A Closer Look at Spatiotemporal Convolutions for Action Recognition

CVPR 2018

Du Tran, Heng Wang, Lorenzo Torresani,
Jamie Ray, Yann LeCun, Manohar Paluri

Motivation

3D CNNs are very costly to train and deploy.

Model	Pretrain	K400 Training Time (hours)	K400 Acc.	Inference TFLOPs
I3D 8x8 R50	ImageNet-1K	1440	73.4	1.11
SlowFast R50	ImageNet-1K	3840	75.6	1.97
SlowFast R50	N/A	6336	76.4	1.97

Motivation

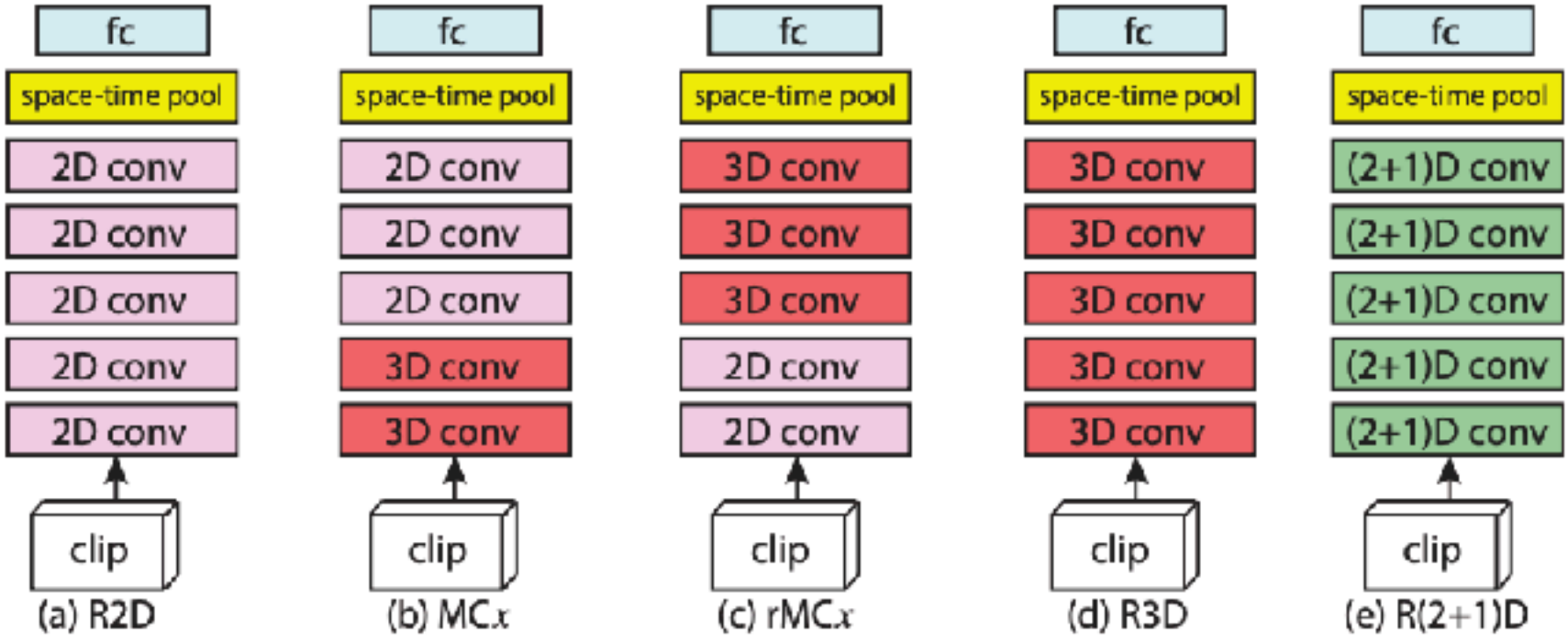
3D CNNs are very costly to train and deploy.

Model	Pretrain	K400 Training Time (hours)	K400 Acc.	Inference TFLOPs
I3D 8x8 R50	ImageNet-1K	1440	73.4	1.11
SlowFast R50	ImageNet-1K	3840	75.6	1.97
SlowFast R50	N/A	6336	76.4	1.97

How can we make video models more efficient?

