NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE

EE6222 Machine Vision

Topics 11/12
Vision Beyond Image 1: Video Analysis with Human Action Recognition 2

Dr. Xu Yuecong

1 Nov 2023

Topics

Topics Outline:

- Characteristic of videos and video analysis. (11)
- Representation-based (Traditional) video feature extraction. (11)
- **(Deep) Learning-based video feature extraction. (12)**
- **Latest developments in video analysis (mostly with deep learning) (12)**

References (non-exhaustive)

- Kong, Y., & Fu, Y. (2022). Human action recognition and prediction: A survey. *International Journal of Computer Vision*, 130(5), 1366-1401.
- Zhou, B., Andonian, A., Oliva, A., & Torralba, A. (2018). Temporal relational reasoning in videos. In *Proceedings of the European conference on computer vision* (ECCV) (pp. 803-818). (**TRN**)
- Dosovitskiy, Alexey, Philipp Fischer, Eddy Ilg, Philip Hausser, Caner Hazirbas, Vladimir Golkov, Patrick Van Der Smagt, Daniel Cremers, and Thomas Brox. “Flownet: Learning optical flow with convolutional networks.” In Proceedings of the *IEEE international conference on computer vision*, pp. 2758-2766. 2015. (**FlowNet**)
- Carreira, J., & Zisserman, A. (2017). Quo vadis, action recognition? a new model and the kinetics dataset. In proceedings of the *IEEE Conference on Computer Vision and Pattern Recognition* (pp. 6299-6308). (**I3D**)
- Feichtenhofer, C., Fan, H., Malik, J., & He, K. (2019). Slowfast networks for video recognition. In Proceedings of the *IEEE/CVF international conference on computer vision* (pp. 6202-6211). (**SlowFast**)
- Girdhar, R., Carreira, J., Doersch, C., & Zisserman, A. (2019). Video action transformer network. In Proceedings of the *IEEE/CVF conference on computer vision and pattern recognition* (pp. 244-253). (**ViT**)
- Liu, Z., Ning, J., Cao, Y., Wei, Y., Zhang, Z., Lin, S., & Hu, H. (2022). Video swin transformer. In Proceedings of the *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 3202-3211). (**Video Swin Transformer**)
- Ju, Chen, Tengda Han, Kunhao Zheng, Ya Zhang, and Weidi Xie. “Prompting visual-language models for efficient video understanding.” In European Conference on Computer Vision, pp. 105-124. Cham: Springer Nature Switzerland, 2022. (**Video LLVM**)

Goal and Scope of Topic 12

- ✓ Basic knowledge in the development of video analysis (action recognition).
- ✓ Basic knowledge of the different categories of deep learning-based video analysis (action recognition) methods.
- ✓ Develop critical thinking in learning-based video analysis (action recognition).

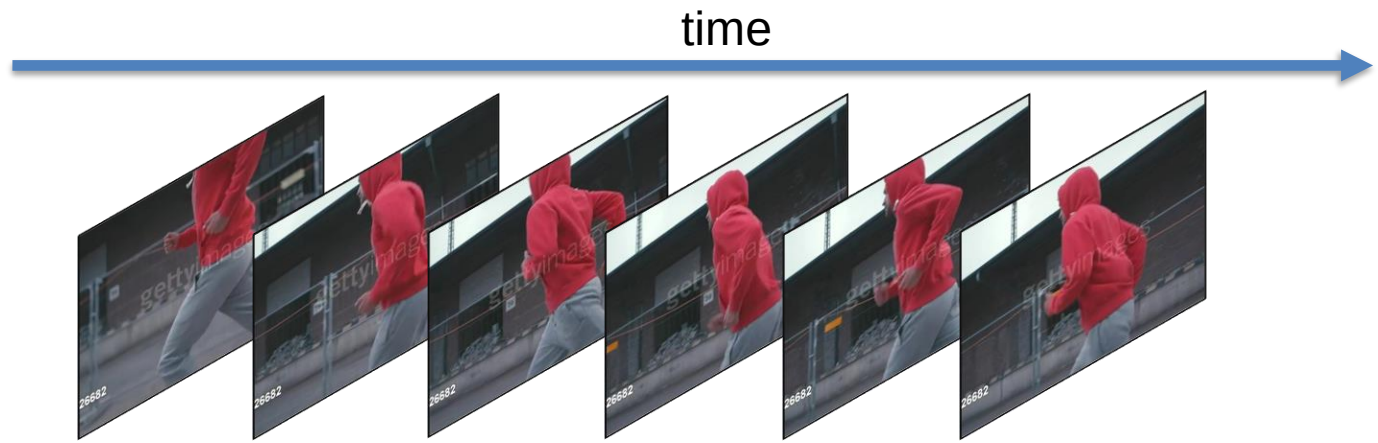
- ✓ **Understand and Analyze**. (There is NO perfect method!)
- ✓ **Think and Apply**.

- ✓ We will be limiting our scope of discussion to "**Deep Learning-based Action Recognition**".
- ✓ For other video-based tasks (e.g., spatial/temporal segmentation, captioning), please read through papers/blogs/github at your own interests.

Additional Notes for CA

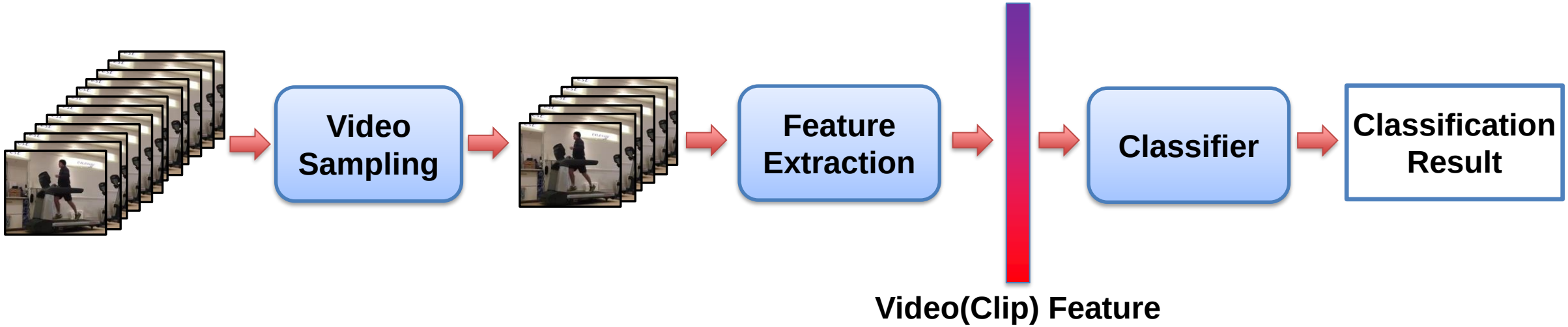
- Focus on the exploration of HAR with different options (e.g., different sampling methods, different structures, different enhancements, etc.).
- Express your understanding on the following question:
  - Why the changes made (does not) help performance improvement?
- Mark will NOT be given based on the performance.
- Mark will be given based on the demonstration of exploration process, and the display of observations and discussions.

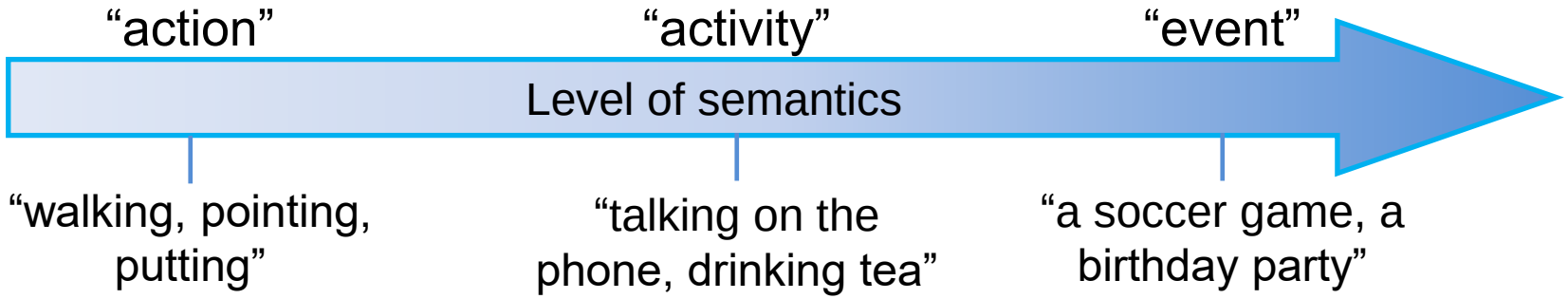
Recap: Video Fundamentals: from Images to Videos



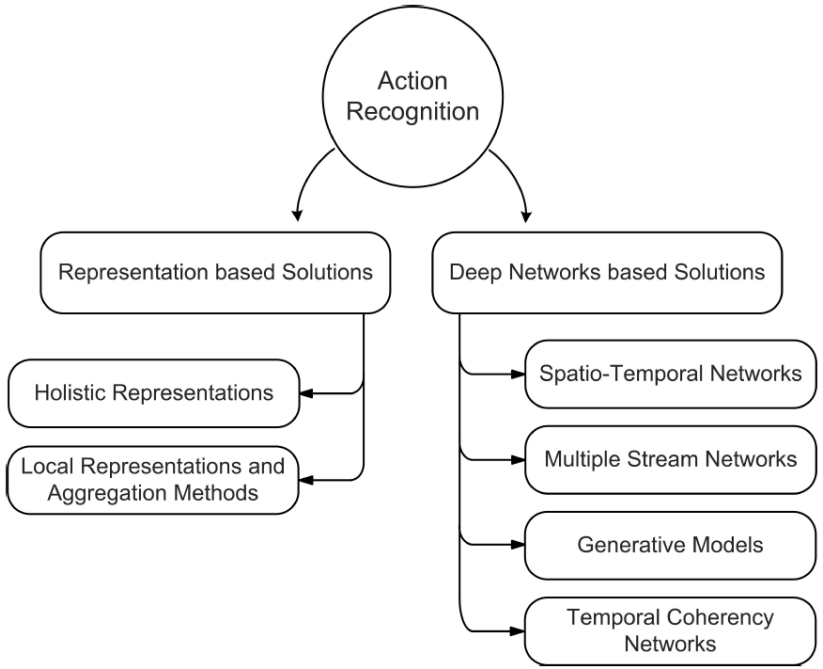
- A video can be viewed as a **sequence** of images.
  - By default, in computer vision the sequence of images are **closely related**.
- A video is more than a sequence of images:
  - Other modalities (types) of information are also included, e.g., audio.

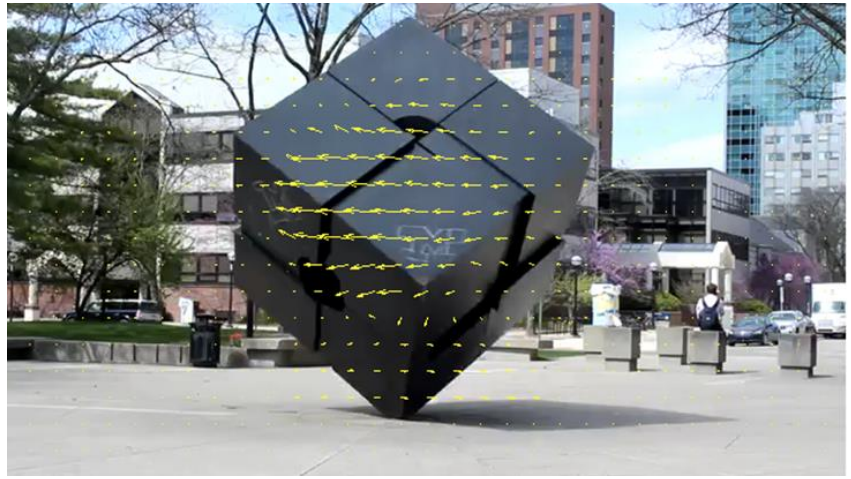
Recap: Video Analysis and Video Feature Extraction

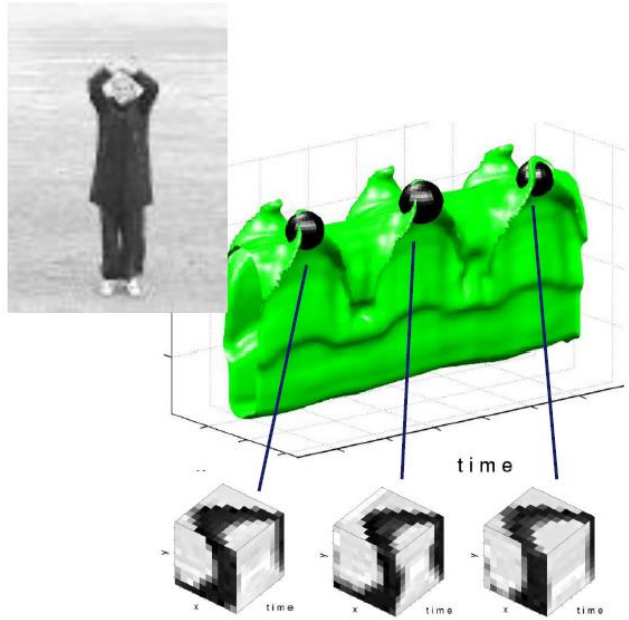




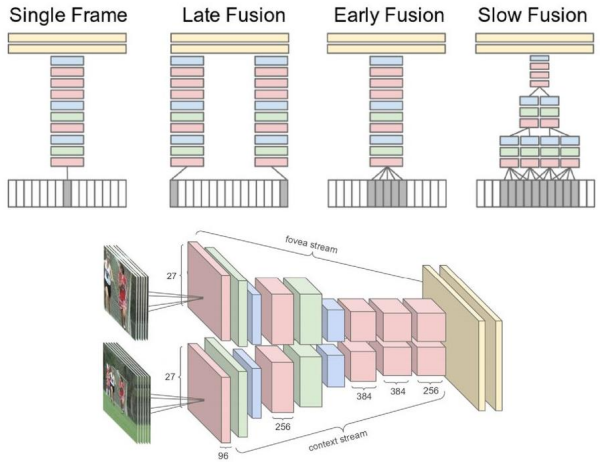
Recap: Video Analysis and Video Feature Extraction





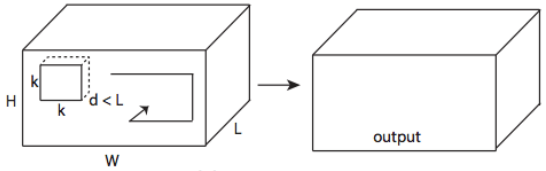


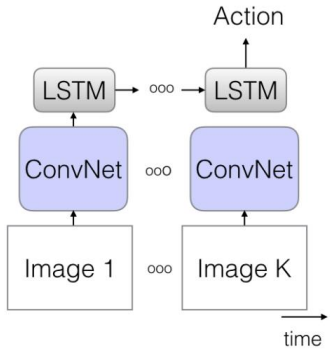
Deep Learning-based Video Analysis: A Brief Outline



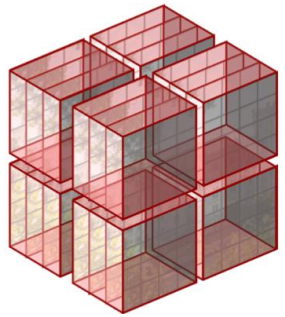
2D CNN

3D CNN

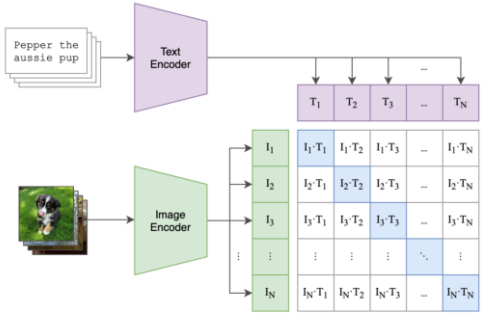




2D CNN with RNN



Video Transformer



Large Language-Vision Model

Before Moving On: How to Learn DL Methods in Brief

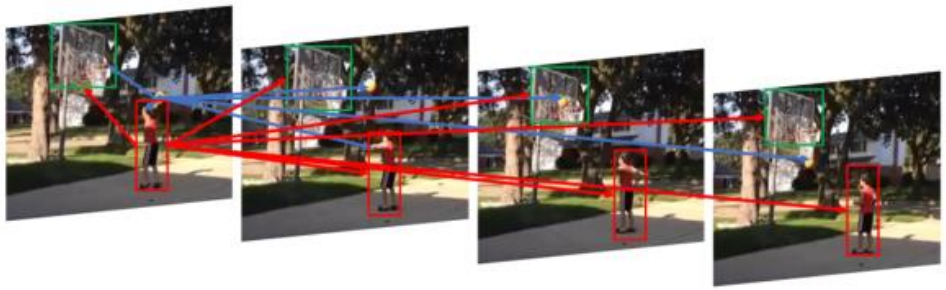


Figure 1: Illustration of utilizing regional correlation for action recognition. The original non-local block captures long-range spatiotemporal dependencies through pixel correlation, shown as blue arrows. The action "playing basketball" could alternatively be recognized through regional correlation between the boy and the backboard, shown as red arrows.

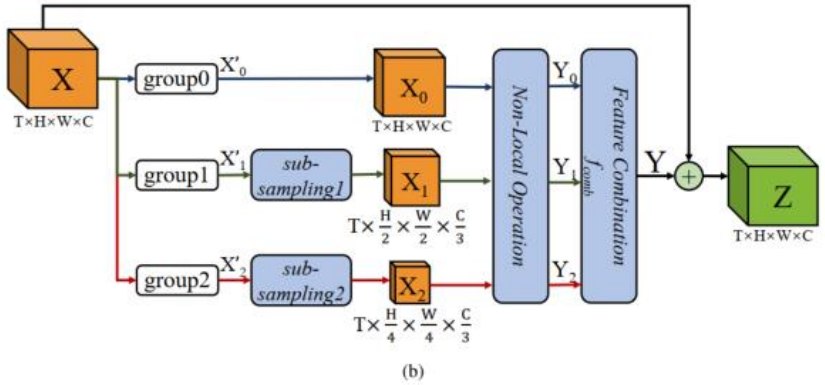


Figure 2: Comparison of the original non-local block (a) with our proposed PNL module (b). We present the case where the embedded Gaussian function is utilized for the non-local operation. The dimension of the input and output features are also presented, with the "batch" dimension ignored.

