Project Proposals

- Next Wednesday (**09/15/2021**).
- A 5 minute presentation + 1-2 minutes for Q/A (time limit will be strictly enforced).
- Your presentation should cover: 1) your research problem, 2) the motivation, 3) basic methodology, 4) datasets that you plan to use, 5) experiments that you want to run.
- Send me the **PDF** slides by **September 14th, 11:59 PM**.
- Project report due **September 15th, 11:59 PM.**

Non-local Neural Networks

CVPR 2018

Xiaolong Wang, Ross Girshick, Abhinav Gupta, Kaiming He

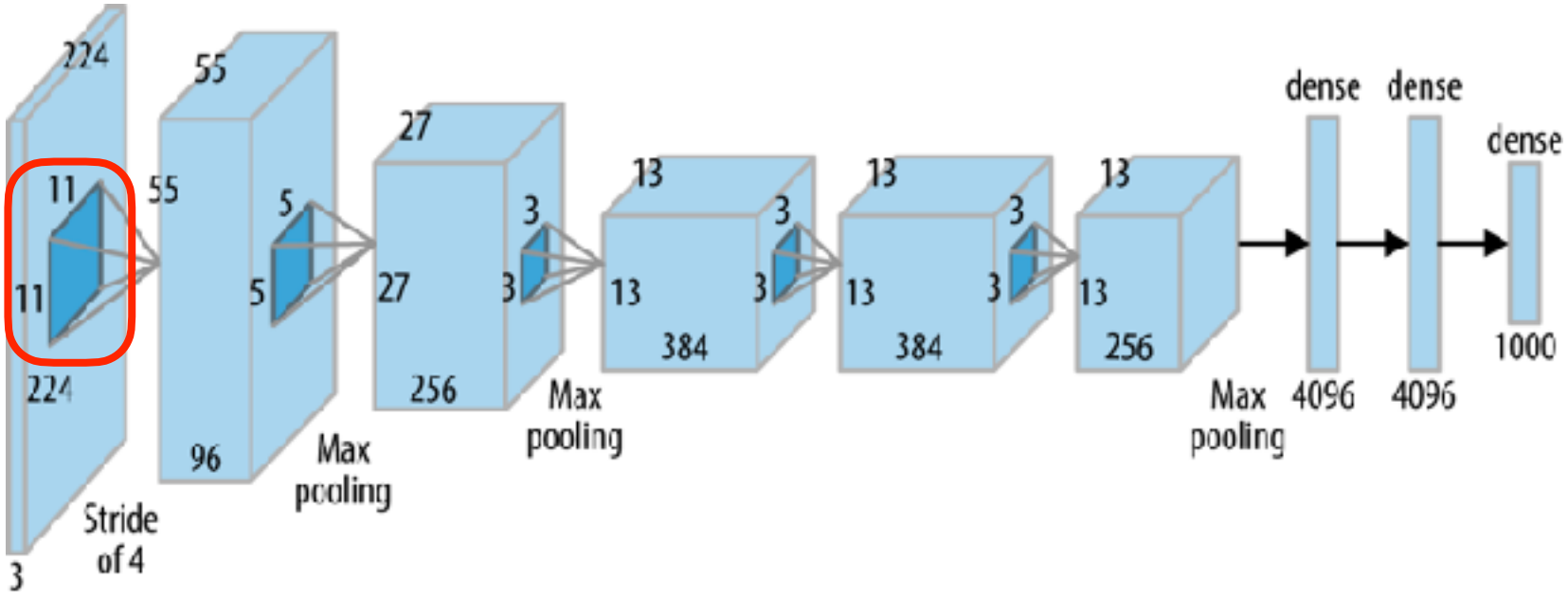
Problem Overview

Standard 2D or 3D convolutional models cannot capture long-range space-time dependencies in the video.

Input Image

Potential Solutions

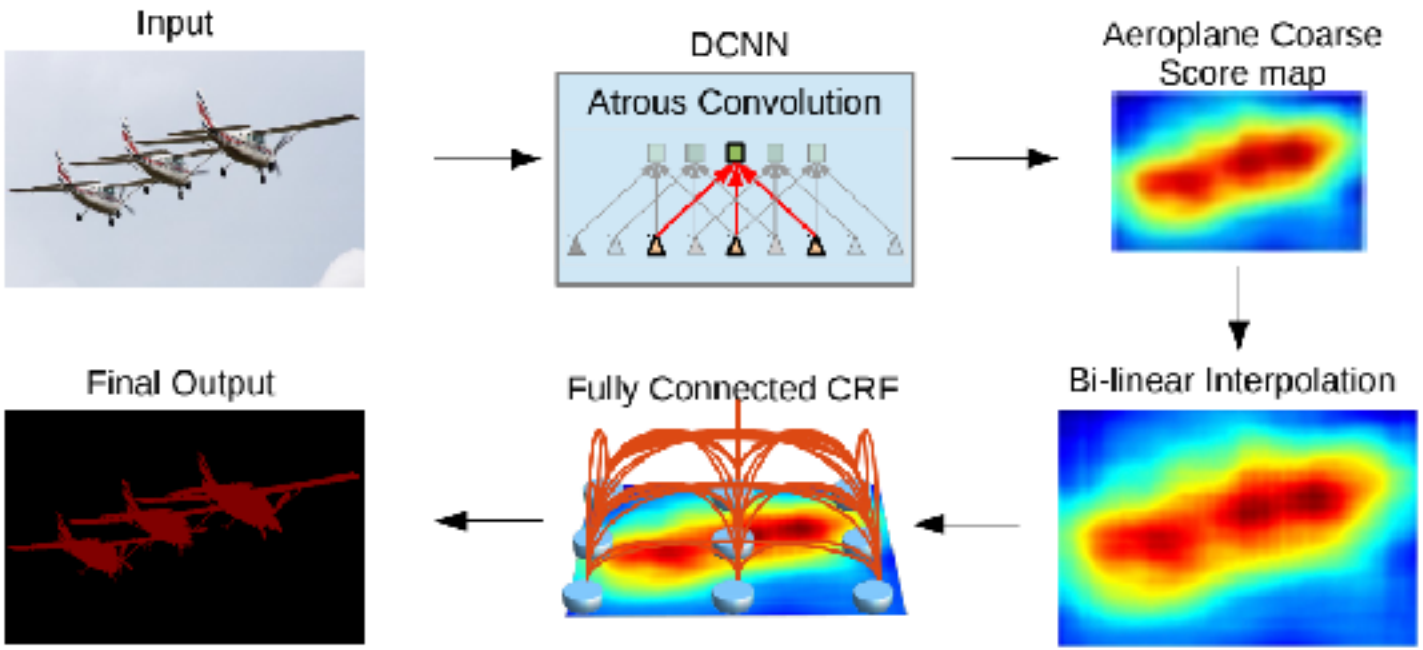
Increase the spatial and temporal kernel size in each or at least some convolutional layers.



