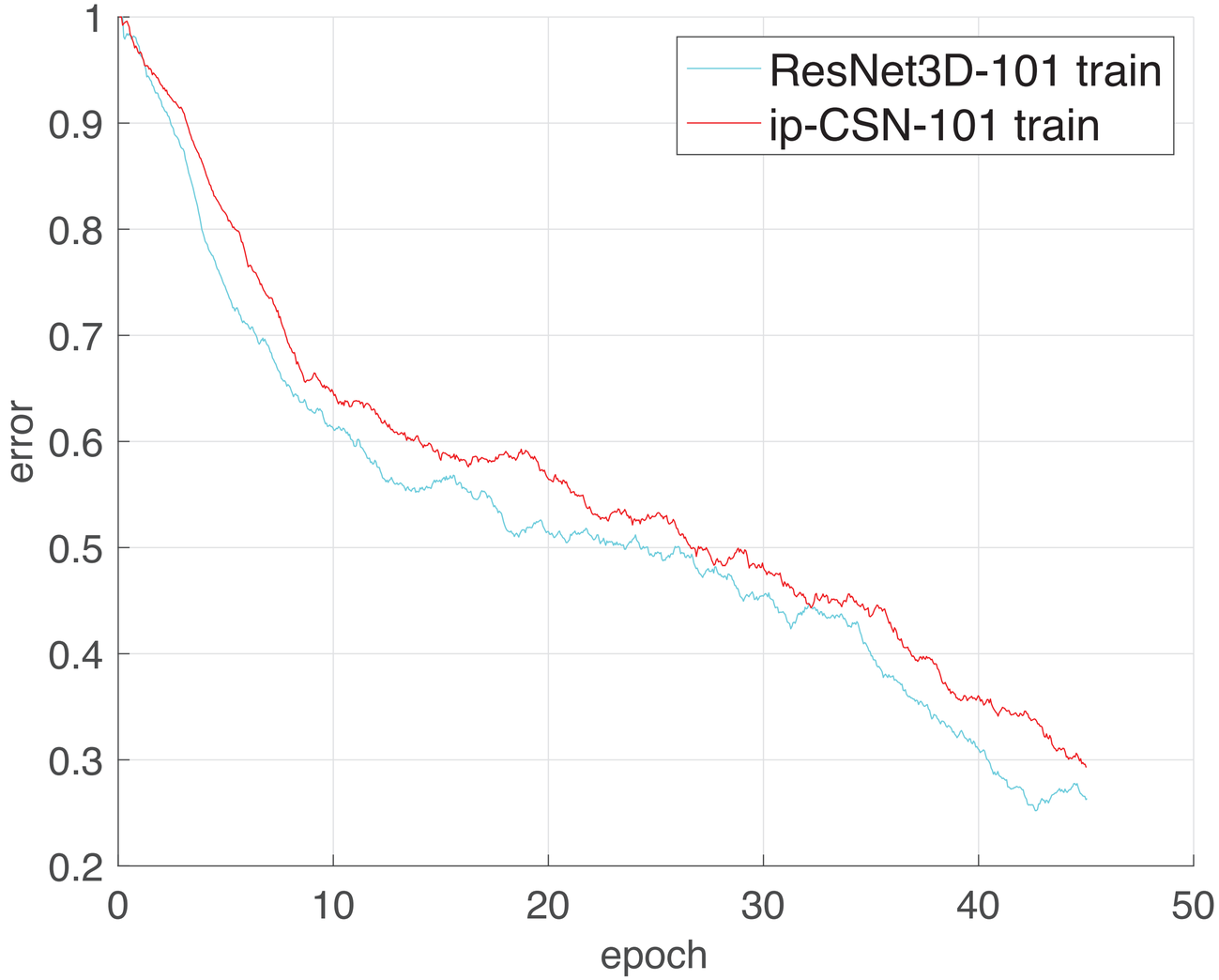
"Channel-Separated 3D Convs > Group 3D Convs > Traditional 3D Convs"



Results on Kinetics-400 using nets of 50 layers

What makes CSNs work better?

- Lower test error and higher training error: channel separation acts as a regularizer



Comparison with the State-of-the-Art

- Results on Kinetics-400:

Method	pretrain	video@1	GFLOPs×crops
ResNeXt	none	65.1	NA
ARTNet(d)	none	69.2	24×250
I3D	ImageNet	71.1	108×dense
TSM	ImageNet	72.5	65×NA
MFNet	ImageNet	72.8	11×NA
Inception-ResNet	ImageNet	73.0	NA
R(2+1)D	Sports1M	74.3	152×dense
A^2-Net	ImageNet	74.6	41×NA
S3D-G	ImageNet	74.7	71×dense
D3D	ImageNet	75.9	NA
GloRe	ImageNet	76.1	55×NA
I3D+NLN	ImageNet	77.7	359×30
SlowFast	none	78.9	213×30
SlowFast+NLN	none	**79.8**	234×30
CSN-152	Sports1M	79.0	96.7×30
ip-CSN-152	Sports1M	**79.2**	108.8×30