	Method	Mini-Kinetics Top-1	# Params	FLOPs
Two-stream CNNs	MARS [44]	73.5%	-	-
	ResFrame TS [45]	73.9%	-	-
	I3D (TS) [46]	78.7%	25.0M	>107.9G
3D CNNs	C3D [19]	66.2%	33.3M	-
	I3D (RGB) [46]	74.1%	12.06M	107.9G
	(2+C1)D [47]	75.74%	**7.3M**	31.9G
	S3D [17]	78.0%	8.77M	43.47G
	MFNet [38]	78.35%	7.84M	**11.17G**
CNN with long-range dependencies	Res50-NL [16]	77.53%	27.66M	19.67G
	Res50-CGD [48]	77.56%	25.58M	17.88G
	Res50-CGNL [26]	77.76%	27.2M	19.16G
	MFNet-NL [26]	79.74%	8.15M	11.66G
Ours	**MFNet-PNL**(×1)	82.16%	7.92M	11.22G
	MFNet-PNL(×5)	**83.09%**	8.12M	11.38G

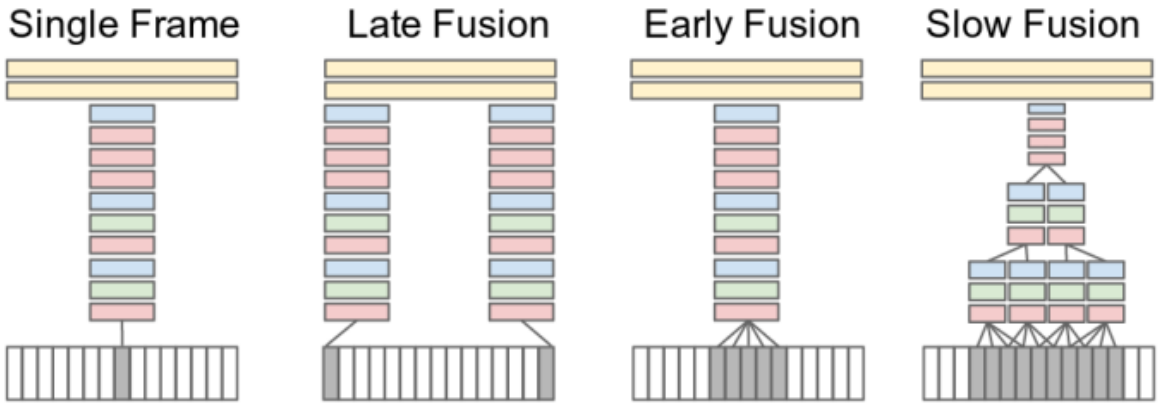
Table 5: Comparison of top-1 and top-5 accuracy, number of parameters and computation cost in FLOPs with state-of-the-art methods on the Mini-Kinetics datasets.

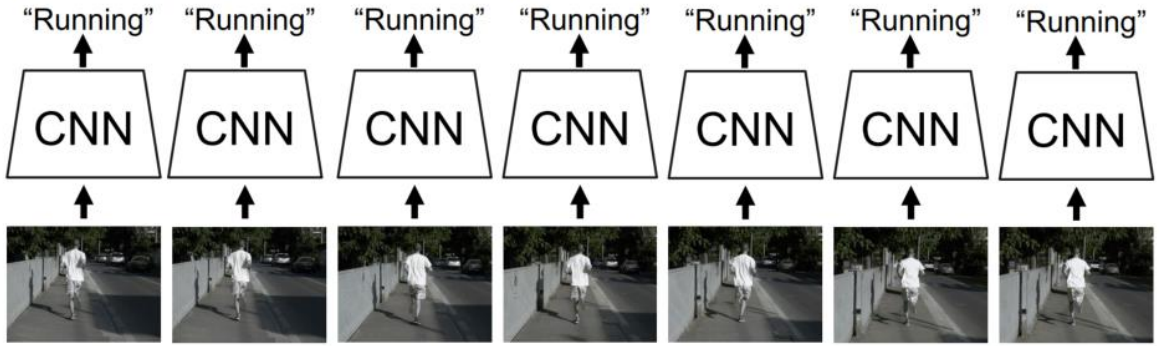
- Introduction: motivation of the proposed method – why this method is proposed?
- Methodology: how is the proposed method designed? Focus on structure
- Experiment: how is the proposed method proven to be effective?

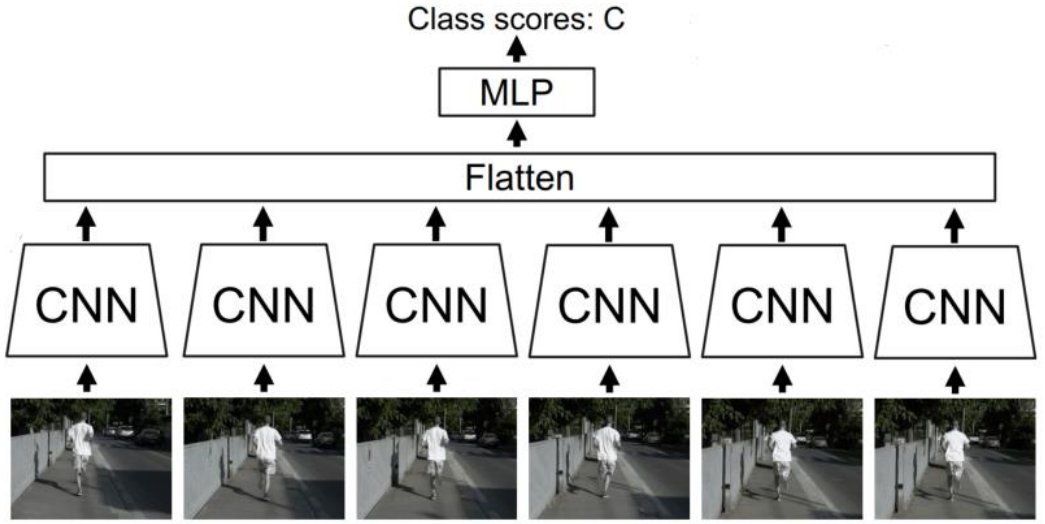
Figure from Xu, Yuecong, Haozhi Cao, Jianfei Yang, Kezhi Mao, Jianxiong Yin, and Simon See. "PNL: Efficient long-range dependencies extraction with pyramid non-local module for action recognition." Neurocomputing 447 (2021): 282-293.

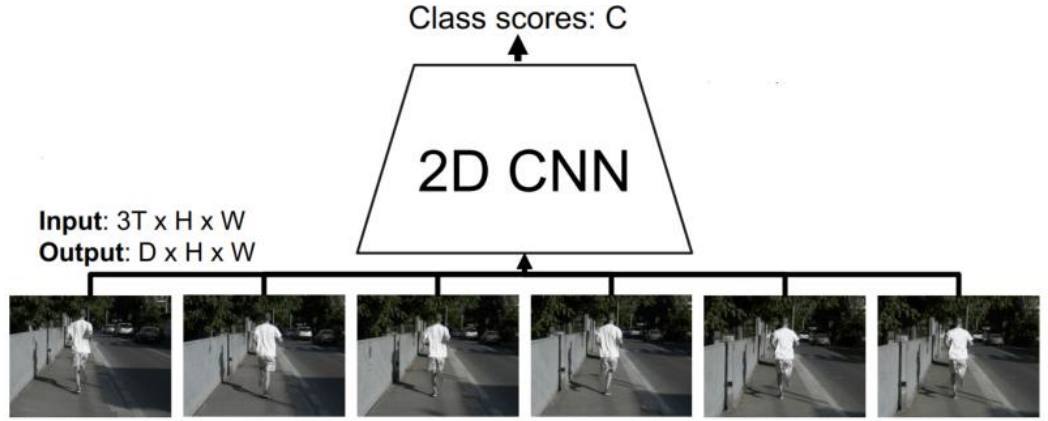
2D CNN: Fusion of Image Features

- Karpathy, Andrej, George Toderici, Sanketh Shetty, Thomas Leung, Rahul Sukthankar, and Li Fei-Fei. "Large-scale video classification with convolutional neural networks." In Proceedings of the IEEE conference on Computer Vision and Pattern Recognition, pp. 1725-1732. 2014.





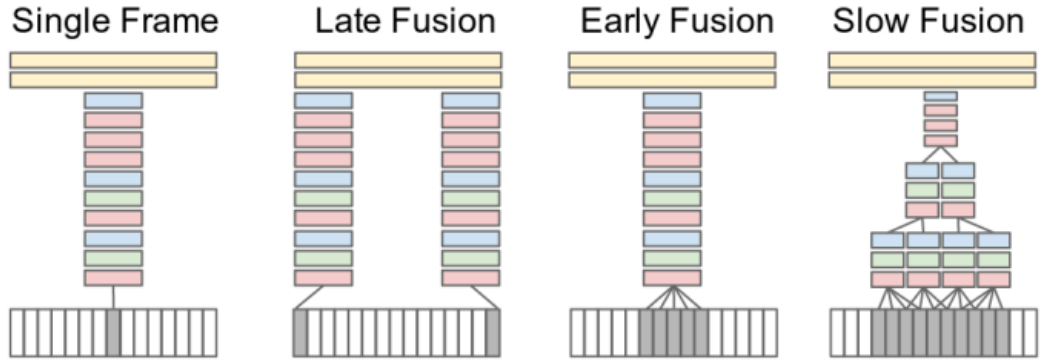


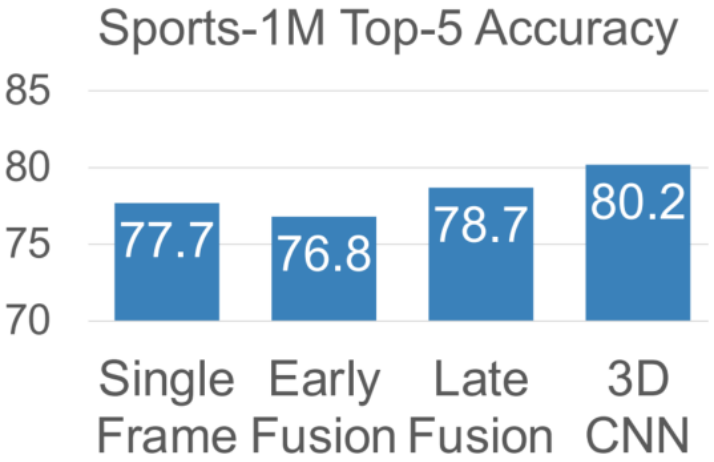


Some figure from CS231N Stanford University, Credit: Prof. Justin Johnson, Prof. Fei-Fei Li

2D CNN: Fusion of Image Features

- Discussion: advantages and limitations of 2D CNN – fusion of image features:
  - Implementation?
  - Accuracy?
  - Scenarios?
  - Computation Cost?
- The core shortcoming of limitations?
  - Cannot deal with **temporal** information explicitly!



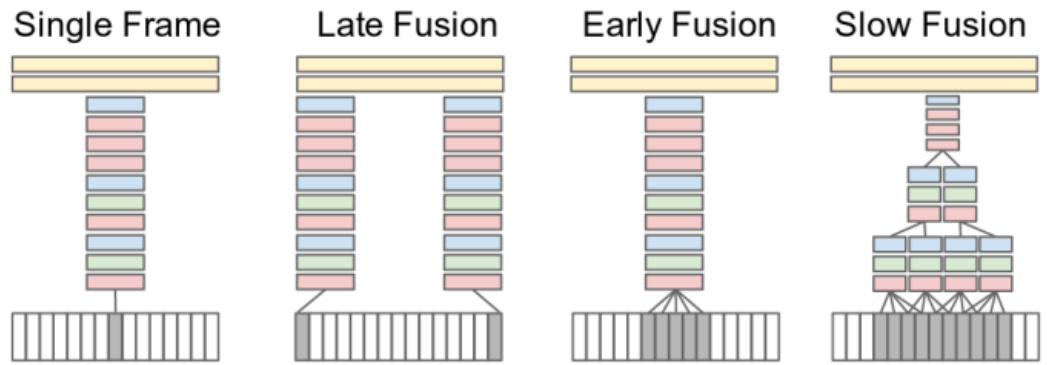


Group	mAP from scratch	mAP fine-tune top 3	mAP fine-tune top
Human-Object Interaction	0.26	0.55	0.52
Body-Motion Only	0.32	0.57	0.52
Human-Human Interaction	0.40	0.68	0.65
Playing Musical Instruments	0.42	0.65	0.46
Sports	**0.57**	**0.79**	**0.80**
All groups	0.44	0.68	0.66

Table 4: Mean Average Precision of the Slow Fusion network on UCF-101 classes broken down by category groups.

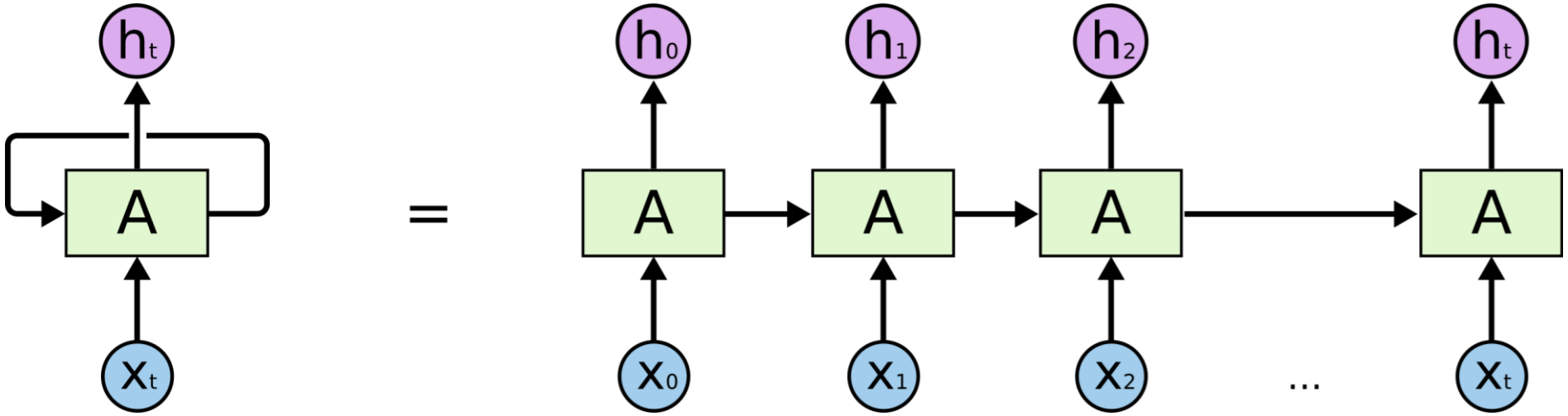
2D CNN: Fusion of Image Features

- The core shortcoming of disadvantages?
  - Cannot deal with **temporal** information explicitly!