"ImageNet Classification with Deep Convolutional Neural Networks", NIPS 2012

Potential Solutions

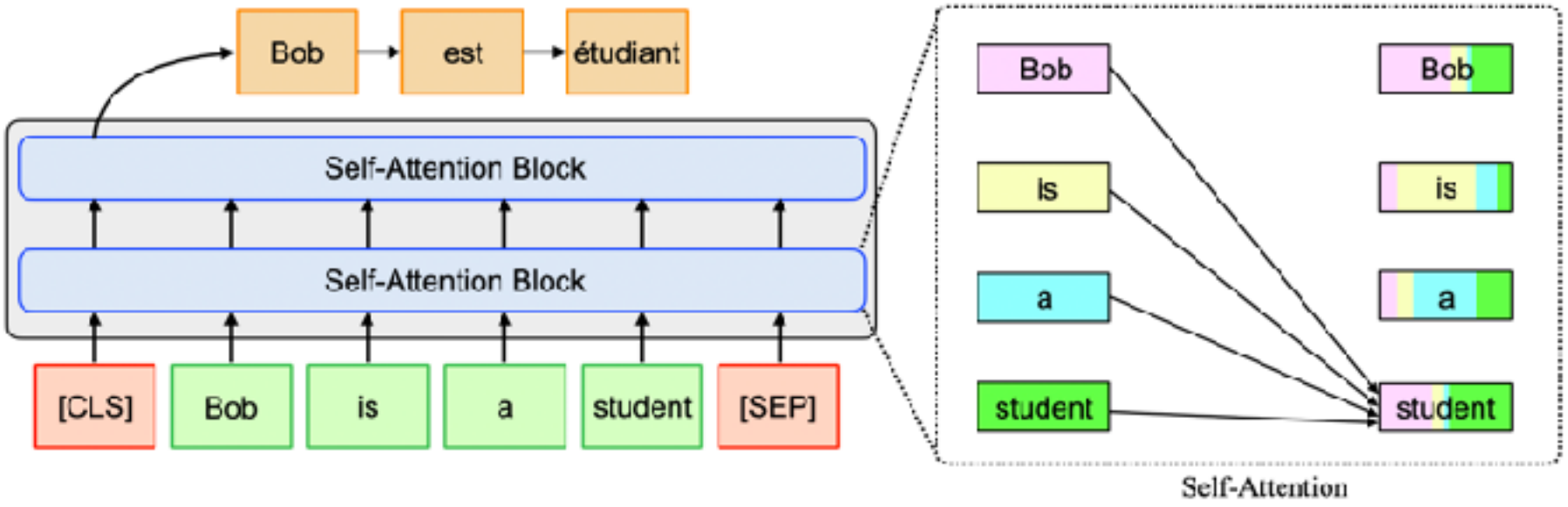
Attach a graphical model on top of the CNN.



"DeepLab: Semantic Image Segmentation with Deep Convolutional Nets, Atrous Convolution, and Fully Connected CRFs", TPAMI 2017

Potential Solutions

Incorporate operations that can model long-range dependencies in the network.



"Attention is All You Need", Vaswani et al., NIPS, 2017

Non-Local Operation

- A non-local operation computes the response at a position as a weighted sum of the features at all positions in the input feature maps.
- The set of positions can be in space, time, or spacetime.

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

Non-Local Operation

- A non-local operation computes the response at a position as a weighted sum of the features at all positions in the input feature maps.
- The set of positions can be in space, time, or spacetime.

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

Input signal at position i

Input signal at position j

Non-Local Operation

- A non-local operation computes the response at a position as a weighted sum of the features at all positions in the input feature maps.
- The set of positions can be in space, time, or spacetime.

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

A pairwise similarity function

Non-Local Operation

- A non-local operation computes the response at a position as a weighted sum of the features at all positions in the input feature maps.
- The set of positions can be in space, time, or spacetime.

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

The unary function that computes a representation of the input signal at the position j

Non-Local Operation

- A non-local operation computes the response at a position as a weighted sum of the features at all positions in the input feature maps.
- The set of positions can be in space, time, or spacetime.

