CVPR 2018

Non-local Neural Networks

2022.07.27

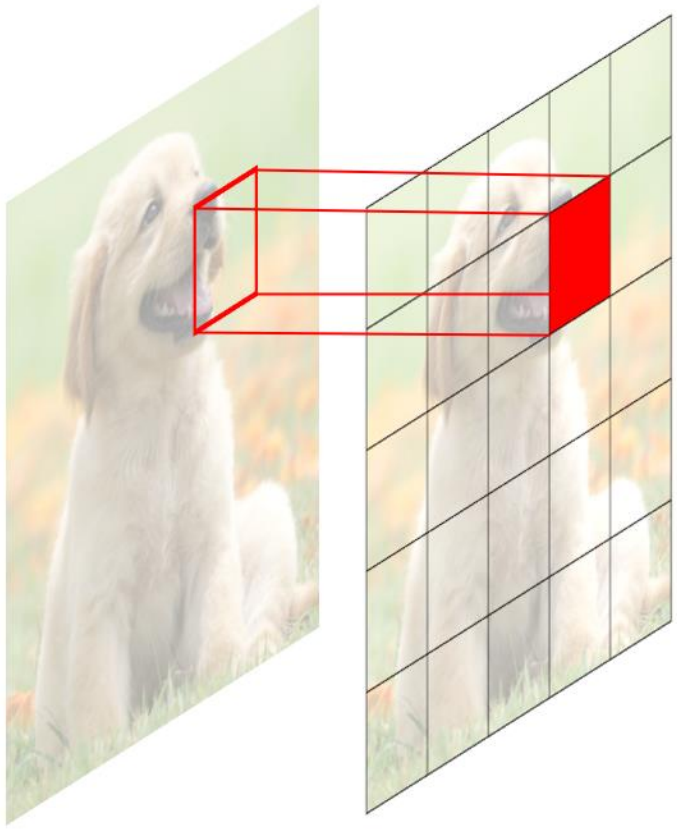
논문 리뷰

배성훈

Non-local Neural Networks (CVPR 2018)

- **Research Background**:
  - Long-range dependencies를 포착하는 것에 있어 RNN, CNN은 근본적인 한계를 가짐
  - local neighborhood만 processing
  - Long-range dependencies를 더 잘 포착하기 위해 **non-local operation**을 응용