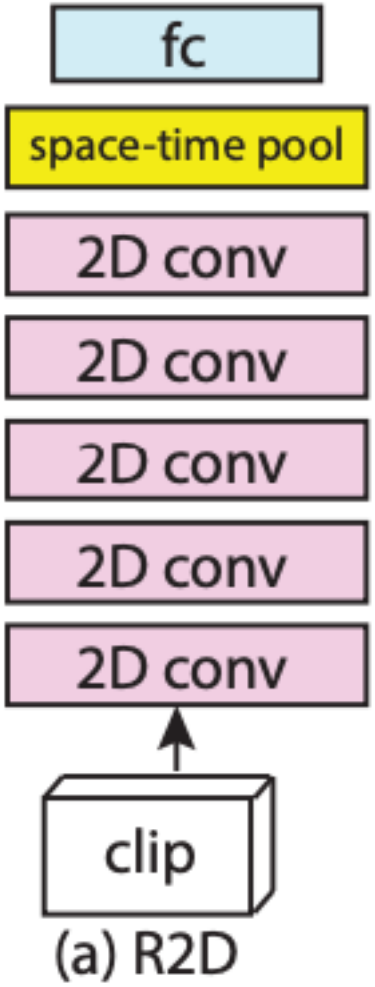
Architecture Variants

The authors conduct an empirical study of several different network architectures for video modeling.



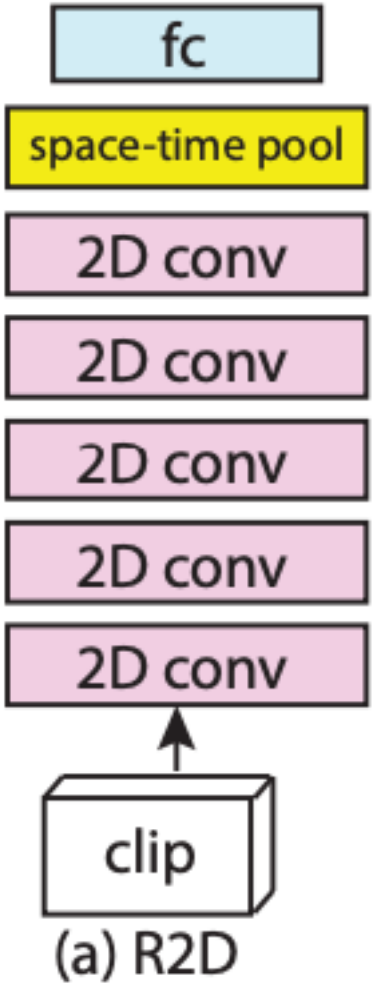
R2D

- A 2D ResNet.
- Temporal information is stacked into a channel dimension.
- Temporal information collapsed in the first convolutional layer.



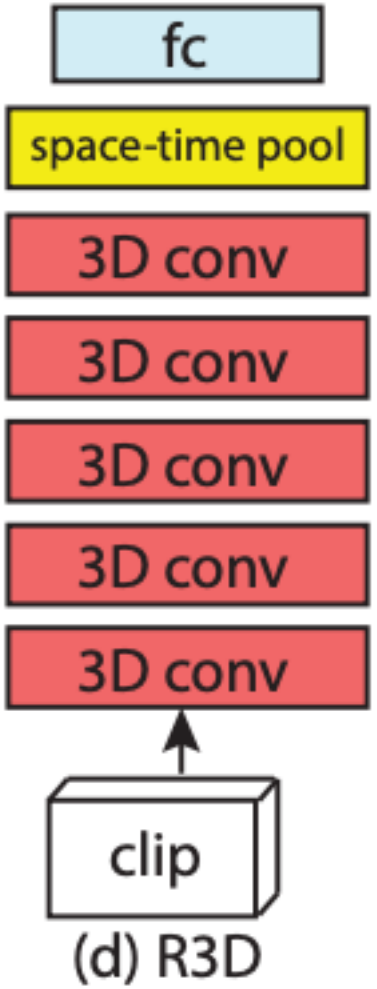
f-R2D

- A frame-based 2D ResNet.
- Frames are processed independently.
- A spatiotemporal pooling layer at the end fuses the information extracted independently from each frame.



