Special Networks

Hung-yi Lee
李宏毅

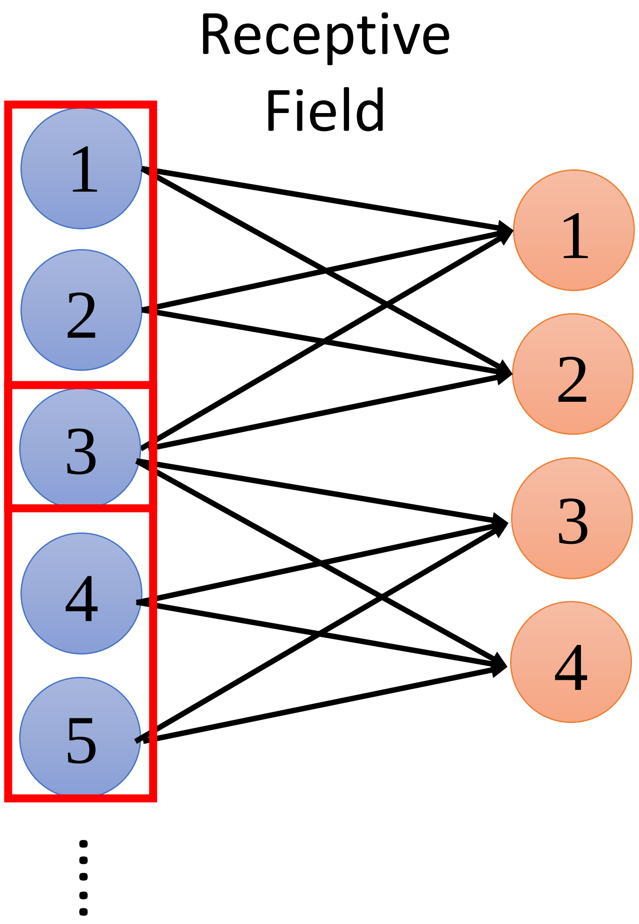
Announcement

- 11/13 (下週一) 14:00 ~ 17:00 台灣微軟參訪
  - 地址：台北市信義區忠孝東路五段68號19樓 (捷運市政府站3號出口)
  - 14:00：在捷運市政府站3號出口集合
  - 報名表單：
    - https://docs.google.com/forms/d/e/1FAIpQLSfs2zloGanjWjJvVkJu8DUe9BlVZ5ugLPIs3FUmMbR9VkF8Fw/viewform?fbzx=-8767653761190698000

Outline

- Convolutional Neural Network (Review)
- Spatial Transformer
- Highway Network & Grid LSTM
- Pointer Network
- External Memory

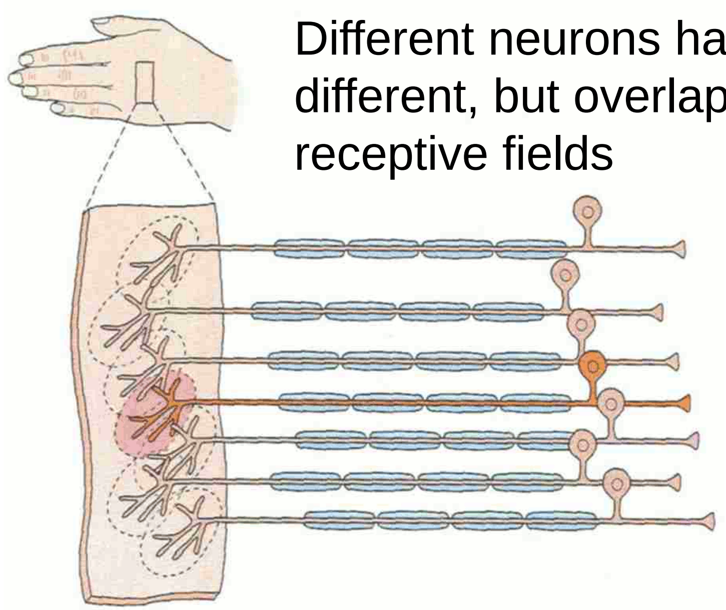
Convolutional Layer



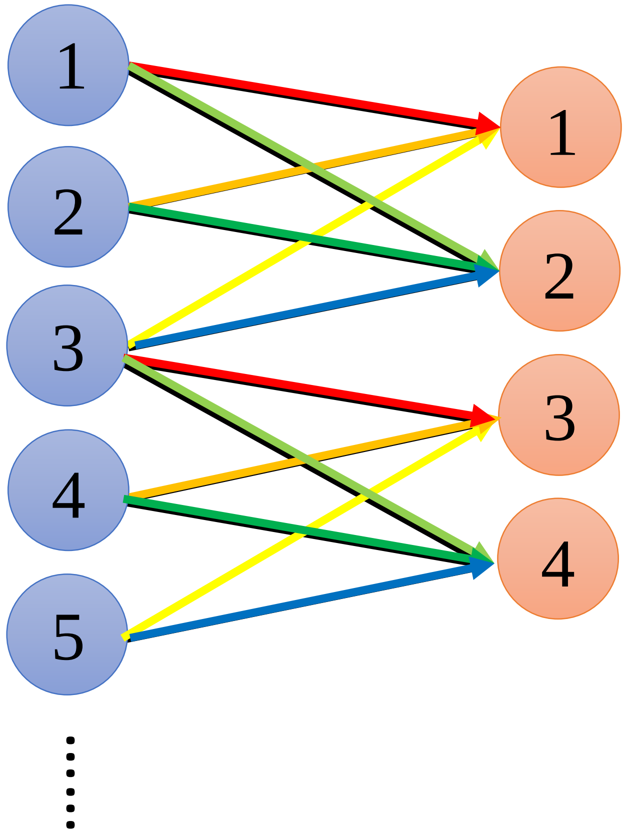
Sparse Connectivity

Each neural only connects to part of the output of the previous layer

Different neurons have different, but overlapping, receptive fields



Convolutional Layer



Sparse Connectivity

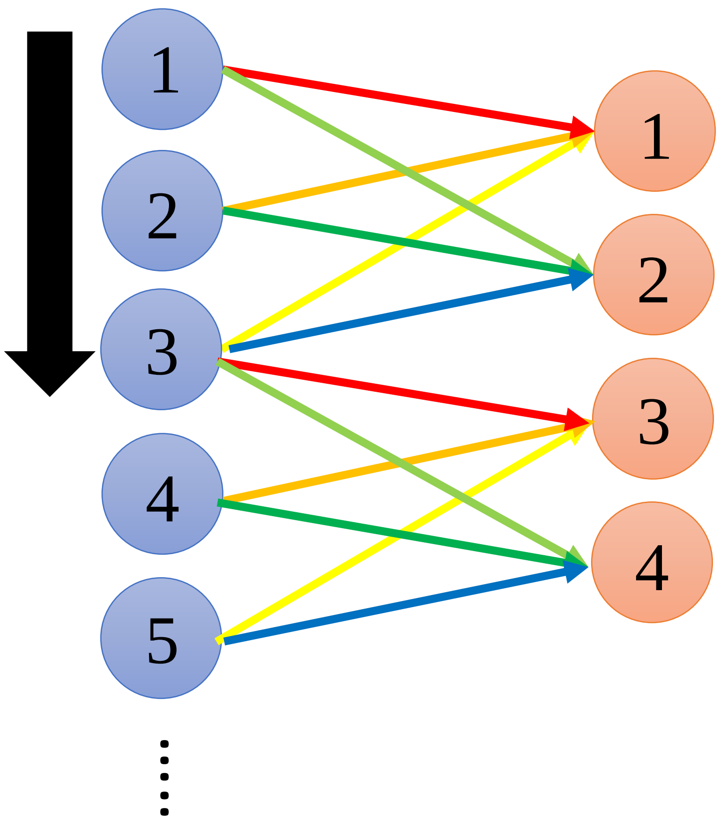
Each neural only connects to part of the output of the previous layer

Parameter Sharing

The neurons with different receptive fields can use the same set of parameters.

Less parameters than fully connected layer

Convolutional Layer



Considering neuron 1 and 3 as "filter 1" (kernel 1)

filter (kernel) size: size of the receptive field of a neuron

Stride = 2

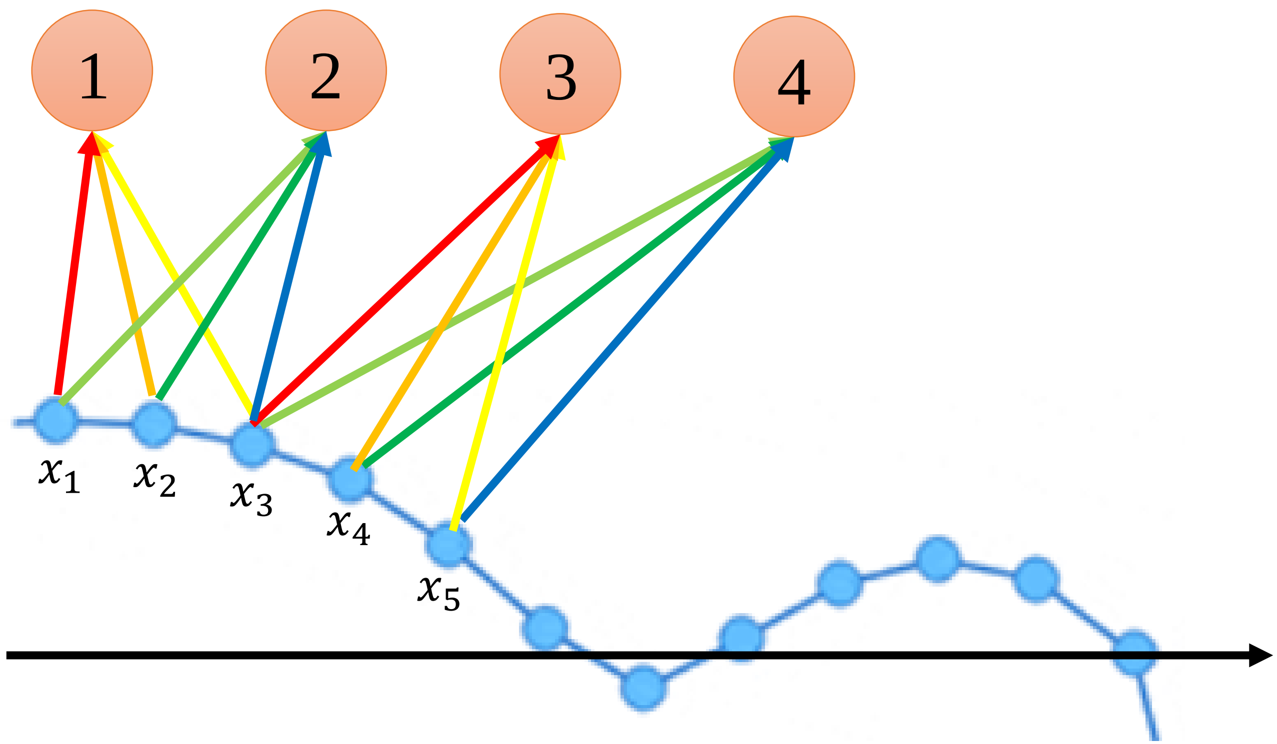
Considering neuron 2 and 4 as "filter 2" (kernel 2)

Kernel size, no. of filter, stride are all designed by the developers.

Example –
1D Signal + Single Channel

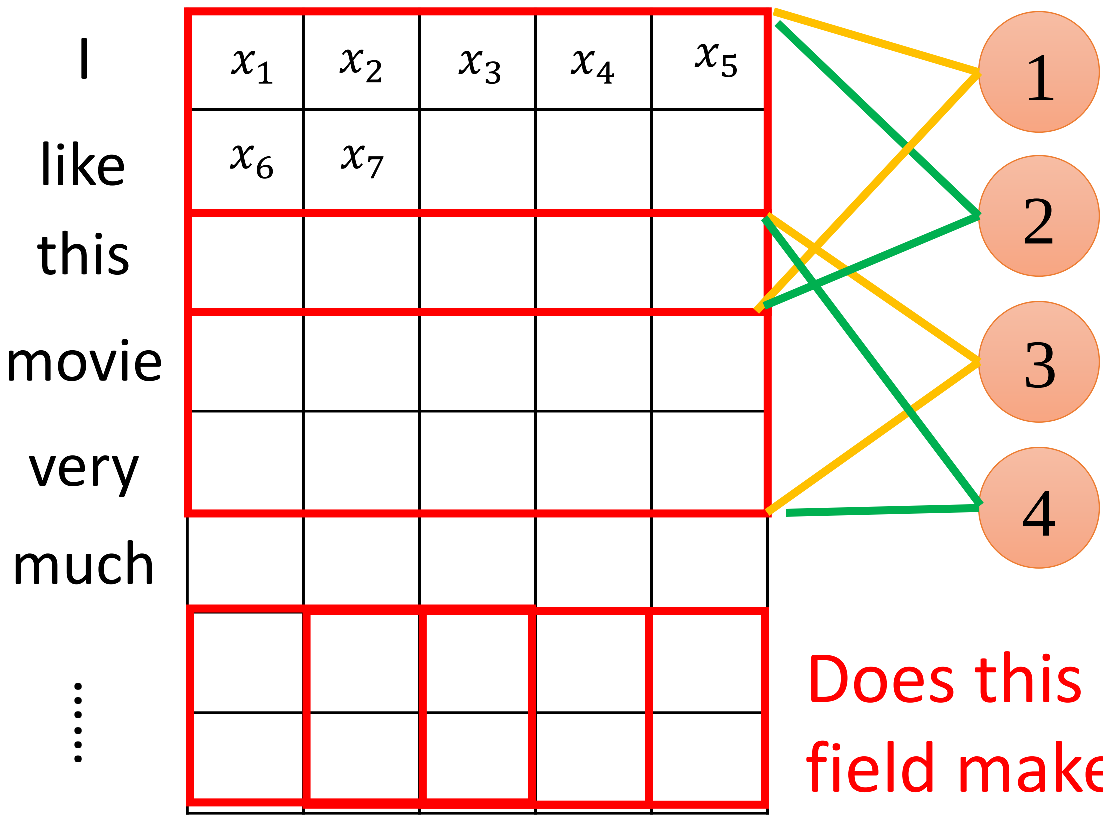
Classification, Predict the future …

Audio Signal,
Stock Value …



Example –
1D Signal + Multiple Channel

A document: each word is a vector



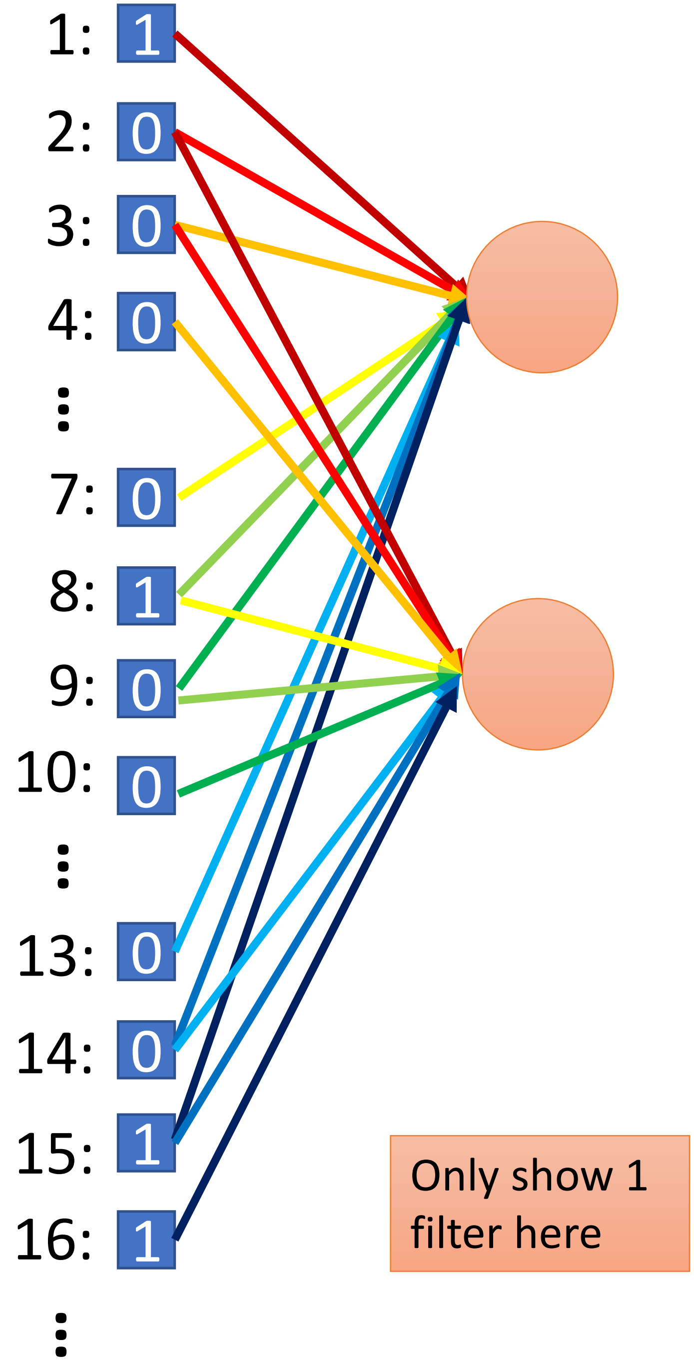
Does this kind of receptive field make sense?

Example – 2D Signal + Single Channel

Size of Receptive field is 3x3, Stride is 1

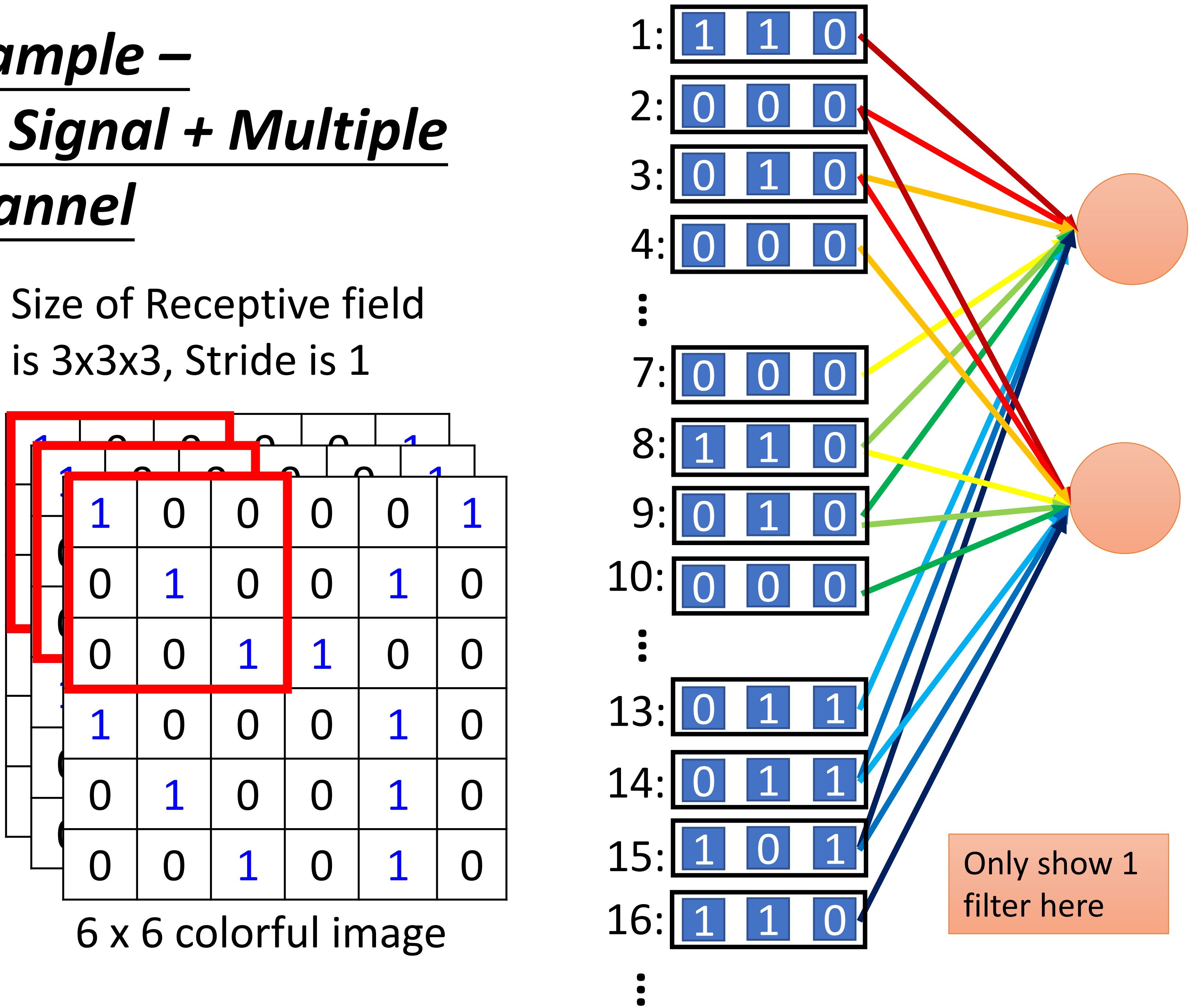
1	0	0	0	0	1
0	1	0	0	1	0
0	0	1	1	0	0
1	0	0	0	1	0
0	1	0	0	1	0
0	0	1	0	1	0