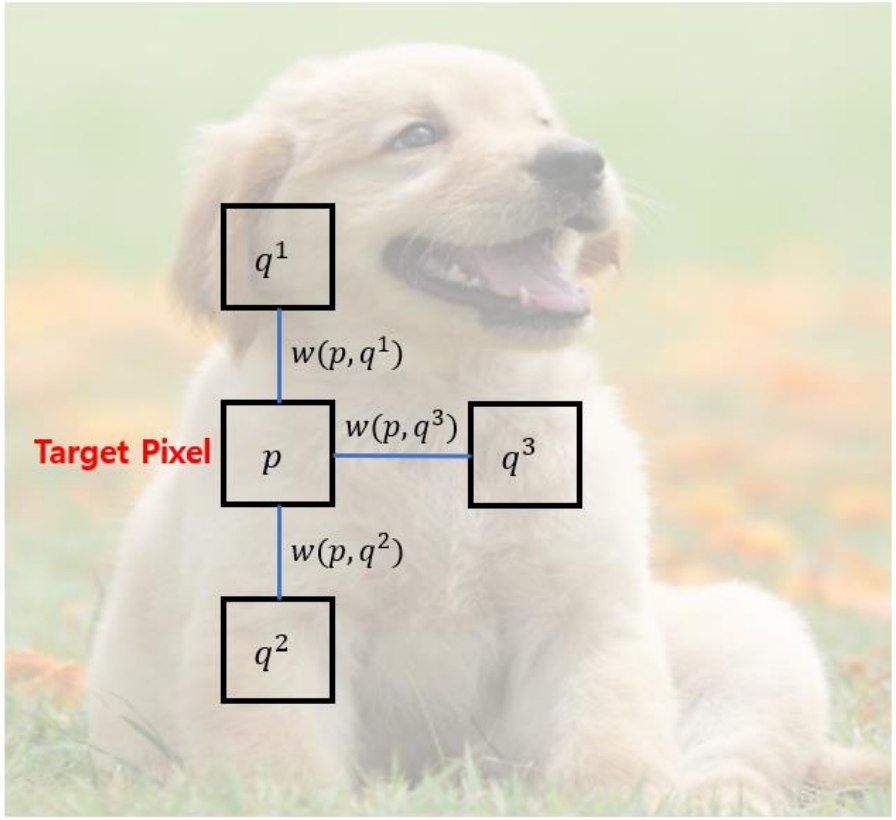
CNN
filter가 한번에 볼 수 있는 영역이 제한적



Non-local Neural Networks (CVPR 2018)

- **Method**:
  - 기존의 denoising에 사용된 Non-local filter로부터 Motivate
  - Non-local Means Filter
  - 한 장의 이미지 내에서 Target Pixel과 유사한 영역(q^1, q^2, q^3)을 찾아 평균을 취하는 방식
  - 영상 전체 영역 (Non-local) 활용

Non-local Mean Filter => Large Receptive Field



Non-local Neural Networks (CVPR 2018)

- **Method**:
  - Non-local mean filter의 영상 전체 영역을 활용하는 방법은 long-range dependency 해결
  - 이를 응용한 **Non-local block** 제안

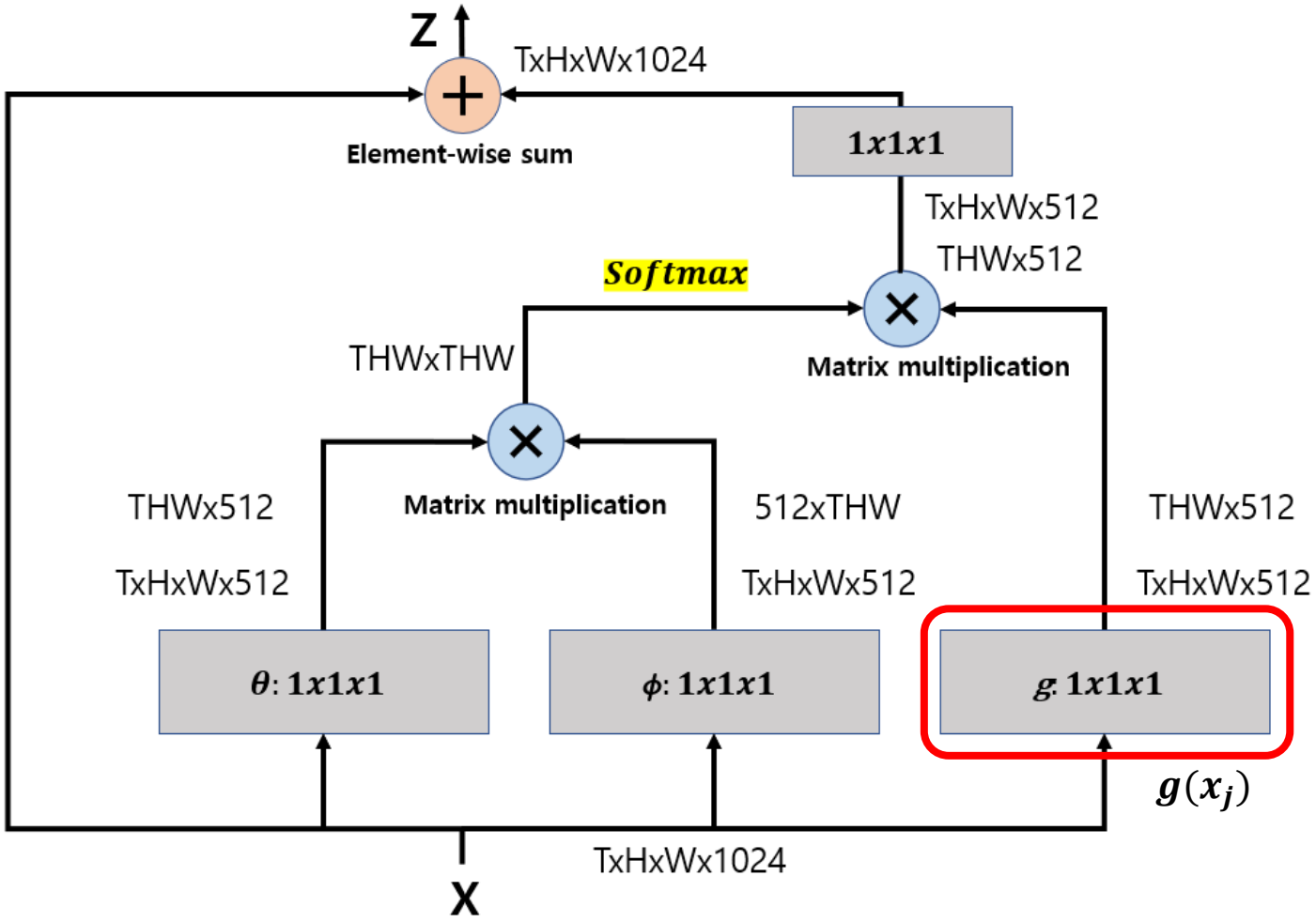
  - **Non-local operation**:

- Input x에서 feature(i)와 전체 영역(j)에 대한 feature의 similarity를 계산하고, similarity가 큰 영역의 embedding된 feature의 값을 더 크게 activate

$$\boldsymbol{y_i} = \frac{\mathbf{1}}{\boldsymbol{C(x)}} \sum_{\forall_j} \underbrace{\boldsymbol{f(x_i, x_j)}}_{\text{Similarity 계산}} \underbrace{\boldsymbol{g(x_j)}}_{\text{Linear embedding}}$$

Non-local Neural Networks (CVPR 2018)

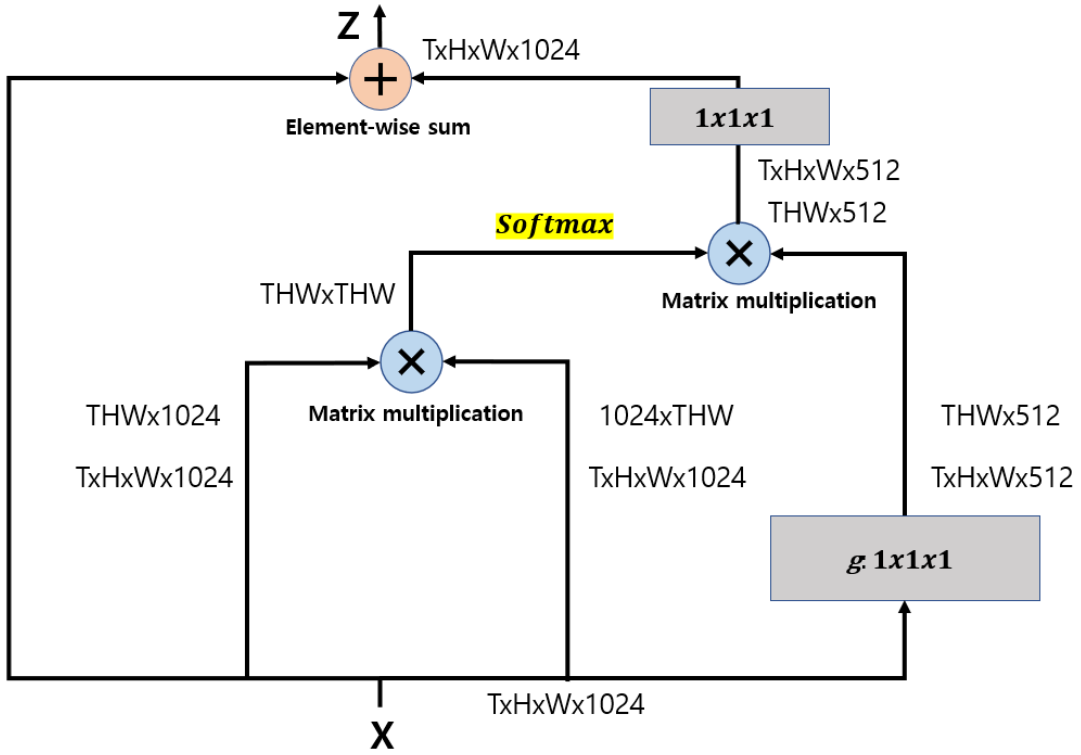
- **Method**:
  - $g(x_j)$: linear embedding
  - $g(x_j)=W_g x_j$ W_g: Weight matrix



