Quo Vadis, Action Recognition? A New Model and the Kinetics Dataset

João Carreira, Andrew Zisserman

Action recognition

Input: Video sequence

Output: Prediction of the action

Introduction

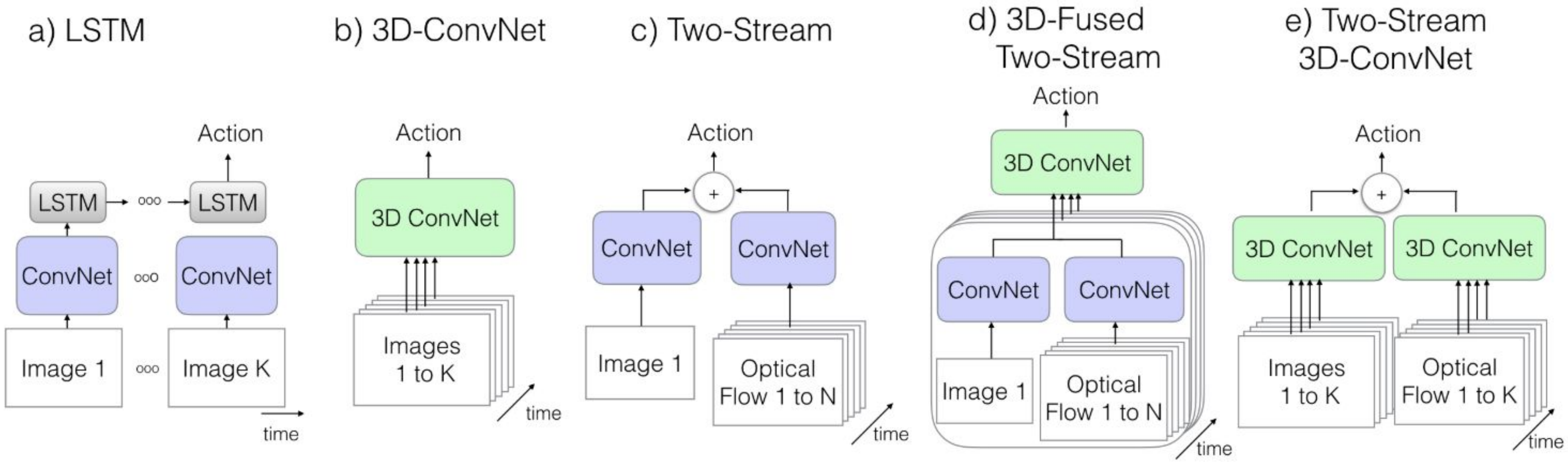
- Benefits of pre-training on ImageNet
  - Same task different data (classification -> classification)
  - Different task (classification -> segmentation/depth prediction etc.)
- In video domain, benefits of pre-training is an open question

Experimental strategy:

- Reimplement action classification DNNs from the literature
- Analyze their transfer behavior
- Introduce of a new model

Question: Is there a benefit in transfer learning using a large scale video dataset?

Background



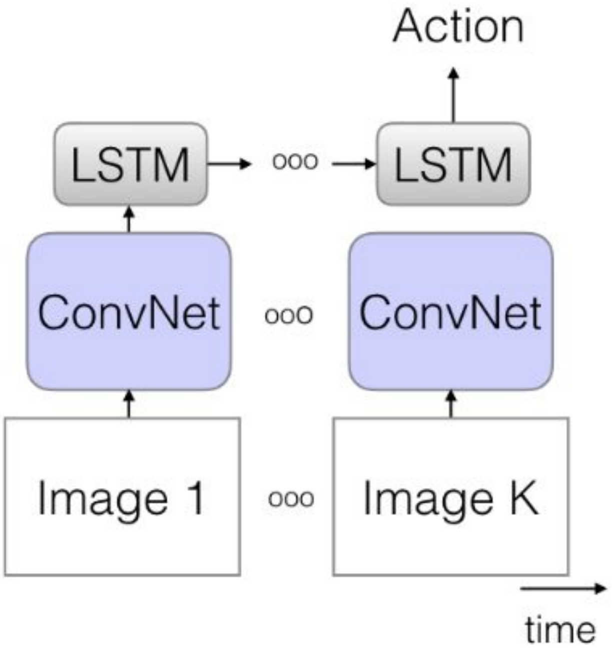
a) LSTM
Action
LSTM
ooo
LSTM
ConvNet
ooo
ConvNet
Image 1
ooo
Image K
time
b) 3D-ConvNet
Action
3D ConvNet
Images
1 to K
time
c) Two-Stream
Action
+
ConvNet
ConvNet
Image 1
Optical
Flow 1 to N
time
d) 3D-Fused
Two-Stream
Action
3D ConvNet
ConvNet
ConvNet
Image 1
Optical
Flow 1 to N
time
e) Two-Stream
3D-ConvNet
Action
+
3D ConvNet
3D ConvNet
Images
1 to K
Optical
Flow 1 to K
time