6 x 6 black & white picture image



Example – 2D Signal + Multiple Channel

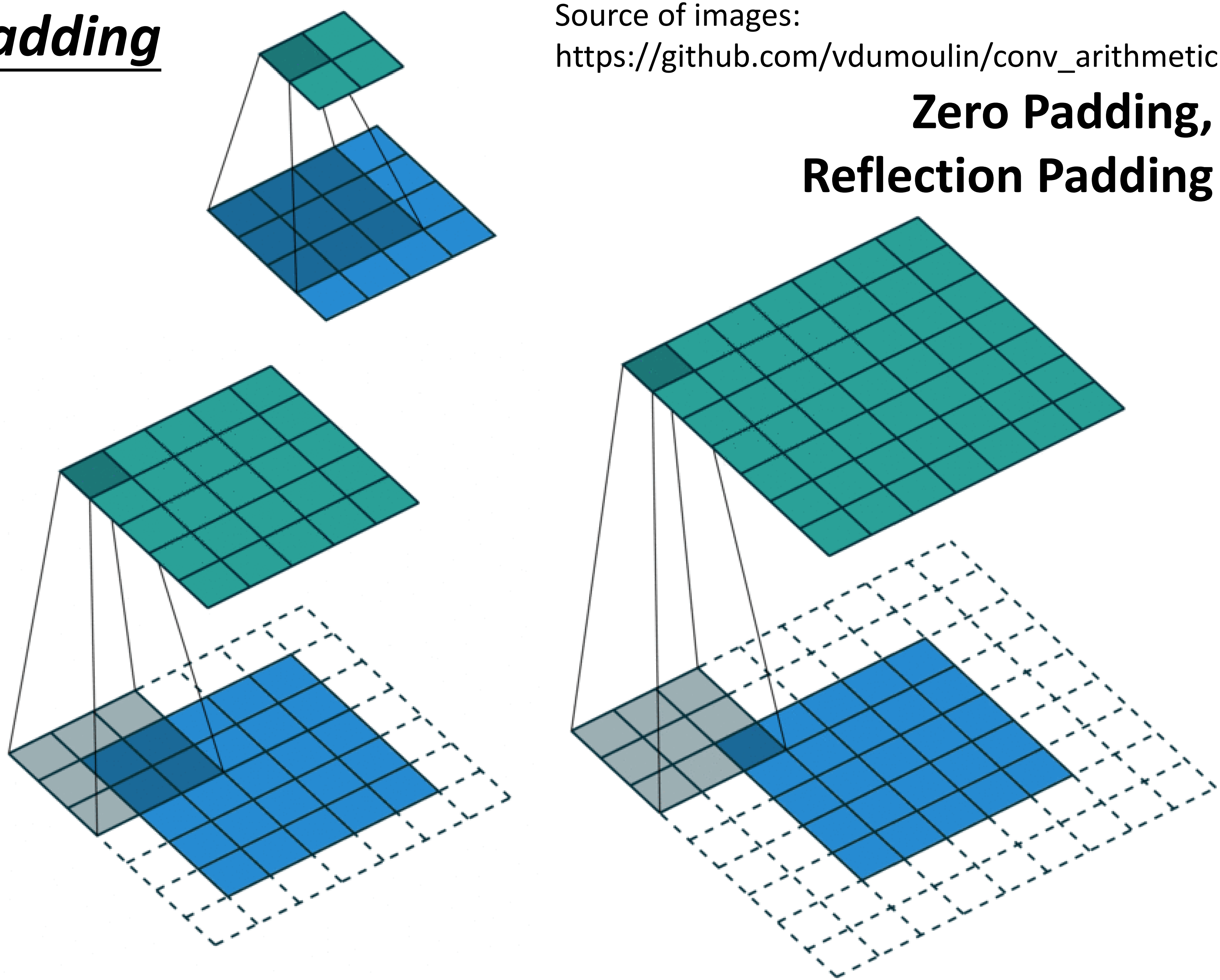
Size of Receptive field is 3x3x3, Stride is 1



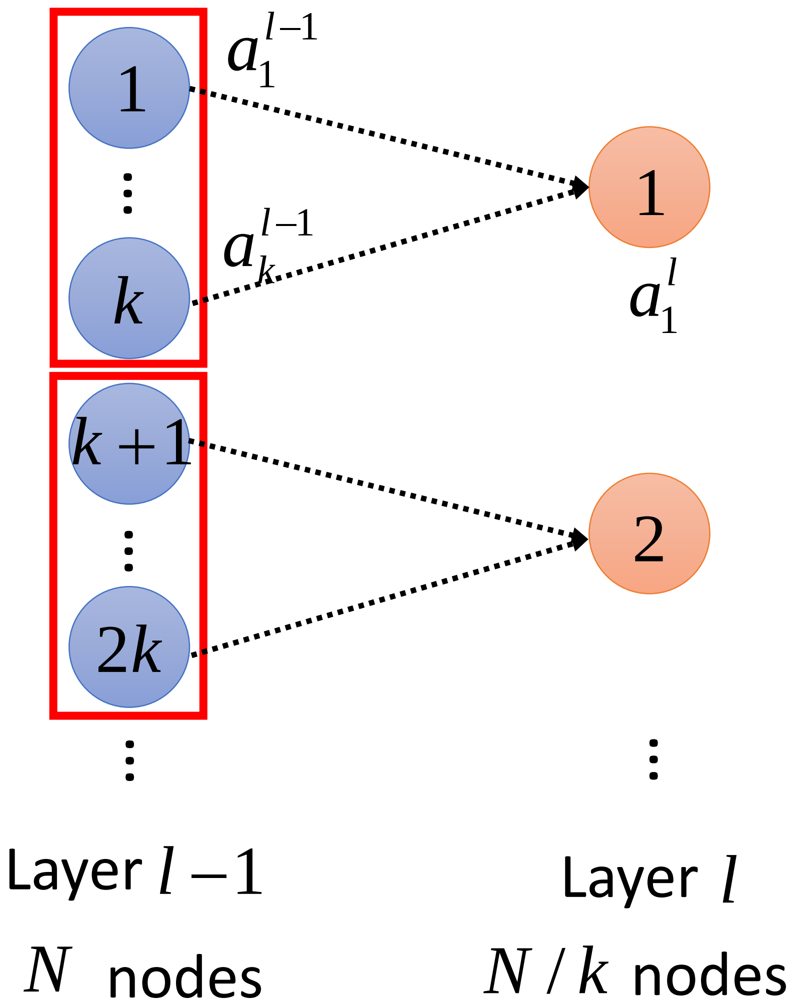
Padding

Source of images:
https://github.com/vdumoulin/conv_arithmetic

Zero Padding,
Reflection Padding



Pooling Layer



k outputs in layer $l-1$ are grouped together

Each output in layer l "summarizes" k inputs

Average Pooling:

$$a_1^l = \frac{1}{k}\sum_{j=1}^{k} a_j^{l-1}$$

Max Pooling:

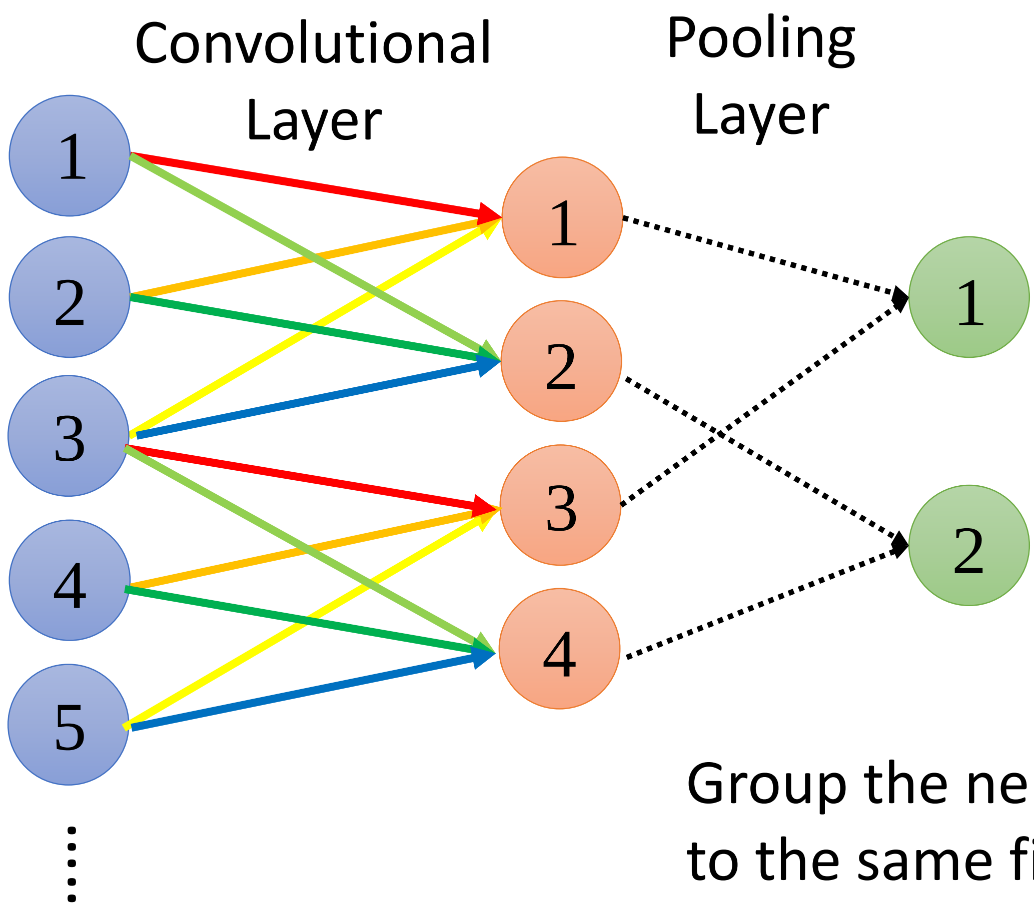
$$a_1^l = \max\left(a_1^{l-1}, a_2^{l-1}, \cdots, a_k^{l-1}\right)$$

L2 Pooling:

$$a_1^l = \frac{1}{k}\sqrt{\sum_{j=1}^{k}\left(a_j^{l-1}\right)^2}$$

Pooling Layer

Which outputs should be grouped together?

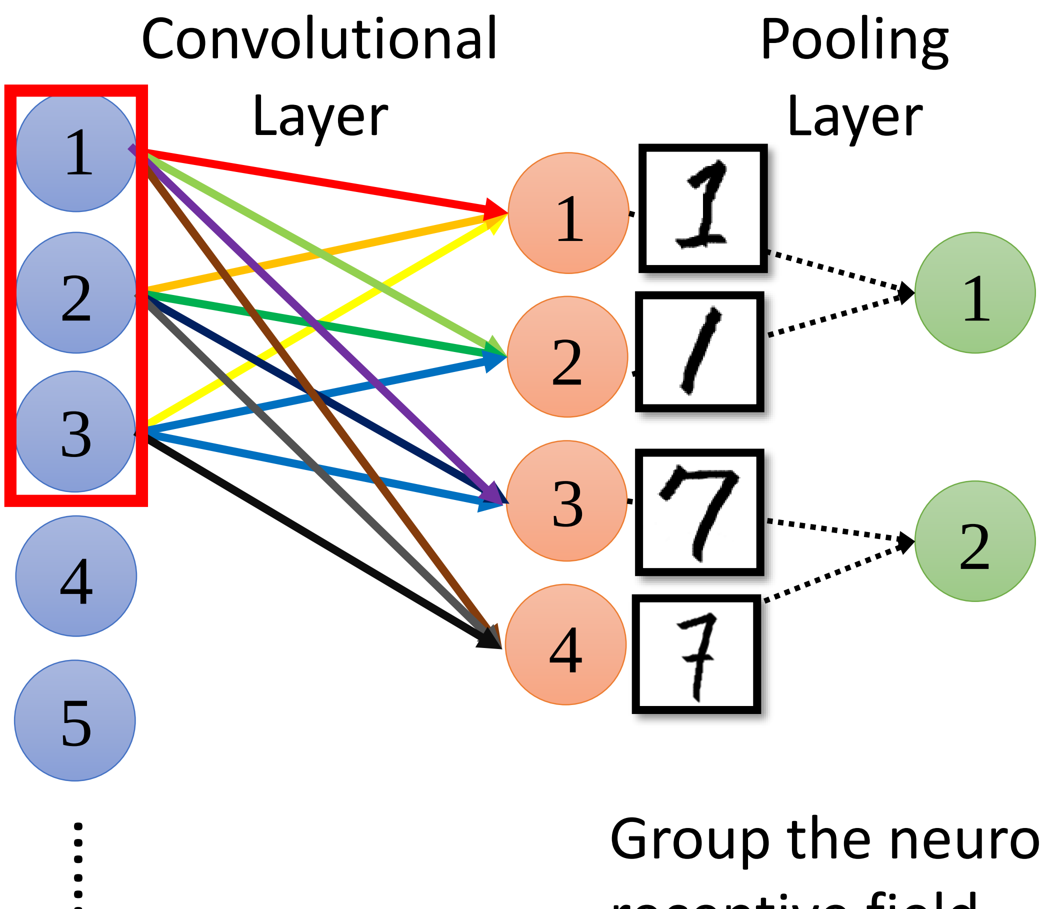


Subsampling

Group the neurons corresponding to the same filter with nearby receptive fields

Pooling Layer

Which outputs should be grouped together?

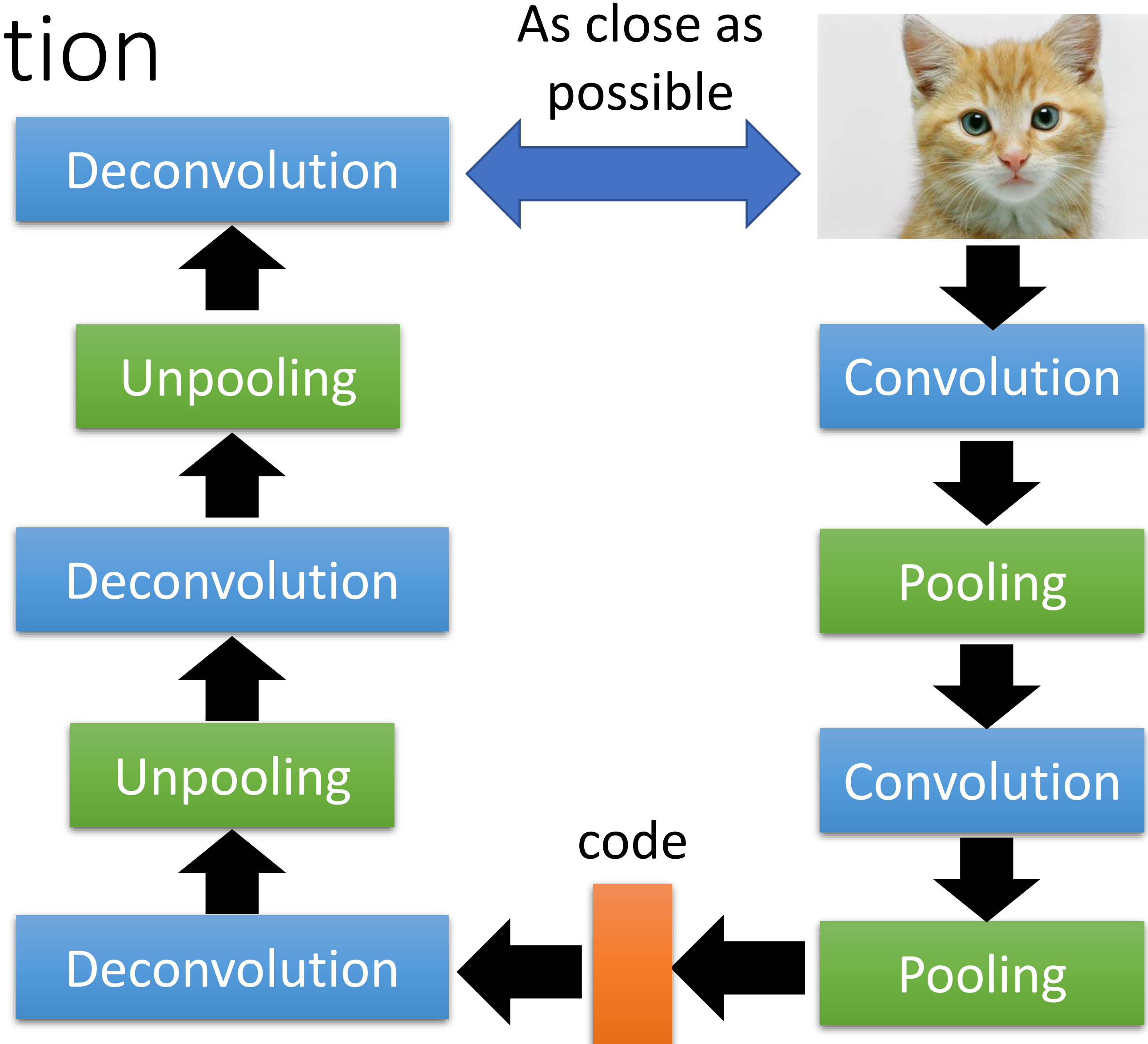


Maxout Network

How can you know whether the neurons detect the same pattern?

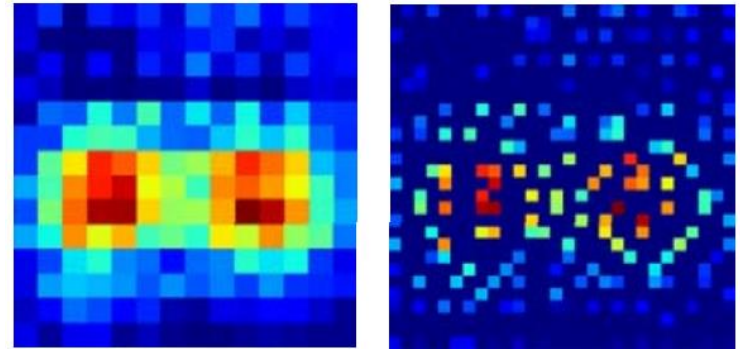
Group the neurons with the same receptive field

Unpooling & Deconvolution

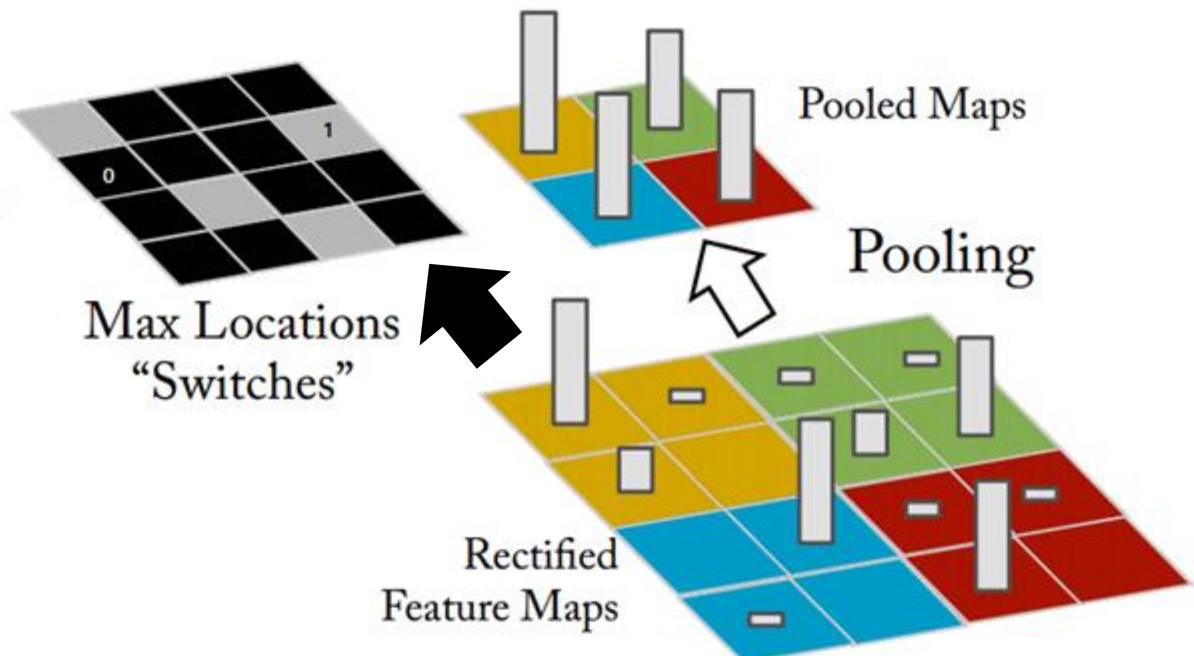


Auto-encoder for CNN

Unpooling



14 x 14 28 x 28

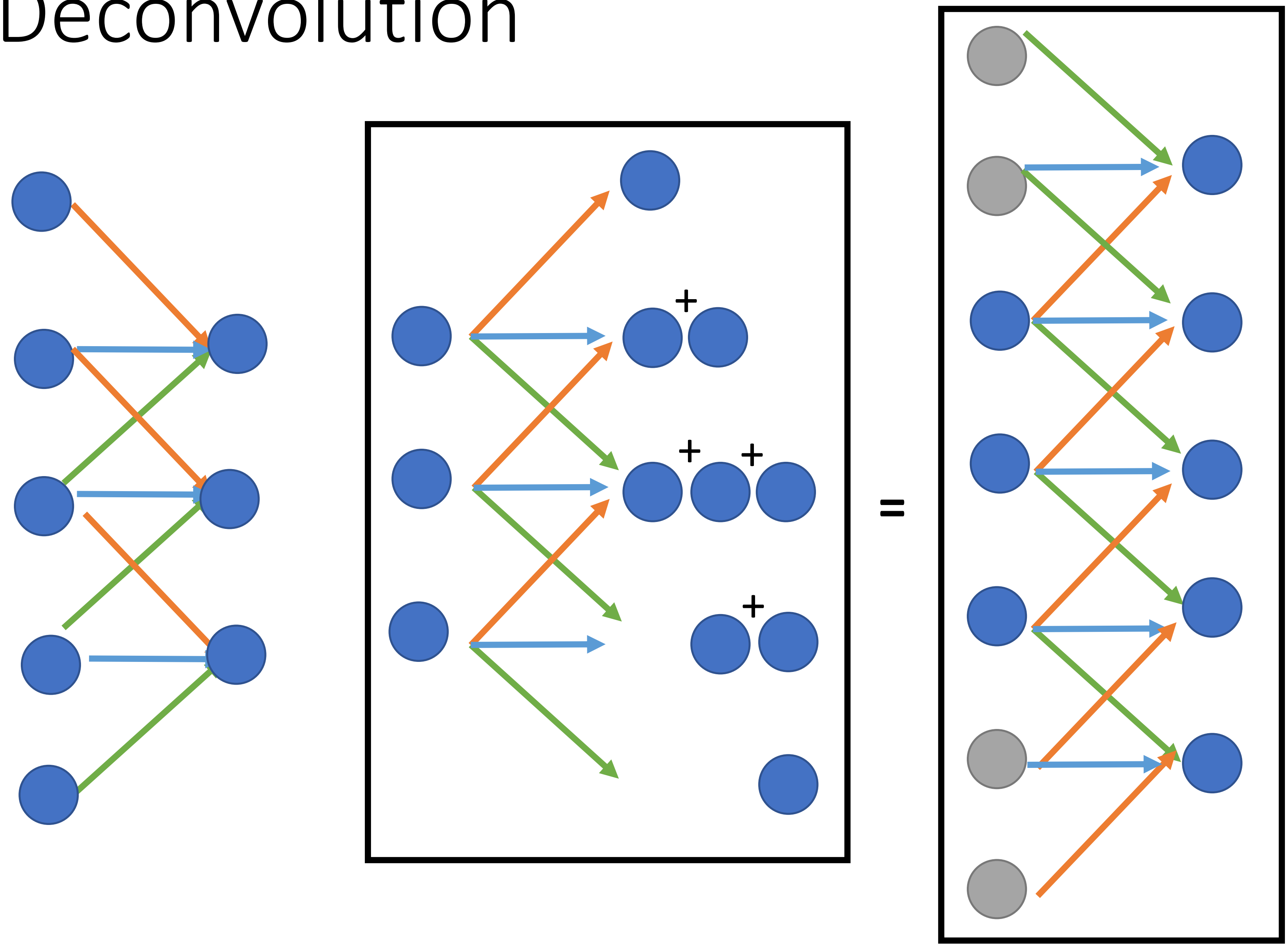


Alternative: simply repeat the values

Source of image : https://leonardoaraujosantos.gitbooks.io/artificial-inteligence/content/image_segmentation.html