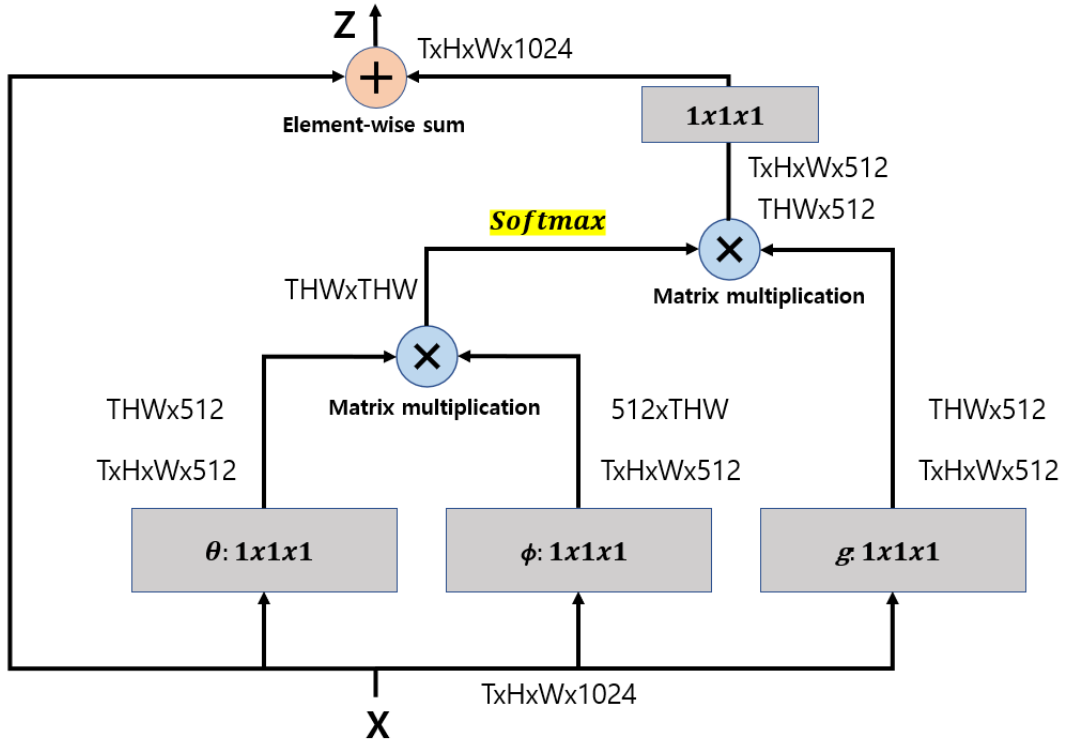
Non-local Neural Networks (CVPR 2018)

- **Method**:
  - $f(x_i, x_j)$: 특정 i번째 feature에서 다른 위치의 feature(j)와 similarity 계산
  - 4가지 version 제안, 성능 향상은 모두 비슷함
  - Gaussian: $f(x_i, x_j) = e^{x_i^T x_j}$
  - Embedded Gaussian: $f(x_i, x_j) = e^{\theta(x_i)^T \varphi(x_j)}$ $\quad \theta(x_j) = W_\theta x_i$, $\varphi(x_j) = W_\varphi x_j$

Non-local block (f=Gaussian)