Comparison with the State-of-the-Art

- Results on Kinetics-400:

Method	pretrain	video@1	GFLOPs×crops
ResNeXt	none	65.1	NA
ARTNet(d)	none	69.2	24×250
I3D	ImageNet	71.1	108×dense
TSM	ImageNet	72.5	65×NA
MFNet	ImageNet	72.8	11×NA
Inception-ResNet	ImageNet	73.0	NA
R(2+1)D	Sports1M	74.3	152×dense
A^2-Net	ImageNet	74.6	41×NA
S3D-G	ImageNet	74.7	71×dense
D3D	ImageNet	75.9	NA
GloRe	ImageNet	76.1	55×NA
I3D+NLN	ImageNet	77.7	359×30
SlowFast	none	78.9	213×30
SlowFast+NLN	none	**79.8**	234×30
CSN-152	Sports1M	79.0	96.7×30
ip-CSN-152	Sports1M	**79.2**	108.8×30

better than
I3D + Non-Local Net

Comparison with the State-of-the-Art

- Results on Kinetics-400:

Method	pretrain	video@1	GFLOPs×crops
ResNeXt	none	65.1	NA
ARTNet(d)	none	69.2	24×250
I3D	ImageNet	71.1	108×dense
TSM	ImageNet	72.5	65×NA
MFNet	ImageNet	72.8	11×NA
Inception-ResNet	ImageNet	73.0	NA
R(2+1)D	Sports1M	74.3	152×dense
A^2-Net	ImageNet	74.6	41×NA
S3D-G	ImageNet	74.7	71×dense
D3D	ImageNet	75.9	NA
GloRe	ImageNet	76.1	55×NA
I3D+NLN	ImageNet	77.7	359×30
SlowFast	none	78.9	213×30
SlowFast+NLN	none	**79.8**	234×30
CSN-152	Sports1M	79.0	96.7×30
ip-CSN-152	Sports1M	**79.2**	108.8×30

a bit better than SlowFast

a bit worse than SlowFast + Non-Local Net

Comparison with the State-of-the-Art

- Results on Kinetics-400:

Method	pretrain	video@1	GFLOPs×crops
ResNeXt	none	65.1	NA
ARTNet(d)	none	69.2	24×250
I3D	ImageNet	71.1	108×dense
TSM	ImageNet	72.5	65×NA
MFNet	ImageNet	72.8	11×NA
Inception-ResNet	ImageNet	73.0	NA
R(2+1)D	Sports1M	74.3	152×dense
A^2-Net	ImageNet	74.6	41×NA
S3D-G	ImageNet	74.7	71×dense
D3D	ImageNet	75.9	NA
GloRe	ImageNet	76.1	55×NA
I3D+NLN	ImageNet	77.7	359×30
SlowFast	none	78.9	213×30
SlowFast+NLN	none	**79.8**	234×30
CSN-152	Sports1M	79.0	96.7×30
ip-CSN-152	Sports1M	**79.2**	108.8×30

3x faster than I3D+NLN

2x faster than SlowFast

Comparison with the State-of-the-Art

- State-of-the-art numbers on both Sports1M and Something-Something:

Sports1M

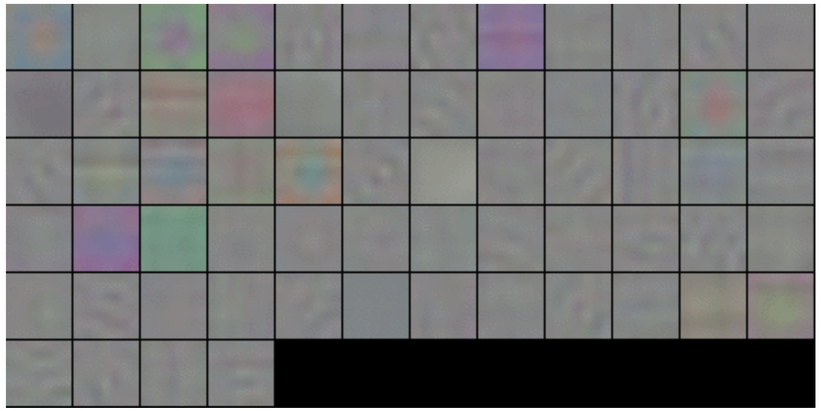
Method	input	video@1	video@5	GFLOPs×crops
C3D	RGB	61.1	85.2	N/A
P3D	RGB	66.4	87.4	N/A
Conv Pool	RGB+OF	71.7	90.4	N/A
R(2+1)D	RGB	73.0	91.5	152×dense
R(2+1)D	RGB+OF	73.3	91.9	305×dense
ip-CSN-101	RGB	74.9	92.6	63.6×10
ip-CSN-152	RGB	**75.5**	**92.8**	83.3×10