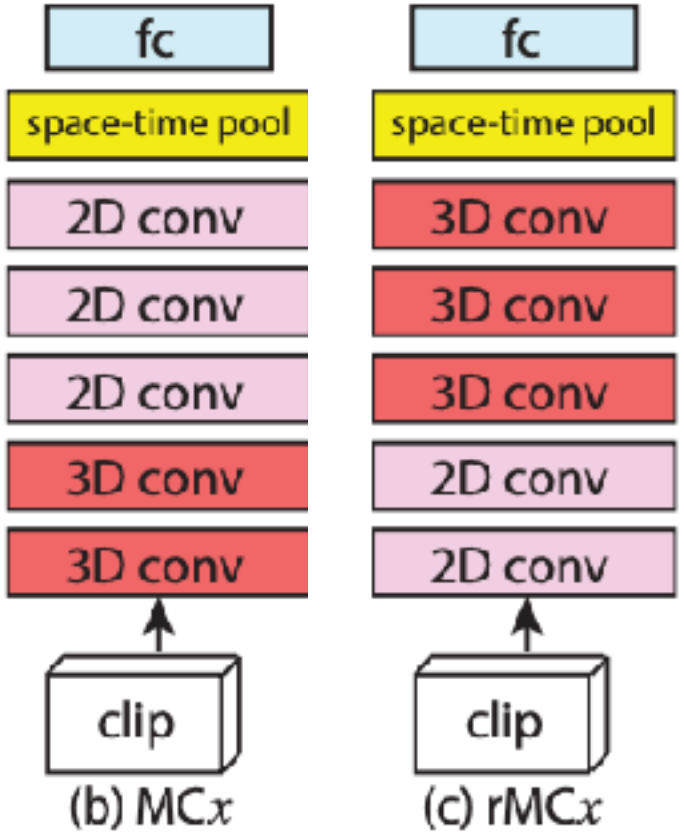
R3D

- A 3D ResNet built using 3D convolutional kernels.
- Temporal information is preserved in every convolutional layer of the network.
- Similar to C3D but with residual connections.



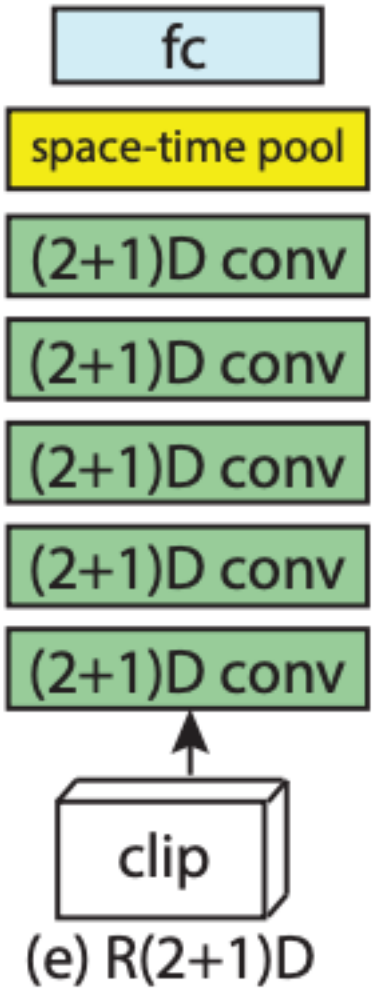
Mixed 3D-2D CNNs

- A network that uses mixed 3D or 2D convolutions.
- MC3 uses 2D convolutions in the last 3 layers and 3D convolutions in the first 2 layers.
- rMC3 uses 3D convolutions in the last 3 layers and 2D convolutions in the first 2 layers.



R(2+1)D

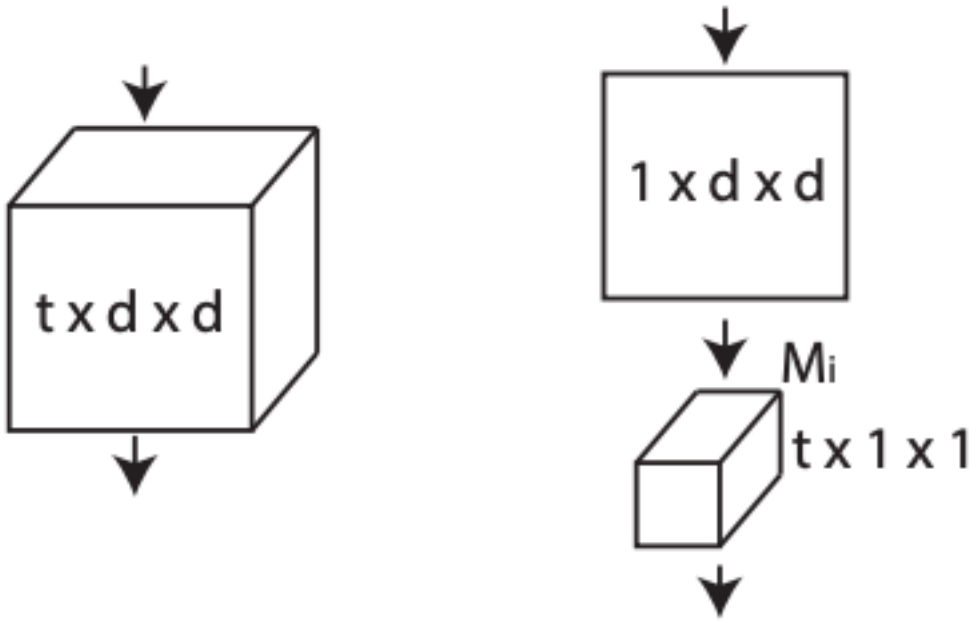
- Uses factorized 3D convolutions, i.e., 2D convolutions followed by 1D convolutions.
- Decomposes spatial and temporal modeling aspects.
- A more efficient approximation of 3D convolutions.