The Old I

- Feature extractor (2D ConvNet)
  - Pool features from individual frame (Ignoring temporal structure) [15]
- Add recurency (LSTM) [5, 34]
  - Encode state and temporal ordering

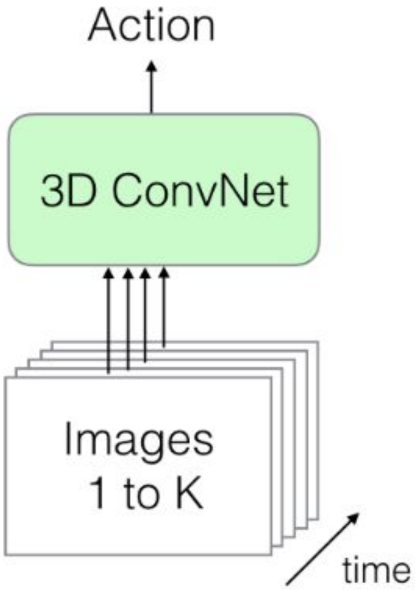
Pros
- Reuse of image classification networks

Cons
- Disjoint/late modeling of spatial and temporal information



The Old II

- 3D ConvNet [14, 15, 28, 29]
  - Frames are stacked in 3rd dimension



Pros
- Directly create spatio-temporal representation

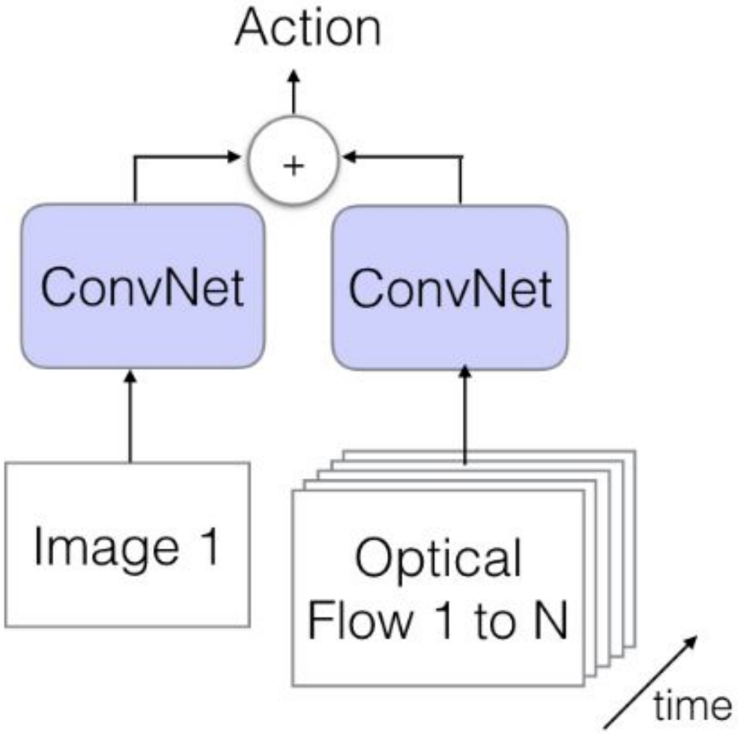
Cons
- Many more parameters than 2D ConvNets (Harder to train)
- Preclude benefit of ImageNet pre-training

The Old III

- Two stream network [8, 25]
  - RGB (2D ConvNet)
  - Optical Flow (2D ConvNet)