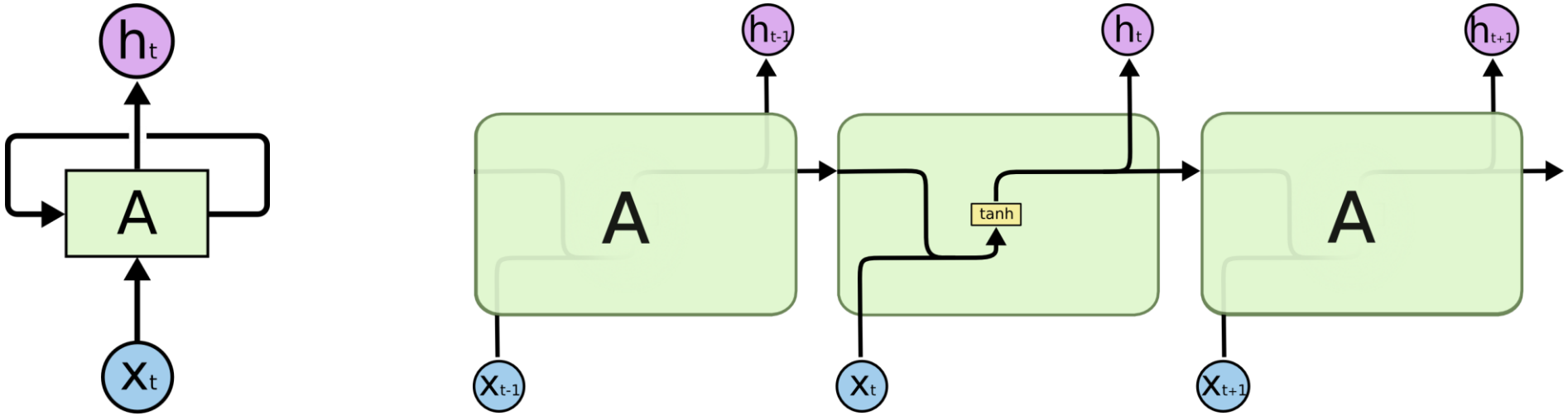


2D CNN: Dealing With Temporal Information via RNN

- Discussion: what temporal information/motion information contains essentially?
  - A **sequence** of **change** of appearance.
  - Both “sequence” and “change of appearance” can be monitored **globally** or **locally**.
- What network can model “sequence” directly?
  - Recurrent Neural Network (A brief introduction)



2D CNN: Dealing With Temporal Information via RNN



$$h_t = f_W(h_{t-1}, x_t)$$

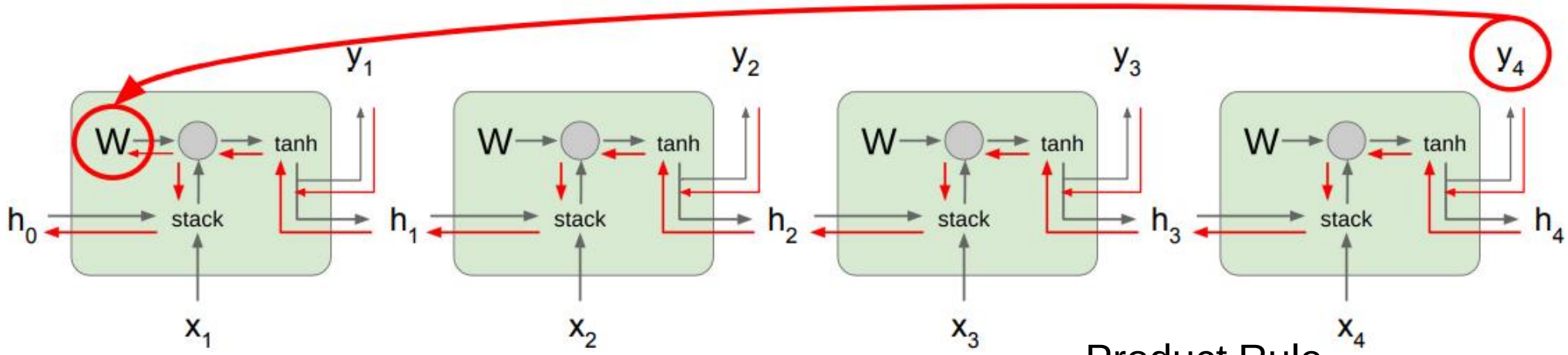
new state — some function with parameters W — old state — input vector at some time step

$$y_t = f_{W_{hy}}(h_t)$$

output — another function with parameters W_o — new state

2D CNN: Dealing With Temporal Information via RNN

- Any problem with such structure?



$$h_t = \tanh(W_{hh}h_{t-1} + W_{xh}x_t)$$

$$= \tanh\left(\begin{pmatrix} W_{hh} & W_{hx} \end{pmatrix} \begin{pmatrix} h_{t-1} \\ x_t \end{pmatrix} \right)$$

$$= \tanh\left(W \begin{pmatrix} h_{t-1} \\ x_t \end{pmatrix} \right)$$

Product Rule

$$\frac{\partial L}{\partial W} = \sum_{t=1}^{T} \frac{\partial L_t}{\partial W} \qquad \frac{\partial h_t}{\partial h_{t-1}} = tanh'(W_{hh}h_{t-1} + W_{xh}x_t)W_{hh}$$

$$\frac{\partial L_T}{\partial W} = \frac{\partial L_T}{\partial h_T}\frac{\partial h_t}{\partial h_{t-1}} \cdots \frac{\partial h_1}{\partial W} = \frac{\partial L_T}{\partial h_T}\left(\prod_{t=2}^{T} \frac{\partial h_t}{\partial h_{t-1}}\right)\frac{\partial h_1}{\partial W}$$

$$= \frac{\partial L_T}{\partial h_T}\left(\prod_{t=2}^{T} tanh'(W_{hh}h_{t-1} + W_{xh}x_t)\right)W_{hh}^{T-1}\frac{\partial h_1}{\partial W}$$

Almost always < 1

Vanishing gradients

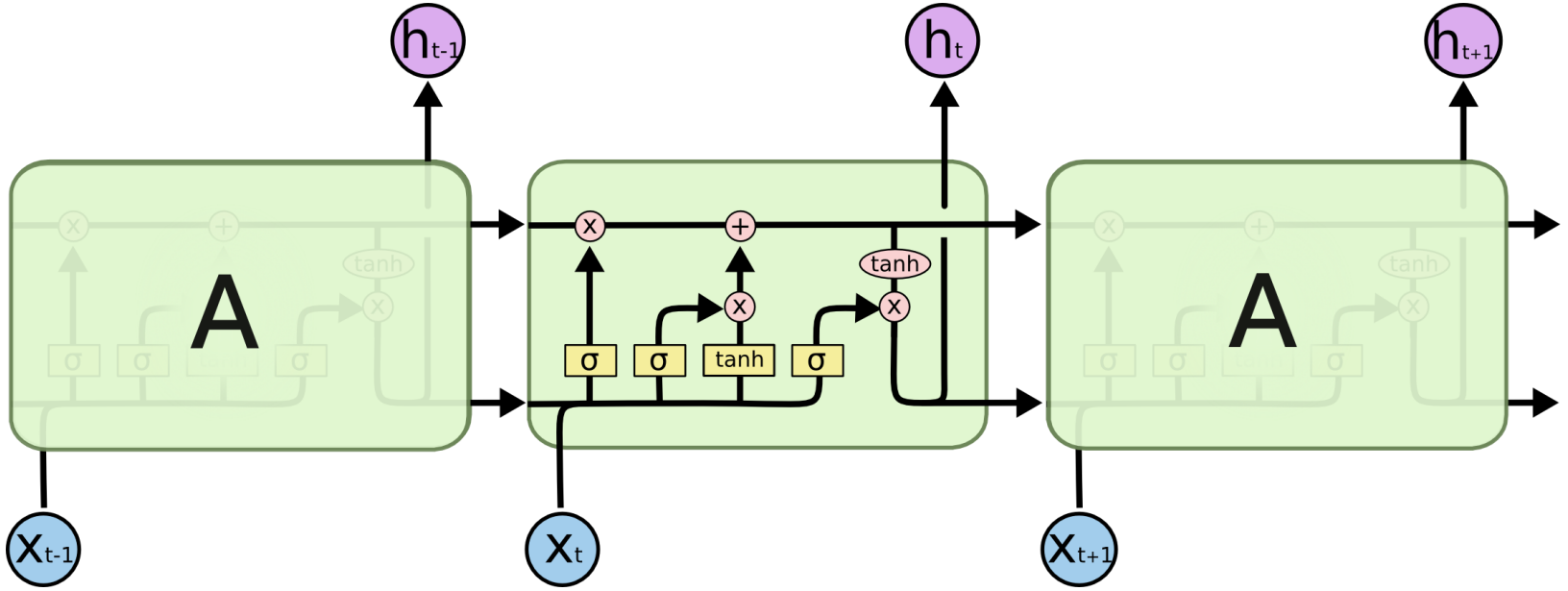
- Vanishing Gradient Problem!

Refer to: Bengio et al, “Learning long-term dependencies with gradient descent is difficult”, IEEE Transactions on Neural Networks, 1994; Pascanu et al, “On the difficulty of training recurrent neural networks”, ICML 2013

Figure from CS231N Stanford University, Credit: Prof. Justin Johnson, Prof. Fei-Fei Li

2D CNN: Dealing With Temporal Information via RNN

- Mitigate vanishing gradient problem by adding residual connections within RNN block – forming the Long Short-Term Memory block (LSTM), a variant of RNN.



- A large portion of video analysis/action recognition methods that leverage RNN defaults to the LSTM variant (some others to the GRU variant) thanks to its performance. Vanilla RNNs tend to also encounter vanishing gradient problem in videos due to data complexity.

LSTM: Hochreiter, Sepp, and Jürgen Schmidhuber. "Long short-term memory." *Neural computation* 9, no. 8 (1997): 1735-1780.

Figure from https://colah.github.io/posts/2015-08-Understanding-LSTMs/, Credit: Christopher Olah (OpenAI)

2D CNN: Dealing With Temporal Information via LSTM

- Karpathy, Andrej, George Toderici, Sanketh Shetty, Thomas Leung, Rahul Sukthankar, and Li Fei-Fei. "Large-scale video classification with convolutional neural networks." In Proceedings of the IEEE conference on Computer Vision and Pattern Recognition, pp. 1725-1732. 2014.

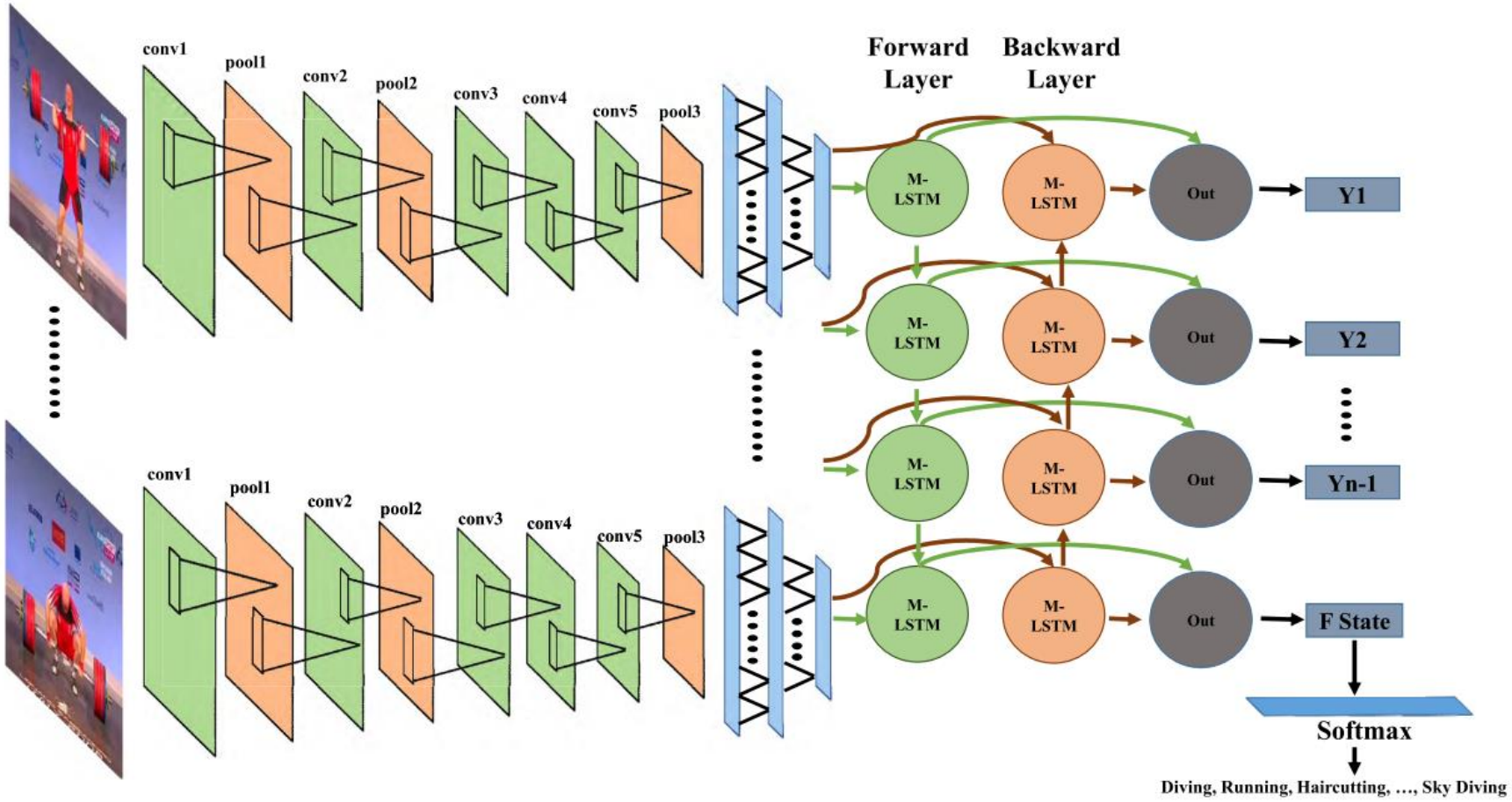
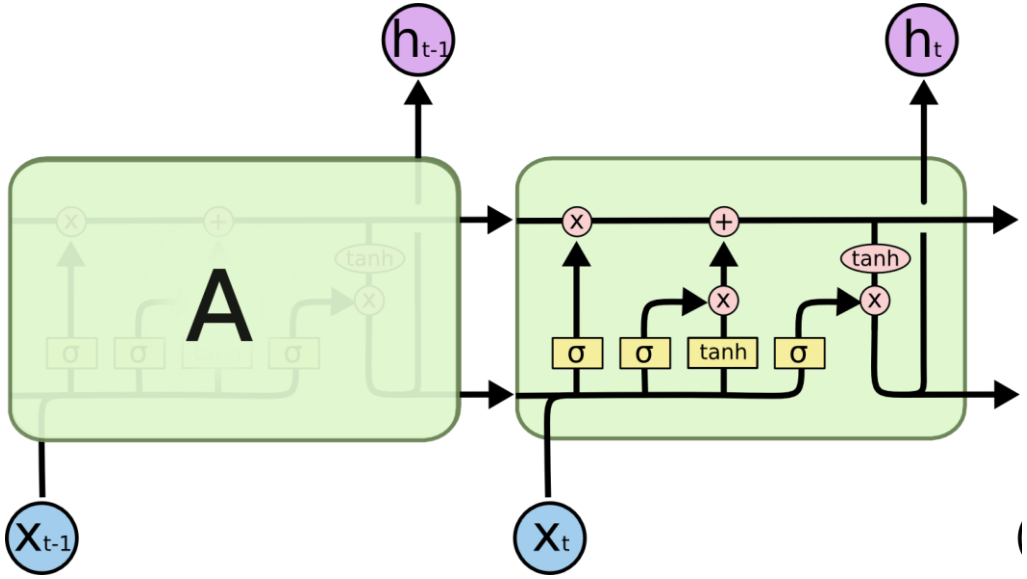


FIGURE 1. Framework of the proposed DB-LSTM for action recognition.

2D CNN: Dealing With Temporal Information via LSTM

- Discussion: advantages and limitations of 2D CNN with LSTM:
  - a) Accuracy? b) Computation Cost? c) Implementation?
- Why LSTM are less commonly used for deep learning task?
  - Cannot proceed to next LSTM block w/o the previous block – Very slow training. (Cannot compute in parallel – GPU advantage)

Method	YouTube	HMDB51	UCF101
Multiresolution CNNs [21]	-	-	65.4%
Proposed DB-LSTM	**92.84**	**87.64**	**91.21**



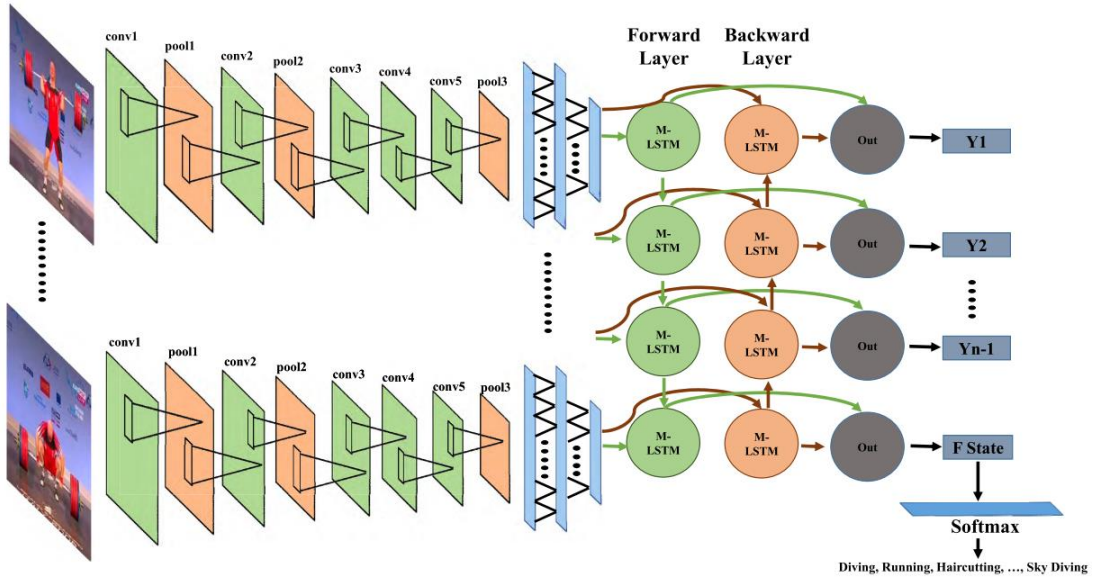
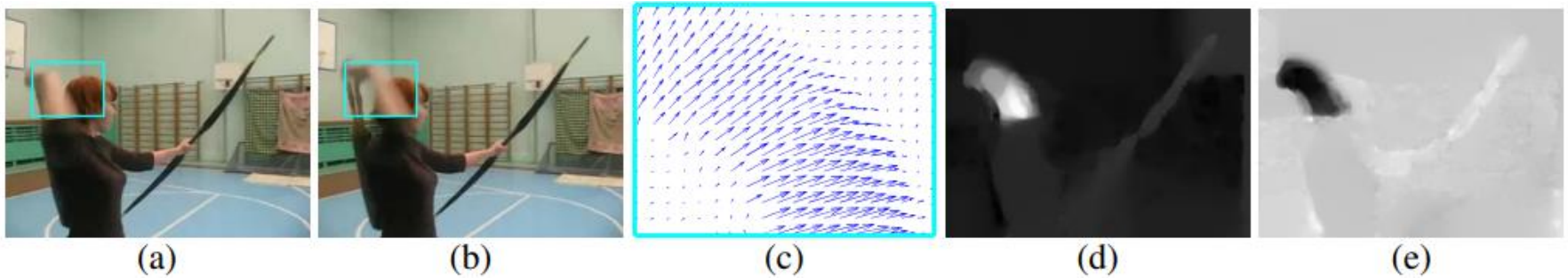
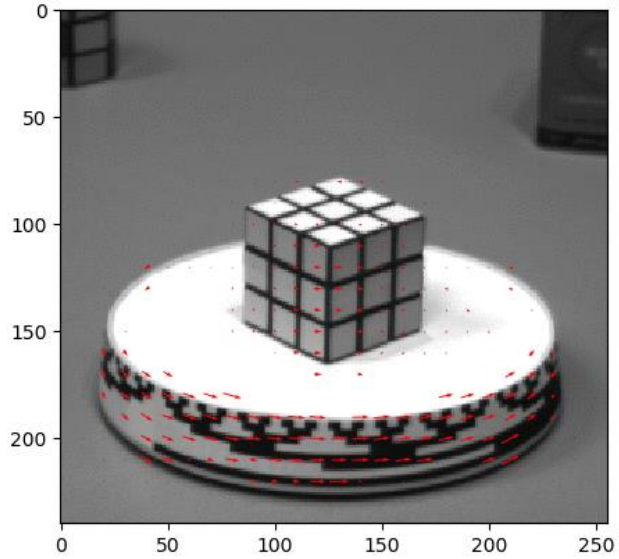


FIGURE 1. Framework of the proposed DB-LSTM for action recognition.