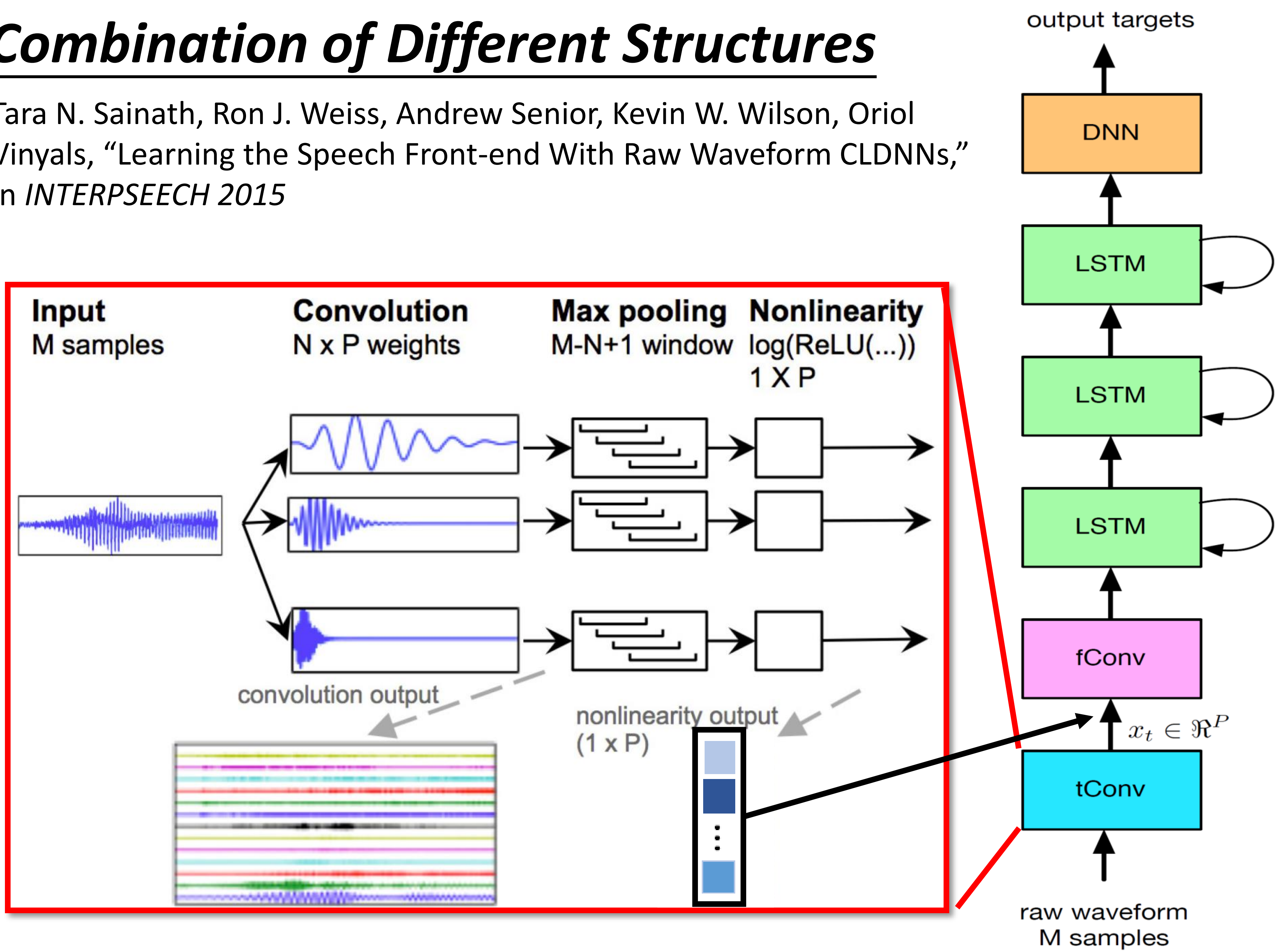
Actually, deconvolution is convolution.

Deconvolution



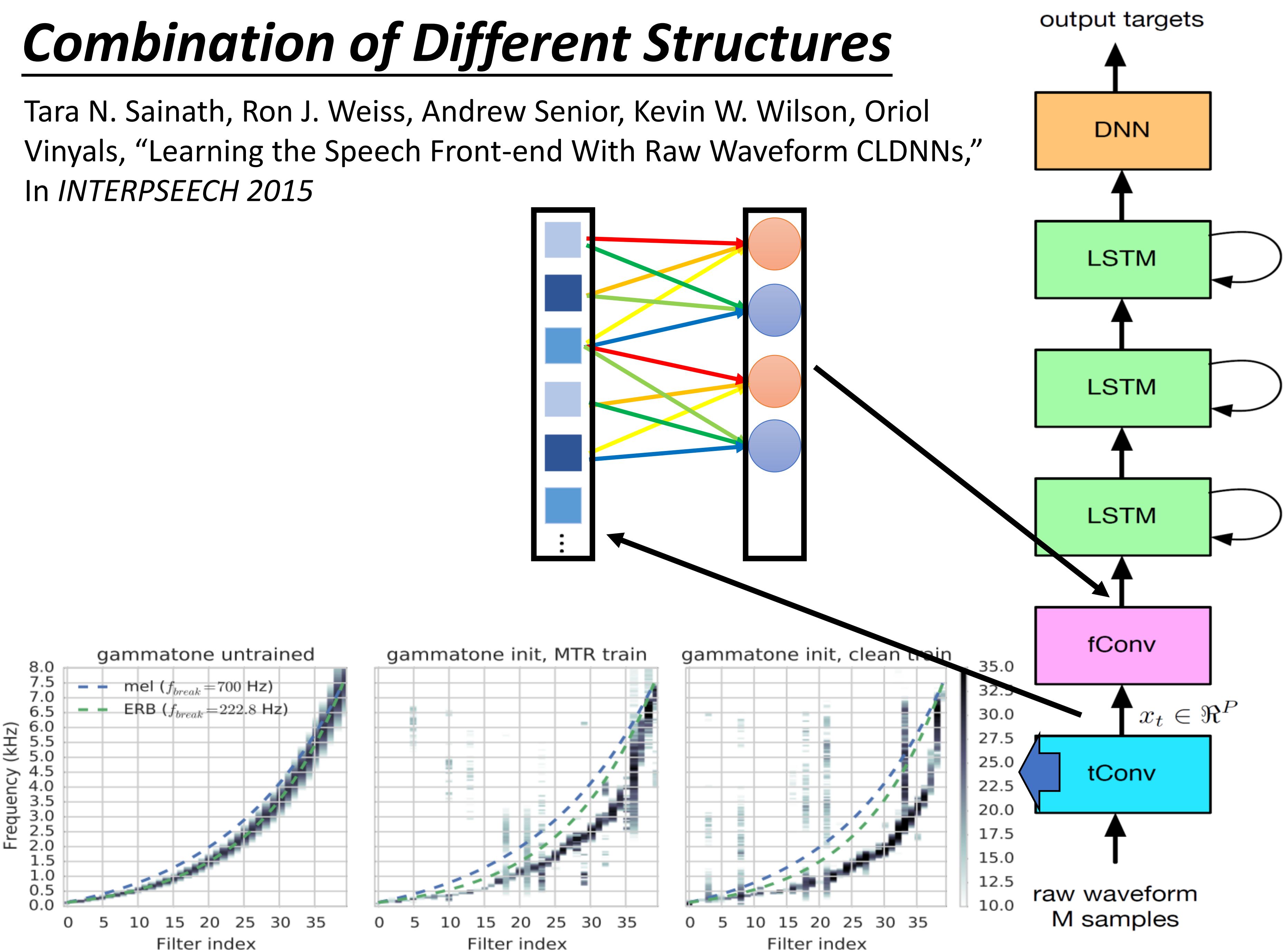
Combination of Different Structures

Tara N. Sainath, Ron J. Weiss, Andrew Senior, Kevin W. Wilson, Oriol Vinyals, "Learning the Speech Front-end With Raw Waveform CLDNNs," In *INTERPSEECH 2015*



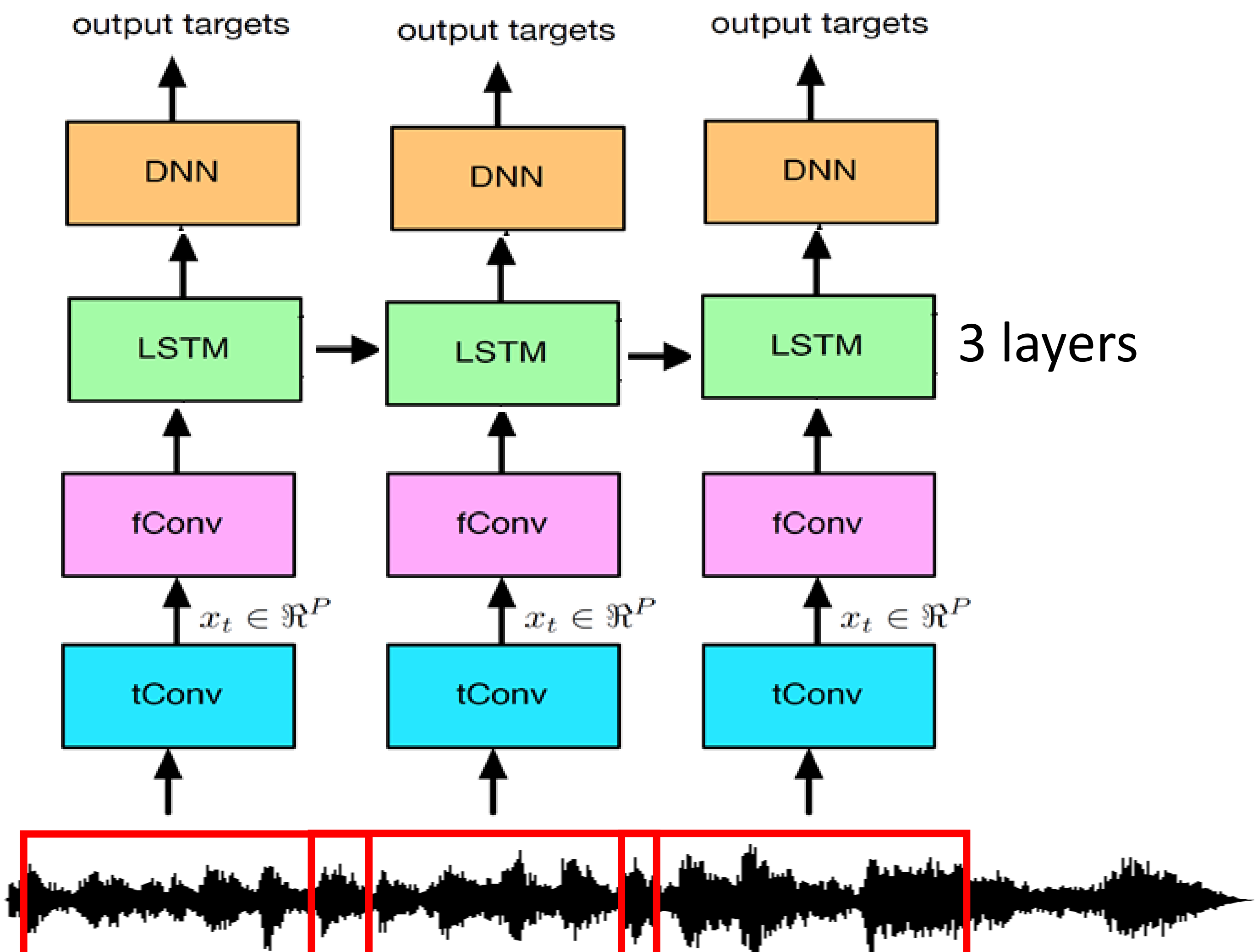
Combination of Different Structures

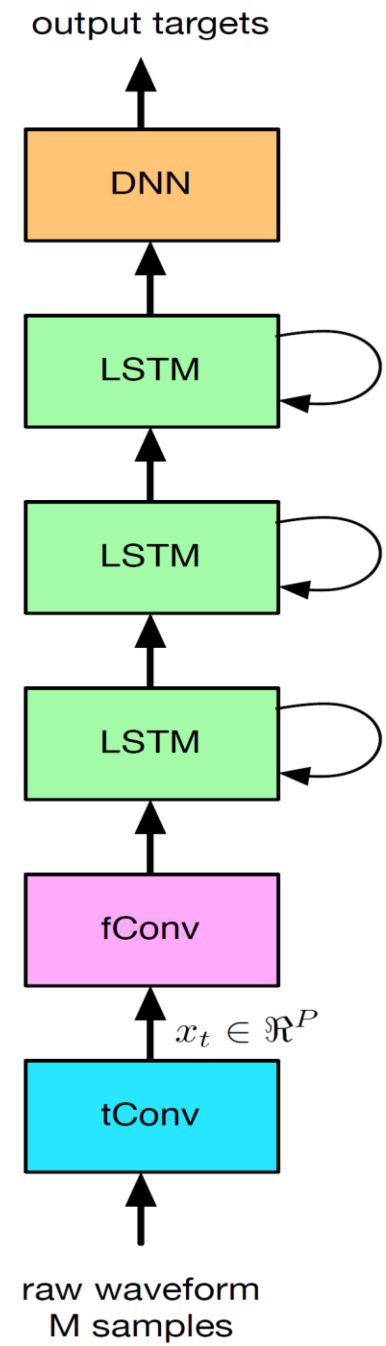
Tara N. Sainath, Ron J. Weiss, Andrew Senior, Kevin W. Wilson, Oriol Vinyals, “Learning the Speech Front-end With Raw Waveform CLDNNs,” In *INTERPSEECH 2015*



Combination of Different Structures

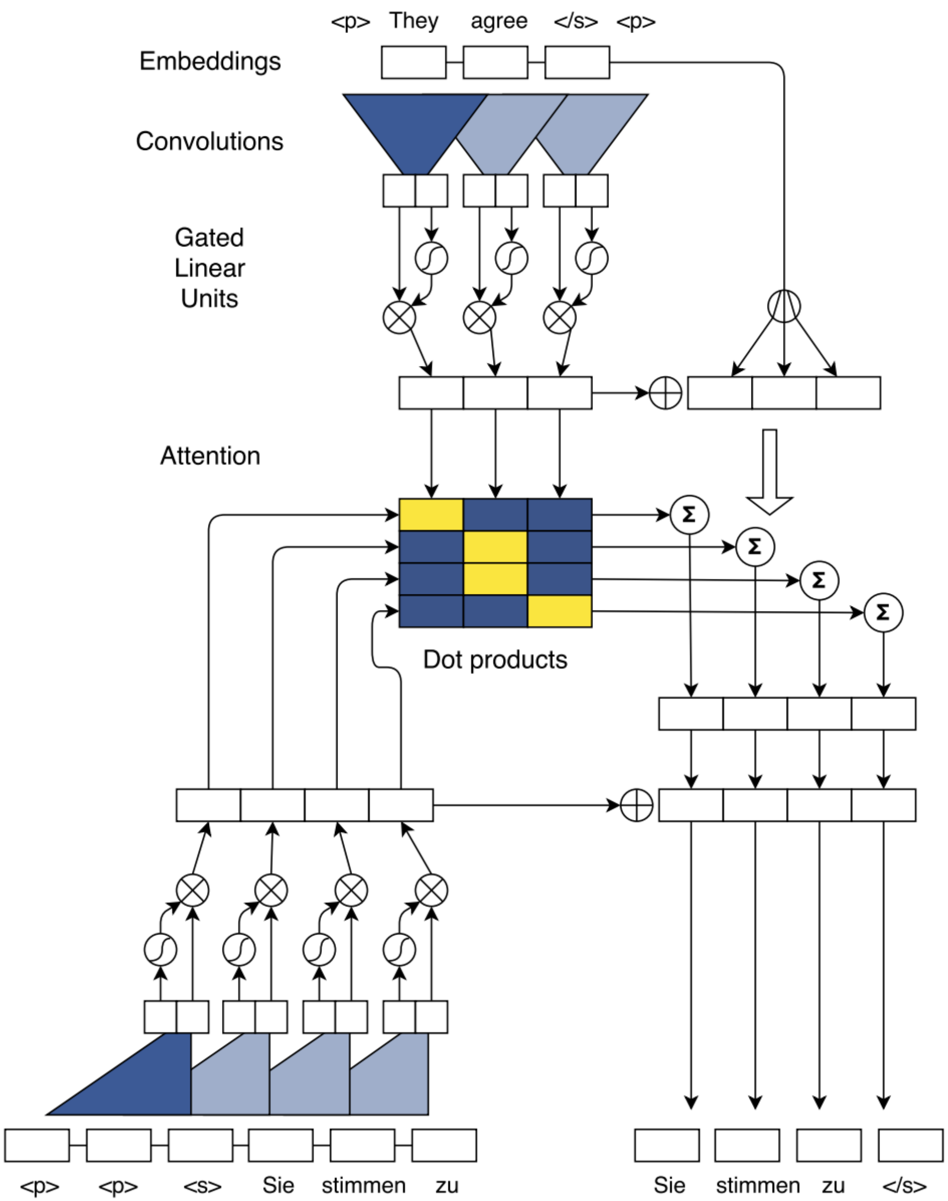
Tara N. Sainath, Ron J. Weiss, Andrew Senior, Kevin W. Wilson, Oriol Vinyals, “Learning the Speech Front-end With Raw Waveform CLDNNs,” In *INTERPSEECH 2015*