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

Normalization factor

Non-Local Operation

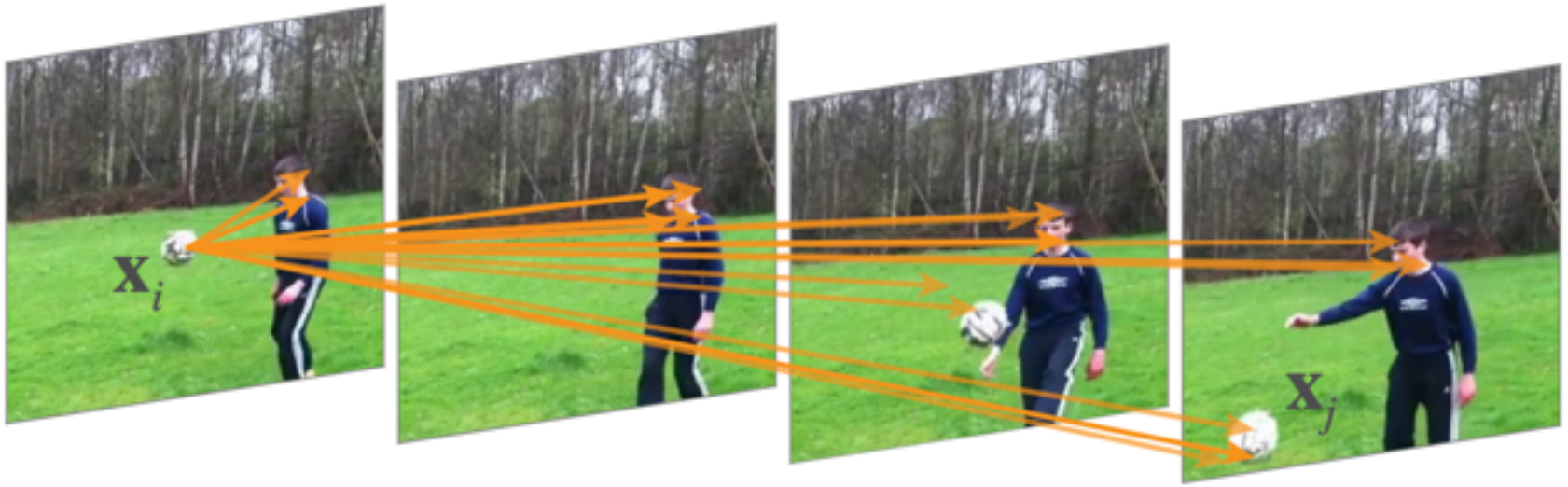
- A non-local operation computes the response at a position as a weighted sum of the features at all positions in the input feature maps.
- The set of positions can be in space, time, or spacetime.

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

Output signal at position i

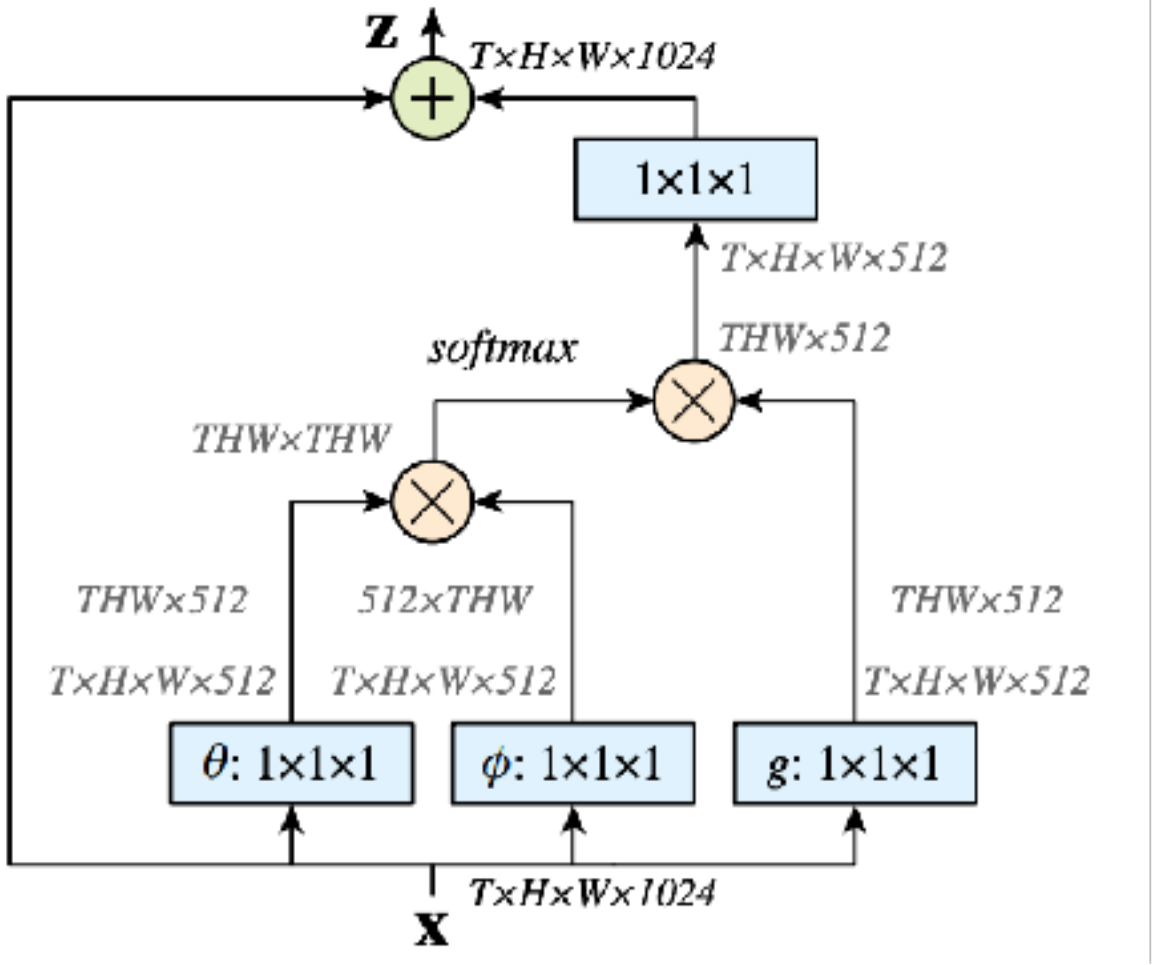
Non-Local Operation

A non-local operation computes the response at a position as a weighted sum of the features at all positions in the input feature maps.