2D CNN: Temporal Information from Optical Flow

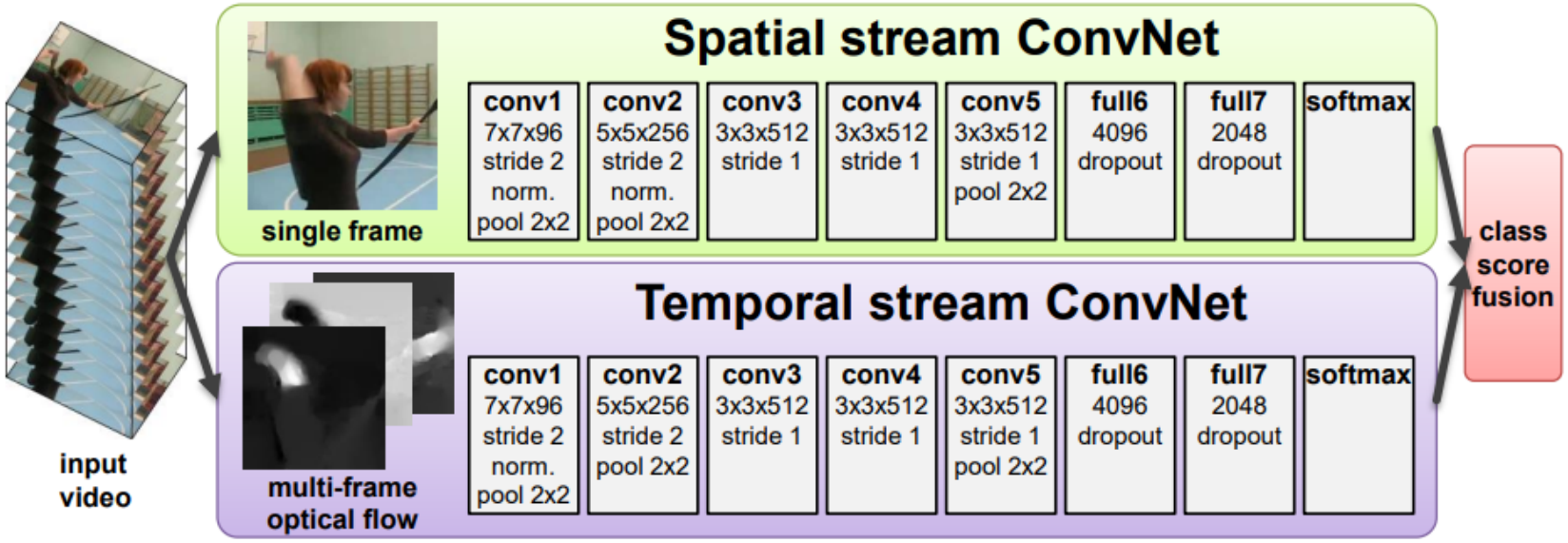
- Other methods to obtain temporal information?
  - Recall: motion information, which is essentially temporal information, can be obtained from **optical flow**.
- Intuitive way to deal with temporal information with 2D CNN:
  - Spatial information from a single frame with 2D CNN; temporal information from optical flow frames with 2D CNN.





2D CNN: Temporal Information from Optical Flow

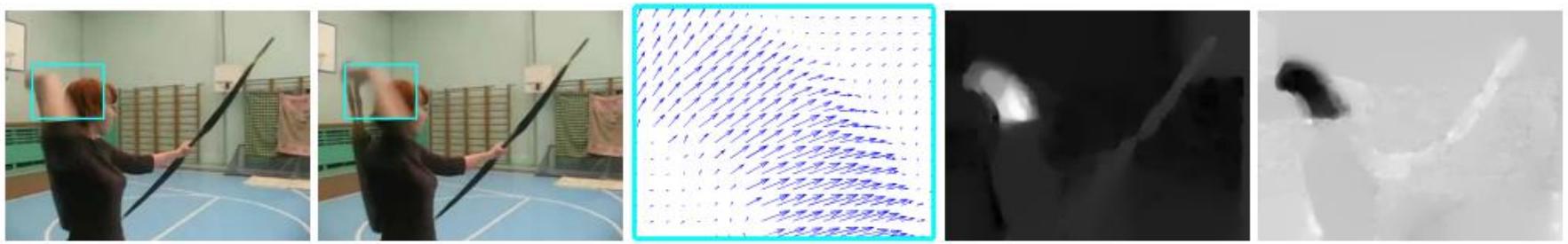
- Simonyan, Karen, and Andrew Zisserman. "Two-stream convolutional networks for action recognition in videos." Advances in neural information processing systems 27 (2014).



Method	UCF-101	HMDB-51
Improved dense trajectories (IDT) [26, 27]	85.9%	57.2%
IDT with higher-dimensional encodings [20]	**87.9%**	61.1%
IDT with stacked Fisher encoding [21] (based on Deep Fisher Net [23])	-	**66.8%**
Spatio-temporal HMAX network [11, 16]	-	22.8%
"Slow fusion" spatio-temporal ConvNet [14]	65.4%	-
Spatial stream ConvNet	73.0%	40.5%
Temporal stream ConvNet	83.7%	54.6%
Two-stream model (fusion by averaging)	86.9%	58.0%
Two-stream model (fusion by SVM)	**88.0%**	**59.4%**

2D CNN: Temporal Information from Optical Flow

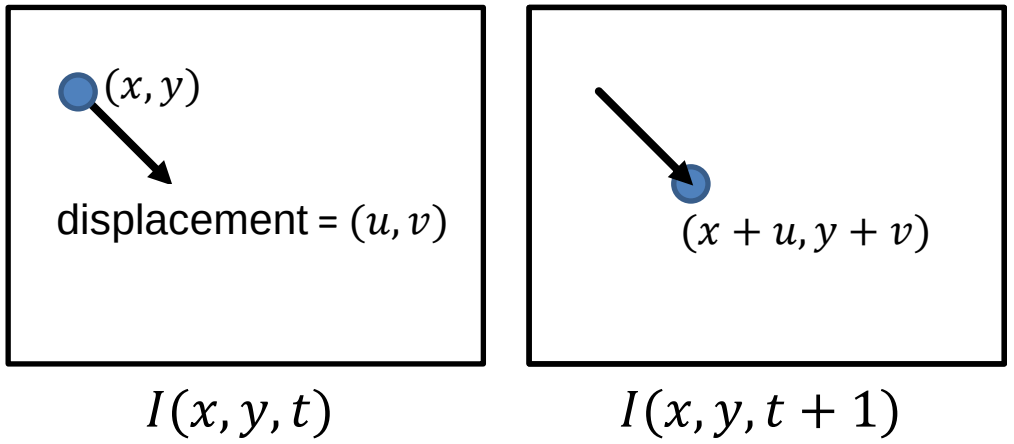
- Discussion: advantages and limitations of 2D CNN with LSTM:
  - a) Accuracy? b) Implementation? c) Computation Cost?
- What do we look forward in Deep Learning?
  - End-to-end process.
  - A balance between computation/memory cost and accuracy.



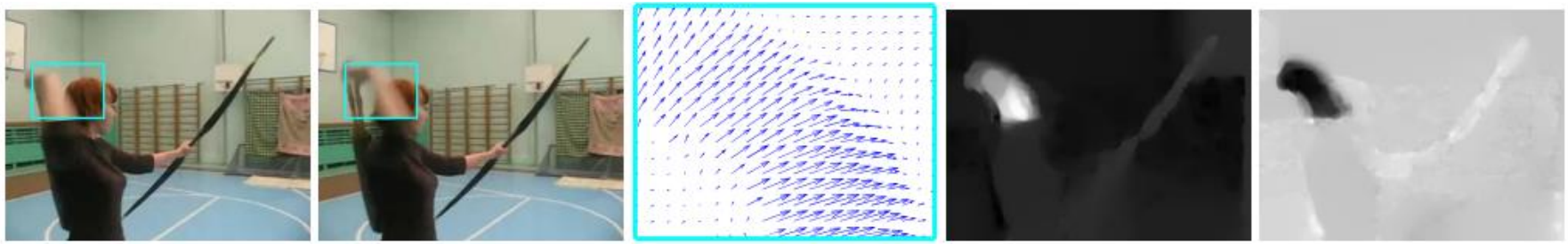
Method	UCF-101	HMDB-51
Improved dense trajectories (IDT) [26, 27]	85.9%	57.2%
IDT with higher-dimensional encodings [20]	**87.9%**	61.1%
IDT with stacked Fisher encoding [21] (based on Deep Fisher Net [23])	-	**66.8%**
Spatio-temporal HMAX network [11, 16]	-	22.8%
"Slow fusion" spatio-temporal ConvNet [14]	65.4%	-
Spatial stream ConvNet	73.0%	40.5%
Temporal stream ConvNet	83.7%	54.6%
Two-stream model (fusion by averaging)	86.9%	58.0%
Two-stream model (fusion by SVM)	**88.0%**	**59.4%**

Recap: Open Questions on Optical Flow

- Why is Optical Flow still leveraged even until today?
- What are the limitations of Optical Flow?



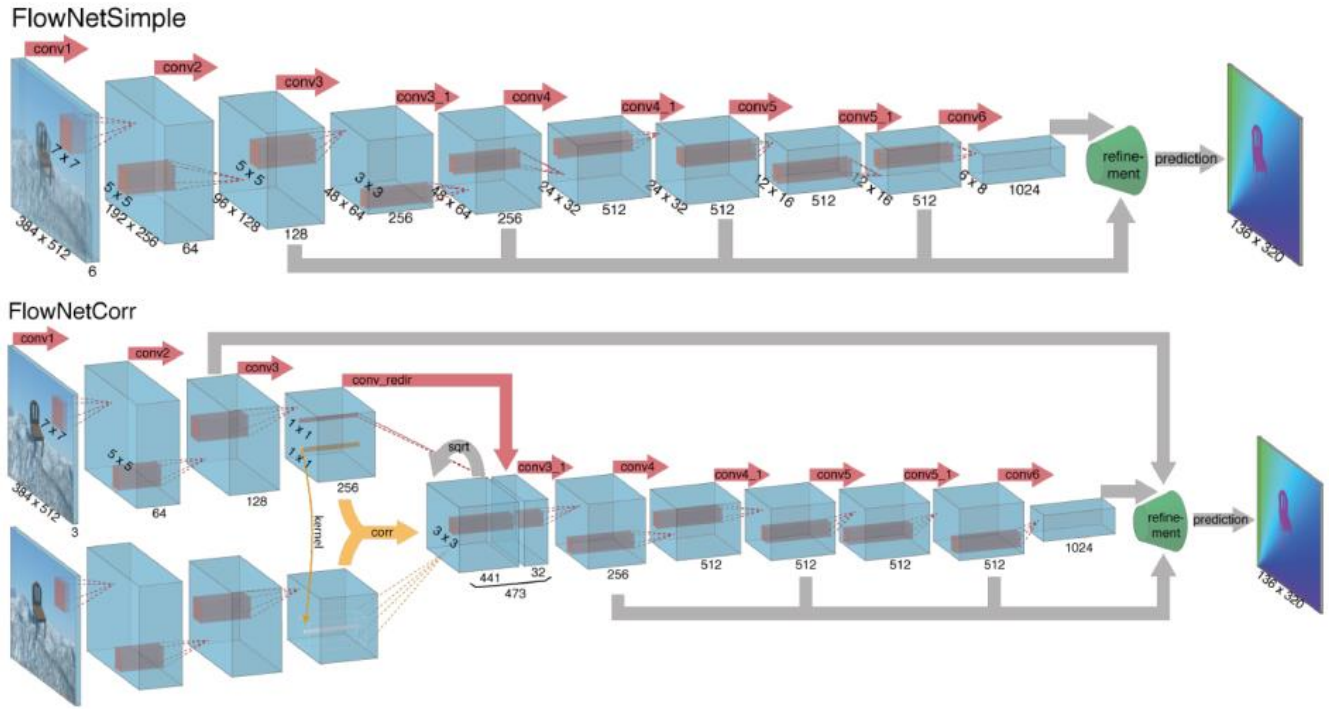
- Core limitations of Optical Flow?
  - NOT end-to-end process.
  - High memory cost. (1MB of RGB → about 3MB of Optical Flow)



Alternative: Estimating Optical Flow with Deep Learning

- Merging optical flow estimation with deep learning structure better:
  - End-to-end optical flow with neural networks
  - Note: optical flow is obtained from AT LEAST two video frames.

- Dosovitskiy, Alexey, Philipp Fischer, Eddy Ilg, Philip Hausser, Caner Hazirbas, Vladimir Golkov, Patrick Van Der Smagt, Daniel Cremers, and Thomas Brox. "Flownet: Learning optical flow with convolutional networks." In Proceedings of the *IEEE international conference on computer vision*, pp. 2758-2766. 2015.

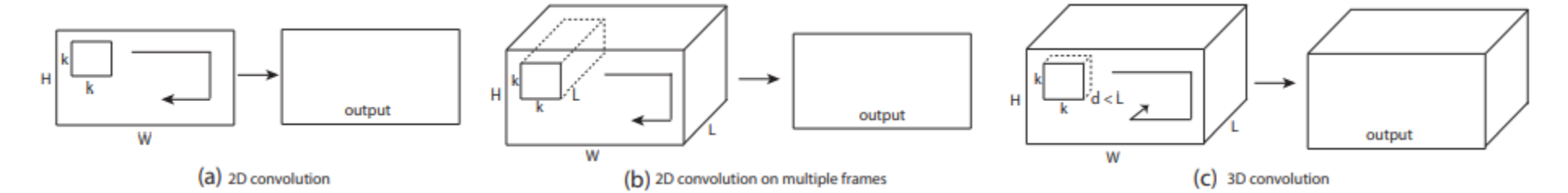


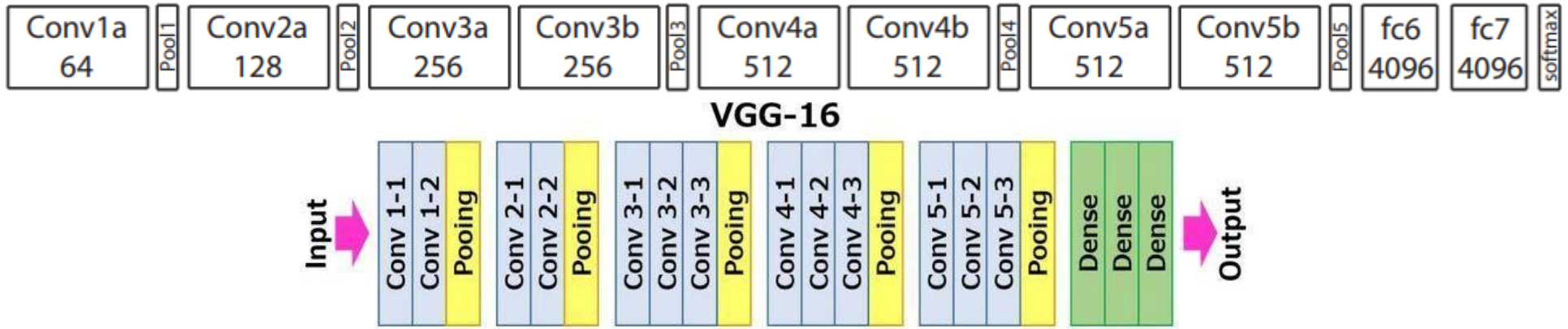
3D CNN: Joint Modeling of Spatial and Temporal Info

- What do we look forward in Deep Learning?
  - **End-to-end process – can we model spatial and temporal information jointly?**
  - A balance between computation/memory cost and accuracy.

- Tran, Du, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. "Learning spatiotemporal features with 3d convolutional networks." In Proceedings of the IEEE international conference on computer vision, pp. 4489-4497. 2015.