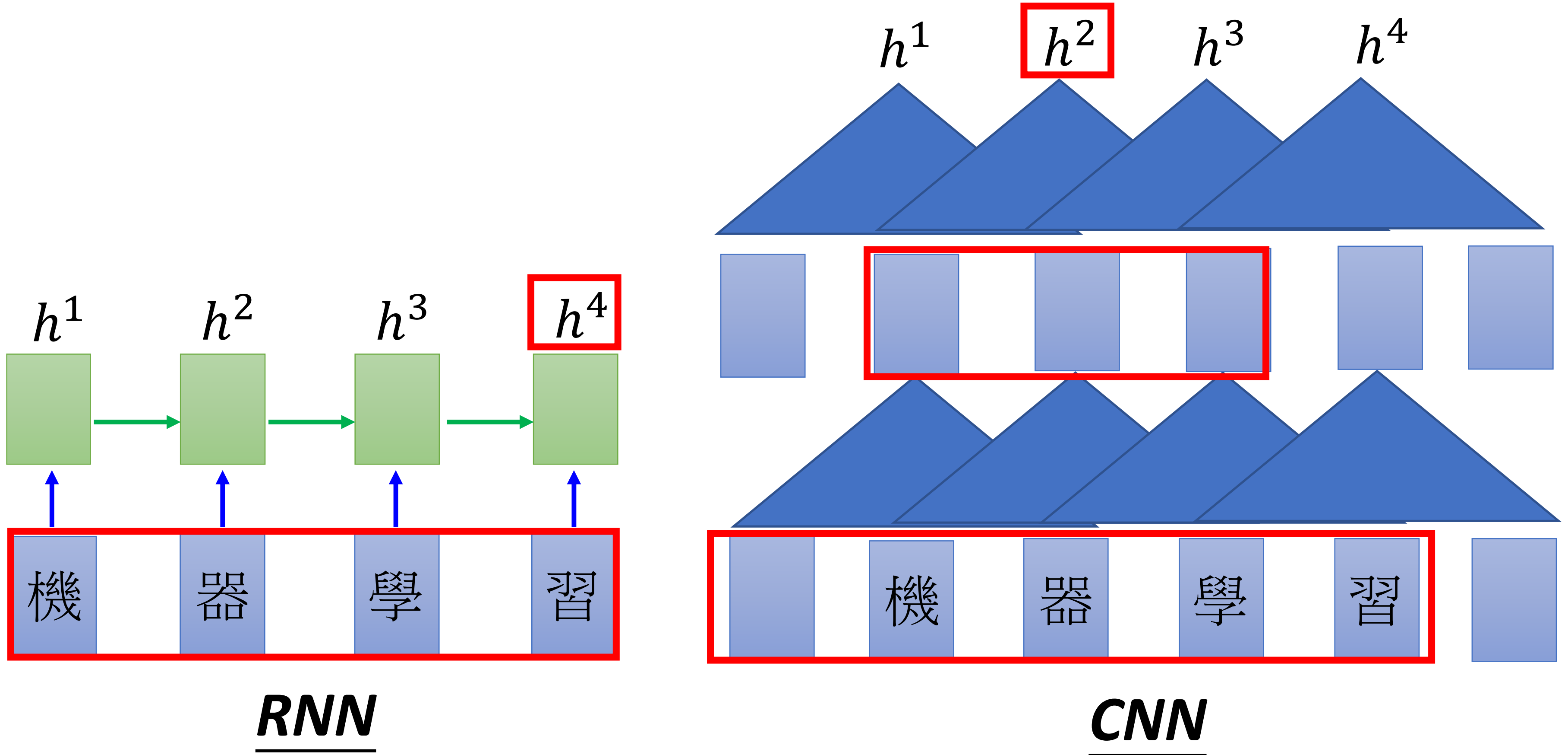
CNN for Sequence-to-sequence

https://arxiv.org/abs/1705.03122



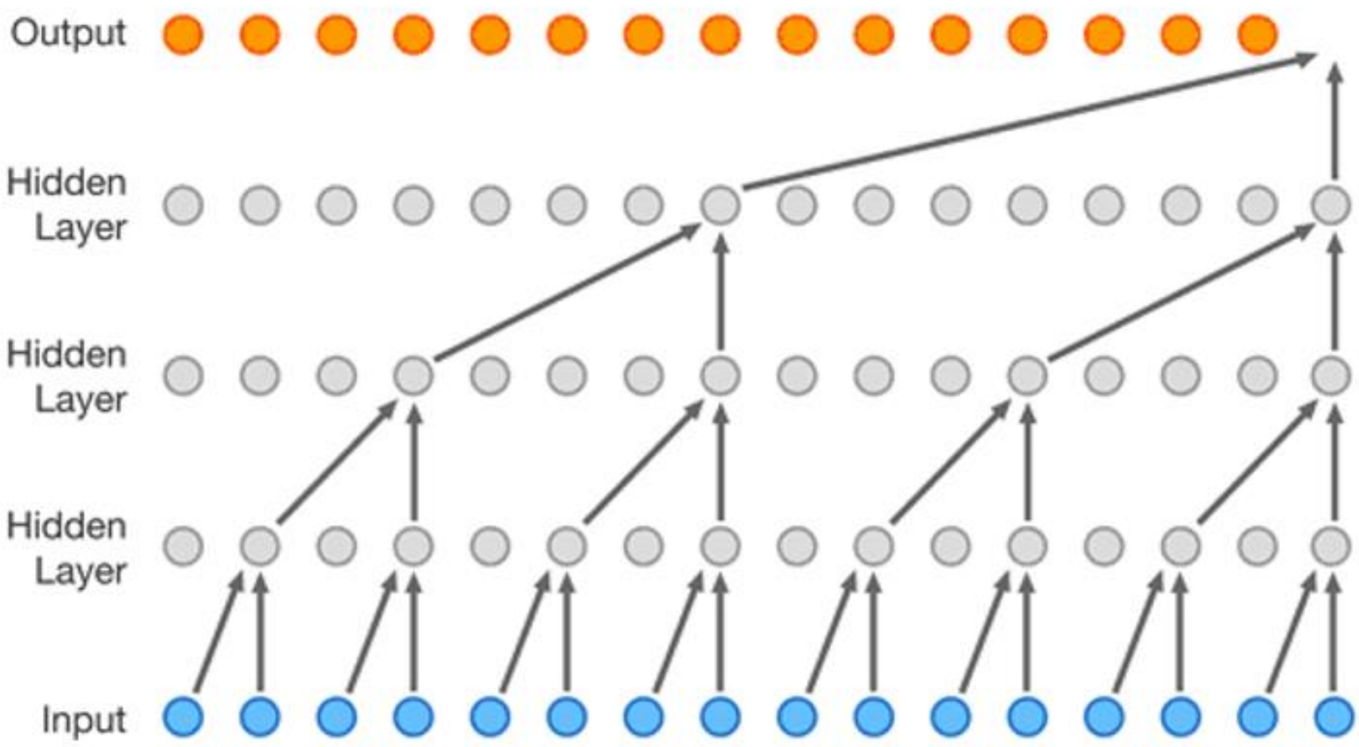
CNN for Sequence-to-sequence

- Encoder



CNN for Sequence-to-sequence

- Decoder - WaveNet

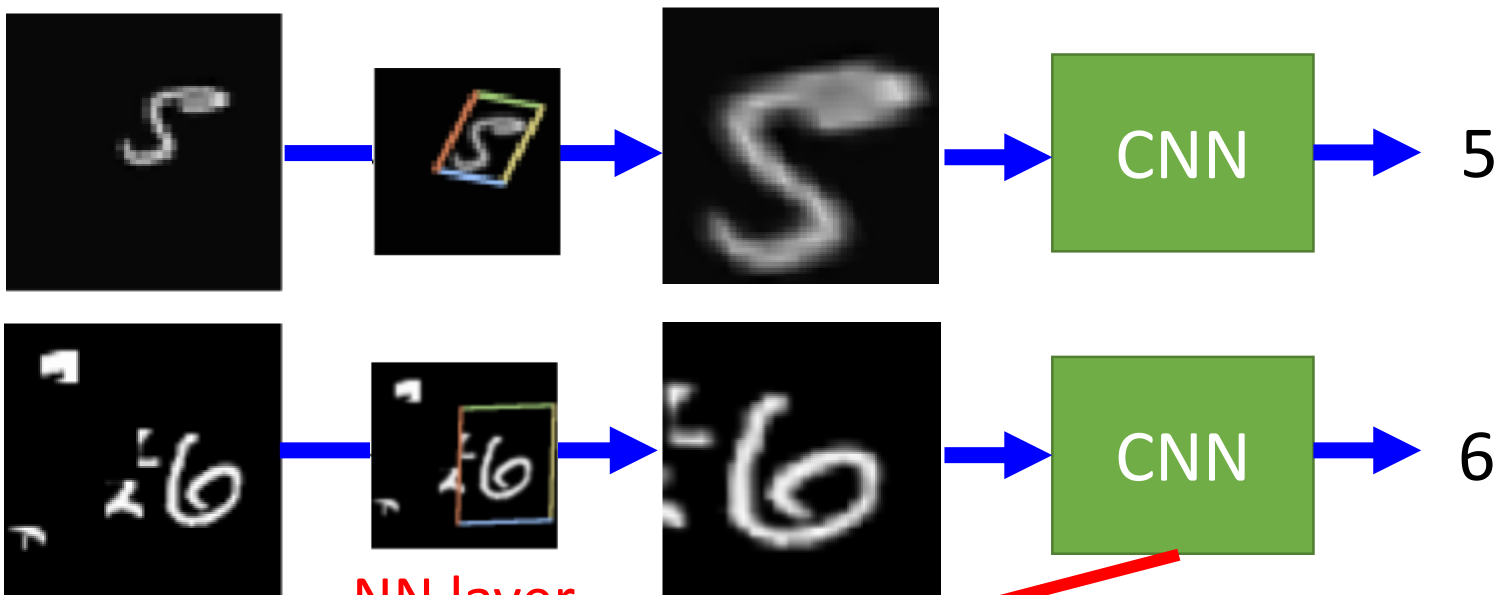


Outline

- Convolutional Neural Network (Review)
- Spatial Transformer
- Highway Network & Grid LSTM
- Pointer Network
- External Memory

Spatial Transformer Layer

- CNN is not invariant to scaling and rotation



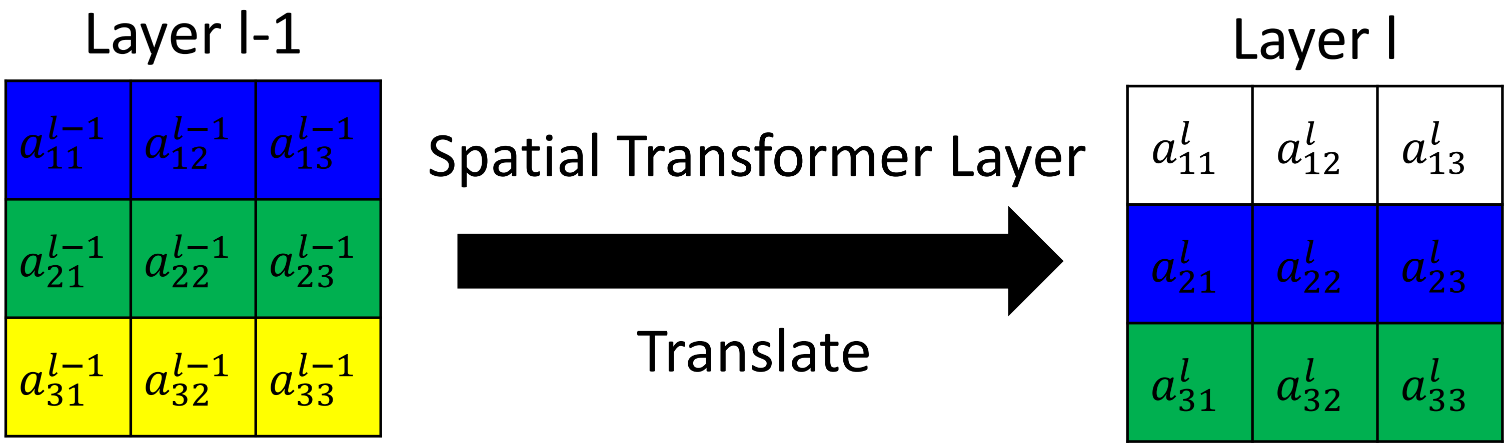
NN layer

End-to-end learn

Can also transform feature map

Spatial Transformer Layer

- How to transform an image/feature map



General layer: $a_{nm}^{l} = \sum_{i=1}^{3} \sum_{j=1}^{3} w_{nm,ij}^{l} a_{ij}^{l-1}$

If we want translate as above: $a_{nm}^{l} = a_{(n-1)m}^{l-1}$

$$w_{nm,ij}^{l} = 1 \quad if\ i = n-1, j = m \qquad w_{nm,ij}^{l} = 0 \quad otherwise$$

Spatial Transformer Layer

- How to transform an image/feature map

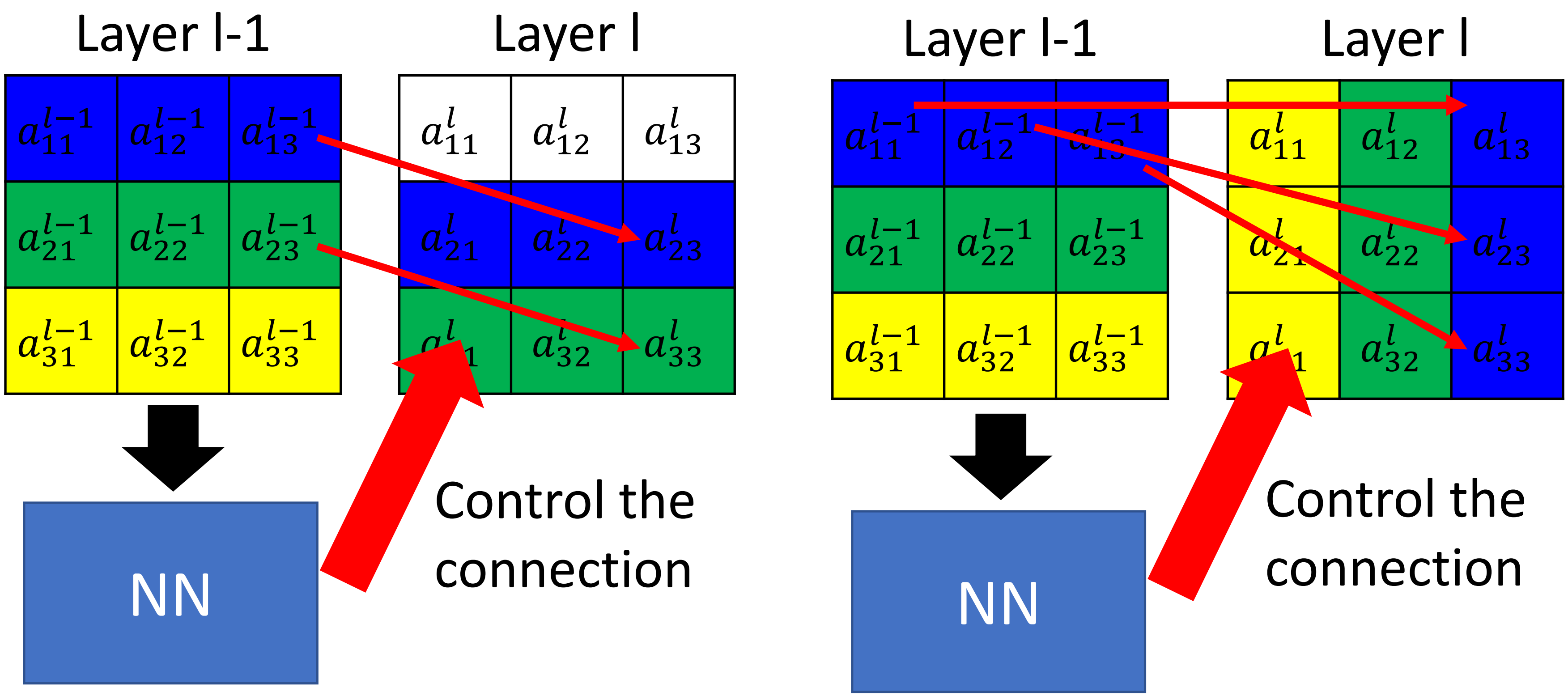
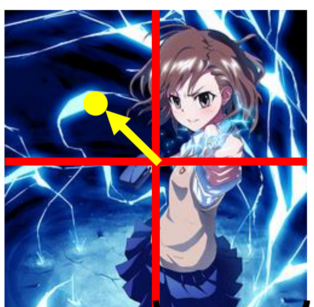


Image Transformation

Expansion, Compression, Translation

$$\begin{bmatrix} x' \\ y' \end{bmatrix} = \begin{bmatrix} 2 & 0 \\ 0 & 2 \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} + \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

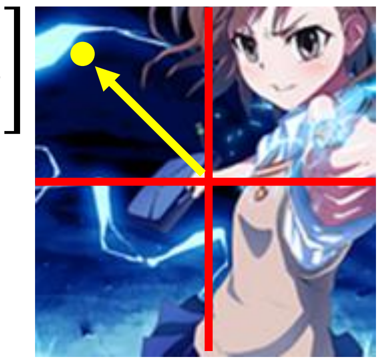
$\begin{bmatrix} x \\ y \end{bmatrix}$



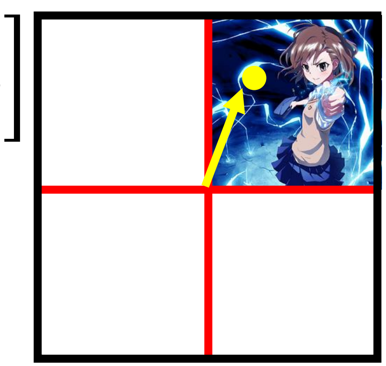
1

1

$\begin{bmatrix} x' \\ y' \end{bmatrix}$



$\begin{bmatrix} x' \\ y' \end{bmatrix}$



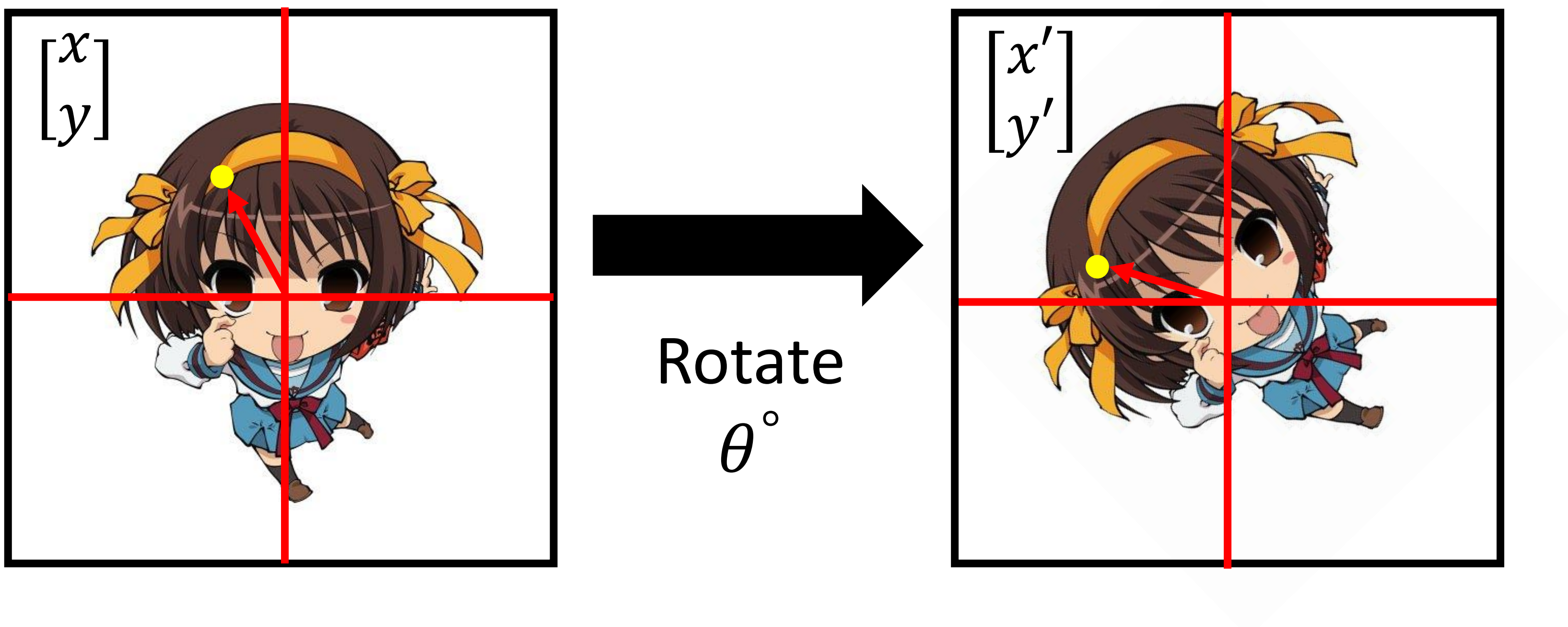
$$\begin{bmatrix} x' \\ y' \end{bmatrix} = \begin{bmatrix} 0.5 & 0 \\ 0 & 0.5 \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} + \begin{bmatrix} 0.5 \\ 0.5 \end{bmatrix}$$

https://home.gamer.com.tw/creationDetail.php?sn=792585

Image Transformation

- ***Rotation***

$$\begin{bmatrix}x'\\y'\end{bmatrix}=\begin{bmatrix}cos\theta & -sin\theta\\ sin\theta & cos\theta\end{bmatrix}\begin{bmatrix}x\\y\end{bmatrix}+\begin{bmatrix}0\\0\end{bmatrix}$$

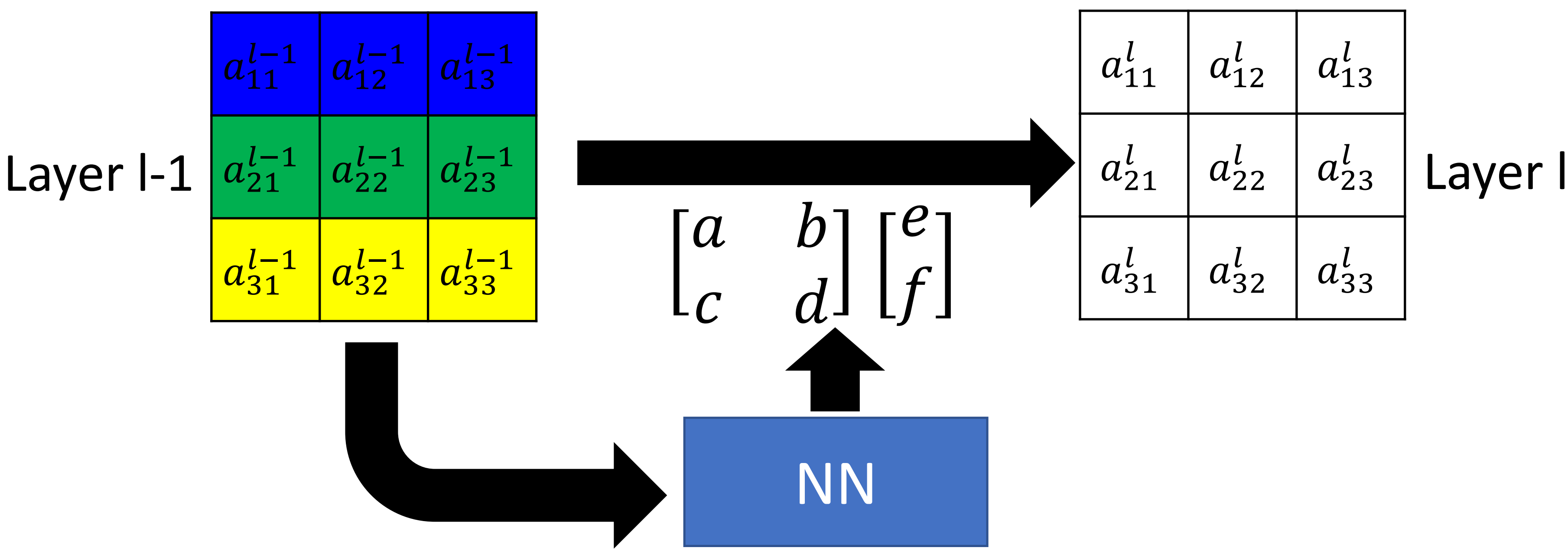


Spatial Transformer Layer

$$\begin{bmatrix} x' \\ y' \end{bmatrix} = \begin{bmatrix} a & b \\ c & d \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} + \begin{bmatrix} e \\ f \end{bmatrix}$$

6 parameters to describe the affine transformation

Index of layer l-1 Index of layer l



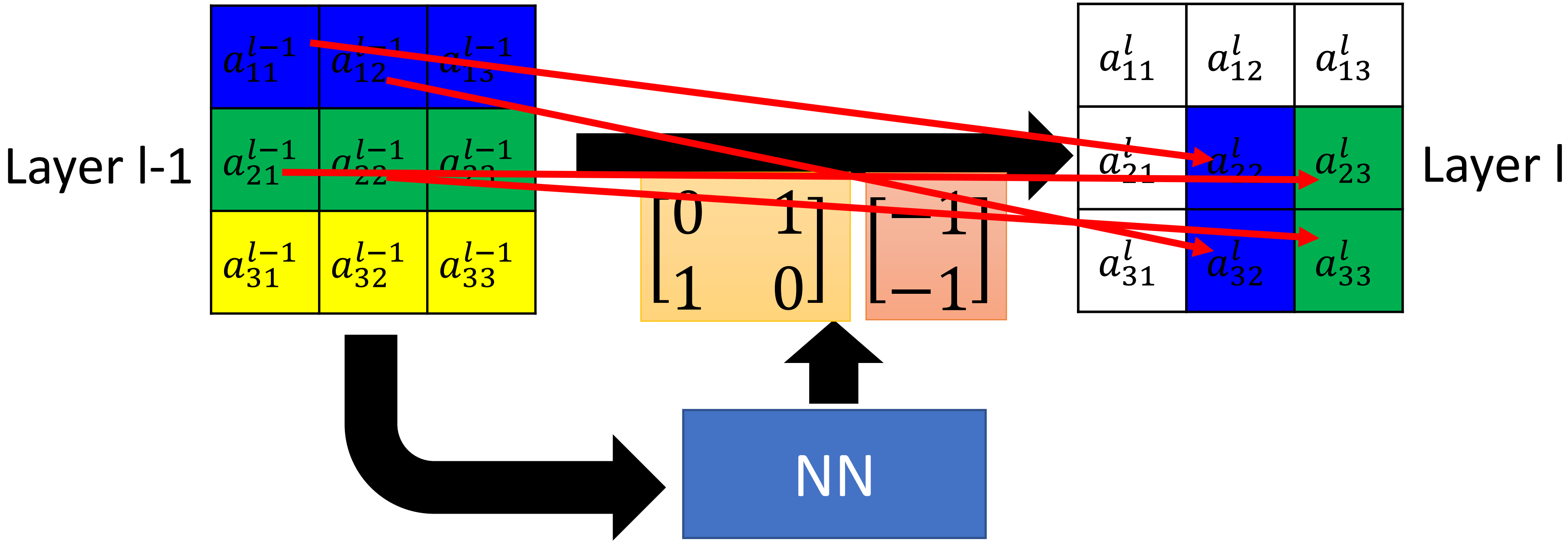
Spatial Transformer Layer

$$\begin{bmatrix} x' \\ y' \end{bmatrix} = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} + \begin{bmatrix} -1 \\ -1 \end{bmatrix}$$