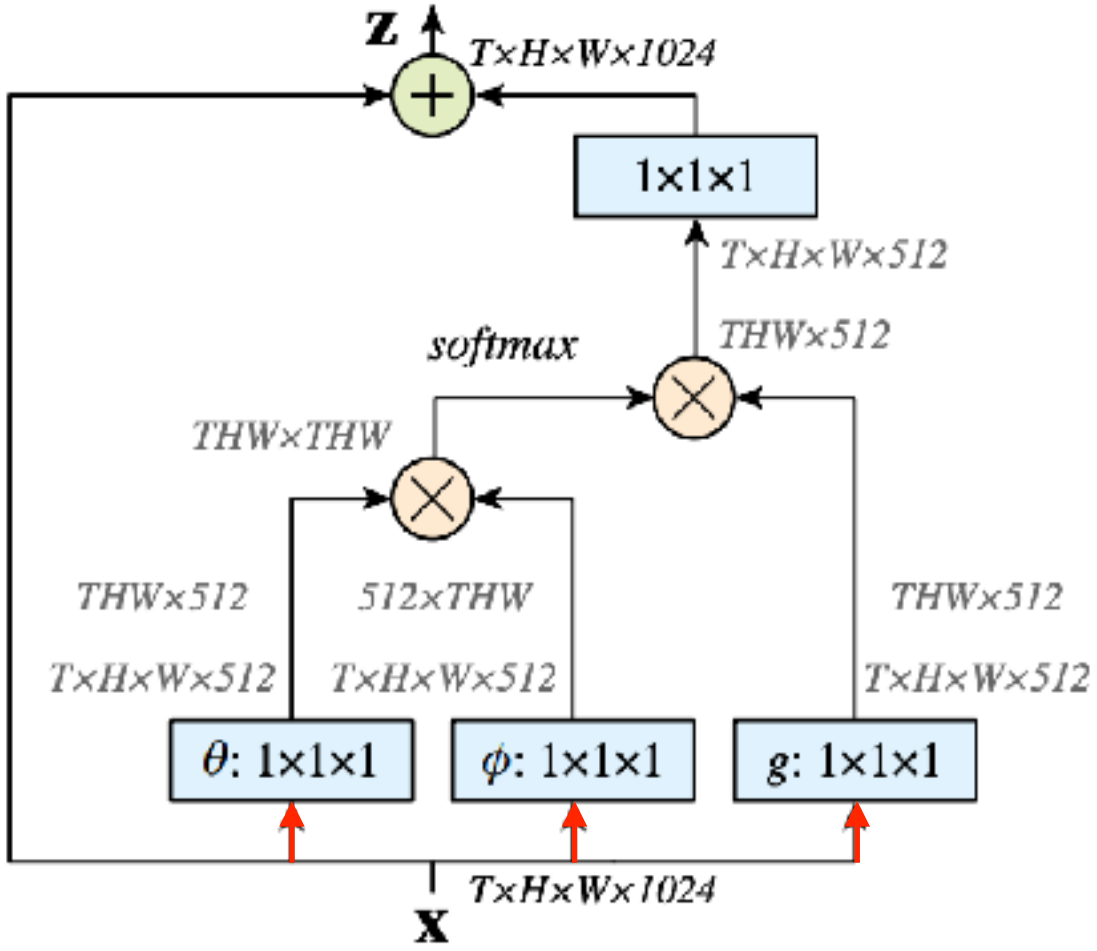


The non-local behavior is due to the fact that all positions are considered in the operation.