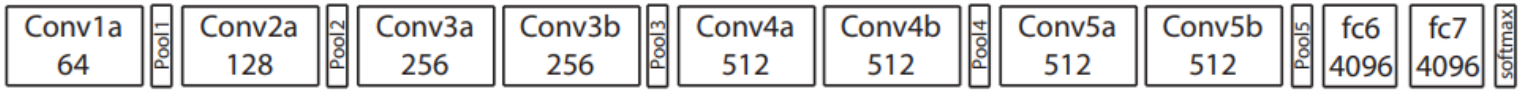
  - 3D CNN: expanding 2D convolution kernel to the temporal axis. (C3D)

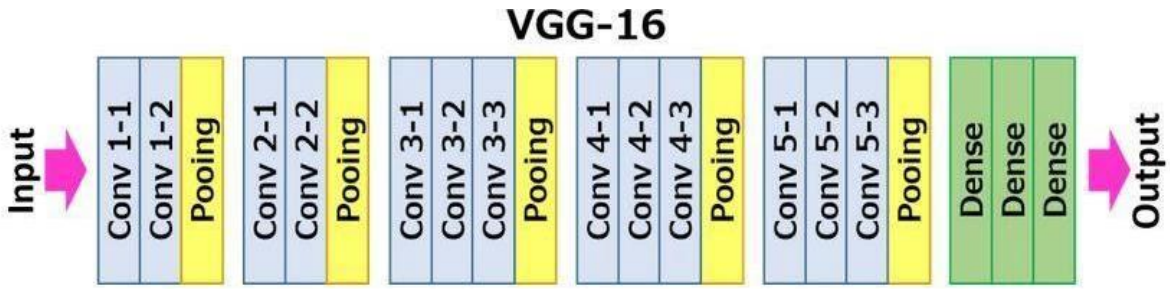




3D CNN: Joint Modeling of Spatial and Temporal Info

- What do we look forward in Deep Learning?
  - End-to-end process.
  - **A balance between computation/memory cost and accuracy.**





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

Layer	Size	MFLOPs
Input	3 x 16 x 112 x 112	
Conv1 (3x3x3)	64 x 16 x 112 x 112	1.04
Pool1 (1x2x2)	64 x 16 x 56 x 56	
Conv2 (3x3x3)	128 x 16 x 56 x 56	11.10
Pool2 (2x2x2)	128 x 8 x 28 x 28	
Conv3a (3x3x3)	256 x 8 x 28 x 28	5.55
Conv3b (3x3x3)	256 x 8 x 28 x 28	11.10
Pool3 (2x2x2)	256 x 4 x 14 x 14	
Conv4a (3x3x3)	512 x 4 x 14 x 14	2.77
Conv4b (3x3x3)	512 x 4 x 14 x 14	5.55
Pool4 (2x2x2)	512 x 2 x 7 x 7	
Conv5a (3x3x3)	512 x 2 x 7 x 7	0.69
Conv5b (3x3x3)	512 x 2 x 7 x 7	0.69
Pool5	512 x 1 x 3 x 3	
FC6	4096	0.51
FC7	4096	0.45
FC8	C	0.05

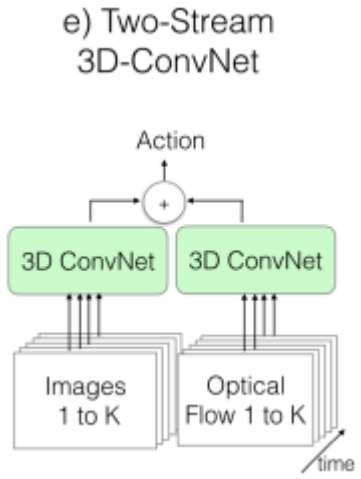
AlexNet: 0.7 GFLOP
VGG-16: 13.6 GFLOP
C3D: **39.5 GFLOP (2.9x VGG!)**

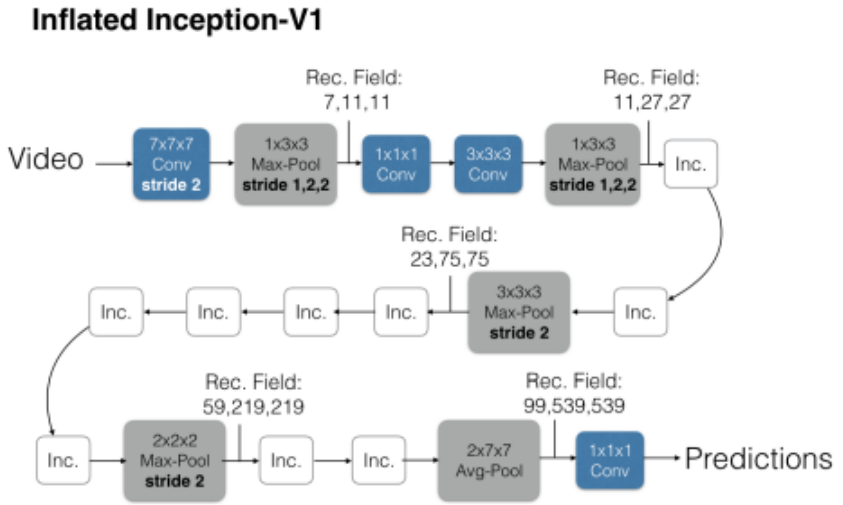
- Essentially, C3D is only a 10/11 layer-CNN (count only convolution and dense/fully-connected layers).
- Its computation cost (in GFLOPs) is 2× that of VGG-16.
- Require initialization of all 3D kernels, which is also computationally costly.

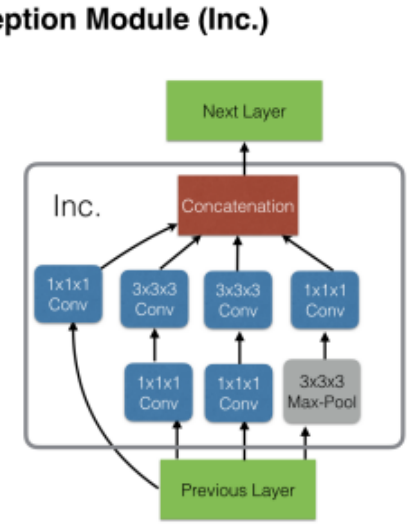
3D CNN: Joint Modeling of Spatial and Temporal Info (2)

- What do we look forward in Deep Learning?
  - A balance between computation/memory cost and accuracy.
  - Can we better leverage the relationship of 2D CNN and 3D CNN for easier initialization and lower computation cost?

- Carreira, Joao, and Andrew Zisserman. "Quo vadis, action recognition? a new model and the kinetics dataset." In proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 6299-6308. 2017.
  - I3D: inflating 2D CNNs to 3D; bootstrapping 3D filters from 2D filters (initialize through copying 2D parameters across temporal dimension.

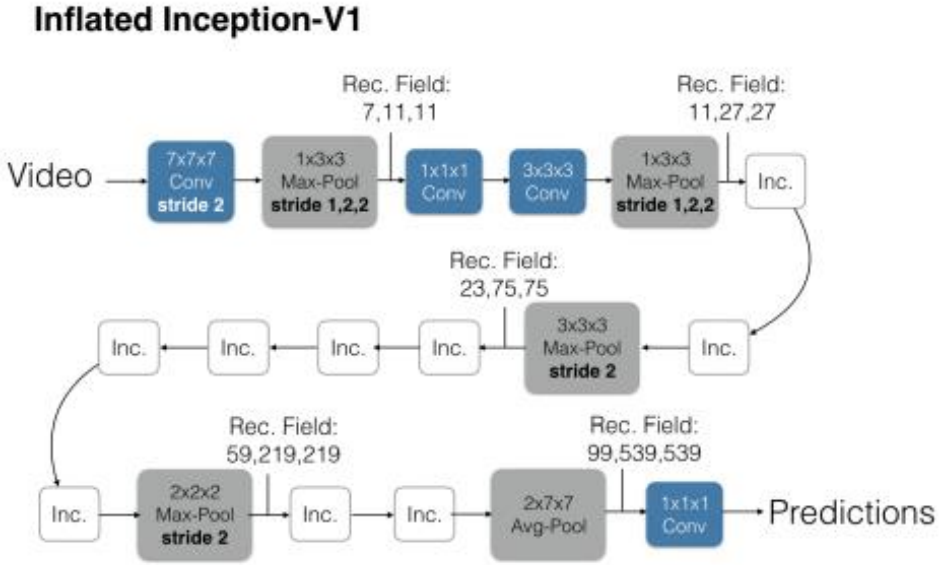


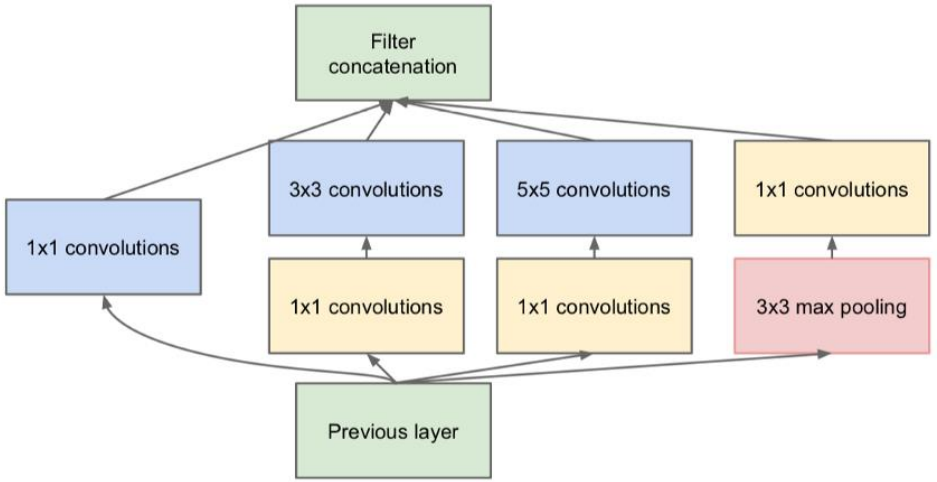




3D CNN: Joint Modeling of Spatial and Temporal Info (2)

- Carreira, Joao, and Andrew Zisserman. "Quo vadis, action recognition? a new model and the kinetics dataset." In *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 6299-6308. 2017.
  - I3D: inflating 2D CNNs to 3D; bootstrapping 3D filters from 2D filters (initialize through copying 2D parameters across temporal dimension.
    - Easy construct of 3D CNNs (Originally Inception-I3D, subsequently ResNet-I3D)
    - Leverage on efficient structure from the original 2D CNNs (Inception is more efficient thanks to its dimension reduction design across channels)





3D CNN: Joint Modeling of Spatial and Temporal Info (2)

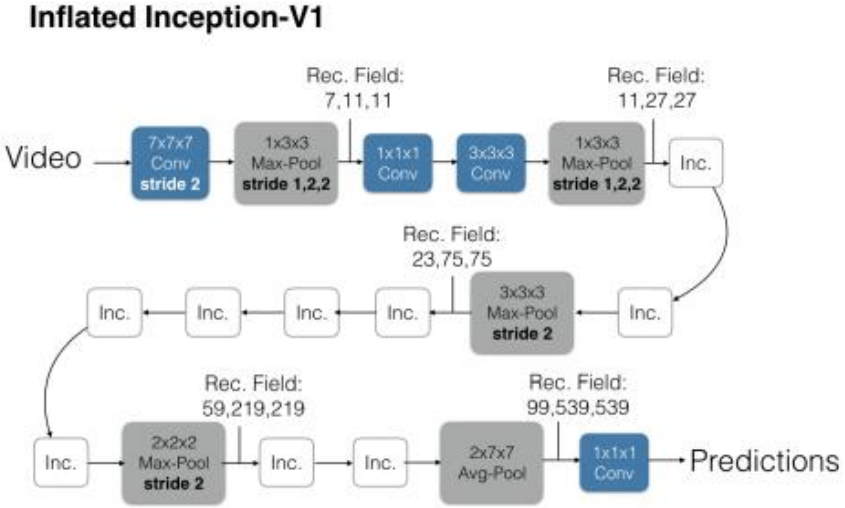
- Carreira, Joao, and Andrew Zisserman. "Quo vadis, action recognition? a new model and the kinetics dataset." In *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 6299-6308. 2017.
  - I3D: inflating 2D CNNs to 3D; bootstrapping 3D filters from 2D filters (initialize through copying 2D parameters across temporal dimension.
    - Easy construct of 3D CNNs (Originally Inception-I3D, subsequently ResNet-I3D)
    - Leverage on efficient structure from the original 2D CNNs (Inception is more efficient thanks to its dimension reduction design across channels)

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

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**

3D CNN: Joint Modeling of Spatial and Temporal Info (3)

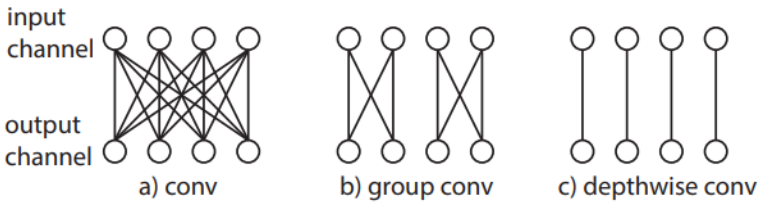
- Discussion: advantages and limitations of I3D:
  - Computation Cost – can we further improve the computation efficiency of 3D CNNs?
  - Accuracy – is 3D CNN truly the best solution for action recognition?
- Computation Cost: what determines the computation cost of 3D CNN?
  - Parameter size; FLOPs; Channel Interactions.
  - How to further reduce computation cost?



$$\#\text{parameters} = C_{out} \cdot \frac{C_{in}}{G} \cdot k^3 \quad (1)$$

$$\#\text{FLOPs} = C_{out} \cdot \frac{C_{in}}{G} \cdot k^3 \cdot THW \quad (2)$$

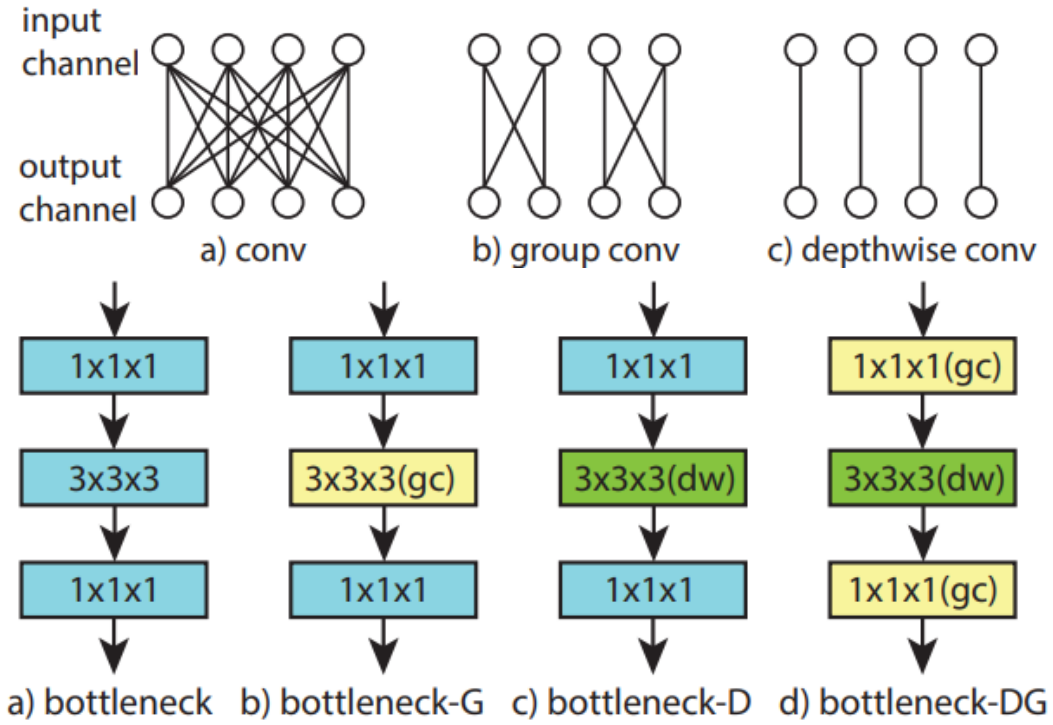
$$\#\text{interactions} = C_{out} \cdot \binom{\frac{C_{in}}{G}}{2} \quad (3)$$



3D CNN: Joint Modeling of Spatial and Temporal Info (3)

- Channel grouping: a efficient and effective way for reducing computation cost while maintain model performance

- Tran, Du, Heng Wang, Lorenzo Torresani, and Matt Feiszli. "Video classification with channel-separated convolutional networks." In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 5552-5561. 2019.