6 parameters to describe the affine transformation

Index of layer l-1

Index of layer l



Spatial Transformer Layer

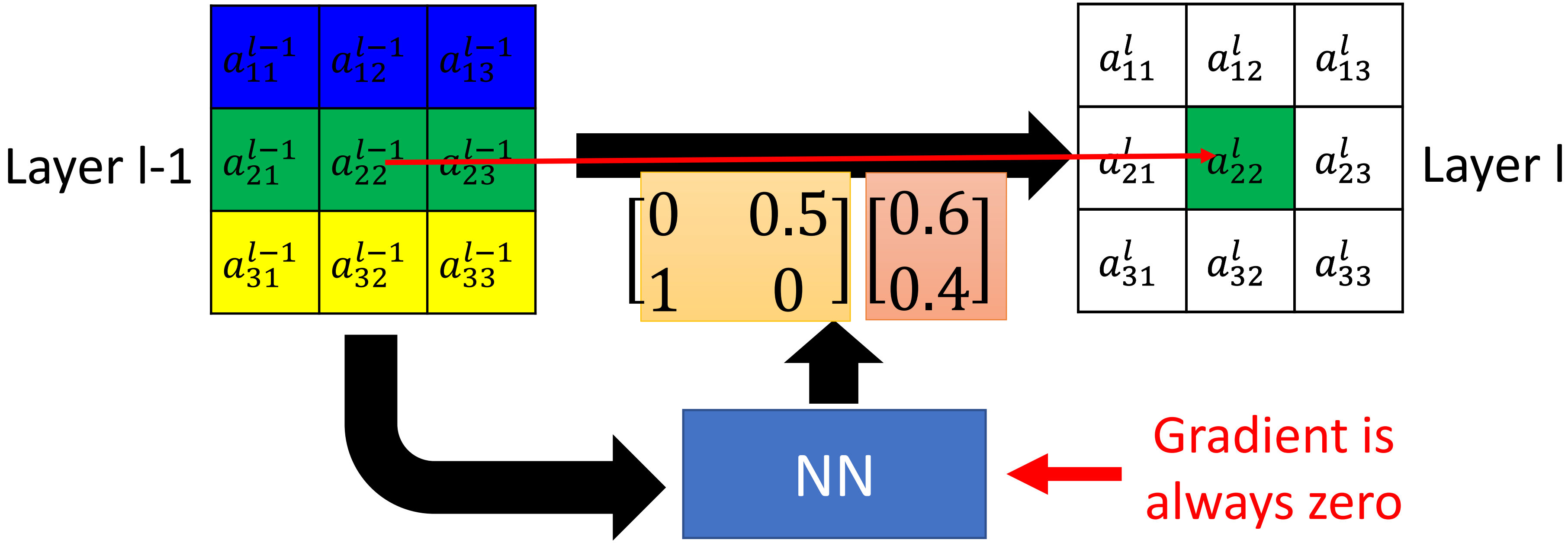
$$\begin{bmatrix}1.6\\2.4\end{bmatrix}=\begin{bmatrix}0 & 0.5\\1 & 0\end{bmatrix}\begin{bmatrix}2\\2\end{bmatrix}+\begin{bmatrix}0.6\\0.4\end{bmatrix}$$

Index of layer l-1

Index of layer l

6 parameters to describe the affine transformation

What is the problem?



Gradient is always zero

Interpolation

Now we can use gradient descent

$$\begin{bmatrix}1.6\\2.4\end{bmatrix} = \begin{bmatrix}0 & 0.5\\1 & 0\end{bmatrix}\begin{bmatrix}2\\2\end{bmatrix} + \begin{bmatrix}0.6\\0.4\end{bmatrix}$$

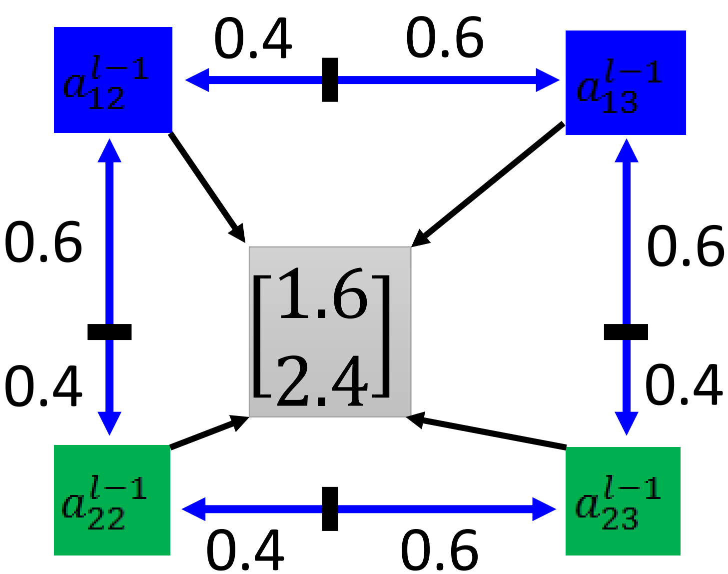
Index of layer l-1

Index of layer l

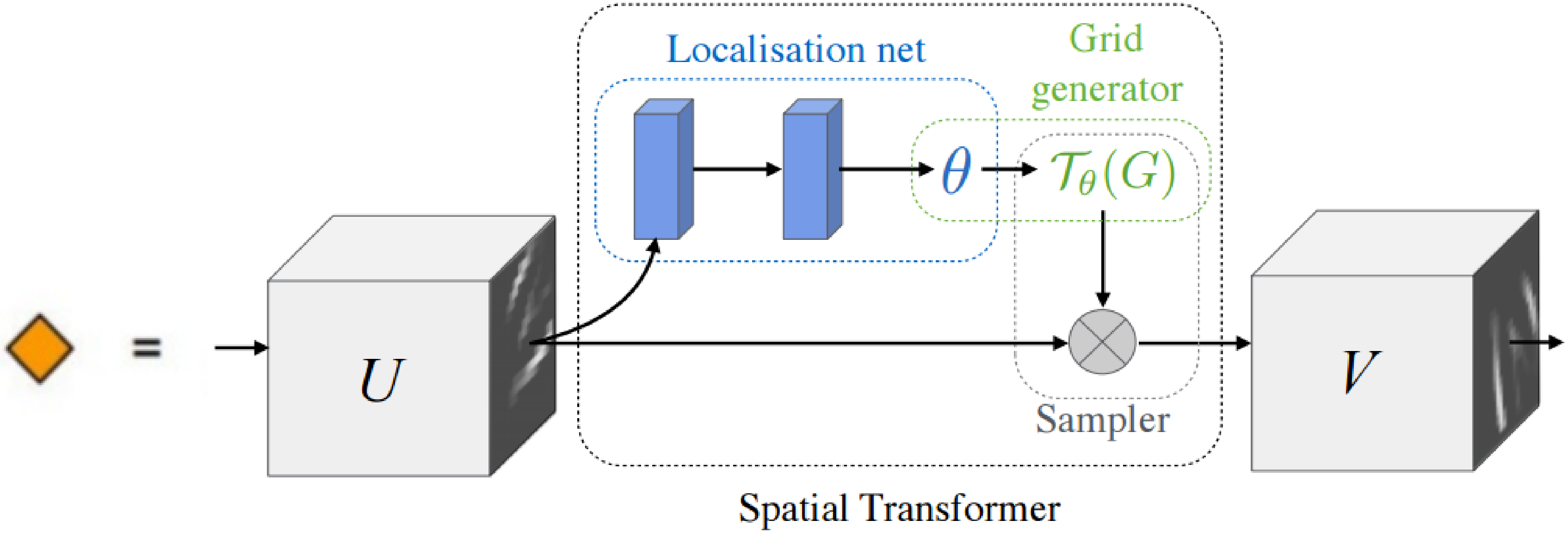
6 parameters to describe the affine transformation

a_{11}^{l}	a_{12}^{l}	a_{13}^{l}
a_{21}^{l}	a_{22}^{l}	a_{23}^{l}
a_{31}^{l}	a_{32}^{l}	a_{33}^{l}

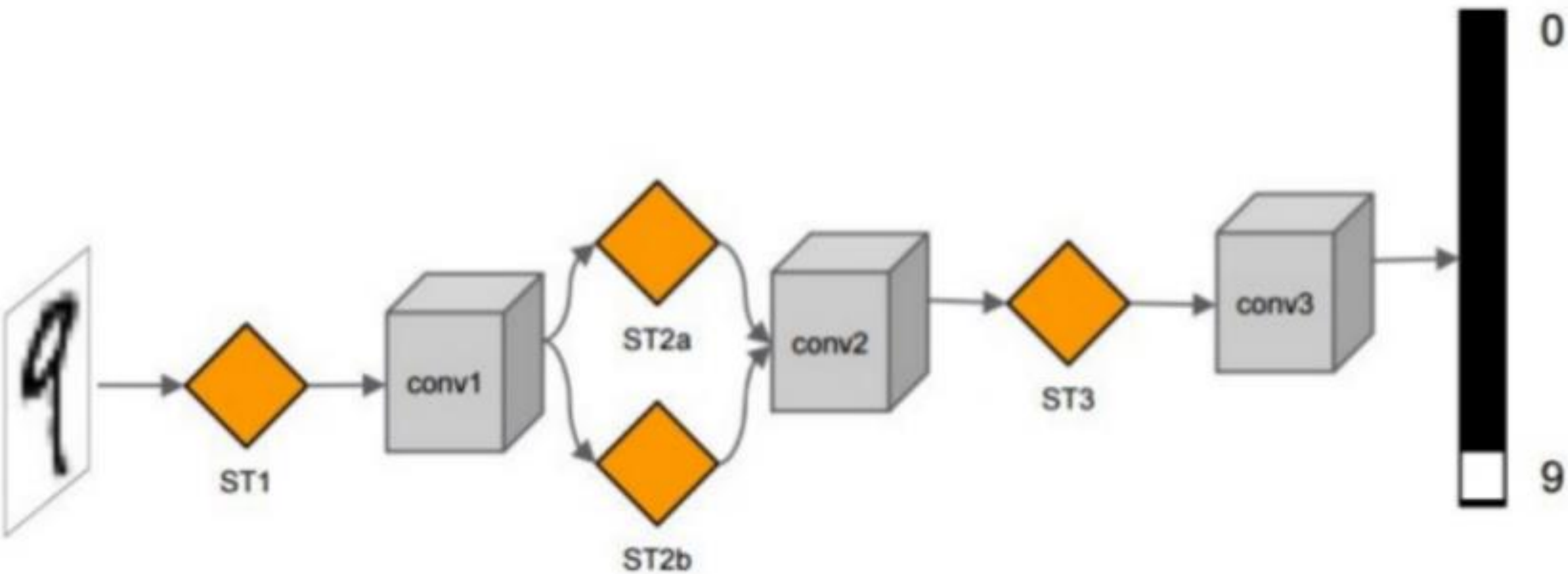
Layer l



$$\begin{aligned} a_{22}^{l} = & (1-0.4)\times(1-0.4)\times a_{22}^{l-1} \\ & +(1-0.6)\times(1-0.4)\times a_{12}^{l-1} \\ & +(1-0.6)\times(1-0.6)\times a_{13}^{l-1} \\ & +(1-0.4)\times(1-0.6)\times a_{23}^{l-1} \end{aligned}$$



Localisation net
Grid generator
θ
$\mathcal{T}_\theta(G)$
U
V
Sampler
Spatial Transformer



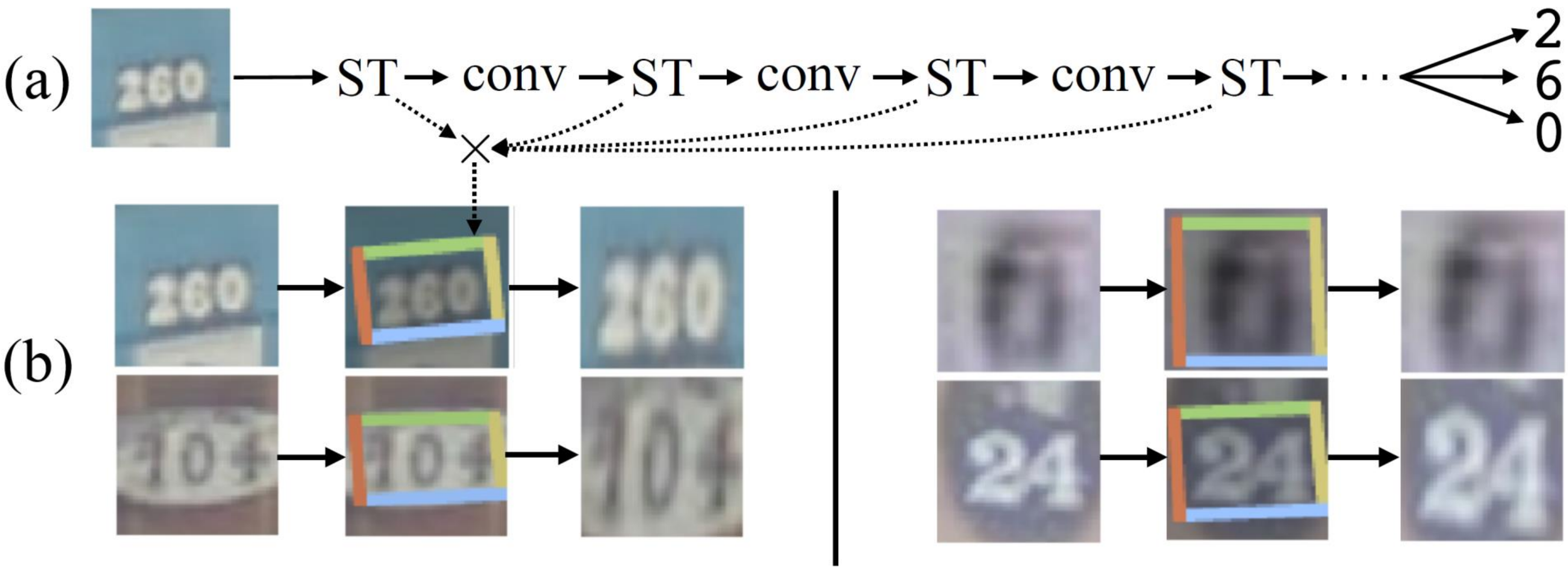
ST1
conv1
ST2a
ST2b
conv2
ST3
conv3
0
9

Street View House Number

Model		Size 64px	128px
Maxout CNN [10]		4.0	-
CNN (ours)		4.0	5.6
DRAM* [1]		3.9	4.5
ST-CNN	Single	3.7	**3.9**
	Multi	**3.6**	**3.9**

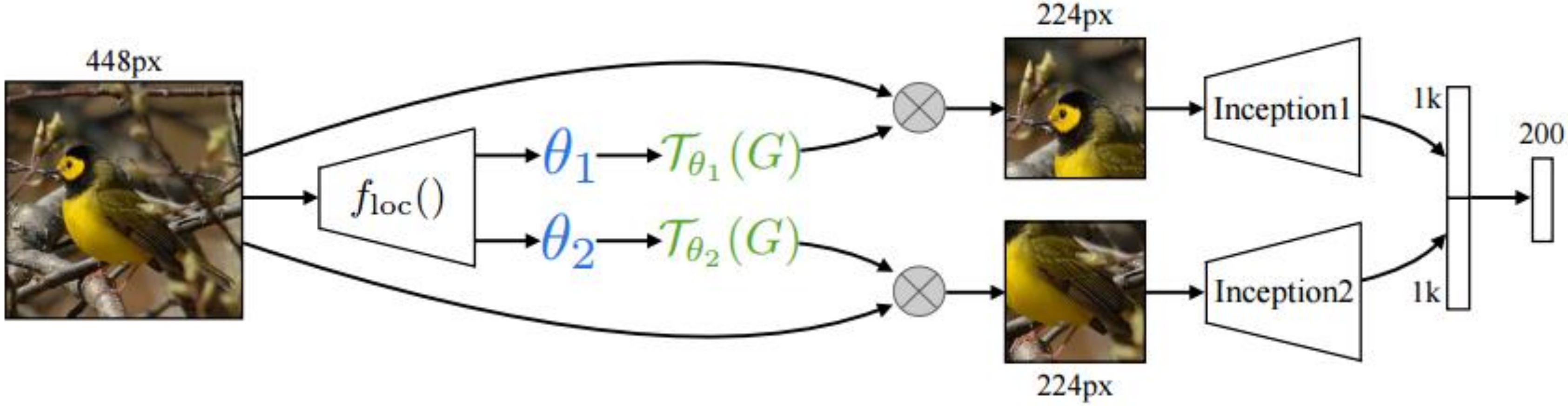
Single: one transformation layer

Multi: many transformation layer

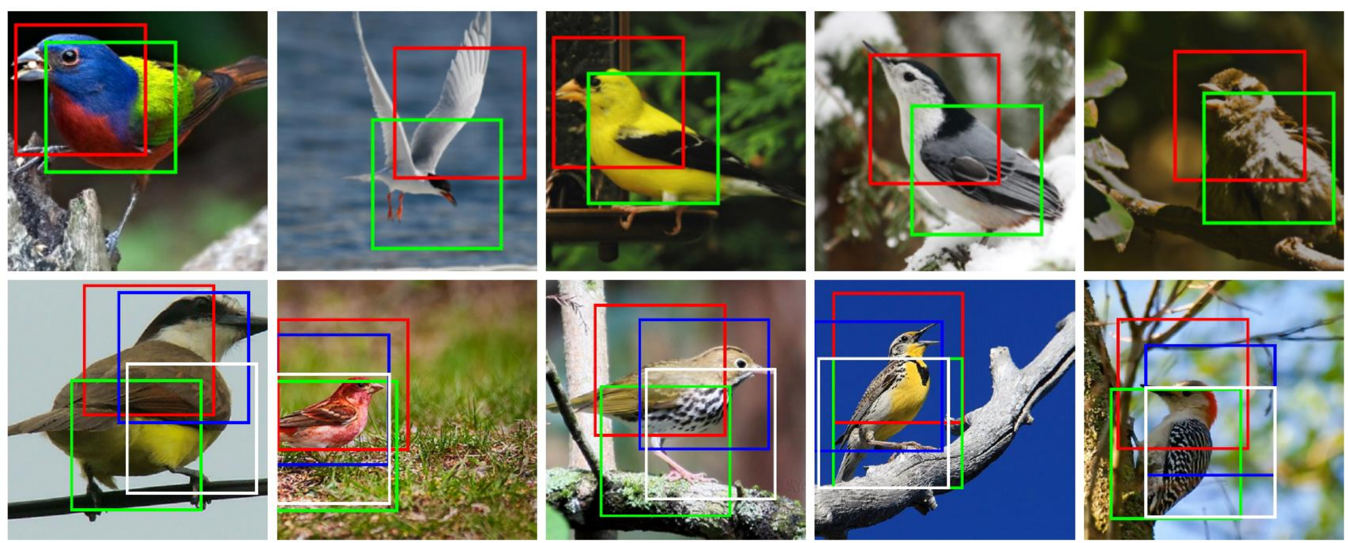


Bird Recognition

$$\begin{bmatrix} a & 0 \\ 0 & d \end{bmatrix} \begin{bmatrix} e \\ f \end{bmatrix}$$

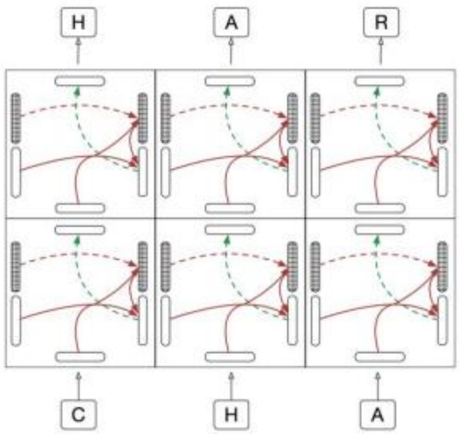


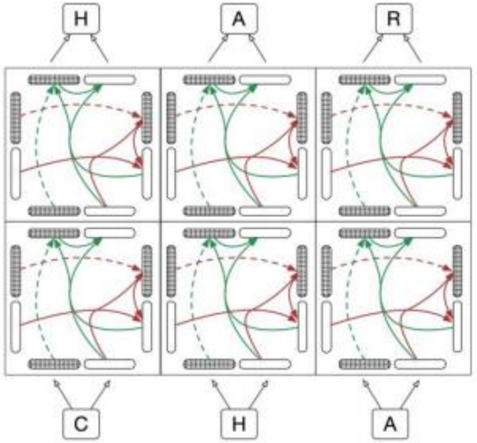
Model		
Cimpoi '15 [4]		66.7
Zhang '14 [30]		74.9
Branson '14 [2]		75.7
Lin '15 [20]		80.9
Simon '15 [24]		81.0
CNN (ours)	224px	82.3
2×ST-CNN	224px	83.1
2×ST-CNN	448px	83.9
4×ST-CNN	448px	**84.1**



Outline

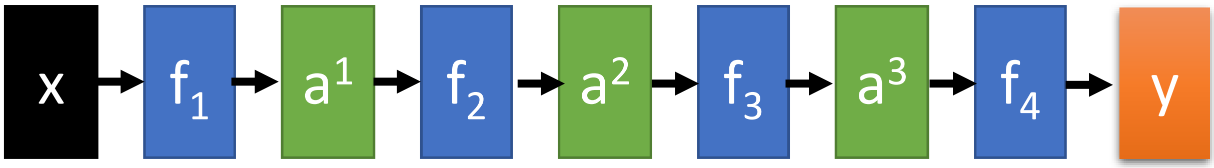
- Convolutional Neural Network (Review)
- Spatial Transformer
- Highway Network & Grid LSTM
- Pointer Network
- External Memory