A Space-Time Non-Local Block

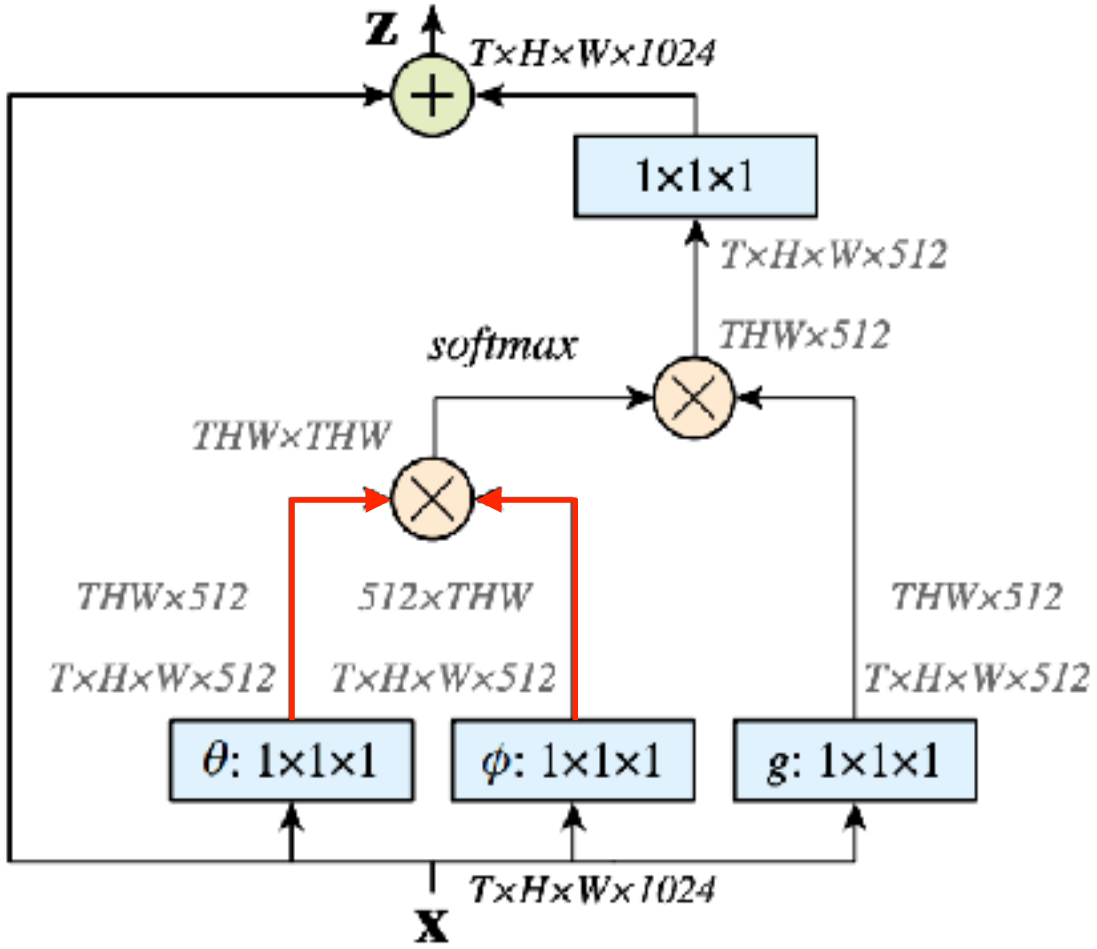


A Space-Time Non-Local Block



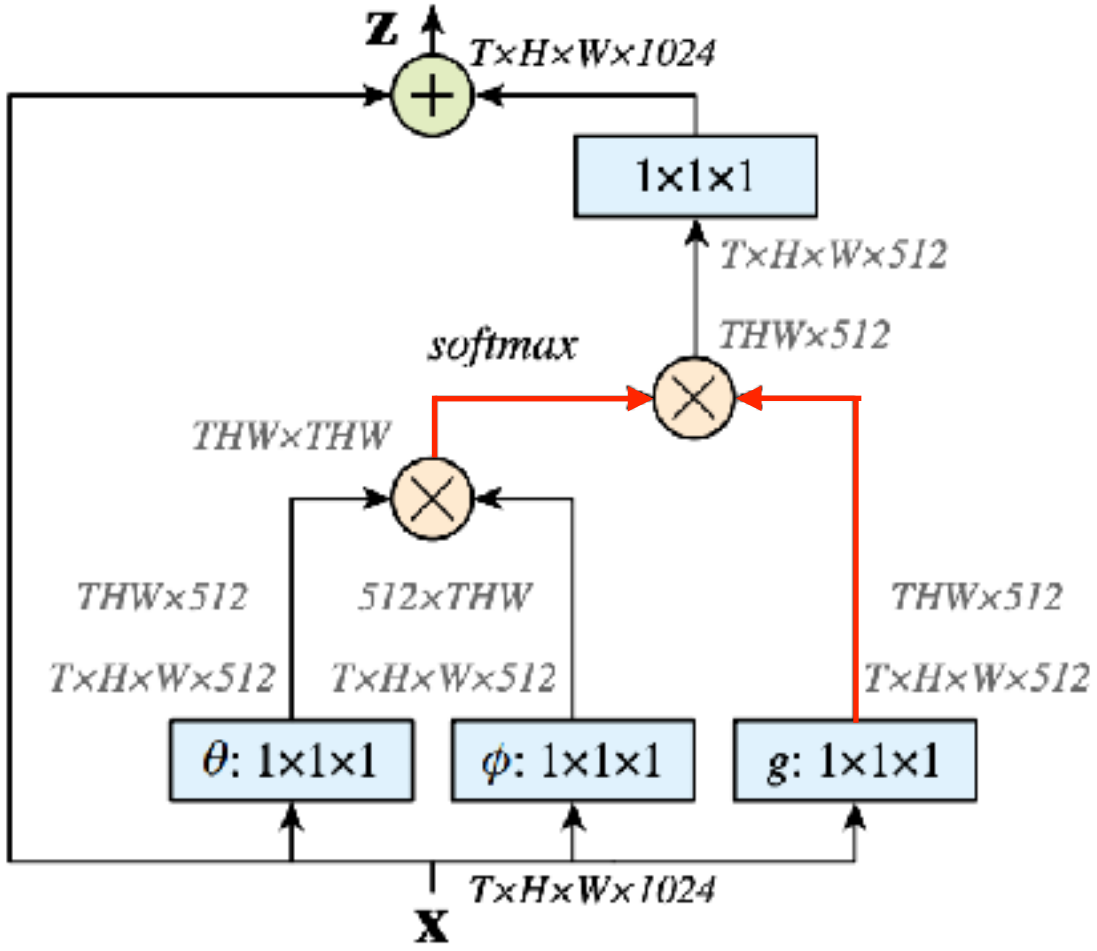
1. The initial feature tensor is mapped into three distinct intermediate feature representations.

A Space-Time Non-Local Block



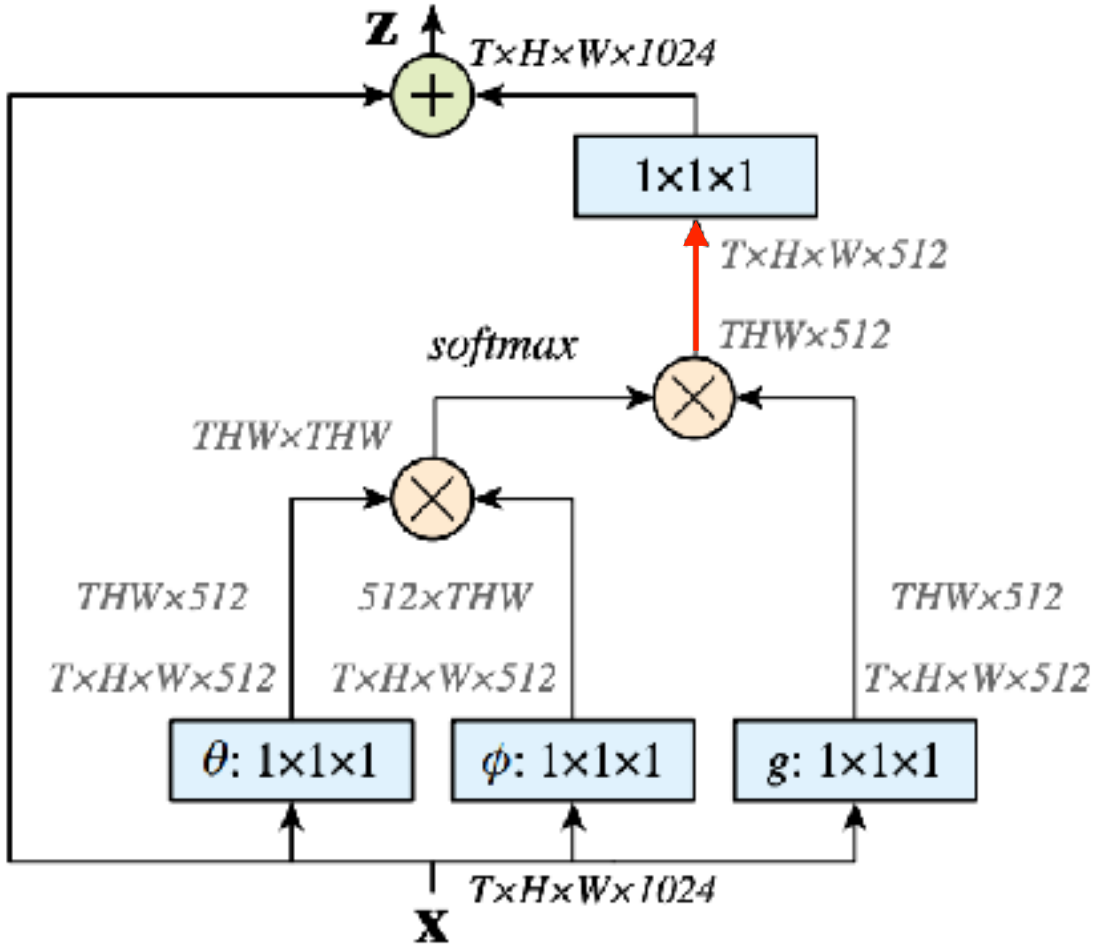
2. We use dot product to compute pairwise similarity between every single pair of features.