Pros
- Reuse of image classification networks

Cons
- Disjoint/late modeling of spatial and temporal information



The New

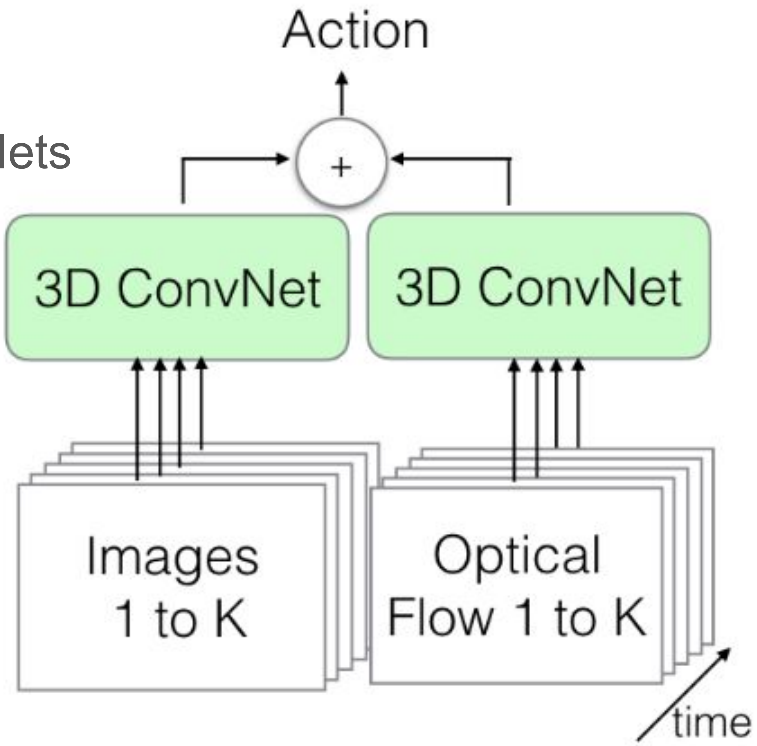
Inflating 2D ConvNets into 3D (This paper)

- Reusing structure of well studied 2D ConvNets
- Inflating an additional dimension to kernels
- Bootstrapping parameters from 2D filters

Pros

- Reuse of image classification networks
- Directly create spatio-temporal representation

Cons



The New

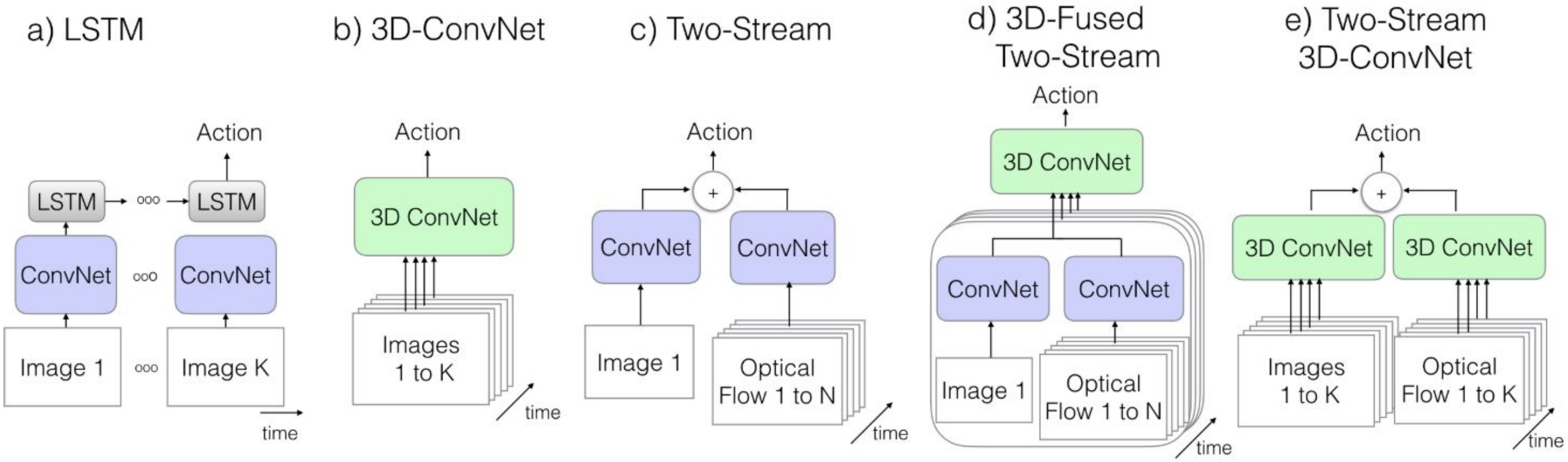
- Bootstrapping parameters by satisfying the boring-video fixed point
    - Copy an image to **convert** it to a “boring-video”
    - The activation from boring-video should be **the same** as from the original image
    - **Achieved by** repeating 2D-filters N times and rescale by dividing by N
- Symmetric receptive field might not be optimal
    - Grow too quickly -> conflate edges from different objects
    - Grow too slow -> not capture the entire scene dynamics