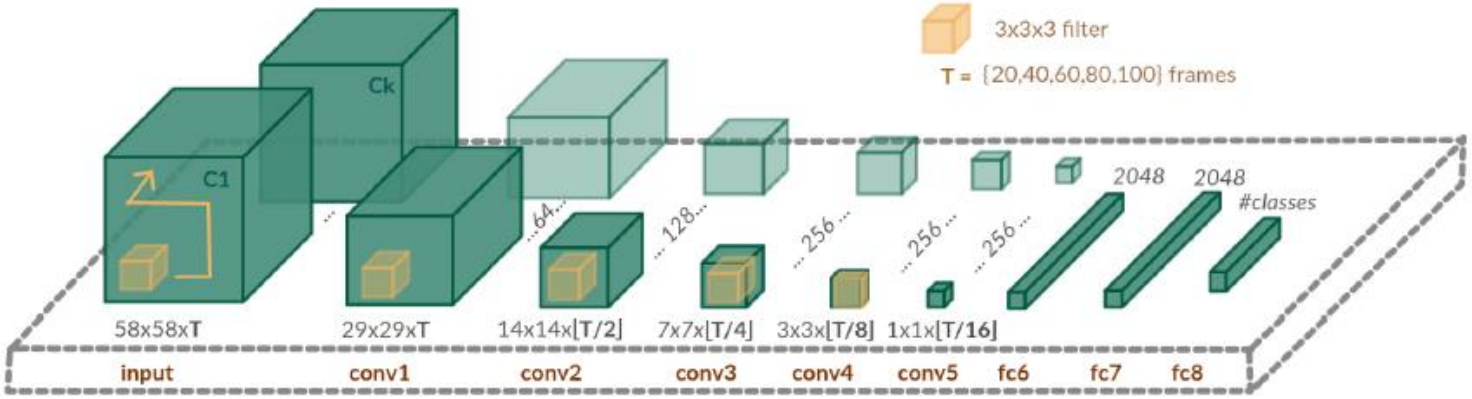


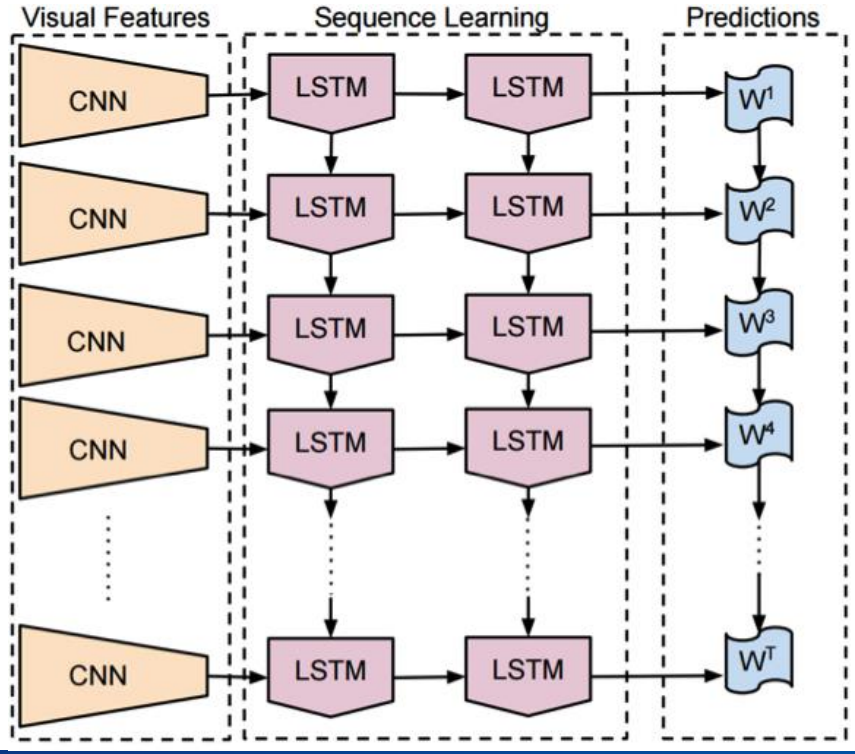
Method	input	video@1	video@5	GFLOPs×crops
C3D [29]	RGB	61.1	85.2	N/A
P3D [23]	RGB	66.4	87.4	N/A
Conv Pool [38]	RGB+OF	71.7	90.4	N/A
R(2+1)D [30]	RGB	73.0	91.5	152×N/A
R(2+1)D [30]	RGB+OF	73.3	91.9	305×N/A
ir-CSN-101	RGB	74.8	92.6	56.5×10
ip-CSN-101	RGB	74.9	92.6	63.6×10
ir-CSN-152	RGB	**75.5**	**92.7**	74.0×10
ip-CSN-152	RGB	**75.5**	**92.8**	83.3×10

Table 4. **Comparison with state-of-the-art architectures on Sports1M**. Our CSNs with 101 or 152 layers outperform all the previous models by good margins while being 2-4x faster.

Global/Long-Term Temporal Information

- Recall: discussion on temporal information/motion information:
  - Temporal information/motion information can be monitored **globally** or **locally**.
- Both RNN/LSTM and 3D CNNs extract temporal information "**locally**".
  - Global temporal information are obtained by stacking multiple RNN/CNN blocks.

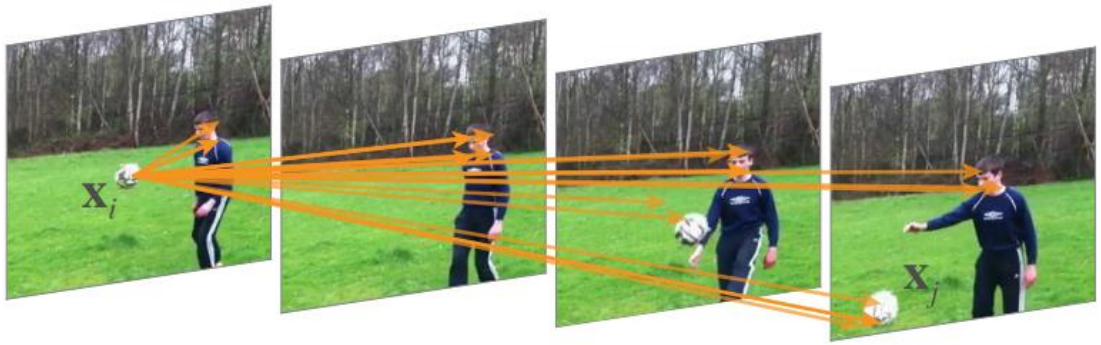




Global/Long-Term Temporal Information

- Both RNN/LSTM and 3D CNNs extract temporal information "**locally**".
  - Can we obtain global temporal information from a single block/module?

- Wang, Xiaolong, Ross Girshick, Abhinav Gupta, and Kaiming He. "Non-local neural networks." In Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 7794-7803. 2018.
  - The Non-Local (NL) module: capturing **long-range** dependencies with a non-local operation that computes the response at a position as a weighted sum of the features at *all positions* in the input feature maps.



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

Gaussian	$f(\mathbf{x}_i, \mathbf{x}_j) = e^{\mathbf{x}_i^T \mathbf{x}_j}.$
Embedded Gaussian	$f(\mathbf{x}_i, \mathbf{x}_j) = e^{\theta(\mathbf{x}_i)^T \phi(\mathbf{x}_j)}.$
Dot Product	$f(\mathbf{x}_i, \mathbf{x}_j) = \theta(\mathbf{x}_i)^T \phi(\mathbf{x}_j).$
Concatenation	$f(\mathbf{x}_i, \mathbf{x}_j) = \mathrm{ReLU}(\mathbf{w}_f^T [\theta(\mathbf{x}_i), \phi(\mathbf{x}_j)]).$

Global/Long-Term Temporal Information

- Wang, Xiaolong, Ross Girshick, Abhinav Gupta, and Kaiming He. "Non-local neural networks." In Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 7794-7803. 2018.
  - The Non-Local (NL) module: capturing **long-range** dependencies with a non-local operation that computes the response at a position as a weighted sum of the features at *all positions* in the input feature maps.

Gaussian $f(\mathbf{x}_i, \mathbf{x}_j) = e^{\mathbf{x}_i^T \mathbf{x}_j}$.

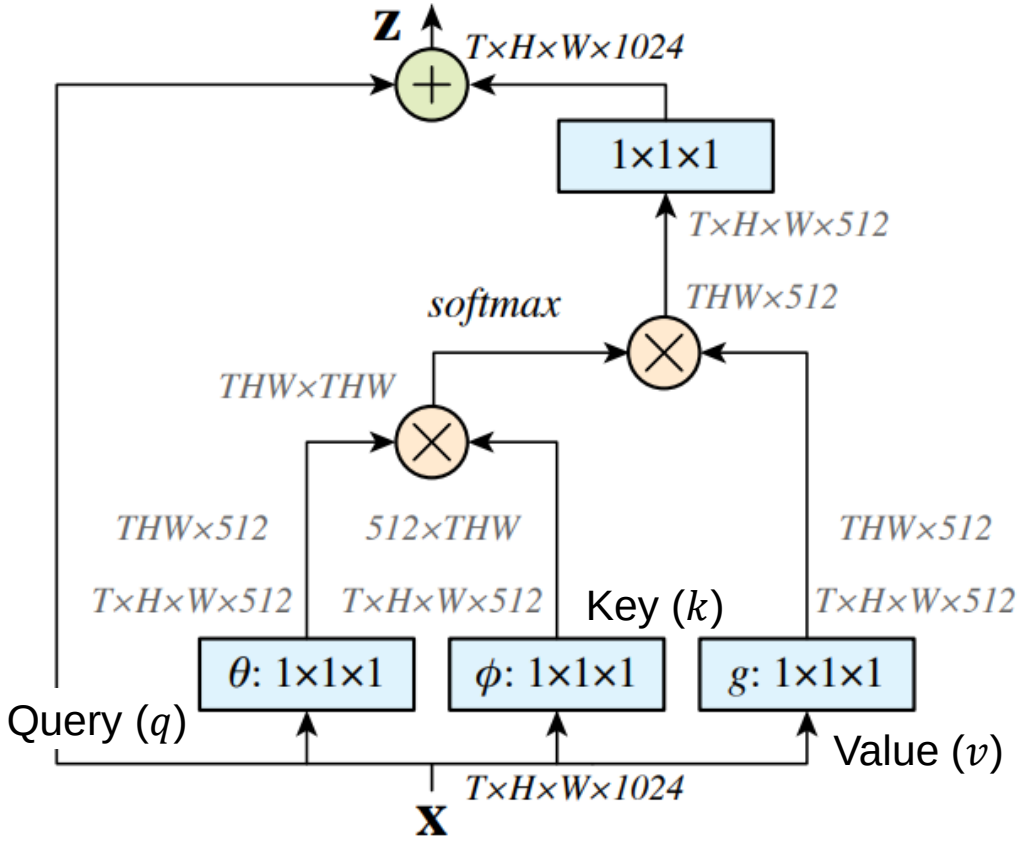
Embedded Gaussian $f(\mathbf{x}_i, \mathbf{x}_j) = e^{\theta(\mathbf{x}_i)^T \phi(\mathbf{x}_j)}$.

Dot Product

- If the pairwise function is in Embedded Gaussian form, $\frac{1}{C(\mathbf{x})} f(\mathbf{x}_i, \mathbf{x}_j)$ is equal to the *softmax* computation along *j*, thus the pairwise function can be re-written as: $\mathbf{y} = softmax(\mathbf{x}^T W_\theta^T W_\phi^T \mathbf{x}) g(\mathbf{x})$, which is exactly the "Transformer"-style self-attention!

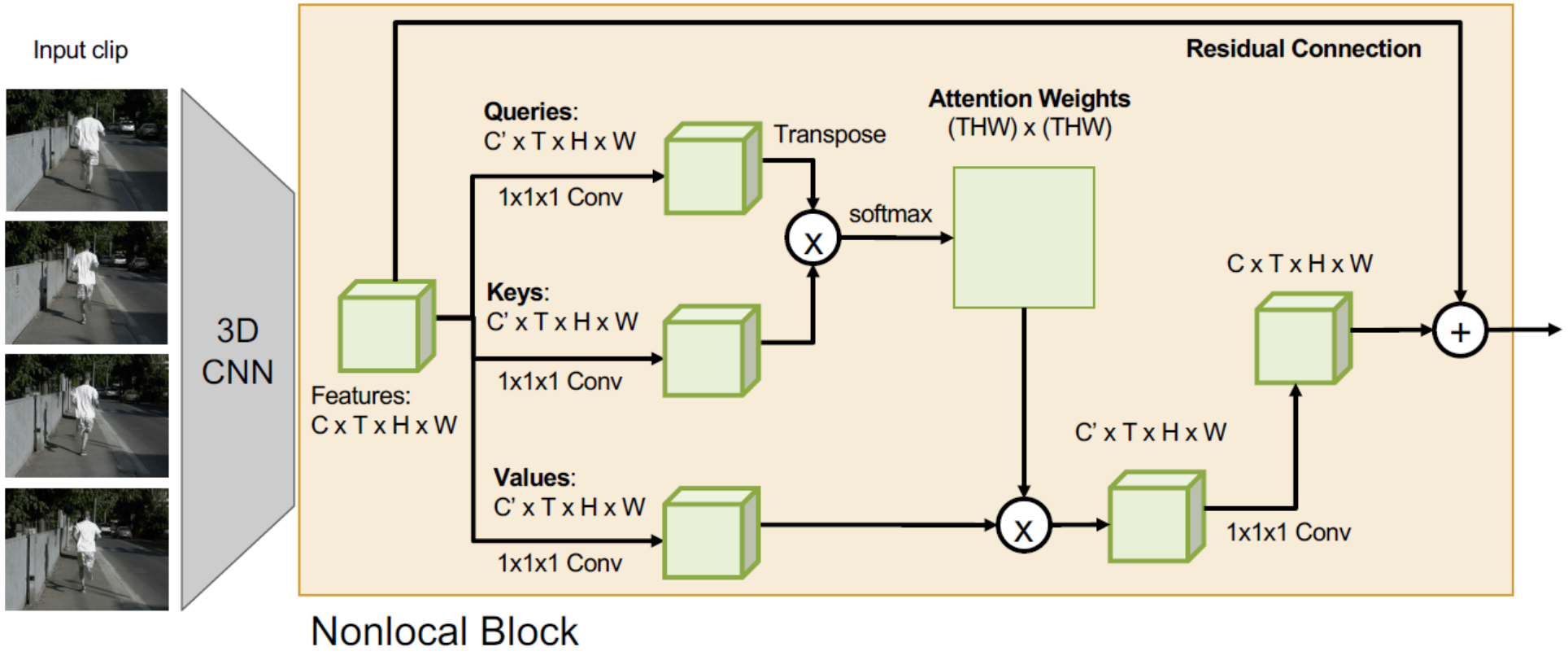
Concatenation $f(\mathbf{x}_i, \mathbf{x}_j) = \mathrm{ReLU}(\mathbf{w}_f^T [\theta(\mathbf{x}_i), \phi(\mathbf{x}_j)])$.

- If the pairwise function is in Embedded Gaussian form, $\frac{1}{C(\mathbf{x})} f(\mathbf{x}_i, \mathbf{x}_j)$ is equal to the *softmax* computation along *j*, thus the pairwise function can be re-written as: $\mathbf{y} = softmax(\mathbf{x}^T W_\theta^T W_\phi^T \mathbf{x}) g(\mathbf{x})$, which is exactly the "Transformer"-style self-attention!



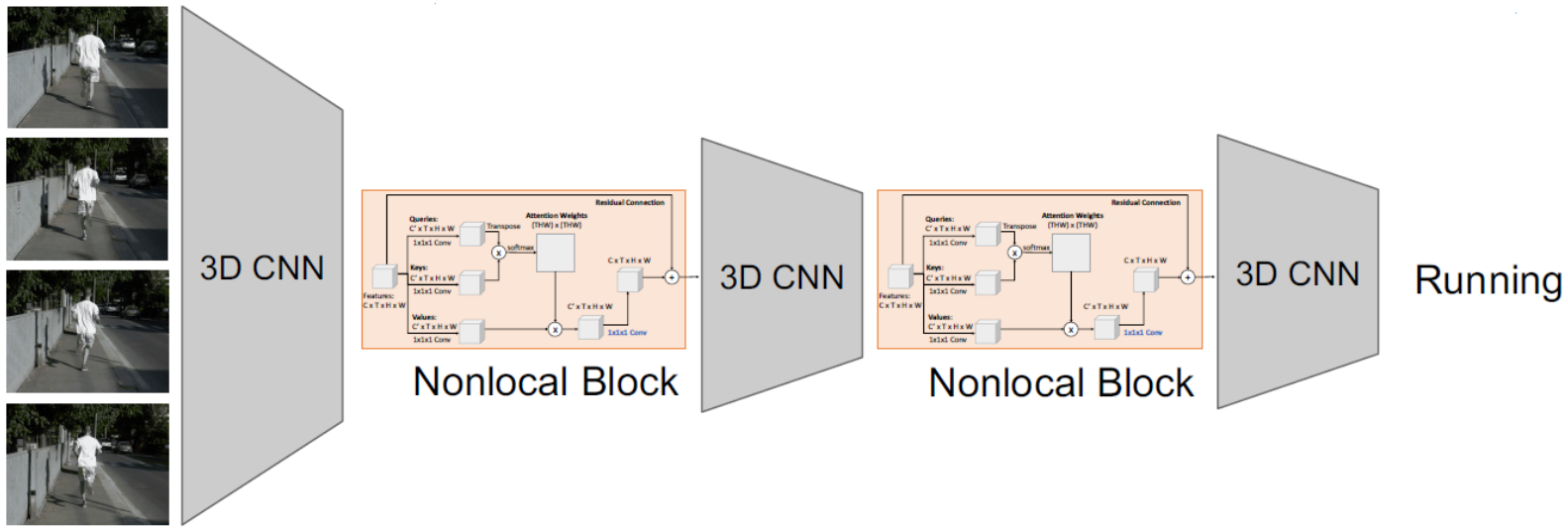
Global/Long-Term Temporal Information

- Wang, Xiaolong, Ross Girshick, Abhinav Gupta, and Kaiming He. "Non-local neural networks." In Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 7794-7803. 2018.



Global/Long-Term Temporal Information

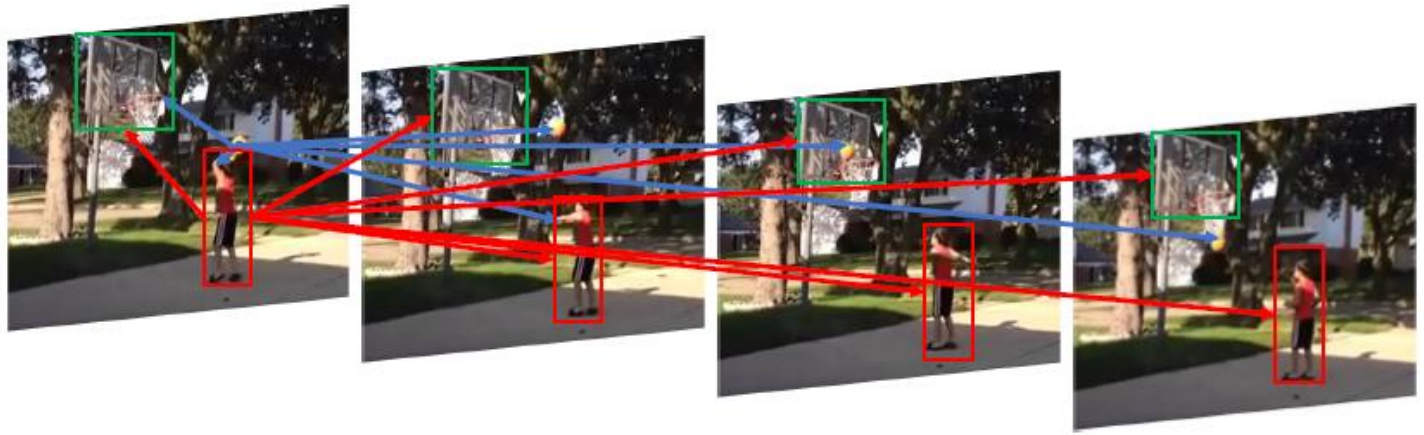
- Wang, Xiaolong, Ross Girshick, Abhinav Gupta, and Kaiming He. "Non-local neural networks." In Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 7794-7803. 2018.