(e) R(2+1)D

3D vs (2+1)D

- Comparison between 3D and (2+1)D convolutions.



a) 3D convolution

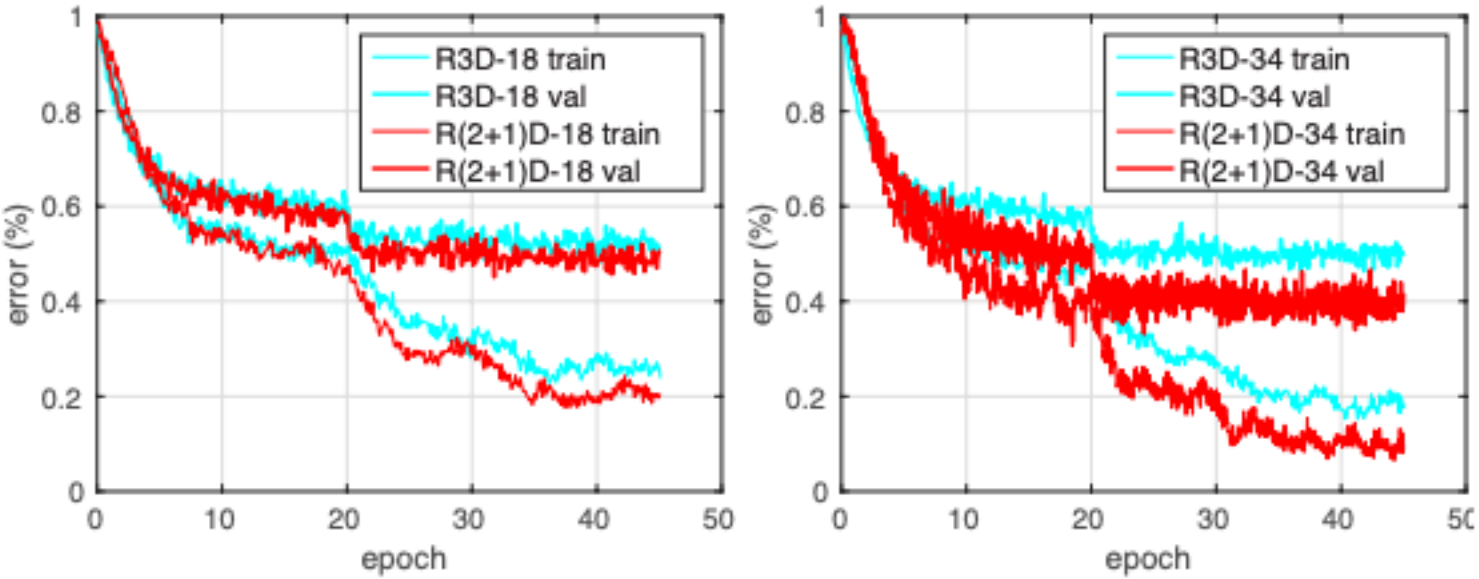
b) (2+1)D convolution

Advantages of (2+1)D Convolutions

- Doubles the number of non-linearities, which increases the complexity of functions that can be represented.
- (2+1)D convolutions generally lead to higher efficiency compared to 3D convolutions.
- Forcing the 3D convolution into separate spatial and temporal components renders the optimization easier.