A Space-Time Non-Local Block



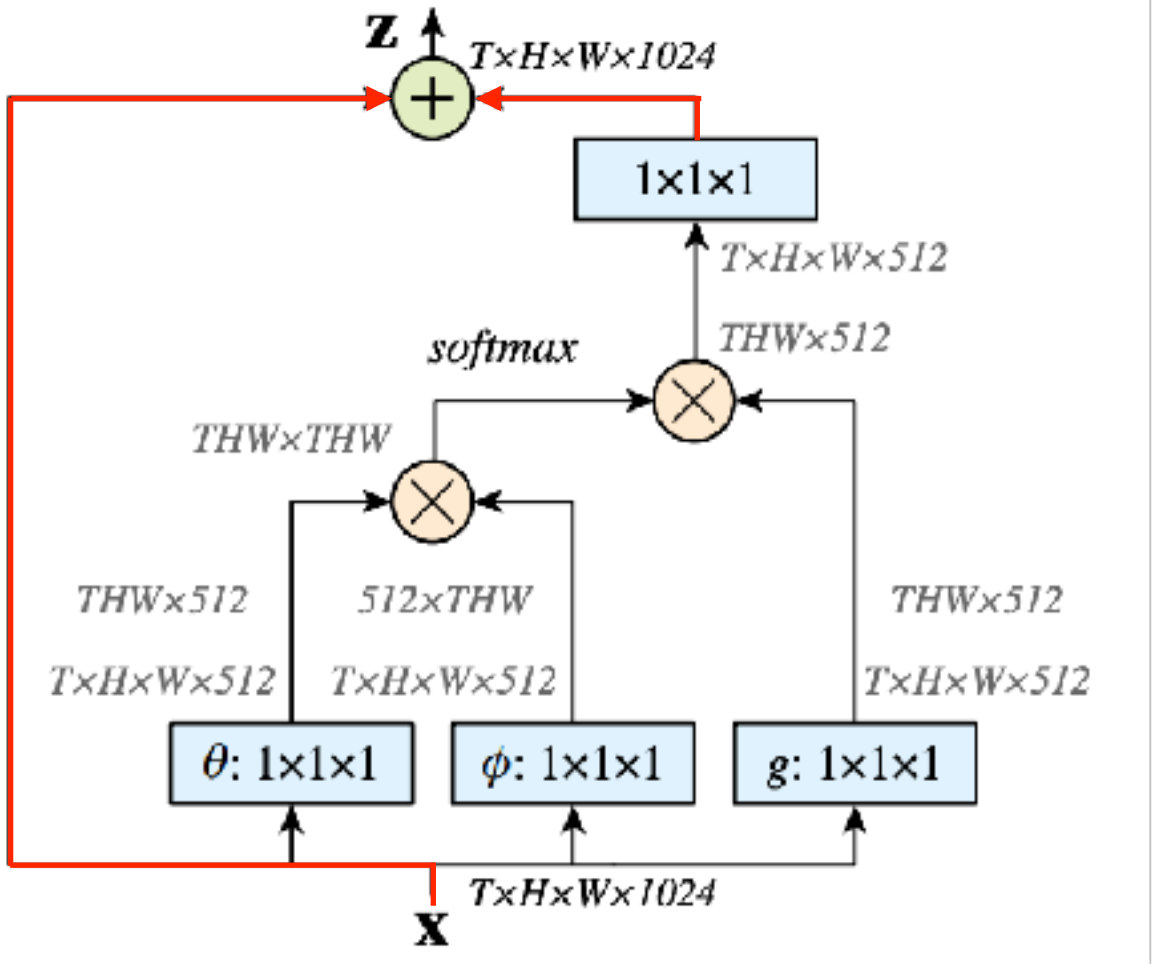
3. We then normalize the resulting similarity matrix, and perform a weighted feature averaging.