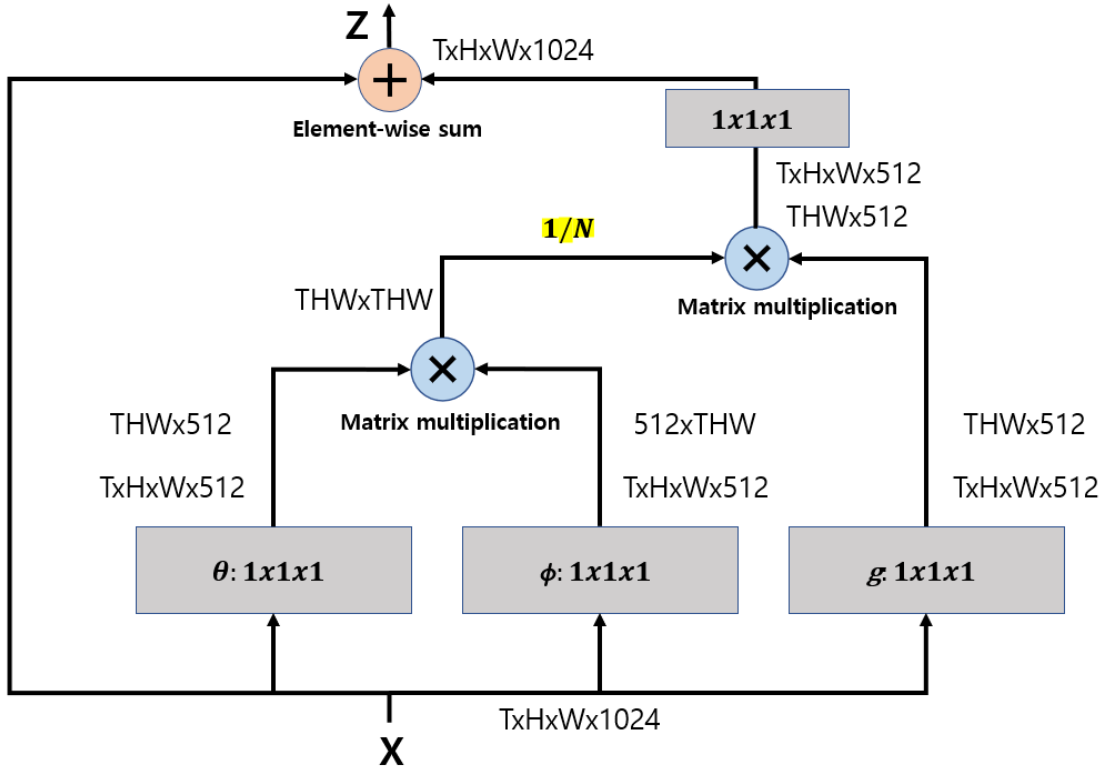
Non-local block (f =Embedded Gaussian)



Non-local Neural Networks (CVPR 2018)

- **Method**:
  - $f(x_i, x_j)$:
  - Dot product: $f(x_i, x_j) = \theta(x_i)^T \varphi(x_j)$
  - Concatenation: $f(x_i, x_j) = \mathbf{ReLU}(w_f^T[\theta(x_i), \varphi(x_j)])$

Non-local block (f=Dot product)