Easier Optimization

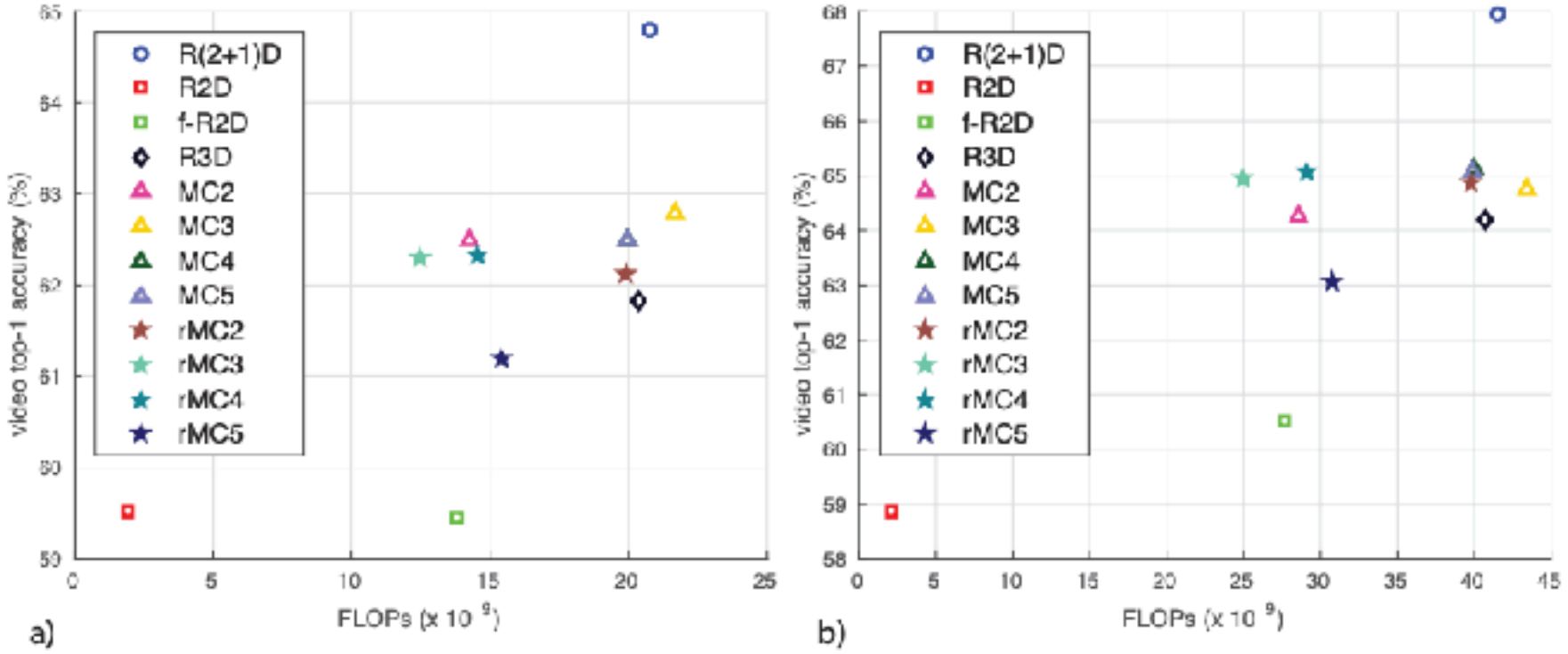
- (2+1)D CNNs are easier to optimize than 3D CNNs.



Comparison of Architectures

Net	# params	Clip@1	Video@1	Clip@1	Video@1
Input		8×112×112		16×112×112	
R2D	11.4M	46.7	59.5	47.0	58.9
f-R2D	11.4M	48.1	59.4	50.3	60.5
R3D	33.4M	49.4	61.8	52.5	64.2
MC2	11.4M	50.2	62.5	53.1	64.2
MC3	11.7M	50.7	62.9	53.7	64.7
MC4	12.7M	50.5	62.5	53.7	65.1
MC5	16.9M	50.3	62.5	53.7	65.1
rMC2	33.3M	49.8	62.1	53.1	64.9
rMC3	33.0M	49.8	62.3	53.2	65.0
rMC4	32.0M	49.9	62.3	53.4	65.1
rMC5	27.9M	49.4	61.2	52.1	63.1
R(2+1)D	33.3M	**52.8**	**64.8**	**56.8**	**68.0**