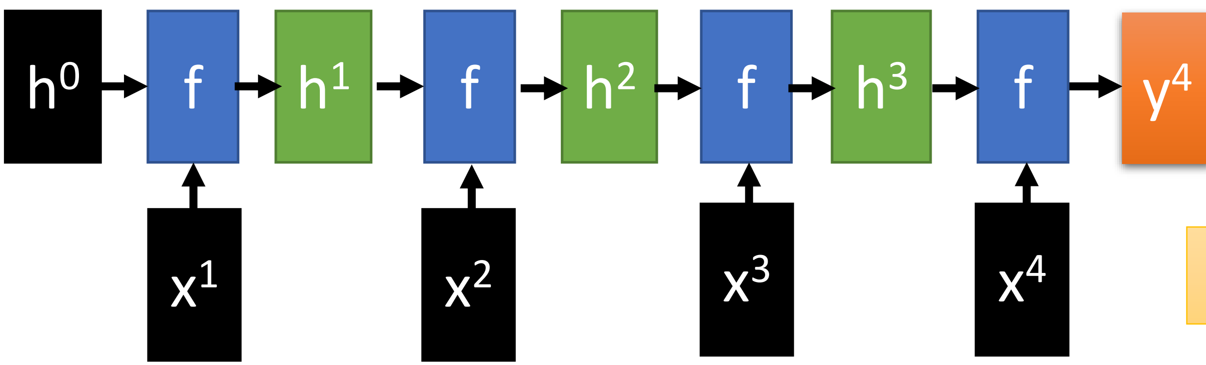
Feedforward v.s. Recurrent

1. Feedforward network does not have input at each step
2. Feedforward network has different parameters for each layer



$$a^t = f_l(a^{t-1}) = \sigma(W^t a^{t-1} + b^t)$$

t is layer



t is time step

$$h^t = f(h^{t-1}, x^t) = \sigma\left(W^h h^{t-1} + W^i x^t + b^i\right)$$

Applying gated structure in feedforward network

GRU → Highway Network

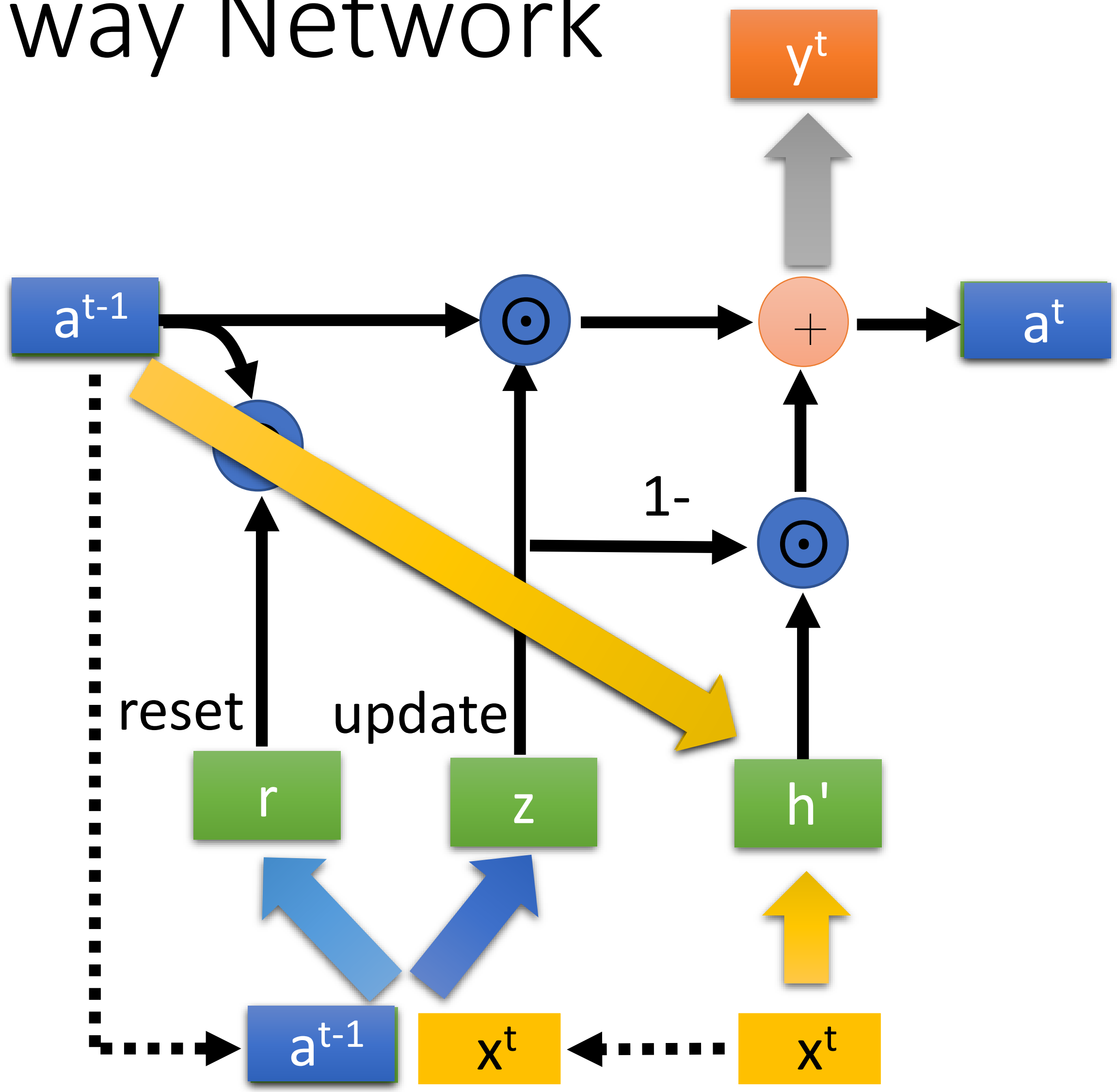
No input x^t at each step

No output y^t at each step

a^{t-1} is the output of the (t-1)-th layer

a^t is the output of the t-th layer

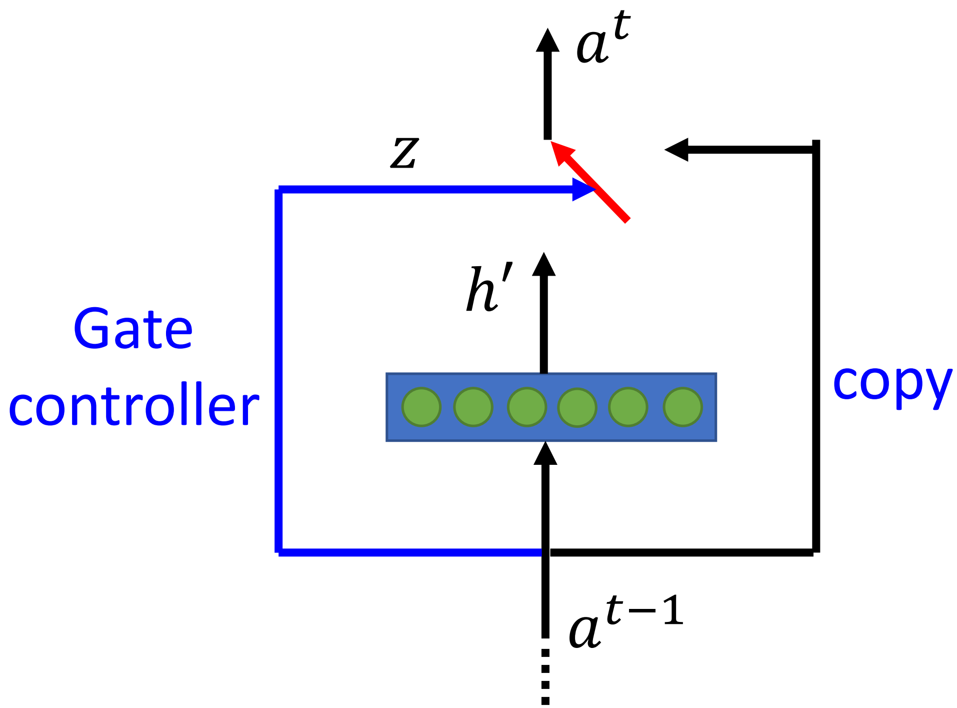
No reset gate



Highway Network

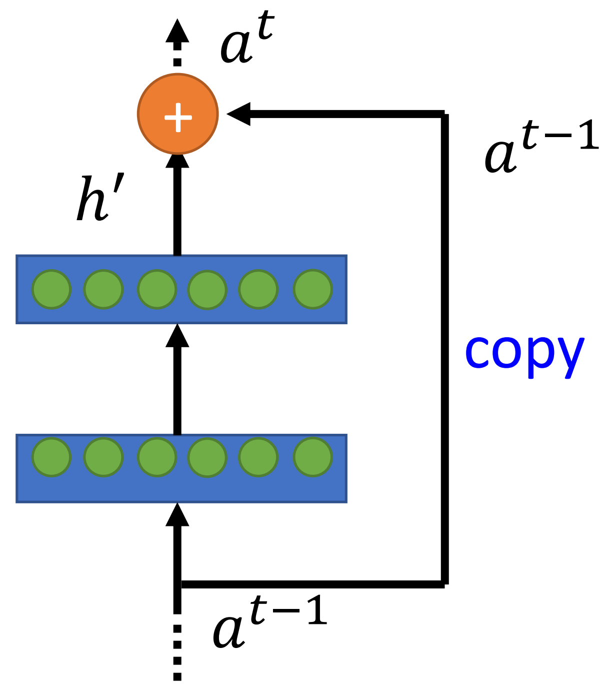
$$h' = \sigma(W a^{t-1})$$
$$z = \sigma(W' a^{t-1})$$
$$a^t = z \odot a^{t-1} + (1-z) \odot h$$

- **Highway Network**

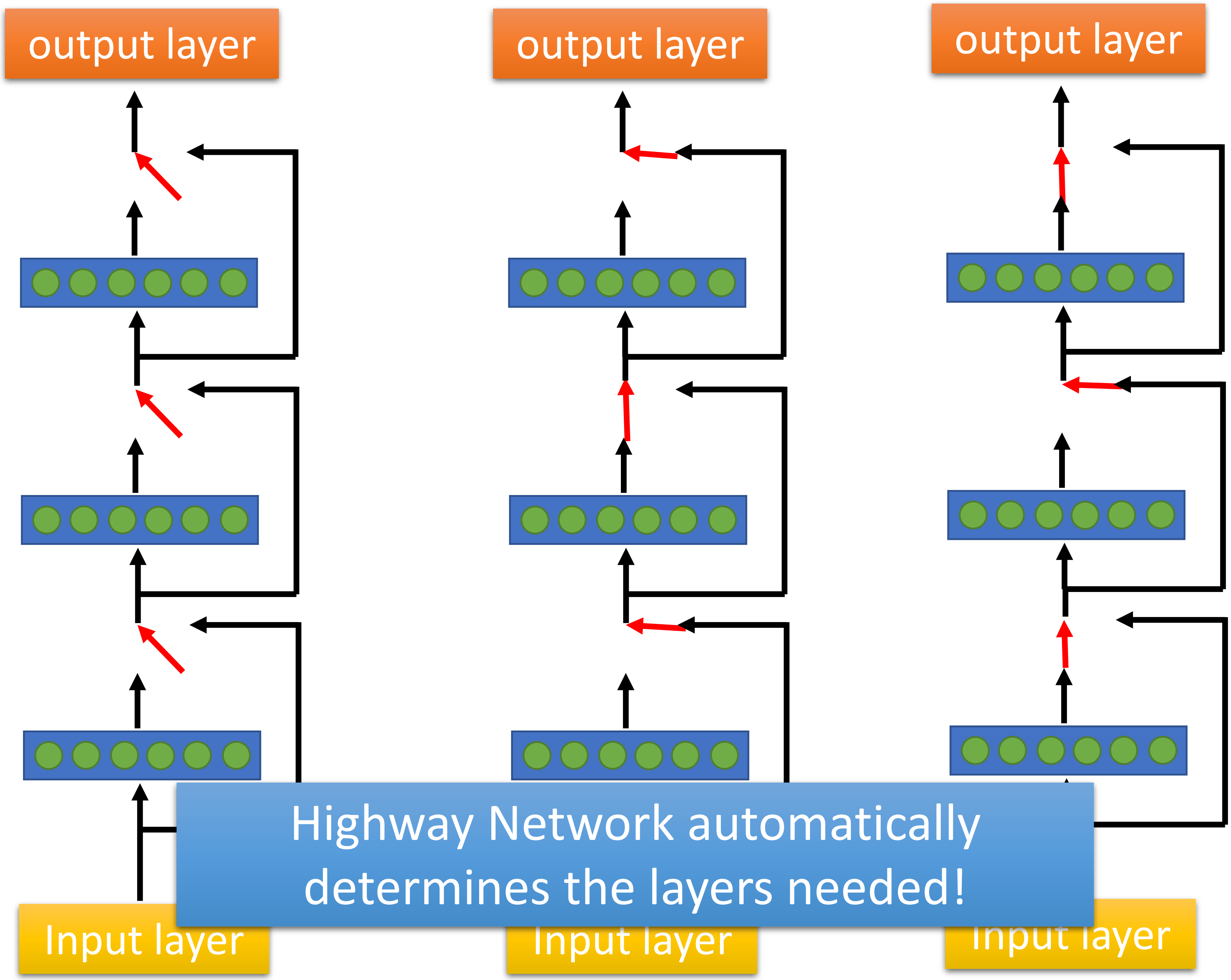


Training Very Deep Networks https://arxiv.org/pdf/1507.06228v2.pdf

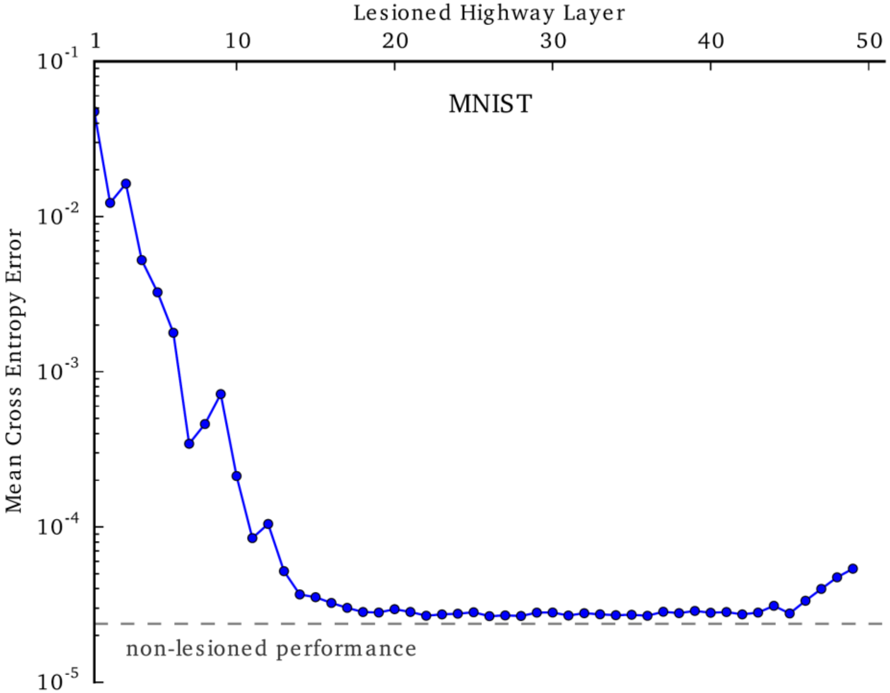
- **Residual Network**

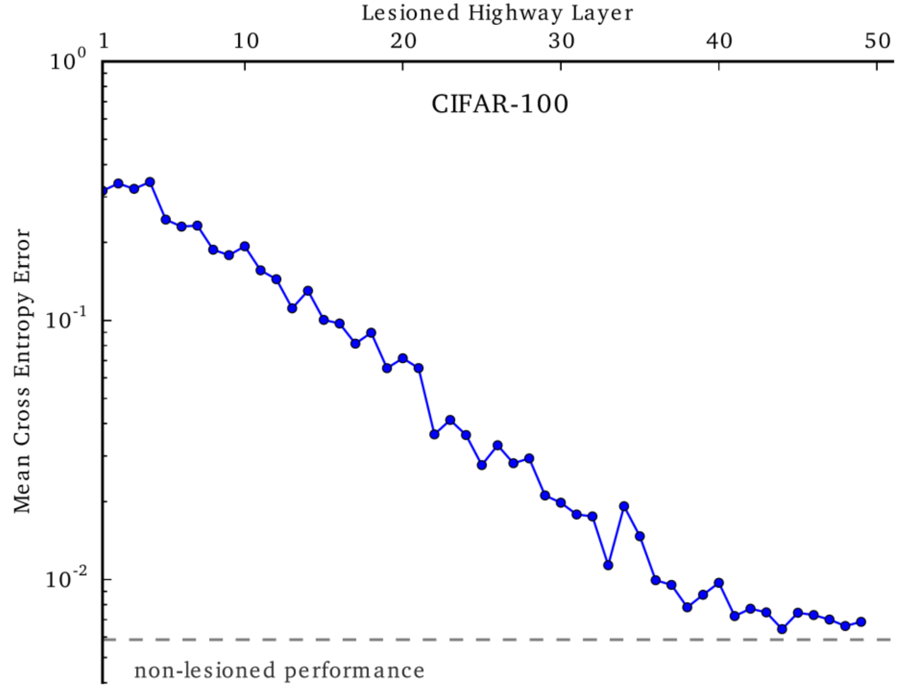


Deep Residual Learning for Image Recognition http://arxiv.org/abs/1512.03385

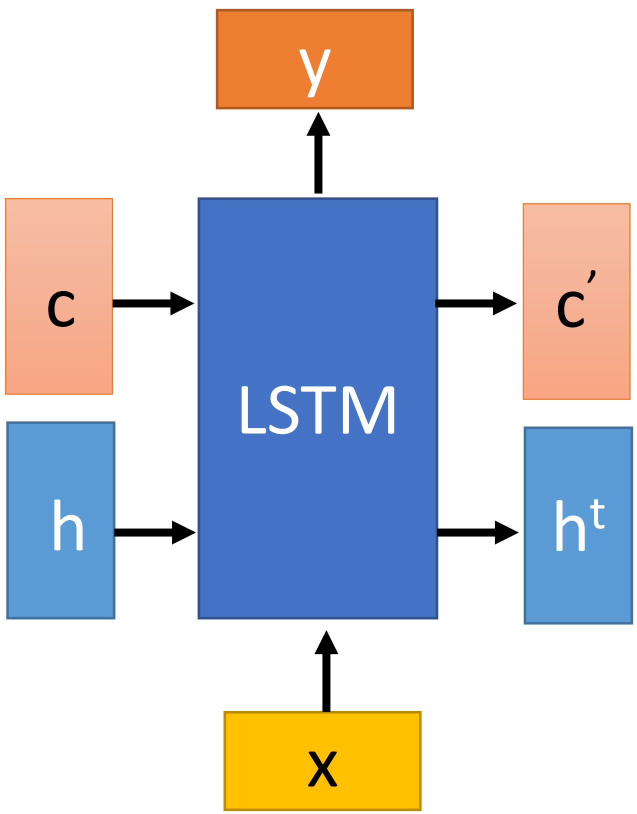

output layer
output layer
output layer
Highway Network automatically
determines the layers needed!
Input layer
Input layer
Input layer