Architectures summary

Common for all: conv+bn+relu,

	a) LSTM	b) 3D ConvNet	c) Two-Stream	d) 3D-fused Two-Stream	e) Two-Stream 3D-ConvNet
ImageNet pretrained (Inception-V1)	yes	no	yes	yes	yes
Resolution	224x224	112x112	224x224	224x224	224x224
Temporal resolution	Sample every fifth frame	16 consecutive frames	*	*	64 consecutive frames

* 5 consecutive frames 10 frames appart + corresponding optical flow frames



Method	#Params	Training		Testing	
		# Input Frames	Temporal Footprint	# Input Frames	Temporal Footprint
ConvNet+LSTM	9M	25 rgb	5s	50 rgb	10s
3D-ConvNet	79M	16 rgb	0.64s	240 rgb	9.6s
Two-Stream	12M	1 rgb, 10 flow	0.4s	25 rgb, 250 flow	10s
3D-Fused	39M	5 rgb, 50 flow	2s	25 rgb, 250 flow	10s
Two-Stream I3D	25M	64 rgb, 64 flow	2.56s	250 rgb, 250 flow	10s

Table 1. Number of parameters and temporal input sizes of the models.

Kinetics dataset [16]

- Covers:
  - Person actions — drawing, drinking
  - Person-Person actions — hugging, kissing
  - Person-Object actions — opening presents, washing dishes
- 400 human action classes
- 400+ clips per class
- miniKinetics: subset of Kinetics:
  - 213 classes
  - Total of 120k clips

Results 1

Architecture	UCF-101			HMDB-51			miniKinetics		
	RGB	Flow	RGB + Flow	RGB	Flow	RGB + Flow	RGB	Flow	RGB + Flow
(a) LSTM	81.0	–	–	36.0	–	–	69.9	–	–
(b) 3D-ConvNet	51.6	–	–	24.3	–	–	60.0	–	–
(c) Two-Stream	83.6	85.6	91.2	43.2	56.3	58.3	70.1	58.4	72.9
(d) 3D-Fused	83.2	85.8	89.3	49.2	55.5	56.8	71.4	61.0	74.0
(e) Two-Stream I3D	**84.5**	**90.6**	**93.4**	**49.8**	**61.9**	**66.4**	**74.1**	**69.6**	**78.7**