Accuracy vs Computational Cost



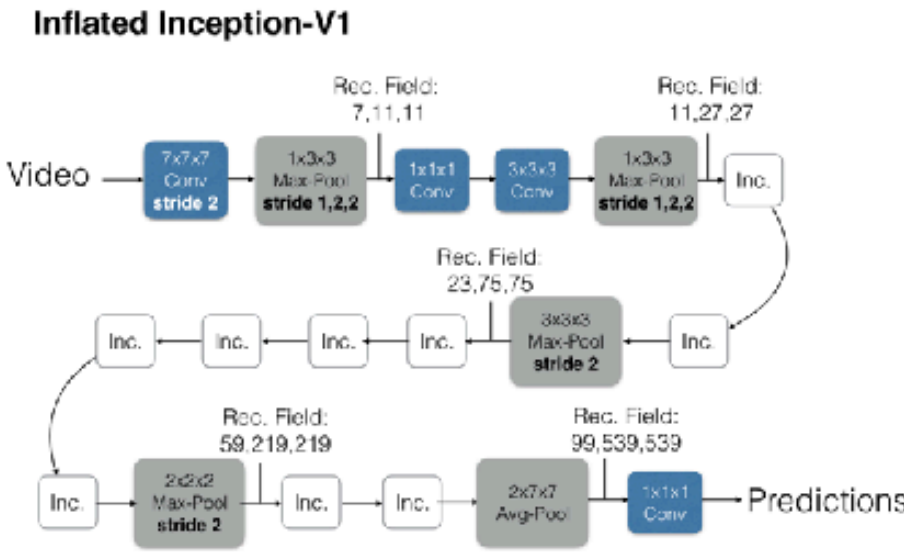
Results on Kinetics

method	pretraining dataset	top1	top5
I3D-RGB [4]	none	67.5	87.2
I3D-RGB [4]	ImageNet	72.1	90.3
I3D-Flow [4]	ImageNet	65.3	86.2
I3D-Two-Stream [4]	ImageNet	**75.7**	**92.0**
R(2+1)D-RGB	none	72.0	90.0
R(2+1)D-Flow	none	67.5	87.2
R(2+1)D-Two-Stream	none	73.9	90.9
R(2+1)D-RGB	Sports-1M	74.3	91.4
R(2+1)D-Flow	Sports-1M	68.5	88.1
R(2+1)D-Two-Stream	Sports-1M	**75.4**	**91.9**

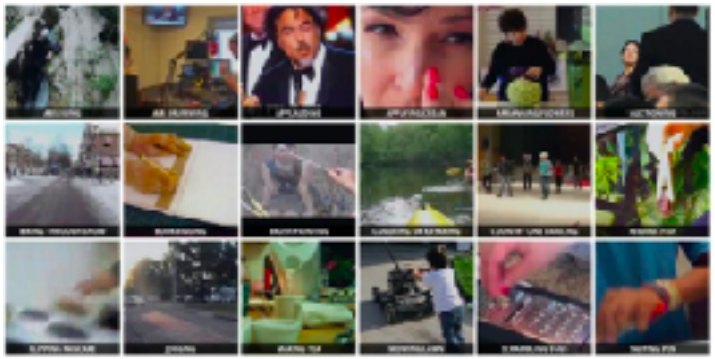
Arguments for I3D

Dataset and Model Contribution

- In addition to proposing a new video model, the paper also introduces a new large-scale dataset.