A Space-Time Non-Local Block



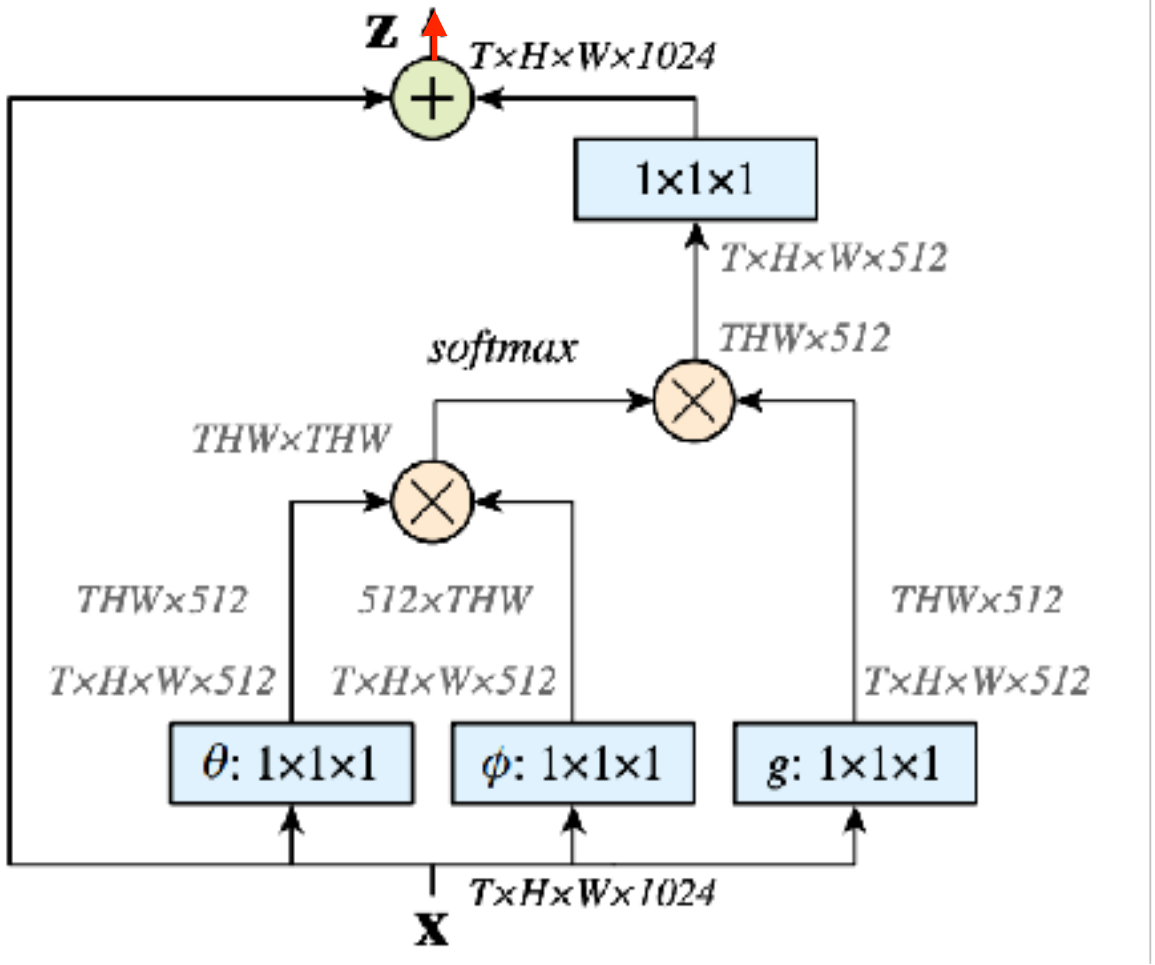
4. The resulting feature tensor is then fed through another 1x1x1 convolutional layer.

A Space-Time Non-Local Block



5. The residual connection is used to aggregate information within the block.

A Space-Time Non-Local Block



6. Finally, the output feature tensor of the same dimensionality is produced.

A Space-Time Non-Local Block

- The residual connection allows inserting a new non-local block into any pre-trained model, without breaking its initial behavior.
- It is typically initialized to be 0.

$$\mathbf{z}_i = W_z \mathbf{y}_i + \mathbf{x}_i,$$

The initial input

A Space-Time Non-Local Block

- The residual connection allows inserting a new non-local block into any pre-trained model, without breaking its initial behavior.
- It is typically initialized to be 0.

Output from the non-local operator

$$\mathbf{z}_i = W_z \mathbf{y}_i + \mathbf{x}_i,$$

A Space-Time Non-Local Block

- The residual connection allows inserting a new non-local block into any pre-trained model, without breaking its initial behavior.
- It is typically initialized to be 0.

$$\mathbf{z}_i = W_z \mathbf{y}_i + \mathbf{x}_i,$$

Residual connection layer

A Space-Time Non-Local Block

- The residual connection allows inserting a new non-local block into any pre-trained model, without breaking its initial behavior.
- It is typically initialized to be 0.

Final output from the non-local block

$$\mathbf{z}_i = W_z \mathbf{y}_i + \mathbf{x}_i,$$

Pairwise Similarity Function

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

Network Architecture

Baseline ResNet-50 C2D architecture for video recognition.

layer		output size
conv_1	7×7, 64, stride 2, 2, 2	16×112×112
pool_1	3×3×3 max, stride 2, 2, 2	8×56×56
res_2	$\begin{bmatrix} 1\times1, 64 \\ 3\times3, 64 \\ 1\times1, 256 \end{bmatrix} \times3$	8×56×56
pool_2	3×1×1 max, stride 2, 1, 1	4×56×56
res_3	$\begin{bmatrix} 1\times1, 128 \\ 3\times3, 128 \\ 1\times1, 512 \end{bmatrix} \times4$	4×28×28
res_4	$\begin{bmatrix} 1\times1, 256 \\ 3\times3, 256 \\ 1\times1, 1024 \end{bmatrix} \times6$	4×14×14
res_5	$\begin{bmatrix} 1\times1, 512 \\ 3\times3, 512 \\ 1\times1, 2048 \end{bmatrix} \times3$	4×7×7
global average pool, fc		1×1×1

Results on Kinetics

Ablating different pairwise similarity functions.

model, R50	top-1	top-5
C2D baseline	71.8	89.7
Gaussian	72.5	90.2
Gaussian, embed	72.7	**90.5**
dot-product	**72.9**	90.3
concatenation	72.8	**90.5**

Results on Kinetics

A single non-local block is added into different ResNet-50 stages.

model, R50	top-1	top-5
baseline	71.8	89.7
res_2	72.7	90.3
res_3	**72.9**	90.4
res_4	72.7	**90.5**
res_5	72.3	90.1

Results on Kinetics

Comparison with 1, 5, and 10 non-local blocks added to the C2D ResNet-50 and ResNet-101 baselines.

model		top-1	top-5
R50	baseline	71.8	89.7
	1-block	72.7	90.5
	5-block	73.8	91.0
	10-block	**74.3**	**91.2**
R101	baseline	73.1	91.0
	1-block	74.3	91.3
	5-block	**75.1**	**91.7**
	10-block	**75.1**	91.6

Results on Kinetics

Comparing non-local operations applied along space, time, and spacetime dimensions respectively.