Something-Something

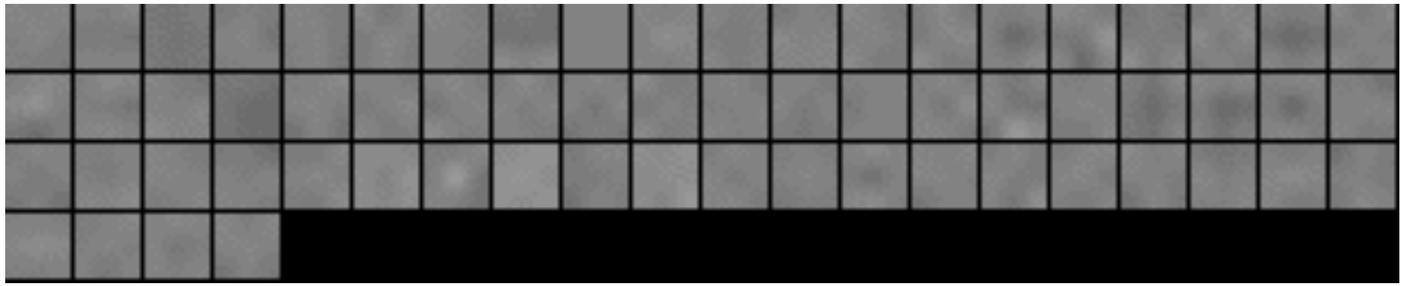
Method	pretrain	video@1
M-TRN	ImageNet	34.4
I3D + NL	ImageNet	44.4
I3D + NL + GCN	ImageNet	46.1
Motion Feature Net	none	43.9
TSM	Kinetics	44.8
TSM (ensemble)	Kinetics	46.8
ECO-Net	ImageNet	46.4
S3D-G	ImageNet	48.2
CSN-101	none	48.4
CSN-152	none	**49.3**

Visualization of CSN 3D filters $3 \times 3 \times \overline{3}$

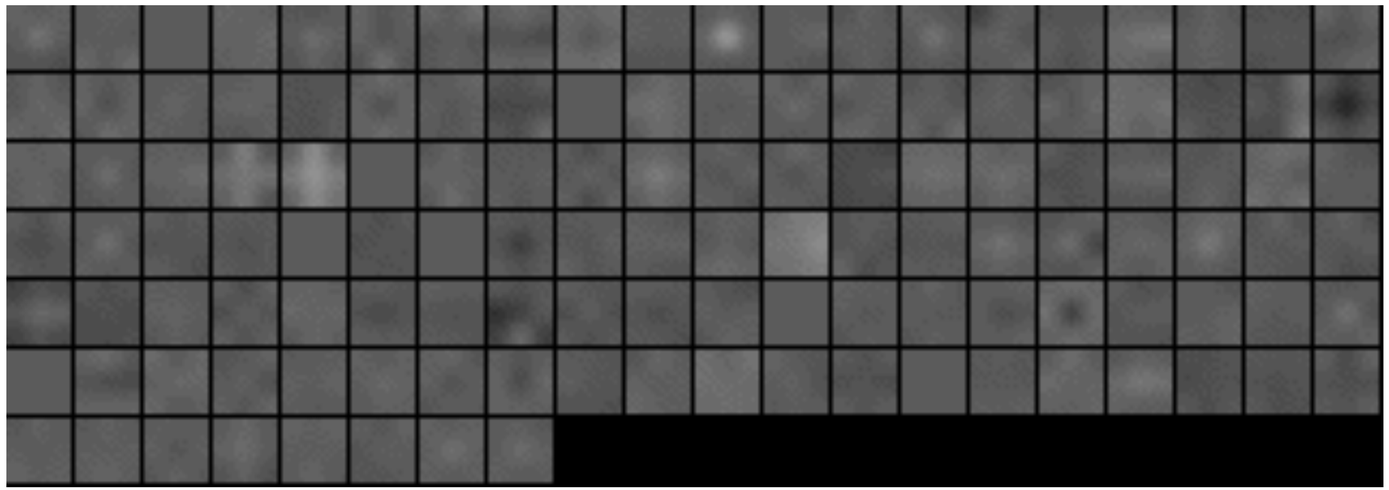
conv1



CSN in 1st group



CSN in 2nd group



Conclusions on spatiotemporal convolutions

✓ *Do we even need 3D convolution?*

Yes, for the same #parameters 3D CNNs provide better accuracy than 2D CNNs

✓ *If so, what layers should we make 3D, and what layers can be 2D?*

Top-heavy mixed convolutional nets perform better than pure 3D CNNs

✓ *Is it beneficial to factorize spatiotemporal filters into disjoint space and time components?*

Factorized space-time kernels lead to easier optimization and better generalization

✓ *Is it useful to factorize spatiotemporal filters across channels?*

Group and channel-separated 3D convolutions have fewer parameters, reduce #FLOPs and yield better action recognition accuracy