a) I3D Model

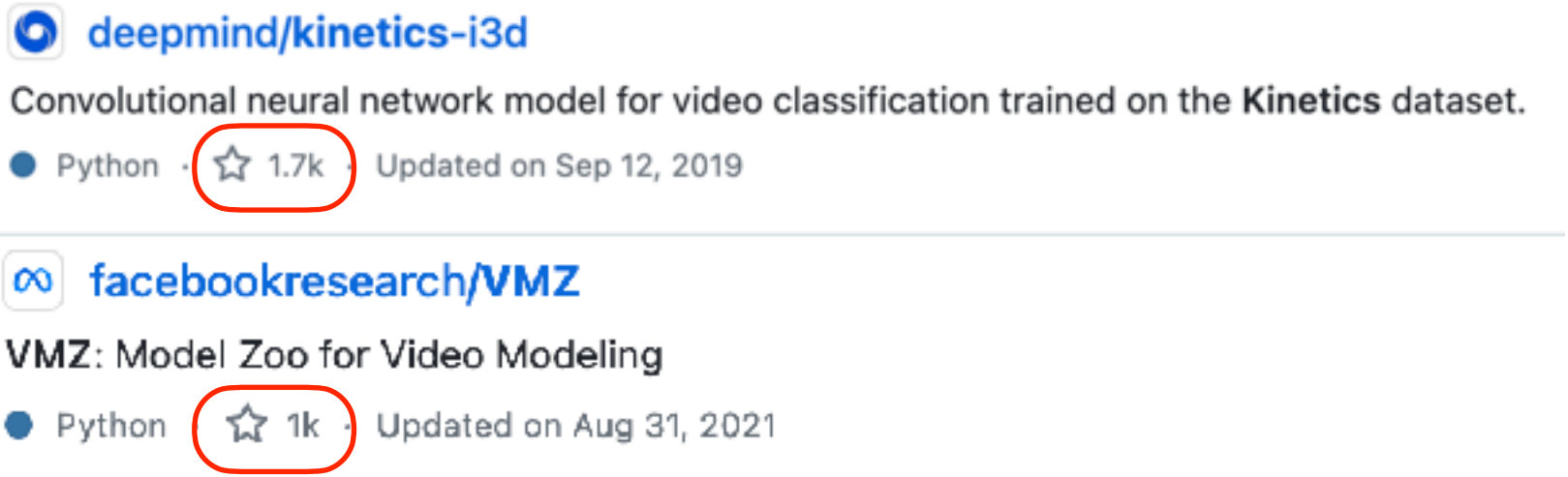


b) Kinetics Dataset

Research Impact

- Arguably, the I3D paper had a larger impact on the video recognition community.

Quo vadis, action recognition? a new model and the kinetics dataset J Carreira, A Zisserman proceedings of the IEEE Conference on Computer Vision and Pattern ...	7233	2017
A closer look at spatiotemporal convolutions for action recognition D Tran, H Wang, L Torresani, J Ray, Y LeCun, M Paluri Proceedings of the IEEE conference on Computer Vision and Pattern ...	2683	2018



Better Results

- Even though I3D was one year older than R(2+1)D, it still achieved better results at the time of R(2+1)D publication.

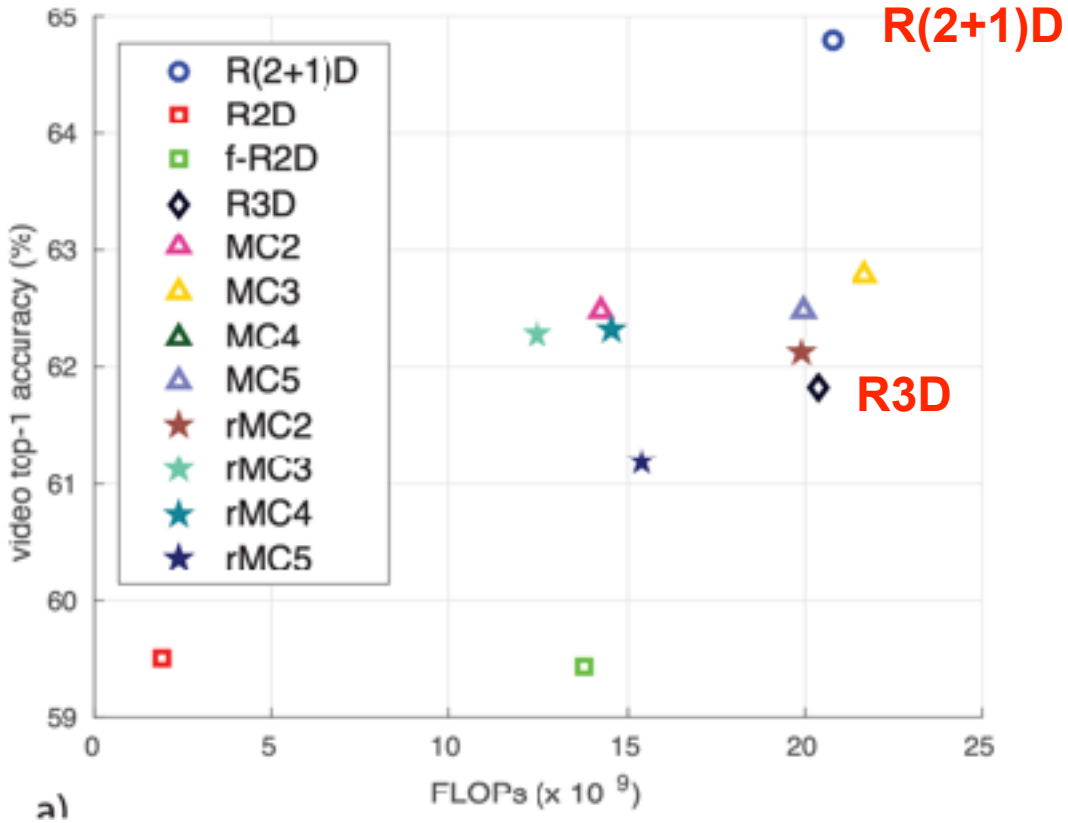
method	pretraining dataset	top1	top5
I3D-RGB [4]	none	67.5	87.2
I3D-RGB [4]	ImageNet	72.1	90.3
I3D-Flow [4]	ImageNet	65.3	86.2
I3D-Two-Stream [4]	ImageNet	**75.7**	**92.0**
R(2+1)D-RGB	none	72.0	90.0
R(2+1)D-Flow	none	67.5	87.2
R(2+1)D-Two-Stream	none	73.9	90.9
R(2+1)D-RGB	Sports-1M	74.3	91.4
R(2+1)D-Flow	Sports-1M	68.5	88.1
R(2+1)D-Two-Stream	Sports-1M	**75.4**	**91.9**