Highway Network

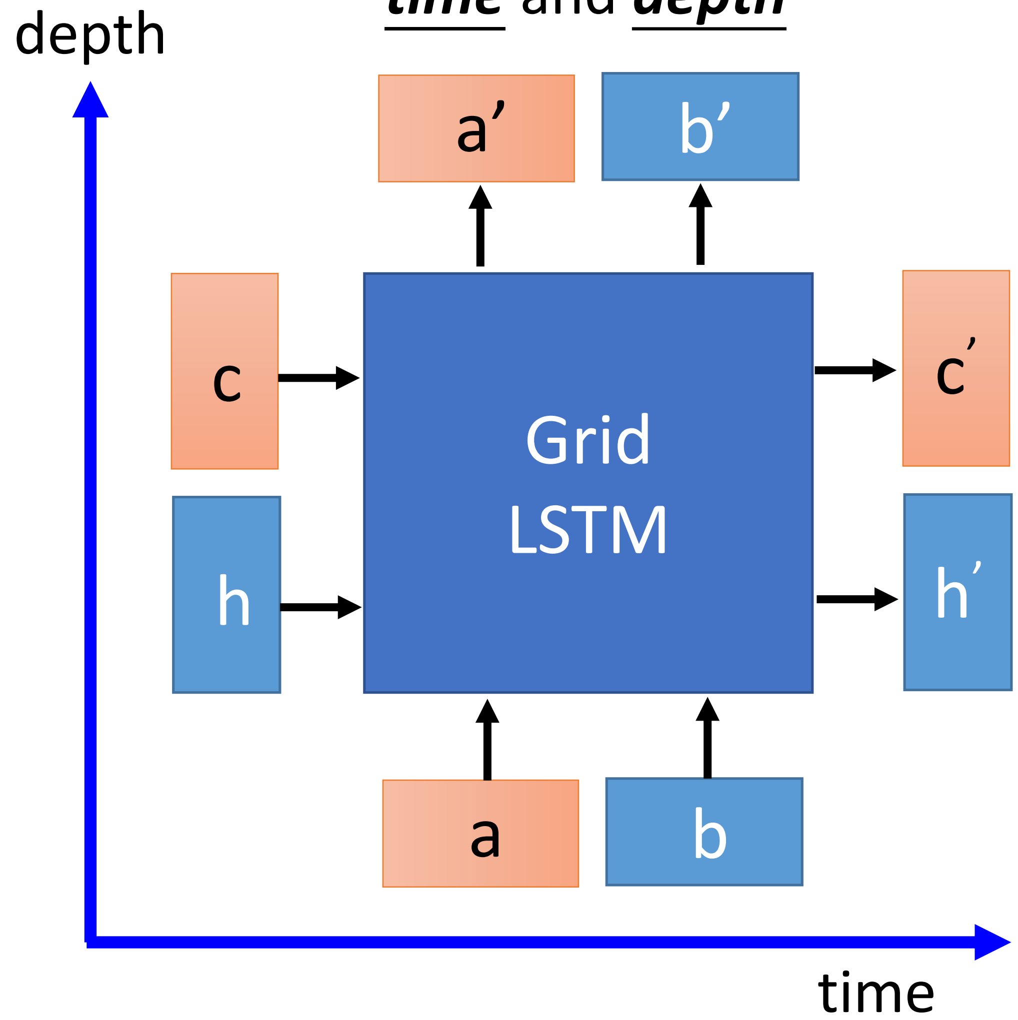


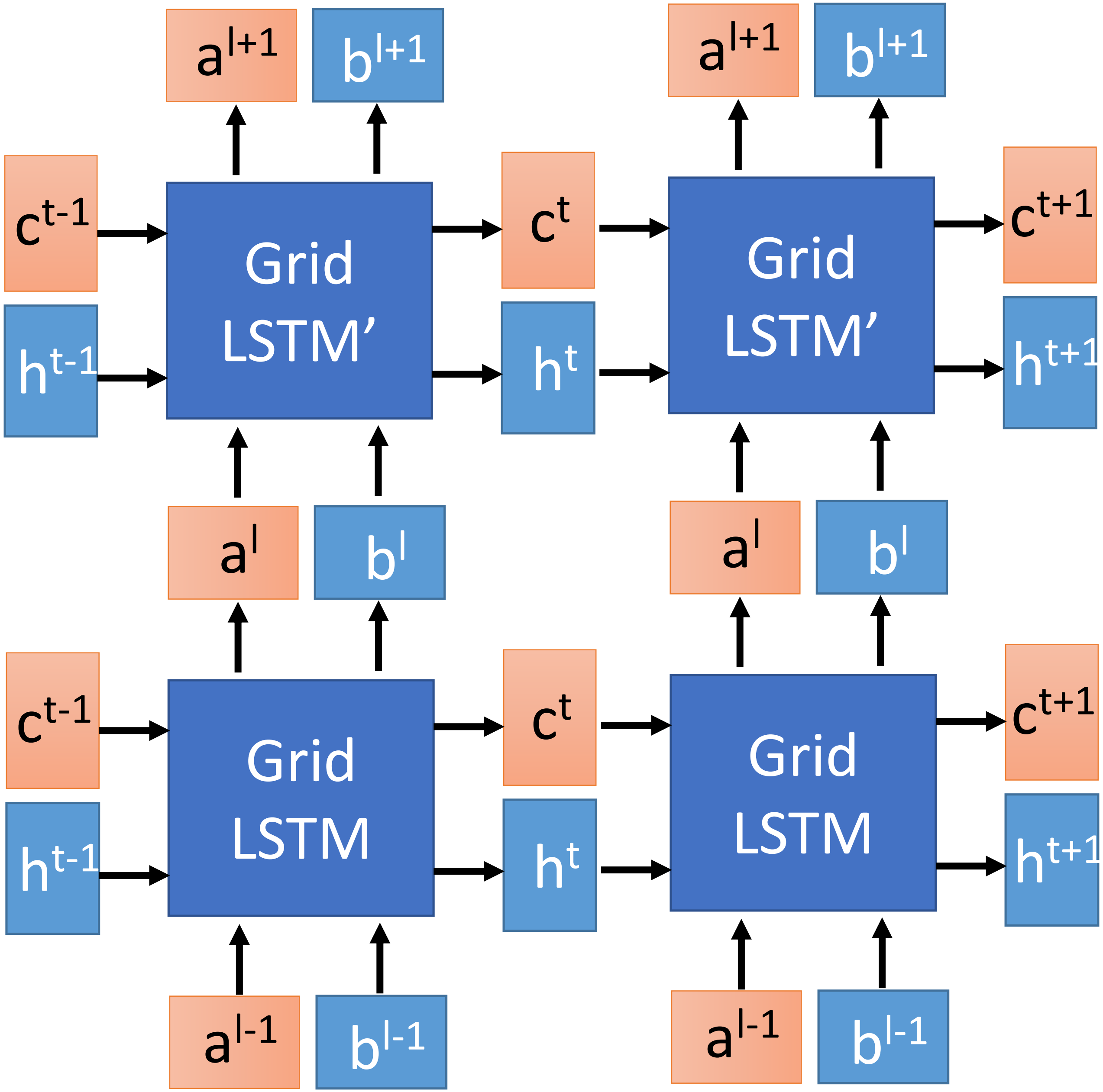


Grid LSTM



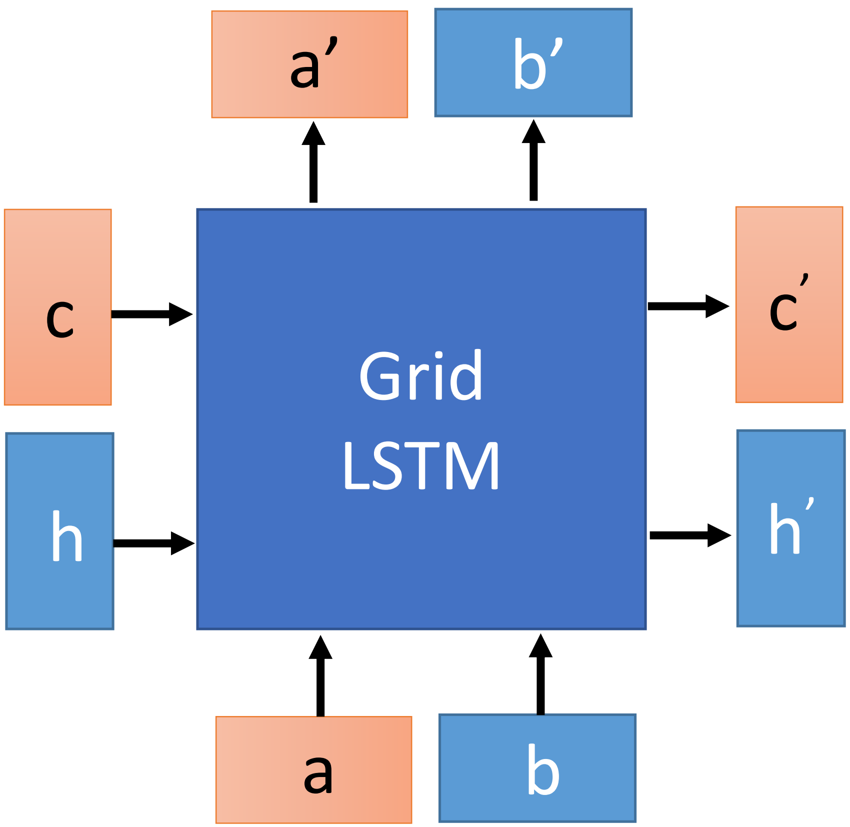
Memory for both
time and ***depth***

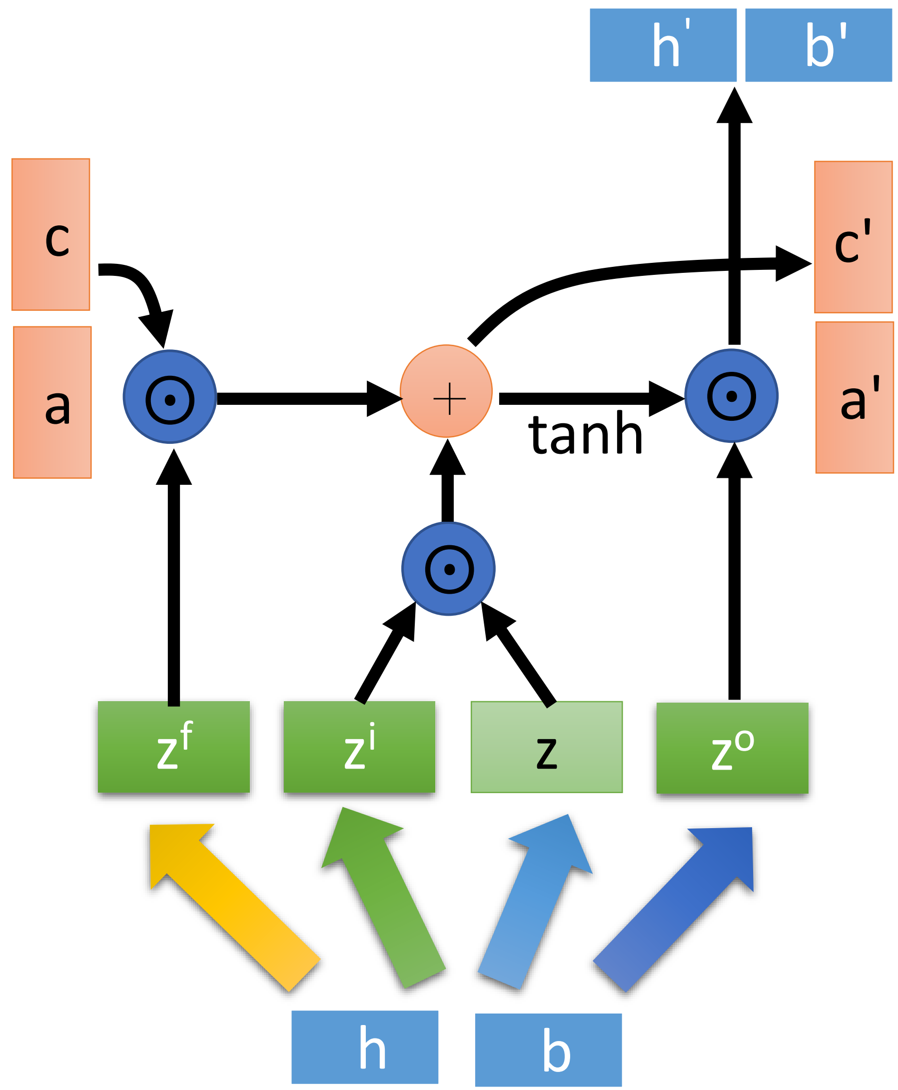




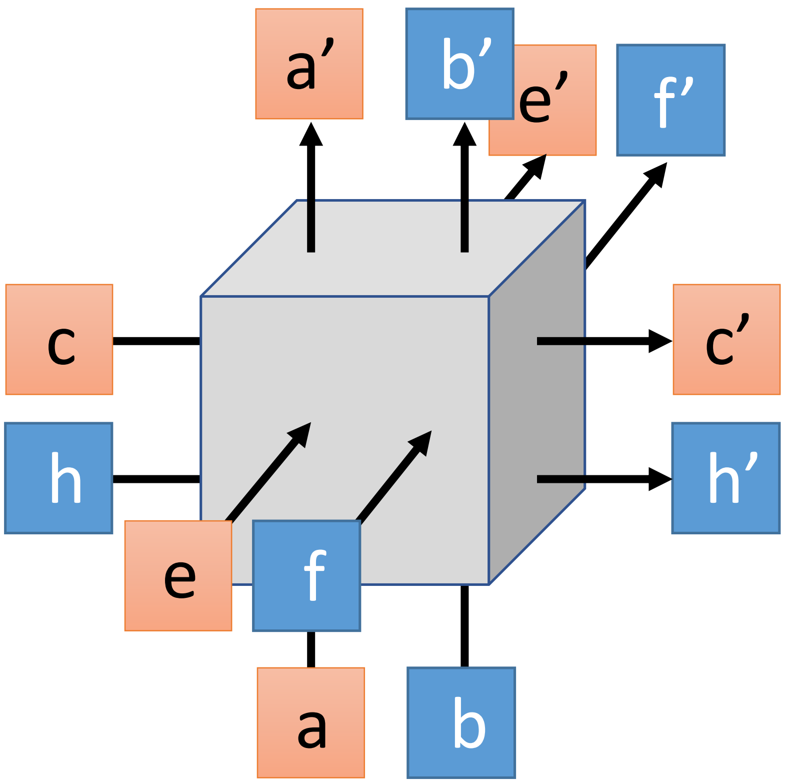
a^{l+1}
b^{l+1}
a^{l+1}
b^{l+1}
c^{t-1}
Grid LSTM'
c^{t}
Grid LSTM'
c^{t+1}
h^{t-1}
h^{t}
h^{t+1}
a^{l}
b^{l}
a^{l}
b^{l}
c^{t-1}
Grid LSTM
c^{t}
Grid LSTM
c^{t+1}
h^{t-1}
h^{t}
h^{t+1}
a^{l-1}
b^{l-1}
a^{l-1}
b^{l-1}