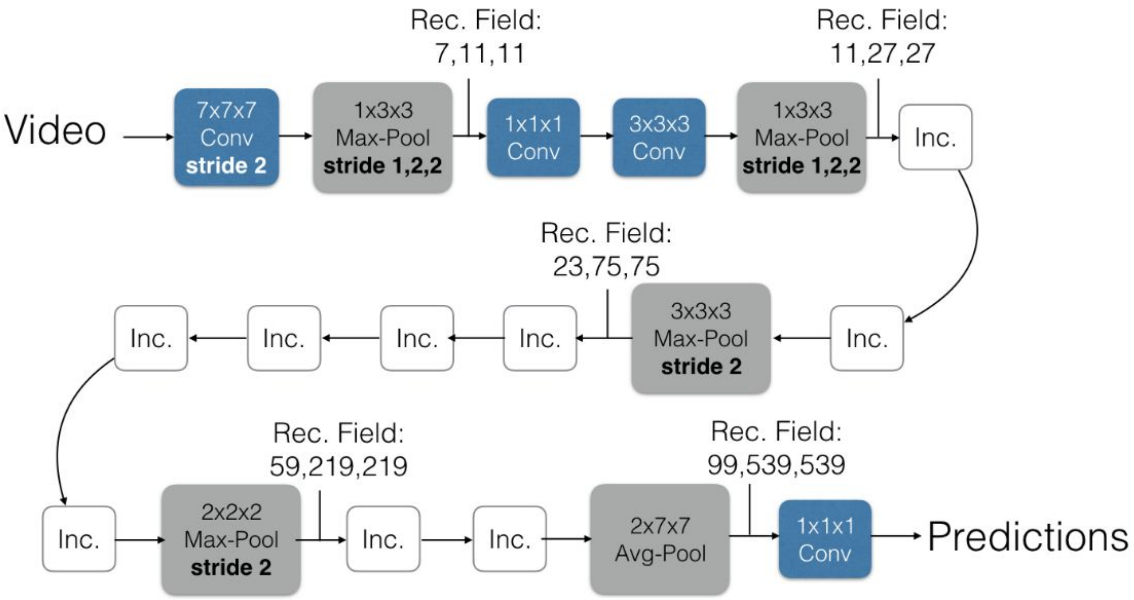
Results 2

Architecture	UCF-101				HMDB-51			
	Original	Fixed	Full-FT	Δ	Original	Fixed	Full-FT	Δ
(a) LSTM	81.0	81.6	82.1	-6%	36.0	46.6	46.4	-16.7%
(b) 3D-ConvNet	49.2	76.0	79.9	**-60.5%**	24.3	47.5	49.4	**-33.1%**
(c) Two-Stream	91.2	90.3	91.5	-3.4%	58.3	64.0	58.7	-13.7%
(d) 3D-Fused	89.3	88.5	90.1	-7.5%	56.8	59.0	61.4	-10.6%
(e) Two-Stream I3D	**93.4**	**95.7**	**96.5**	-47.0%	**66.4**	**74.3**	**75.9**	-28.3%