method	pretraining dataset	UCF101	HMDB51
Two-Stream [29]	ImageNet	88.0	59.4
Action Transf. [40]	ImageNet	92.4	62.0
Conv Pooling [42]	Sports-1M	88.6	-
$F_{ST}CN$ [33]	ImageNet	88.1	59.1
Two-Stream Fusion [10]	ImageNet	92.5	65.4
Spatiotemp. ResNet [9]	ImageNet	93.4	66.4
Temp. Segm. Net [39]	ImageNet	94.2	69.4
P3D [25]	ImageNet+Sports1M	88.6	-
I3D-RGB [4]	ImageNet+Kinetics	95.6	74.8
I3D-Flow [4]	ImageNet+Kinetics	96.7	77.1
I3D-Two-Stream [4]	ImageNet+Kinetics	**98.0**	**80.7**
R(2+1)D-RGB	Sports1M	93.6	66.6
R(2+1)D-Flow	Sports1M	93.3	70.1
R(2+1)D-TwoStream	Sports1M	95.0	72.7
R(2+1)D-RGB	Kinetics	96.8	74.5
R(2+1)D-Flow	Kinetics	95.5	76.4
R(2+1)D-TwoStream	Kinetics	97.3	78.7

Arguments for R(2+1)D

Accuracy-Efficiency Tradeoff

- R(2+1)D has a lot better accuracy-efficiency tradeoff than 3D CNNs (e.g., I3D).



Industry Impact

- R(2+1)D was pre-trained on 65M Instagram videos and deployed internally at Facebook for various use cases.
- This includes flagging cases of violence, pornography, scams, objectionable content, etc.

Internal large-scale computing platform

To support video research and development, Facebook has built an internal platform called Lumos. Lumos provides a simplified process for developers to train AI models on images and videos. First is the data. Many new tools on Lumos around data annotation can do image clustering. Second is the model. Developers can select off-the-shelf deep neural networks from Lumos and integrate particular features, like image feature and text features, into the model.

Lumos runs on billions of images and has more than 400 visual models for purposes of objectionable-content detection and spam fighting to automatic image captioning.

* https://aifrontiers.com/2018/08/17/facebooks-next-ai-adventure-video-understanding/

Scalability

- Due to its efficient design, R(2+1)D is easier to scale to massive datasets (e.g., IG-65M) and larger model sizes.

Method; pre-training	top-1	top-5	Input type
I3D-Two-Stream [11]; ImageNet	75.7	92.0	RGB + flow
R(2+1)D-Two-Stream [14]; Sports-1M	75.4	91.9	RGB + flow
3-stream SATT [69]; ImageNet	77.7	93.2	RGB + flow + audio
NL I3D [65]; ImageNet	77.7	93.3	RGB
R(2+1)D-34; Sports-1M	71.7	90.5	RGB
Ours R(2+1)D-34; IG-Kinetics	79.1	93.9	RGB
Ours R(2+1)D-34; IG-Kinetics; SE	79.6	94.2	RGB
Ours R(2+1)D-152; IG-Kinetics	80.5	94.6	RGB
Ours R(2+1)D-152; IG-Kinetics; SE	**81.3**	**95.1**	RGB