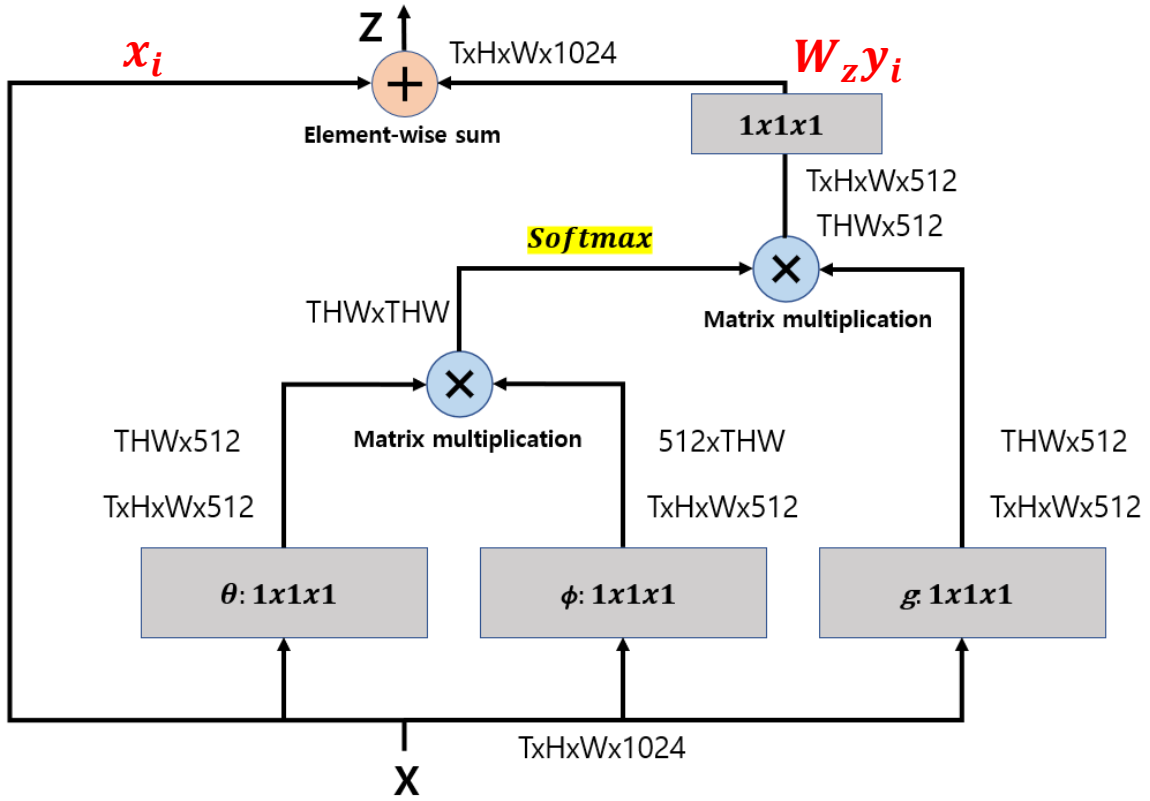
Non-local Neural Networks (CVPR 2018)

- **Method**:
  - Non-local Block:

$$z_i = W_z y_i + x_i$$

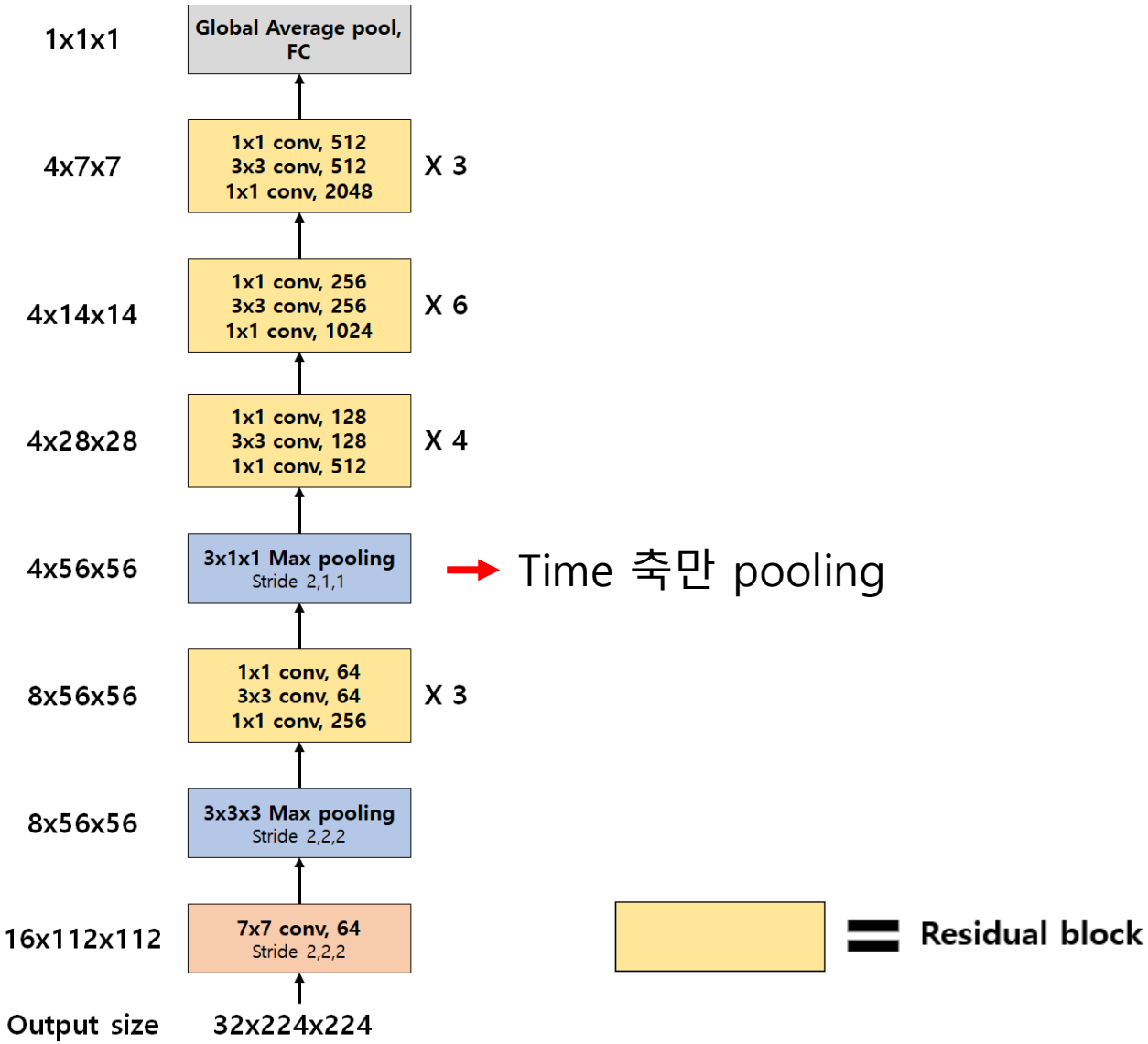
Residual Connection

  - 사전 학습된 모델 내 쉽게 삽입하여 Fine-tune 하는 방식으로 사용 가능



Non-local Neural Networks (CVPR 2018)

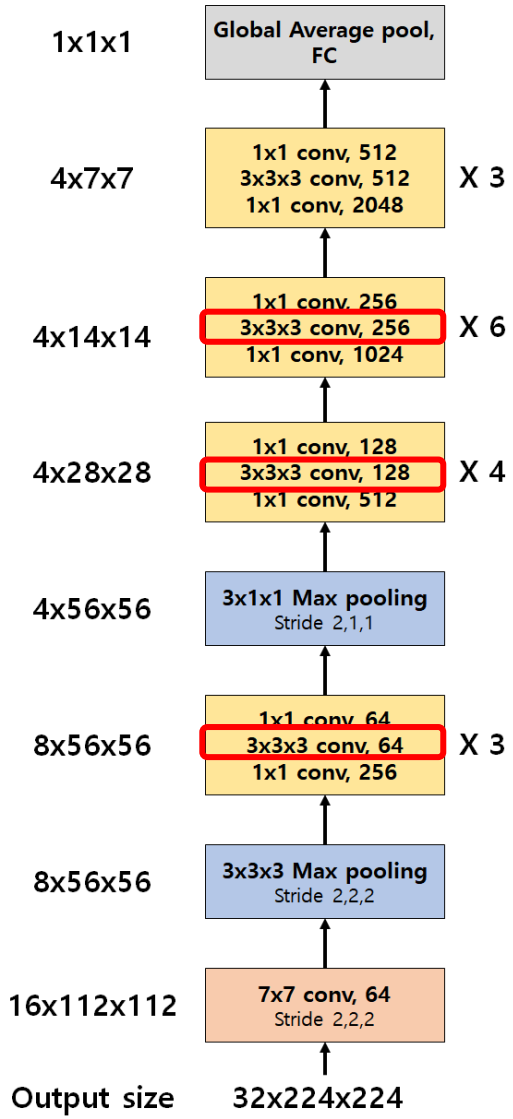
- **Method**:
  - Baseline: ResNet-50 C2D
  - For video