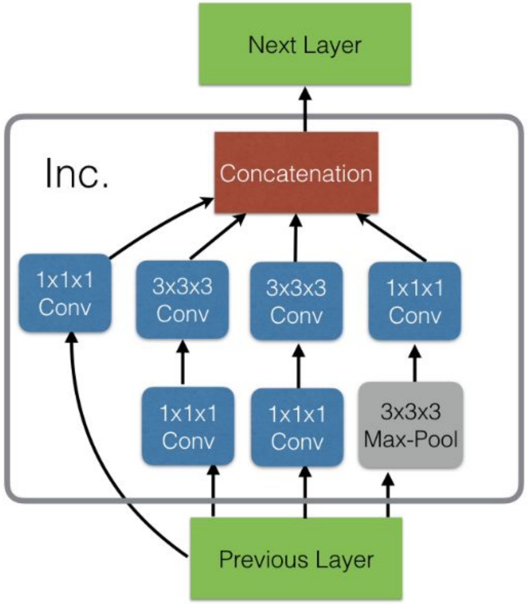
Summary

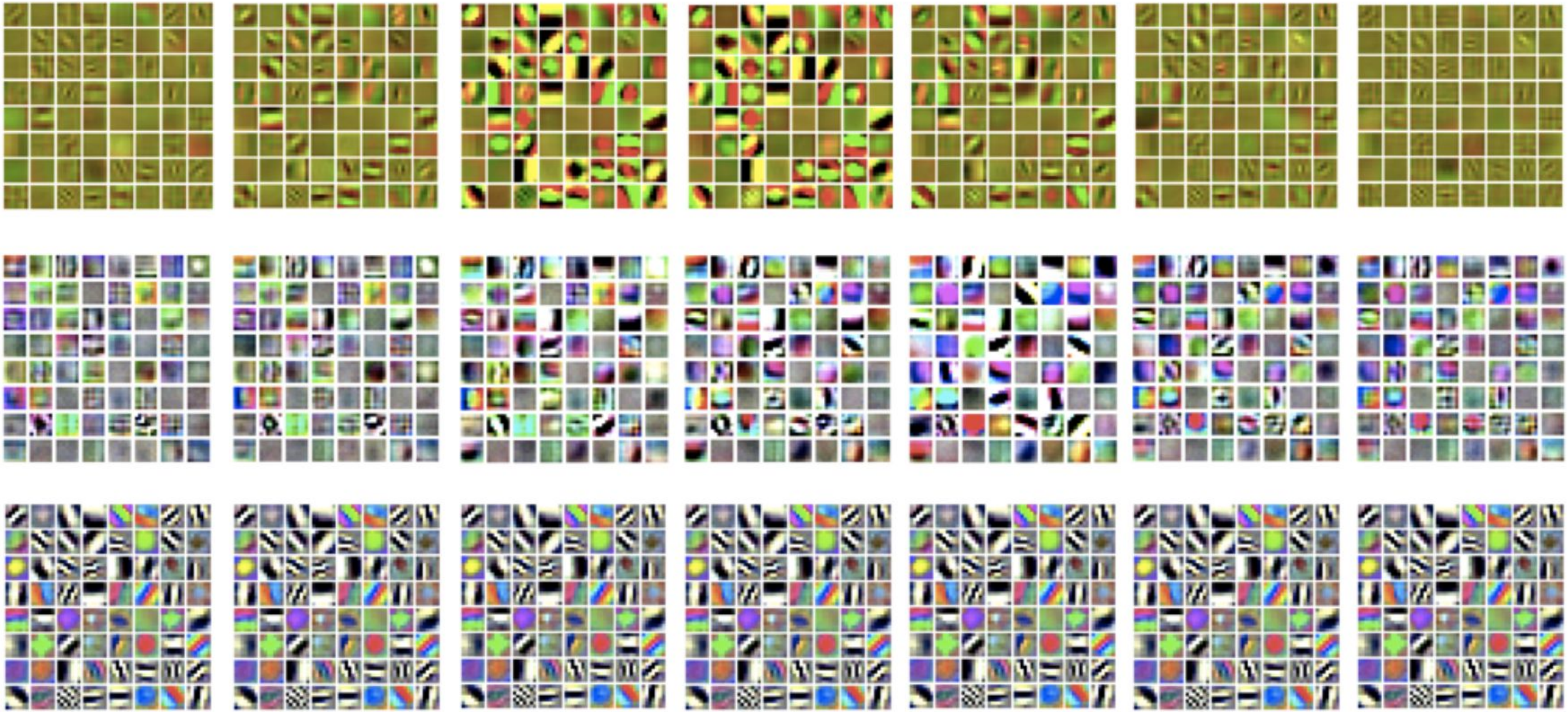
- Demonstrate power of transfer learning in video domain
- Introduce the idea of kernel inflation
- Novel architecture for action recognition

Inflated Inception-V1



Inception Module (Inc.)





References

[5] J. Donahue, L. Anne Hendricks, S. Guadarrama, M. Rohrbach, S. Venugopalan, K. Saenko, and T. Darrell. *Long-term recurrent convolutional networks for visual recognition and description.* In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 2625–2634, 2015.

[8] C. Feichtenhofer, A. Pinz, and A. Zisserman. *Convolutional two-stream network fusion for video action recognition.* In IEEE International Conference on Computer Vision and Pattern Recognition CVPR, 2016.

[14] S. Ji, W. Xu, M. Yang, and K. Yu. *3d convolutional neural networks for human action recognition.* IEEE transactions on pattern analysis and machine intelligence, 35(1):221– 231, 2013.

[15] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, and L. Fei-Fei. *Large-scale video classification with convolutional neural networks.* In Proceedings of the IEEE conference on Computer Vision and Pattern Recognition, pages 1725–1732, 2014.

[16] W. Kay, J. Carreira, K. Simonyan, B. Zhang, C. Hillier, S. Vijayanarasimhan, F. Viola, T. Green, T. Back, P. Natsev, M. Suleyman, and A. Zisserman. The kinetics human action video dataset. arXiv preprint arXiv:1705.06950, 2017.

[25] K. Simonyan and A. Zisserman. *Two-stream convolutional networks for action recognition in videos.* In Advances in Neural Information Processing Systems, pages 568–576, 2014.

[28] G. W. Taylor, R. Fergus, Y. LeCun, and C. Bregler. Convolutional learning of spatio-temporal features. In European conference on computer vision, pages 140–153. Springer, 2010.

[29] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri. Learning spatiotemporal features with 3d convolutional net- works. In 2015 IEEE International Conference on Computer Vision (ICCV), pages 4489–4497. IEEE, 2015.

[32] L. Wang, Y. Xiong, Z. Wang, Y. Qiao, D. Lin, X. Tang, and L. Van Gool. Temporal segment networks: towards good practices for deep action recognition. In European Conference on Computer Vision, 2016.

[34] J. Yue-Hei Ng, M. Hausknecht, S. Vijayanarasimhan, O. Vinyals, R. Monga, and G. Toderici. Beyond short snippets: Deep networks for video classification. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 4694–4702, 2015.