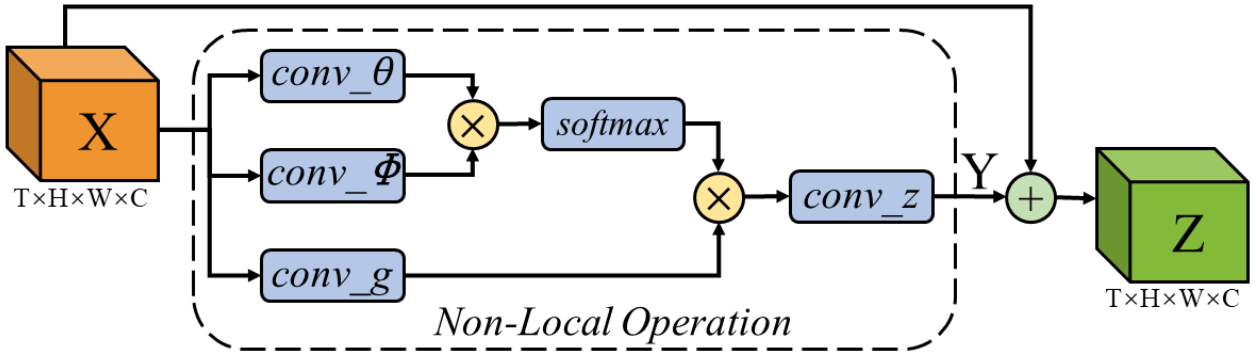


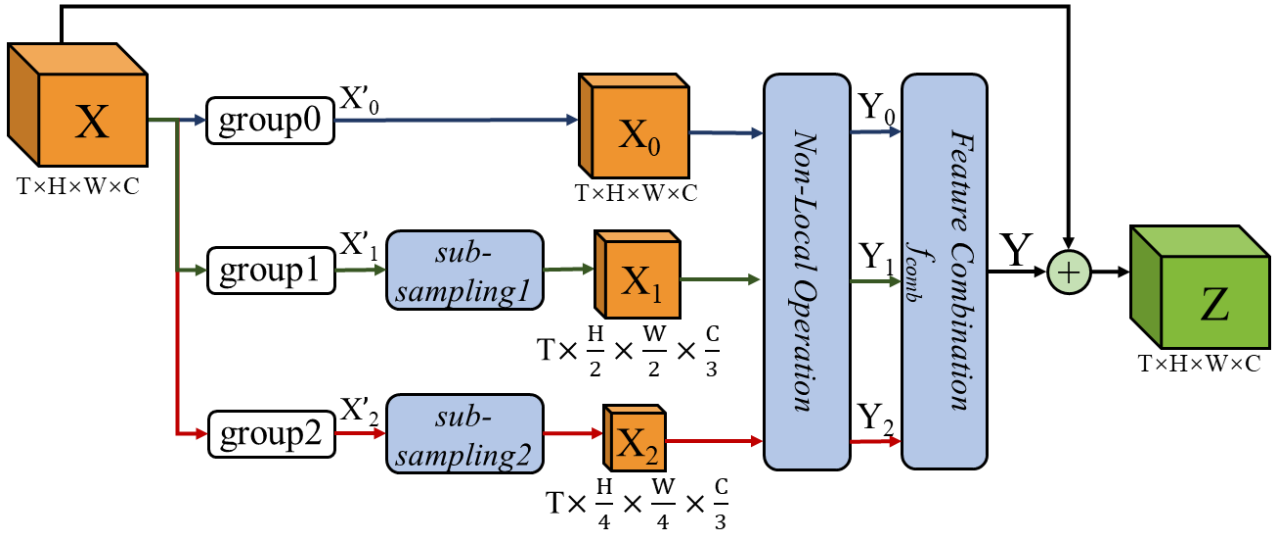
model	backbone	modality	top-1 val	top-5 val	top-1 test	top-5 test	avg test†
I3D in [7]	Inception	RGB	72.1	90.3	71.1	89.3	80.2
2-Stream I3D in [7]	Inception	RGB + flow	75.7	92.0	74.2	91.3	82.8
RGB baseline in [3]	Inception-ResNet-v2	RGB	73.0	90.9	-	-	-
3-stream late fusion [3]	Inception-ResNet-v2	RGB + flow + audio	74.9	91.6	-	-	-
3-stream LSTM [3]	Inception-ResNet-v2	RGB + flow + audio	77.1	93.2	-	-	-
3-stream SATT [3]	Inception-ResNet-v2	RGB + flow + audio	77.7	93.2	-	-	-
NL I3D [ours]	ResNet-50	RGB	76.5	92.6	-	-	-
	ResNet-101	RGB	**77.7**	**93.3**	-	-	**83.8**

Global/Long-Term Temporal Information

- Xu, Yuecong, Haozhi Cao, Jianfei Yang, Kezhi Mao, Jianxiong Yin, and Simon See. "PNL: Efficient long-range dependencies extraction with pyramid non-local module for action recognition." *Neurocomputing* 447 (2021): 282-293.



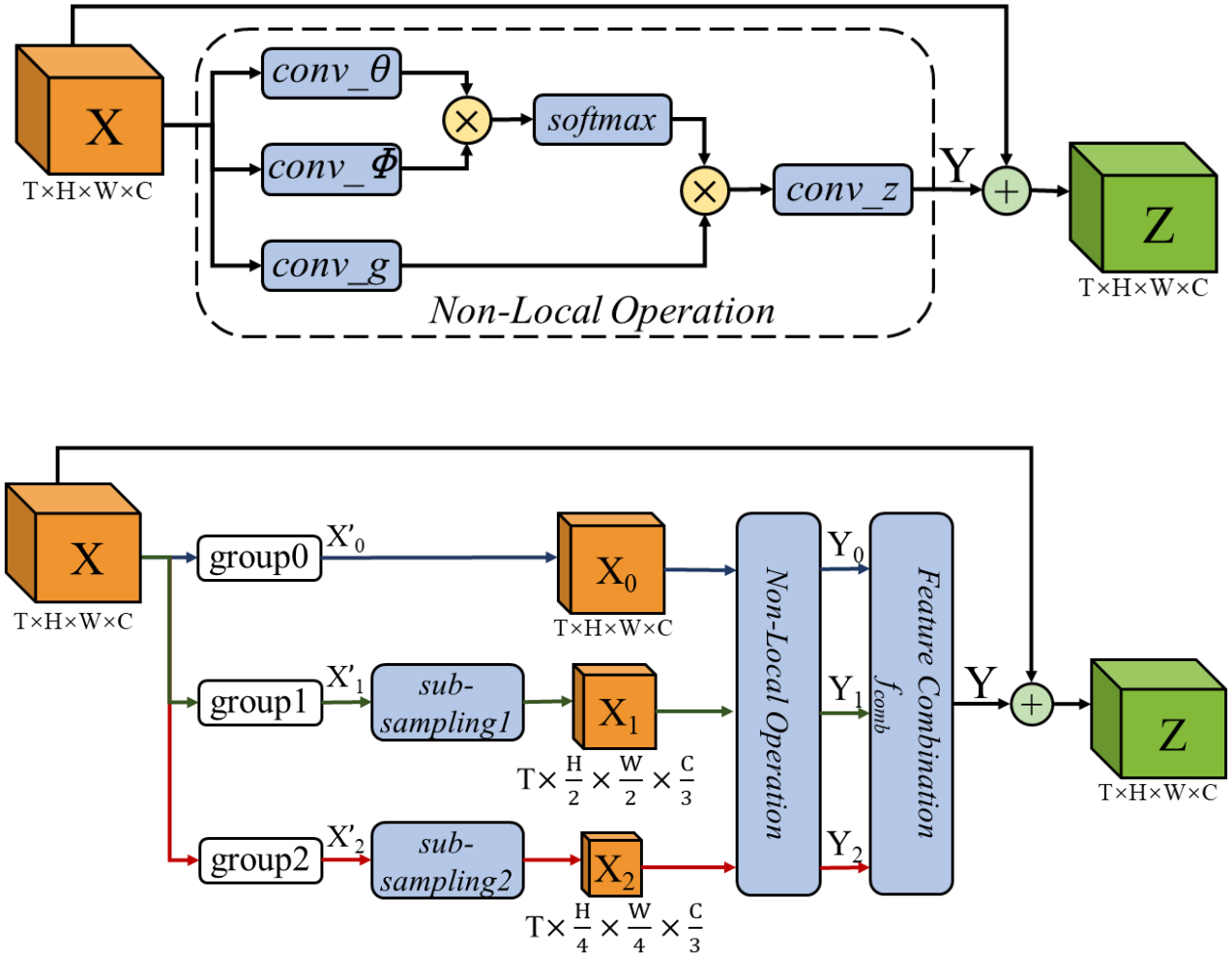




Global/Long-Term Temporal Information

- Xu, Yuecong, Haozhi Cao, Jianfei Yang, Kezhi Mao, Jianxiong Yin, and Simon See. "PNL: Efficient long-range dependencies extraction with pyramid non-local module for action recognition." *Neurocomputing* 447 (2021): 282-293.

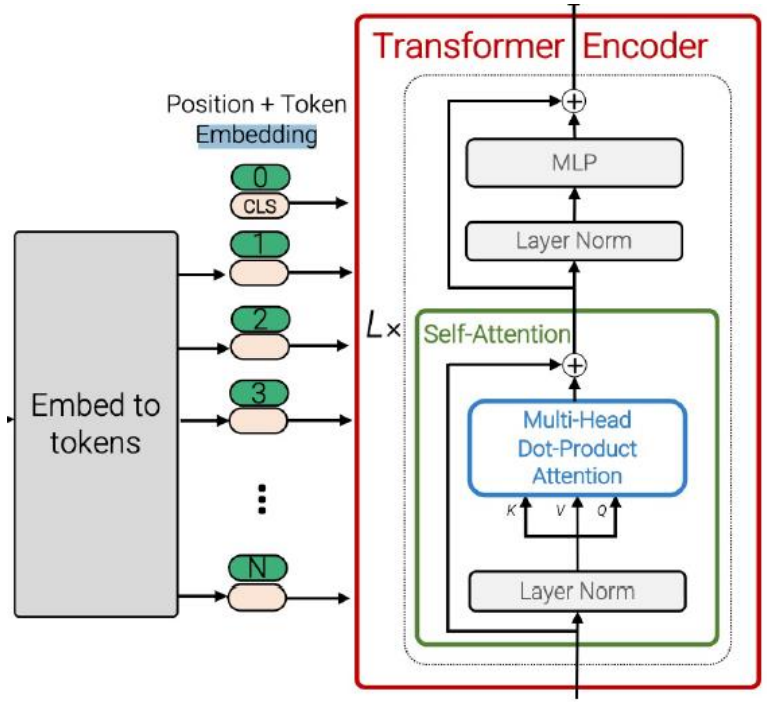
	Method	Mini-Kinetics Top-1	# Params	FLOPs
Two-stream CNNs	MARS [54]	73.5%	-	-
	ResFrame TS [55]	73.9%	-	-
	I3D (TS) [32]	78.7%	25.0M	>107.9G
3D CNNs	C3D [20]	66.2%	33.3M	-
	I3D (RGB) [32]	74.1%	12.06M	107.9G
	(2+C1)D [56]	75.74%	**7.3M**	31.9G
	S3D [17]	78.0%	8.77M	43.47G
	MFNet [48]	78.35%	7.84M	**11.17G**
CNN with long-range dependencies	Res50-NL [16]	77.53%	27.66M	19.67G
	Res50-CGD [57]	77.56%	25.58M	17.88G
	Res50-CGNL [34]	77.76%	27.2M	19.16G
	MFNet-NL [34]	79.74%	8.15M	11.66G
Ours	**MFNet-PNL(×1)**	82.16%	7.92M	11.22G
	MFNet-PNL(×5)	**83.09%**	8.12M	11.38G

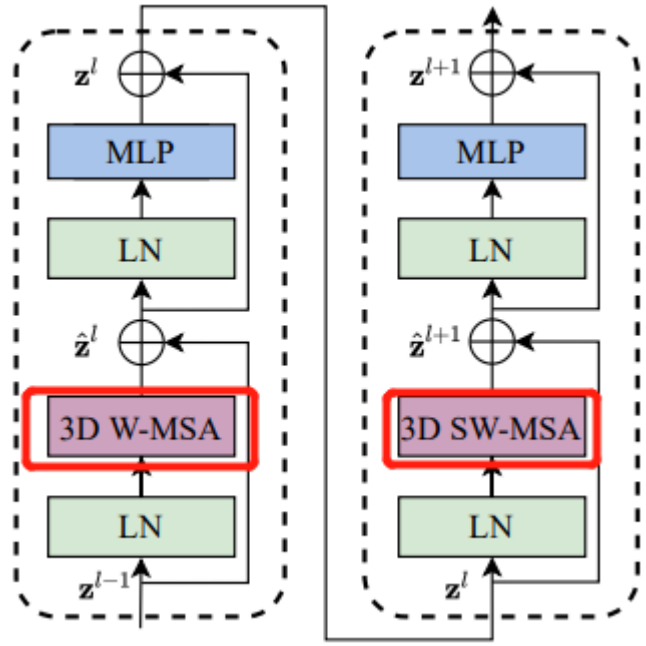


Global/Long-Term Temporal Information

- Since obtaining global/long-term temporal information is useful (and crucial) towards action recognition, can we apply it **without** 3D-CNN (pure global temporal feature)?
  - Video Transformer (Video Swin Transformer as example)

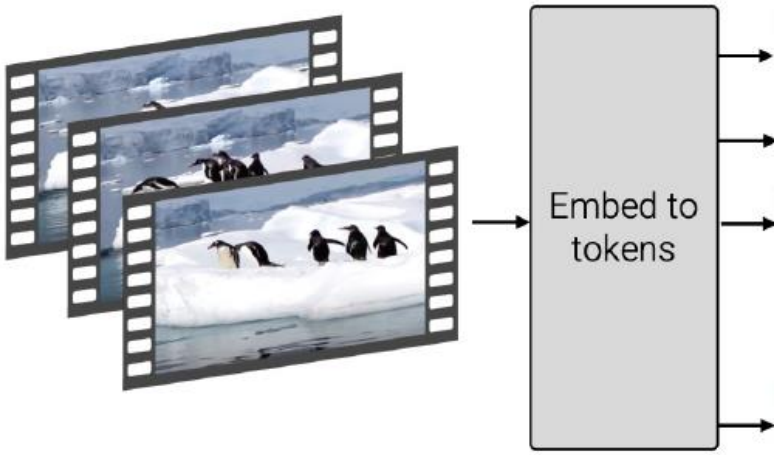
- Liu, Ze, Jia Ning, Yue Cao, Yixuan Wei, Zheng Zhang, Stephen Lin, and Han Hu. "Video swin transformer." In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3202-3211. 2022.

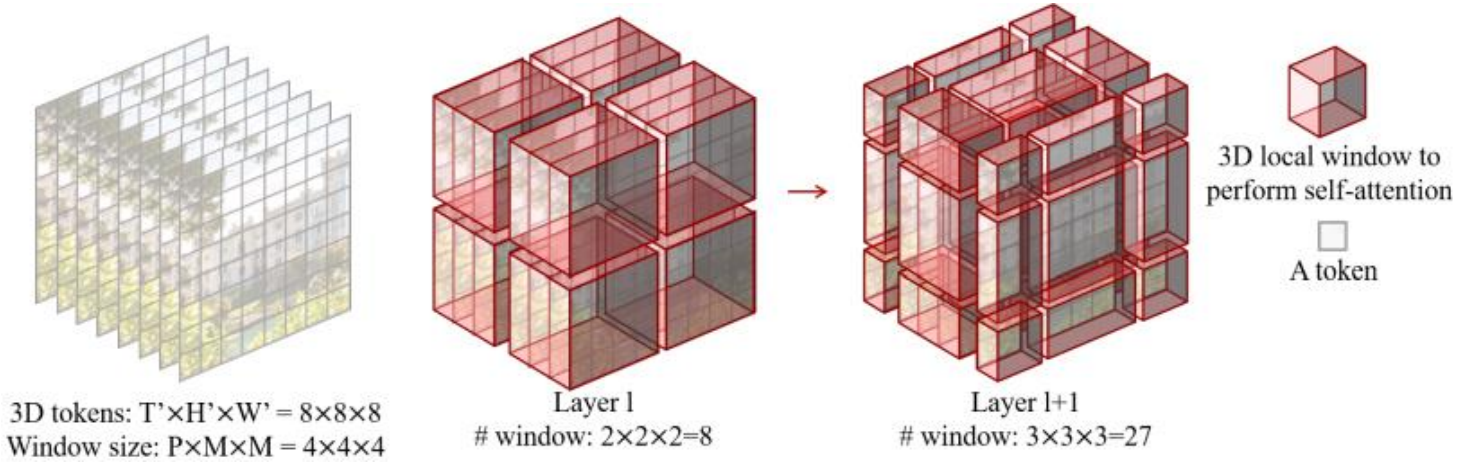




Global/Long-Term Temporal Information

- Liu, Ze, Jia Ning, Yue Cao, Yixuan Wei, Zheng Zhang, Stephen Lin, and Han Hu. "Video swin transformer." In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3202-3211. 2022.