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

Results on Kinetics

Comparing a 5-block non-local C2D vs. inflated 3D ConvNet (I3D).

model, R101	params	FLOPs	top-1	top-5
C2D baseline	1×	1×	73.1	91.0
$I3D_{3\times3\times3}$	1.5×	1.8×	74.1	91.2
$I3D_{3\times1\times1}$	**1.2**×	1.5×	74.4	91.1
NL C2D, 5-block	**1.2**×	**1.2**×	**75.1**	**91.7**

Results on Kinetics

Five non-local blocks are added on top of the best I3D model.

	model	top-1	top-5
R50	C2D baseline	71.8	89.7
	I3D	73.3	90.7
	NL I3D	**74.9**	**91.6**
R101	C2D baseline	73.1	91.0
	I3D	74.4	91.1
	NL I3D	**76.0**	**92.1**

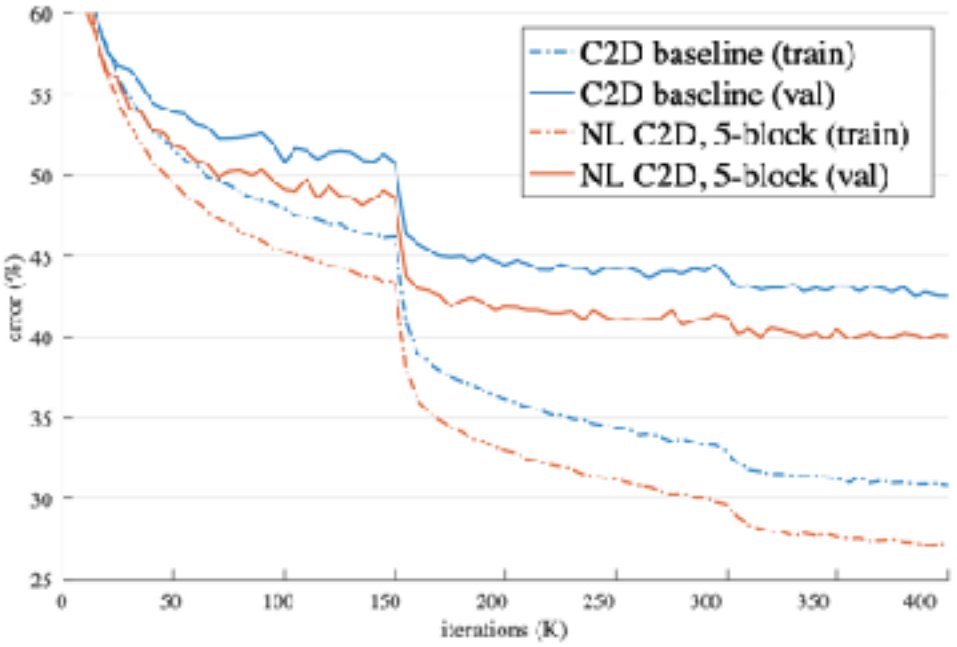
Results on Kinetics

The model is fine-tuned and tested using longer 128-frame clips.

	model	top-1	top-5
R50	C2D baseline	73.8	91.2
	I3D	74.9	91.7
	NL I3D	**76.5**	**92.6**
R101	C2D baseline	75.3	91.8
	I3D	76.4	92.7
	NL I3D	**77.7**	**93.3**

Results on Kinetics

Non-local C2D model produces lower training and validation errors than the C2D baseline.



Results on Kinetics

- Comparison to the state-of-the-art.
- The greyed out approaches use an additional audio modality.

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**

Results on Charades

- Charades is a video dataset with ~8k training, ~1.8k validation, and ~2k testing videos.
- It is a multi-label classification task with 157 action categories.
- The videos are much longer than in Kinetics.

model	modality	*train/val*	*trainval/test*
2-Stream [43]	RGB + flow	18.6	-
2-Stream +LSTM [43]	RGB + flow	17.8	-
Asyn-TF [43]	RGB + flow	22.4	-
I3D [7]	RGB	32.9	34.4
I3D [ours]	RGB	35.5	37.2
NL I3D [ours]	RGB	**37.5**	**39.5**

Object Detection Results on COCO

Adding 1 non-local block to Mask R-CNN for COCO object detection and instance segmentation.

method		AP^{box}	AP^{box}_{50}	AP^{box}_{75}	AP^{mask}	AP^{mask}_{50}	AP^{mask}_{75}
R50	baseline	38.0	59.6	41.0	34.6	56.4	36.5
	+1 NL	**39.0**	**61.1**	**41.9**	**35.5**	**58.0**	**37.4**
R101	baseline	39.5	61.4	42.9	36.0	58.1	38.3
	+1 NL	**40.8**	**63.1**	**44.5**	**37.1**	**59.9**	**39.2**
X152	baseline	44.1	66.4	48.4	39.7	63.2	42.2
	+1 NL	**45.0**	**67.8**	**48.9**	**40.3**	**64.4**	**42.8**

Qualitative Results

Contributions

- Very effective mechanism for aggregating information over long spatial and temporal extents.
- Can be integrated easily with existing pretrained models.
- Works well with a variety of different architectures.
- State-of-the-art results on many tasks / datasets.

Discussion Questions

- What are some of the downsides of the proposed non-local block?

Discussion Questions

- What are some of the downsides of the proposed non-local block?

- How does it differ from the fully connected layer?

Discussion Questions

- What are some of the downsides of the proposed non-local block?
- How does it differ from the fully connected layer?
- Does the non-local model learn to incorporate space-time relationships between objects in the video?