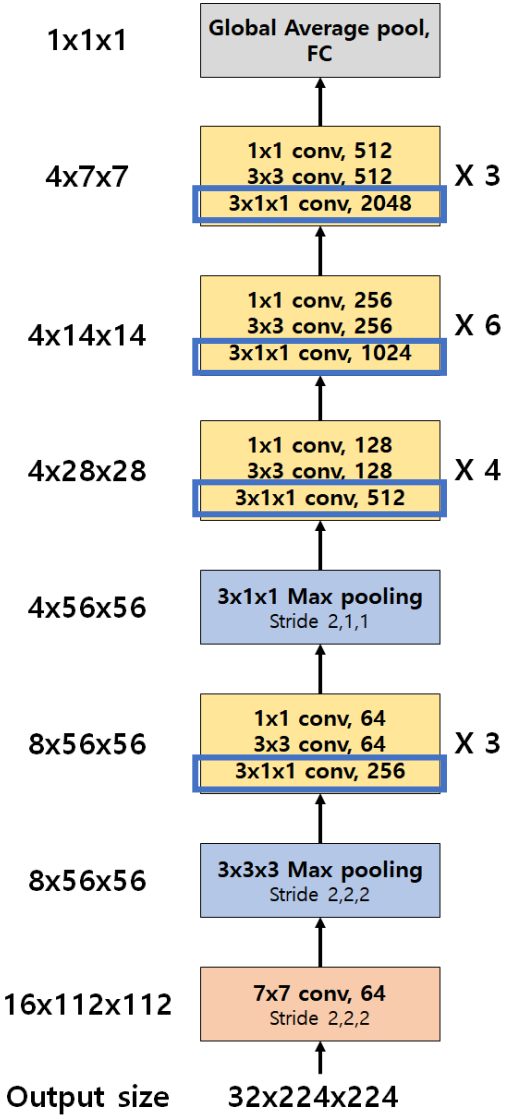


Non-local Neural Networks (CVPR 2018)

- **Method**:
  - Baseline: ResNet-50 $I3D_{3x3x3}$, $I3D_{3x1x1}$

$I3D_{3x3x3}$





$I3D_{3x1x1}$

Non-local Neural Networks (CVPR 2018)

- **Experiments**:
  - Video Classification using Kinetics, Charades dataset
  - Res5 직전에 non-local block 추가하는 것이 제일 좋은 성능

Similarity (f) 계산 방법에 따른 결과 비교

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**

Non-local block이 추가되는 위치에 따른 결과 비교

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

Non-local Neural Networks (CVPR 2018)

- **Experiments**:
  - Non-local block을 사용했을 때 연산량을 줄이면서 성능 향상 (Top-1 error 기준 2% 증가)
  - I3D에 Non-local block 추가하는 실험 또한 최대 3% 성능 향상
  - 시공간축에 non-local block을 추가하는게 가장 좋은 성능을 보임

C2D vs I3D vs A 5 block non-local C2D

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**

I3D + Non local block

	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**

Space vs Time vs Spacetime

	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**

Non-local Neural Networks (CVPR 2018)

- **Experiments**:
  - Kinetics dataset에서 SoTA를 달성
  - 기존의 SoTA를 달성한 모델들과 비교해 적은 입력으로 더 좋은 성능을 보임
  - 이 외에도 다른 분야에서도 의미있는 실험 결과를 보임

Comparisons with state-of-the-art results in Kinetics

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**

한줄평: Denoising에 사용되는 Non-local filter로부터 영감을 받아 효율적이면서 다른 architecture에 적용하기 쉽다는 점과 기존의 video task의 고정 입력으로 쓰인 optical flow를 사용하지 않고 좋은 성능을 낸 점이 상당히 인상적이였다. 특히 이미지의 전체 영역을 활용한다는 점은 다른 분야에서도 충분히 활용할 가치가 있다고 생각한다.