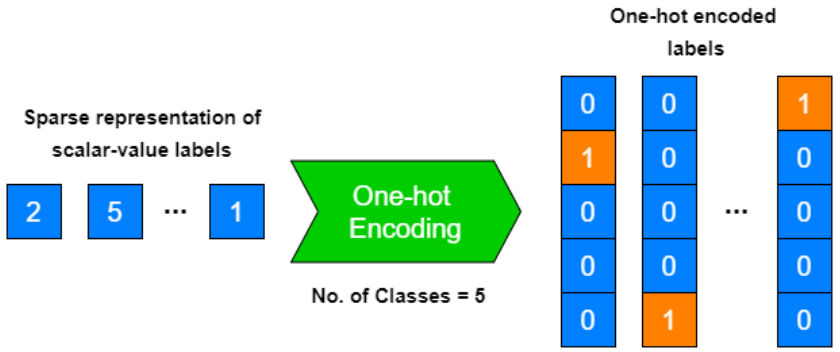


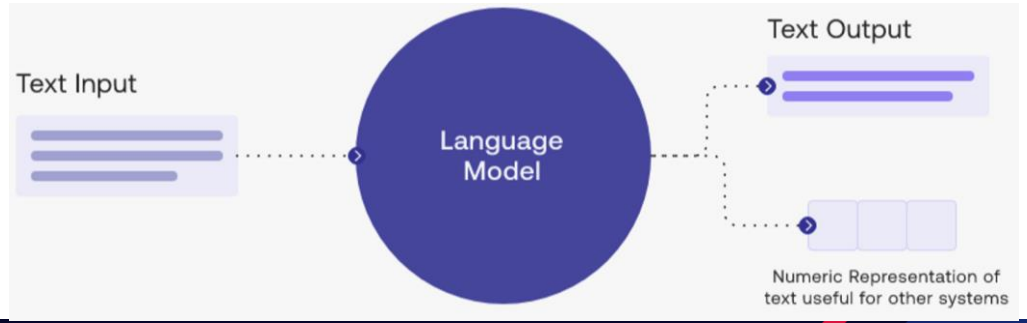


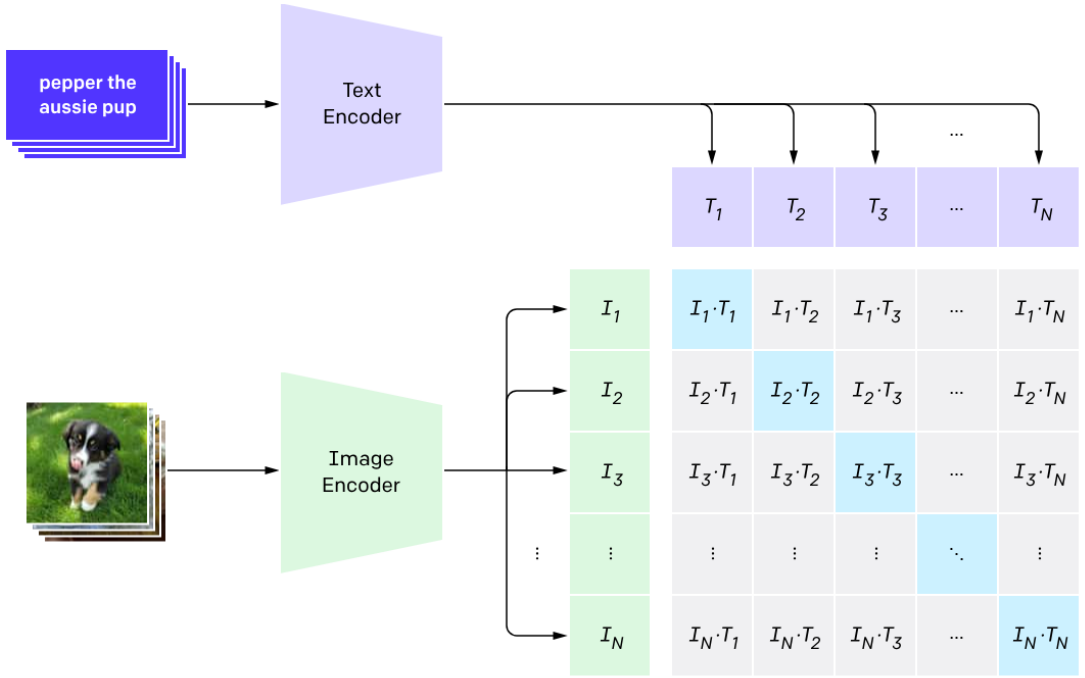
Method	Pretrain	Top-1	Top-5	Views	FLOPs	Param
R(2+1)D [37]	-	72.0	90.0	10 × 1	75	61.8
I3D [6]	ImageNet-1K	72.1	90.3	-	108	25.0
NL I3D-101 [40]	ImageNet-1K	77.7	93.3	10 × 3	359	61.8
ip-CSN-152 [36]	-	77.8	92.8	10 × 3	109	32.8
CorrNet-101 [39]	-	79.2	-	10 × 3	224	-
SlowFast R101+NL [13]	-	79.8	93.9	10 × 3	234	59.9
X3D-XXL [12]	-	80.4	94.6	10 × 3	144	20.3
Swin-T	ImageNet-1K	78.8	93.6	4 × 3	88	28.2
Swin-S	ImageNet-1K	80.6	94.5	4 × 3	166	49.8
Swin-B	ImageNet-1K	80.6	94.6	4 × 3	282	88.1
Swin-B	ImageNet-21K	82.7	95.5	4 × 3	282	88.1
Swin-L	ImageNet-21K	83.1	95.9	4 × 3	604	197.0
Swin-L (384↑)	ImageNet-21K	84.6	96.5	4 × 3	2107	200.0
Swin-L (384↑)	ImageNet-21K	**84.9**	**96.7**	10 × 5	2107	200.0

New Trend: From Numerical to Semantic Classes (LLVM)

- The success of language models: obtaining semantic information from words (text features).
  - What is the difference between one-hot classes and a descriptive sentence?
    - (*Class* “Walking” vs *Description* “A video of a man walking”)
  - How can we leverage semantic texts rather than numerical classes?
    - Visual Prompt + Textual Prompt: Aligning visual information with text information.



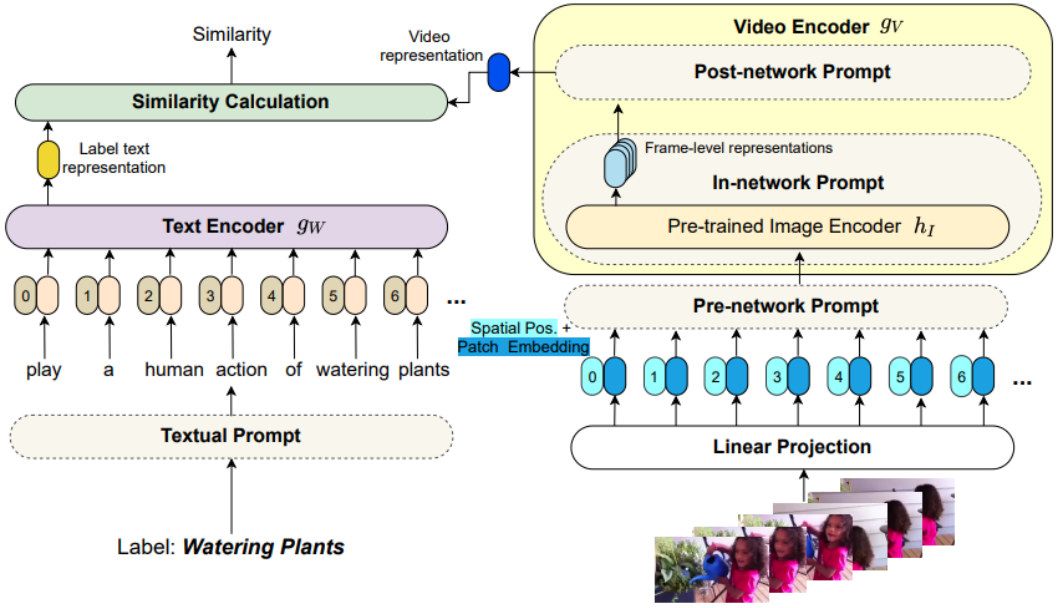


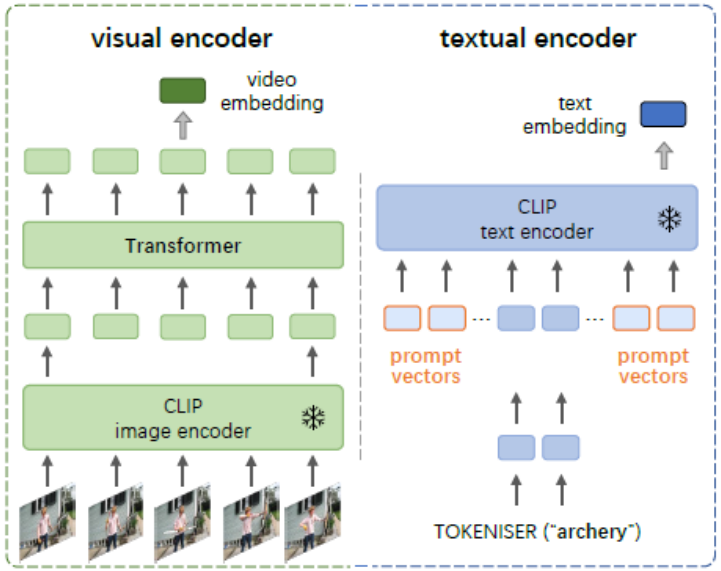


New Trend: From Numerical to Semantic Classes (LLVM)

- The success of language models: obtaining semantic information from words (text features).
  - Intuitively: applying the same strategy towards videos
  - Open question: what is the challenge of adapting this strategy to videos?
    - The preciseness of the descriptive text.
    - The computation cost of training.

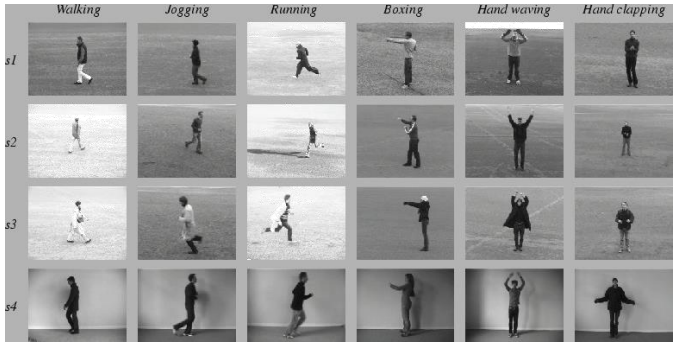
- Ju, Chen, Tengda Han, Kunhao Zheng, Ya Zhang, and Weidi Xie. "Prompting visual-language models for efficient video understanding." In *European Conference on Computer Vision*, pp. 105-124. Cham: Springer Nature Switzerland, 2022.





Action Recognition Datasets

- A man can't walk freely with only one leg – so does the development of deep learning:
  - Two pillars of deep learning – methods + data.

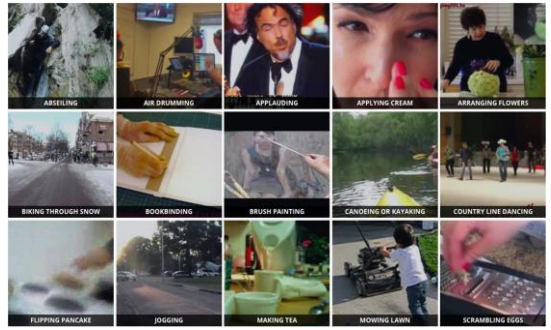


KTH Dataset

- Very small scale, static and easy background.

UCF101 Dataset

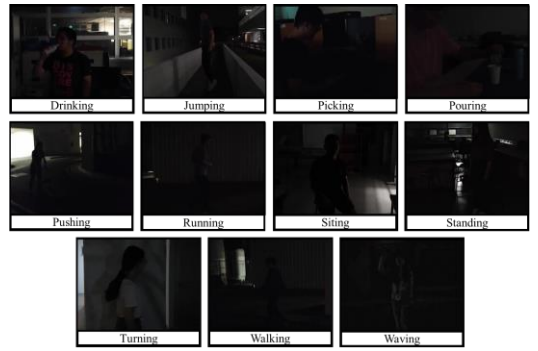
- Small/Medium scale, dynamic but normal background, from Internet.



Kinetics Dataset

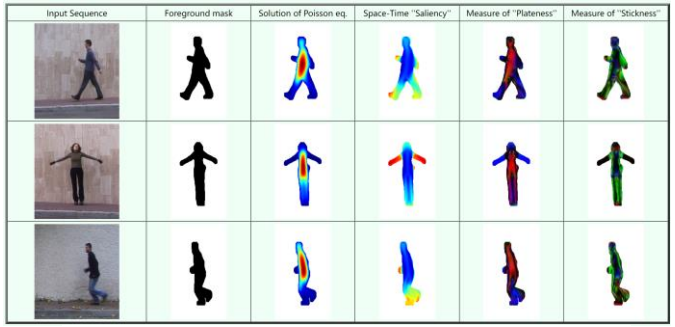
- (Very) Large-scale, dynamic but normal background, from Internet.

Something-Something Dataset



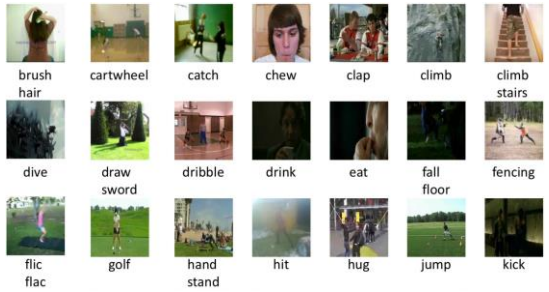
ARID Dataset

- Medium/Large-scale, dynamic and maybe adverse background, fine-grained classes, from Internet or offline.



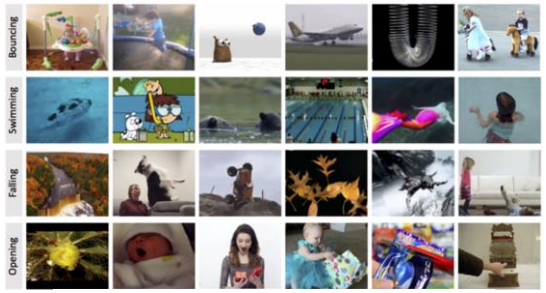
Weizmann Action Dataset

- Very small scale, static and easy background.



HMDB51 Dataset

- Small/Medium scale, dynamic but normal background, from Internet.



Moments-in-Time Dataset

- (Very) Large-scale, dynamic but normal background, from Internet.

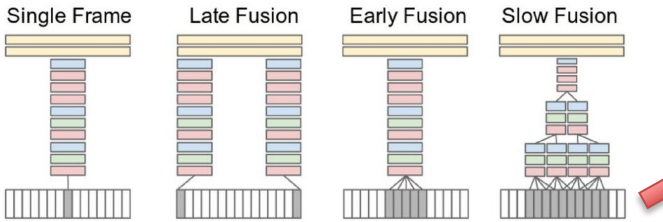
Action Recognition Datasets

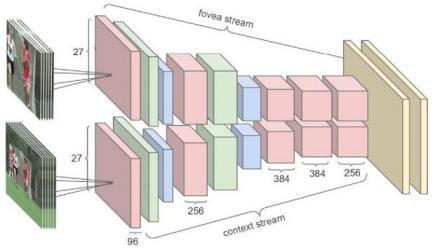
- The trend of dataset development:
  - Earlier development: larger scale – in terms of # of classes and # of videos per classes.
  - To train models with higher generalizability – if the models sees enough examples, it could be more likely generalized to unseen examples.
  - Better and larger models developed – at a cost of huge data annotation cost (Kinetics – developed by Google, Moments-in-Time – developed by big CV group @ MIT, etc.)
  - Recent development: fine-grained actions and actions in adverse environments.
  - More realistic – to train models to be used in real-world applications such as autonomous driving, smart healthcare, etc.

Open Questions on Video Analysis (Action Recognition)

- What would be the next development for action recognition/video analysis models? (Effectiveness, Efficiency, Robustness, Security)

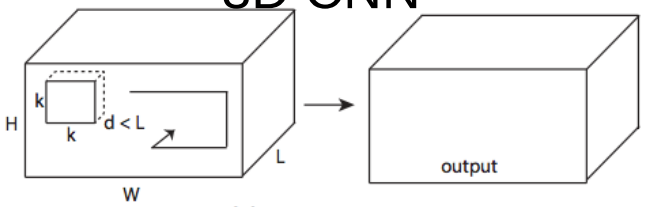
- How would future video datasets develop?

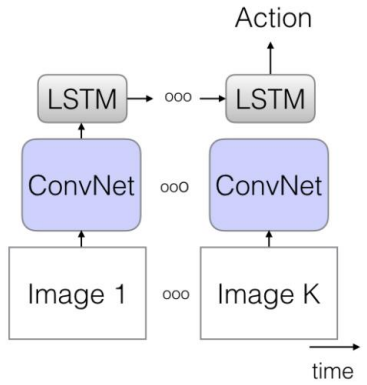




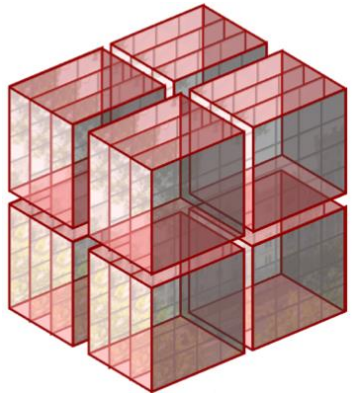
2D CNN

3D CNN

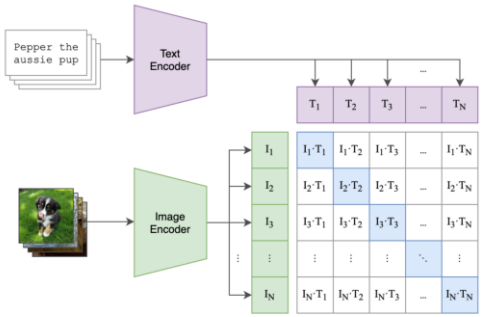




2D CNN with RNN



Video Transformer



Large Language-Vision Model