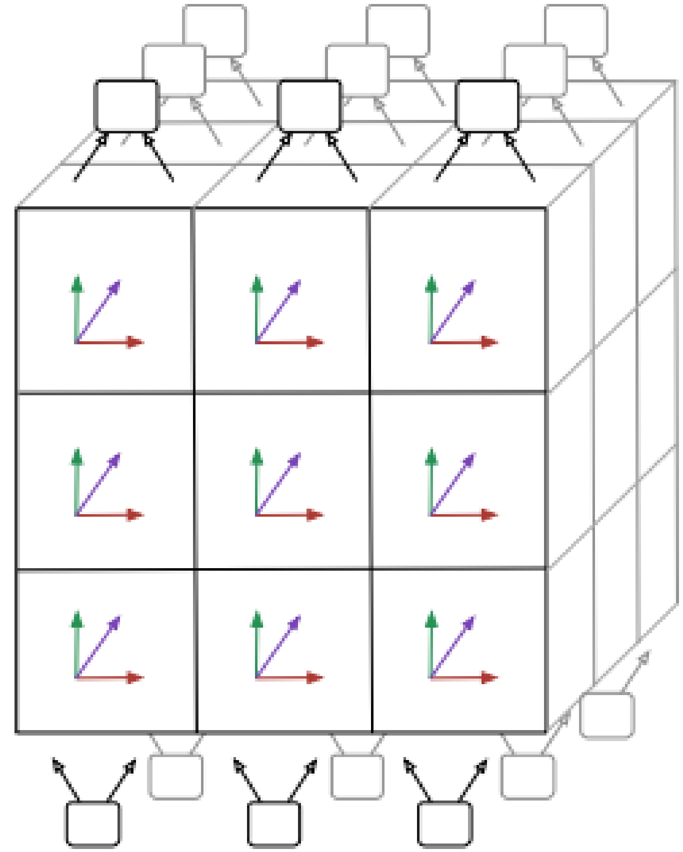
Grid LSTM





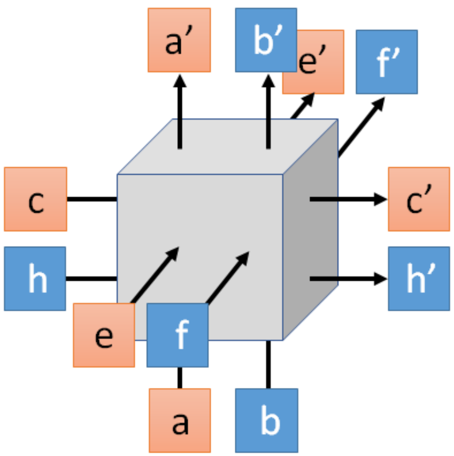
3D Grid LSTM

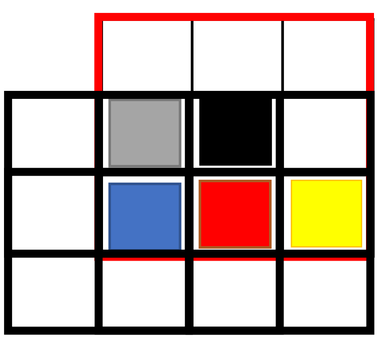


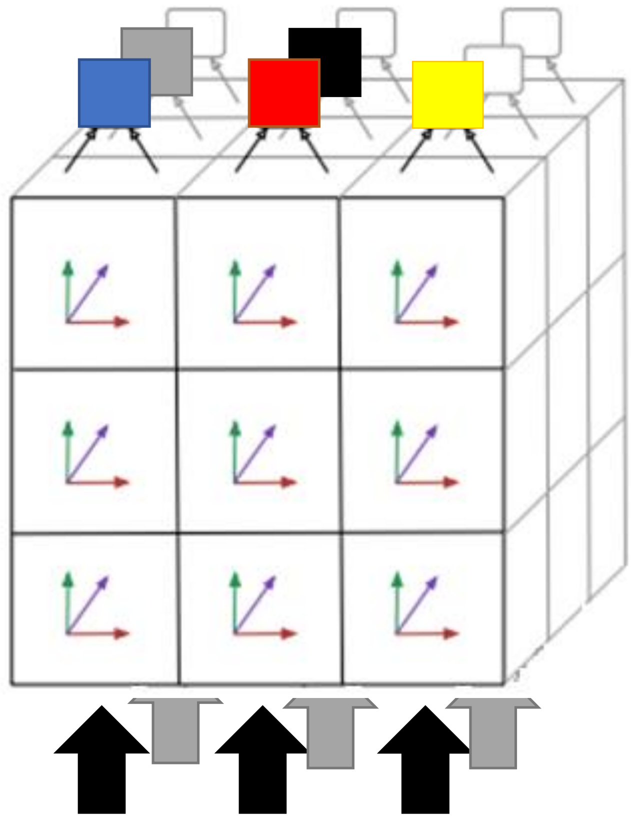


3D Grid LSTM

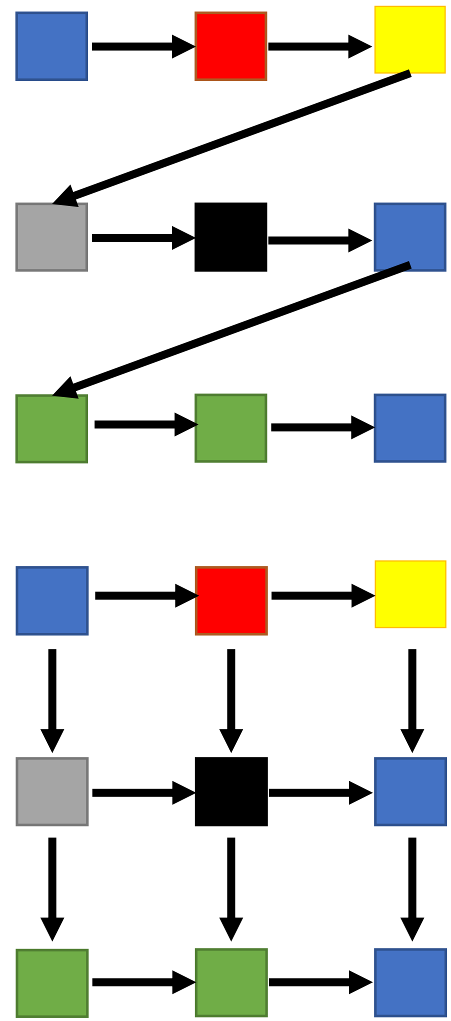
- Images are composed of pixels



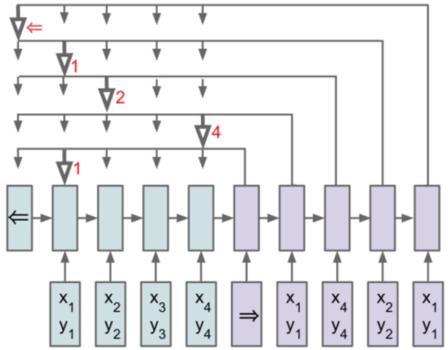




3 x 3 images

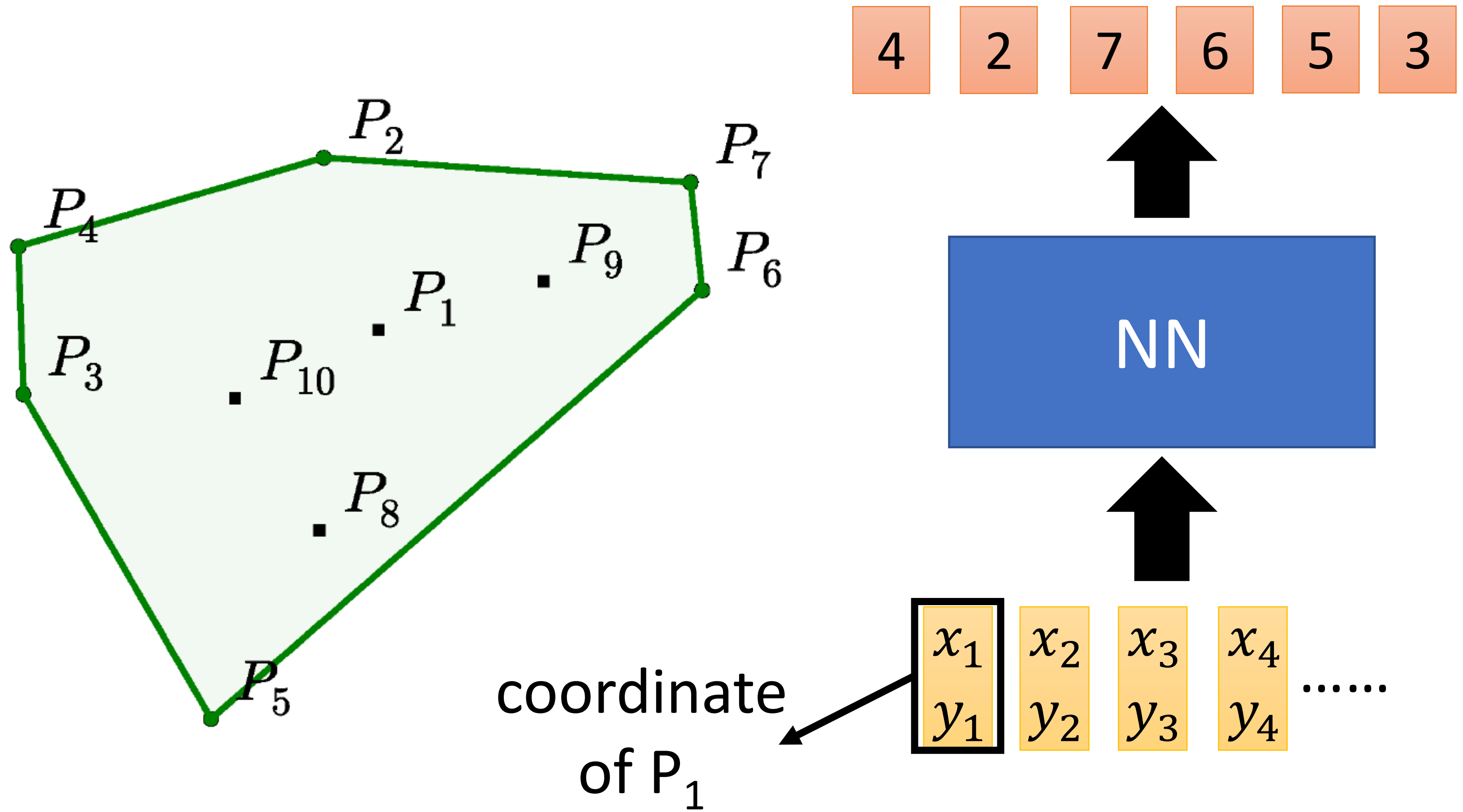


Outline



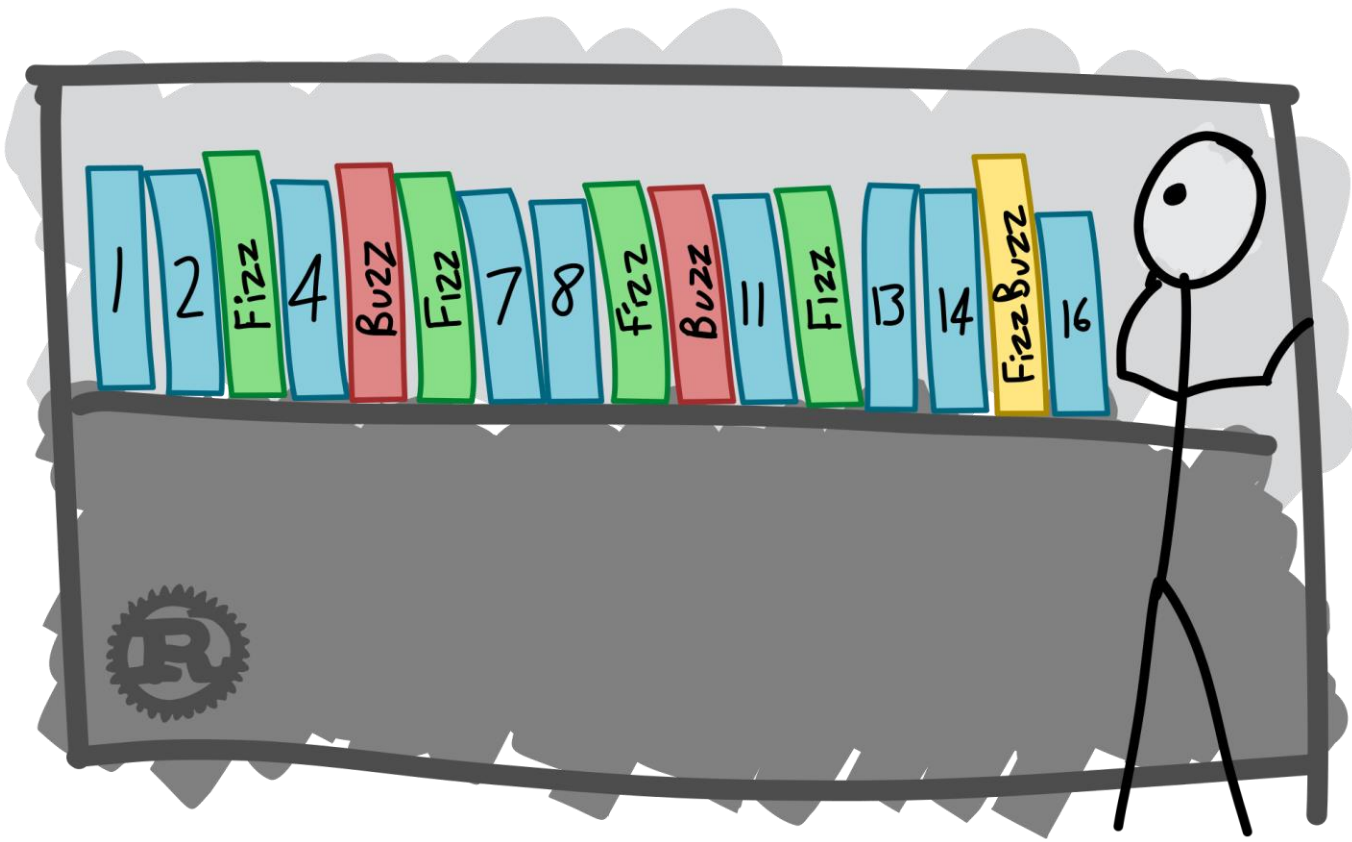
- Convolutional Neural Network (Review)
- Spatial Transformer
- Highway Network & Grid LSTM
- Pointer Network
- External Memory

Pointer Network



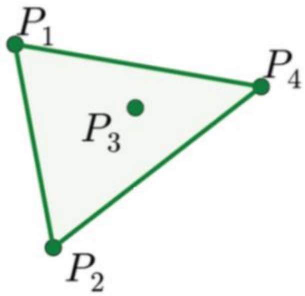
“硬train” 的故事

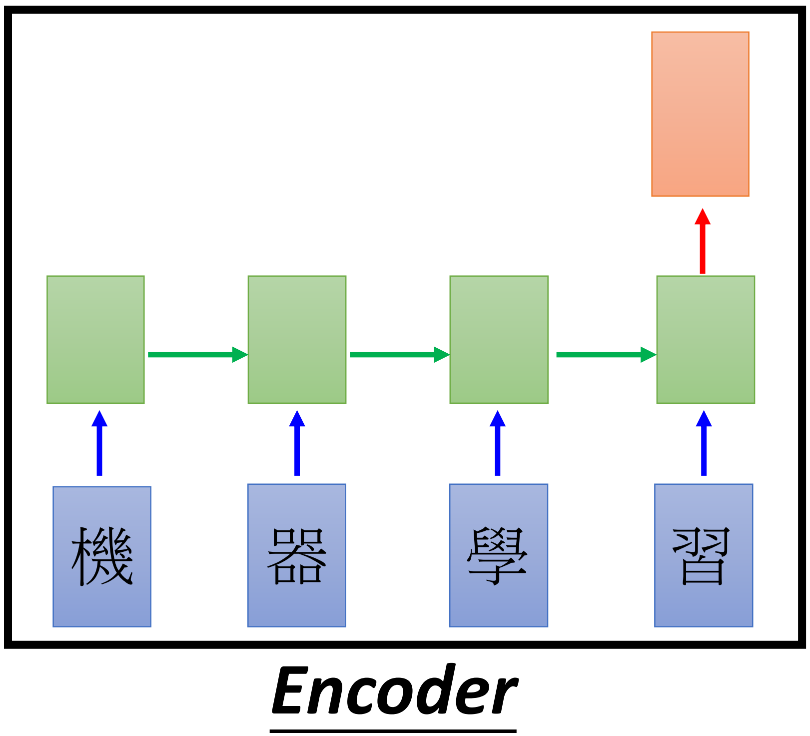
- Fizz Buzz in Tensorflow: http://joelgrus.com/2016/05/23/fizz-buzz-in-tensorflow/

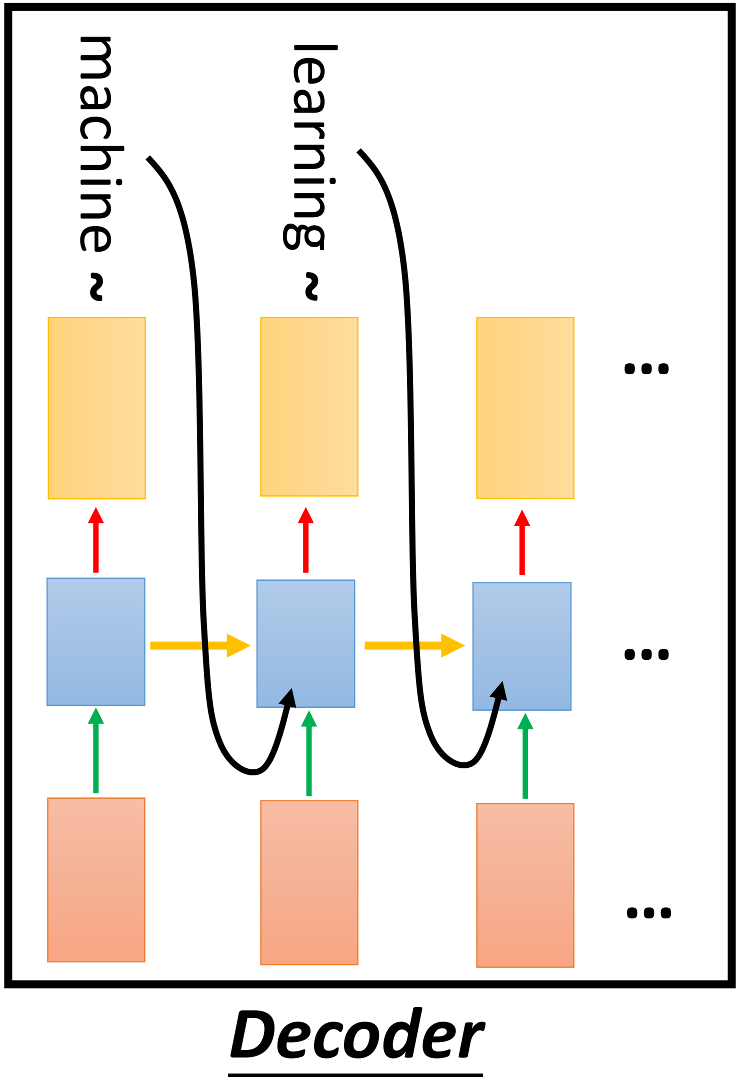


https://ochronus.com/fizzbuzz-in-css/

Sequence-to-sequence?

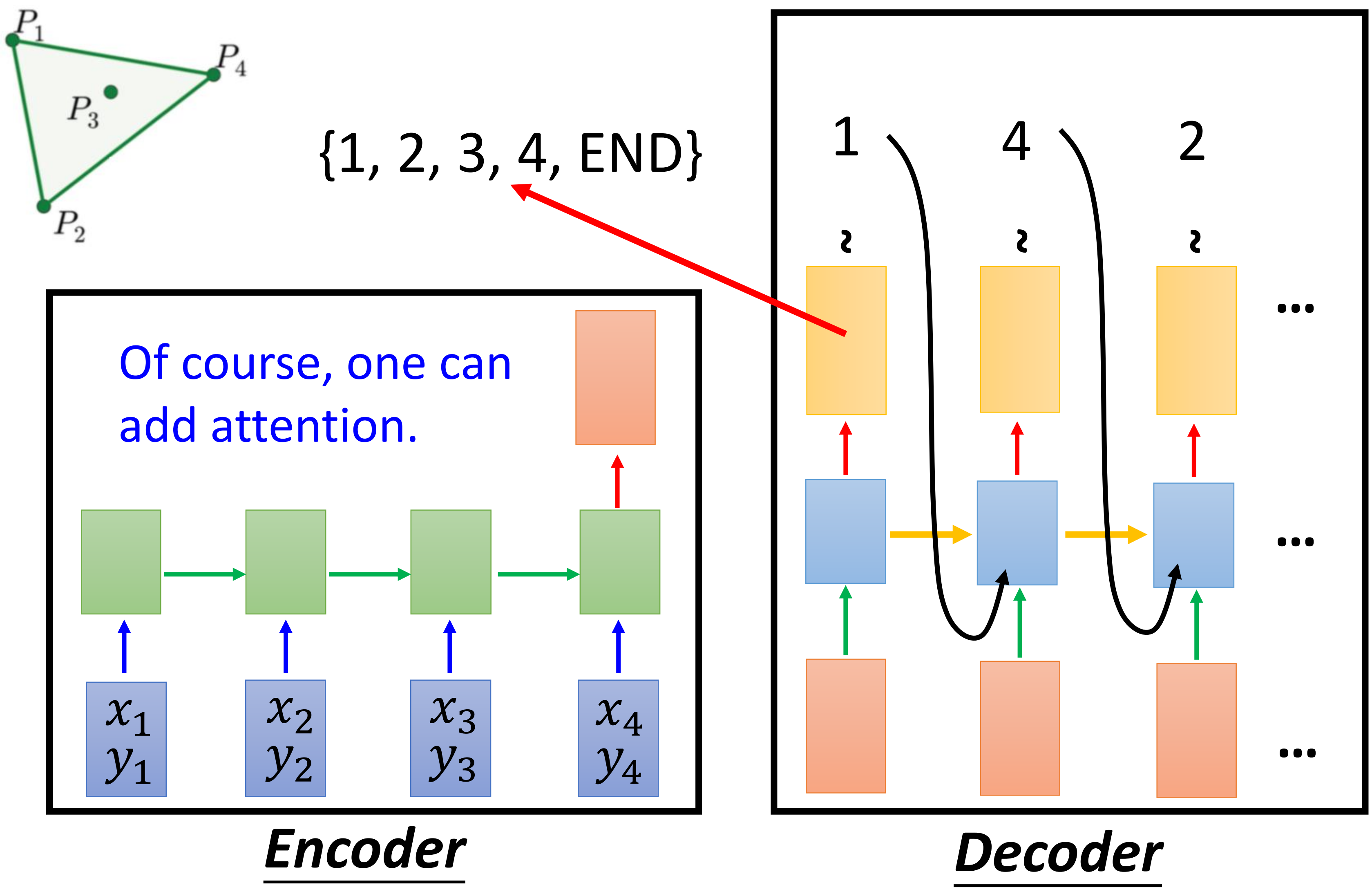






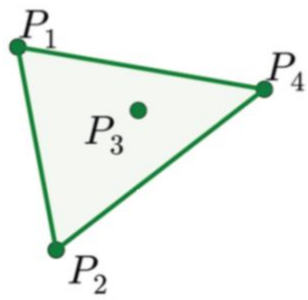
Problem?

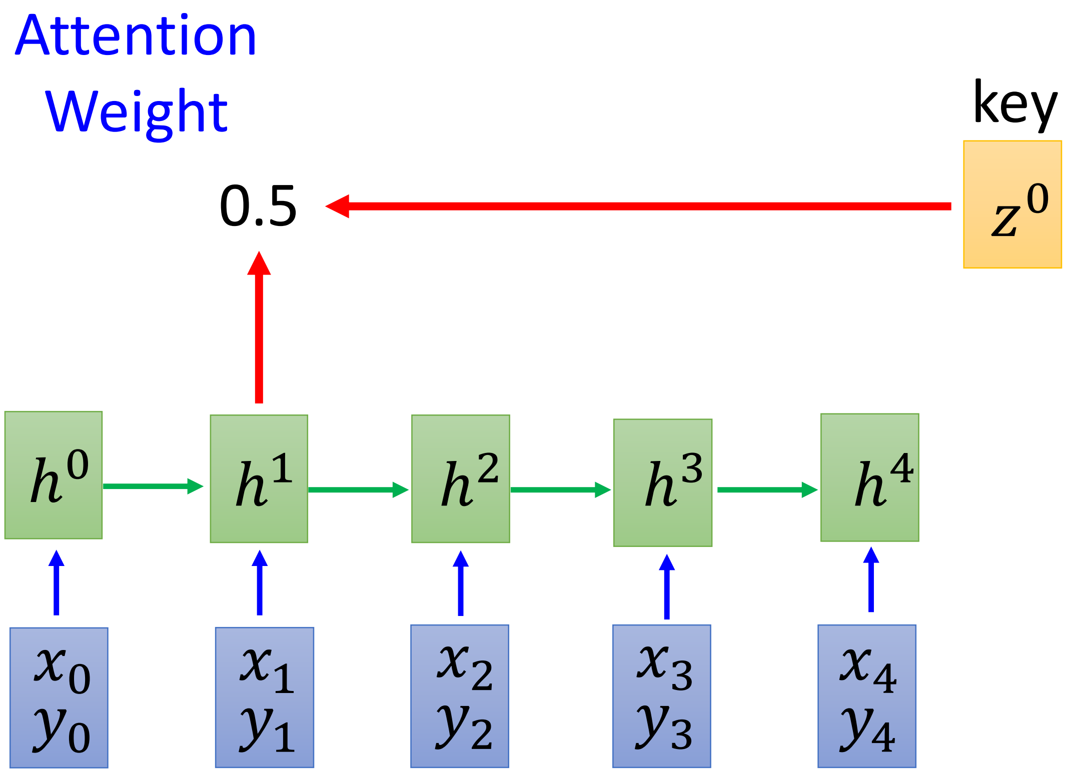
Sequence-to-sequence?



Pointer Network

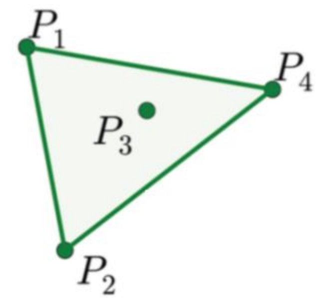
$\begin{matrix} x_0 \\ y_0 \end{matrix}$: END





Pointer Network

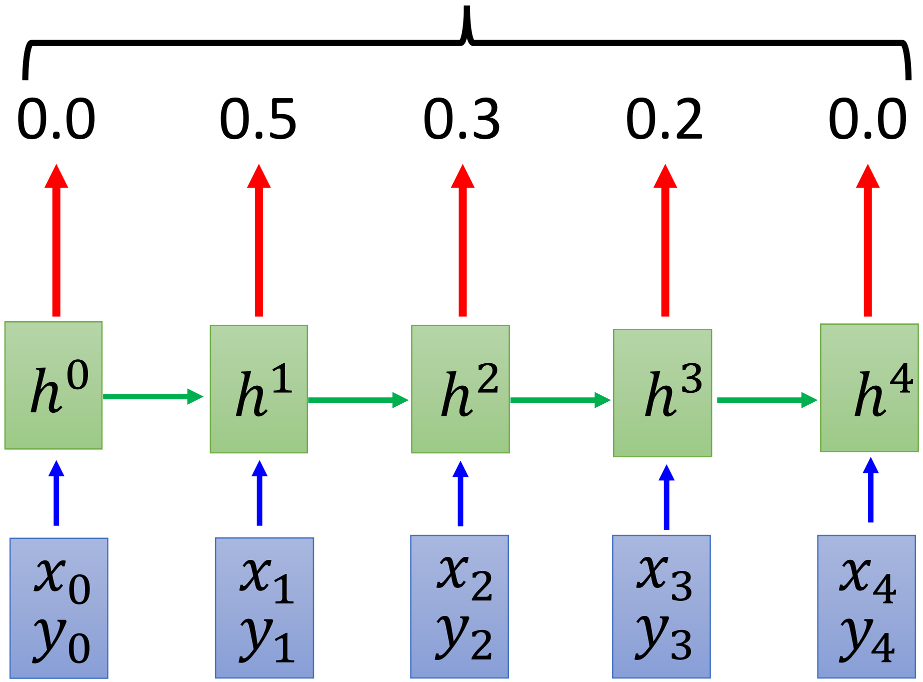
x_0 y_0 : END

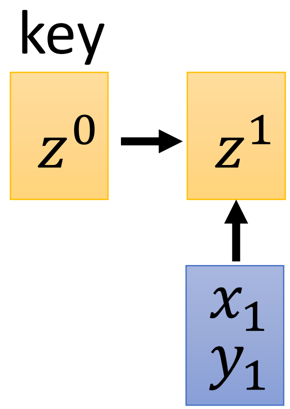


What decoder can output depends on the input.

Output: 1

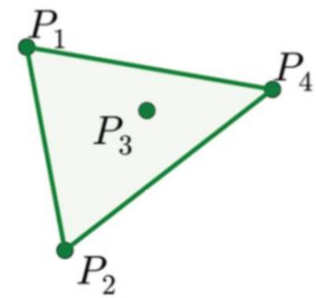
argmax from this distribution





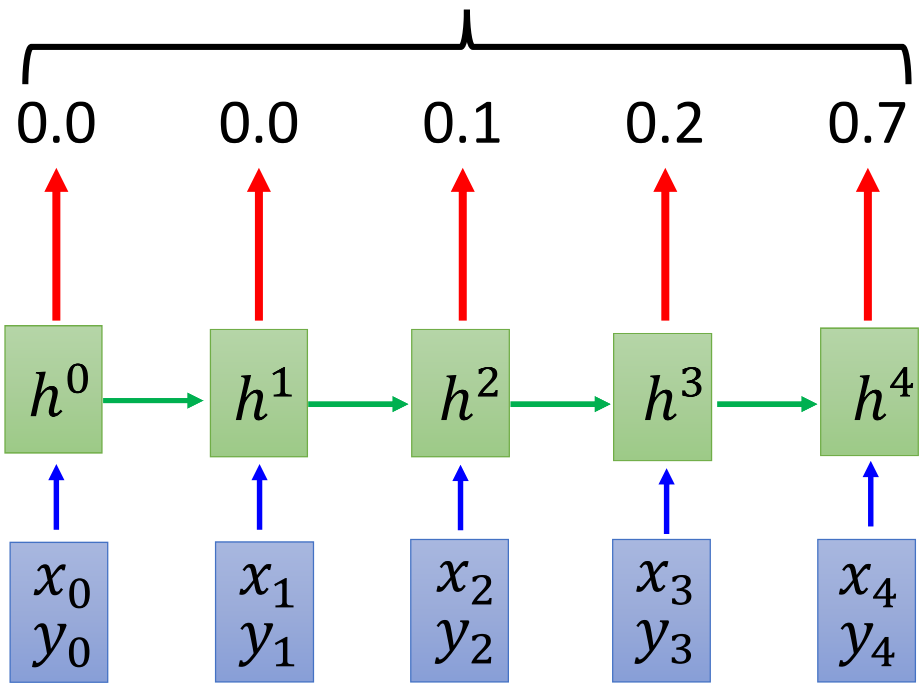
Pointer Network

$\begin{matrix} x_0 \\ y_0 \end{matrix}$: END

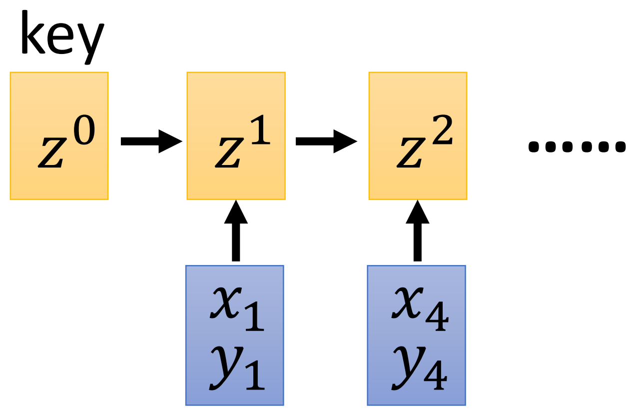


Output: 4

argmax from this distribution



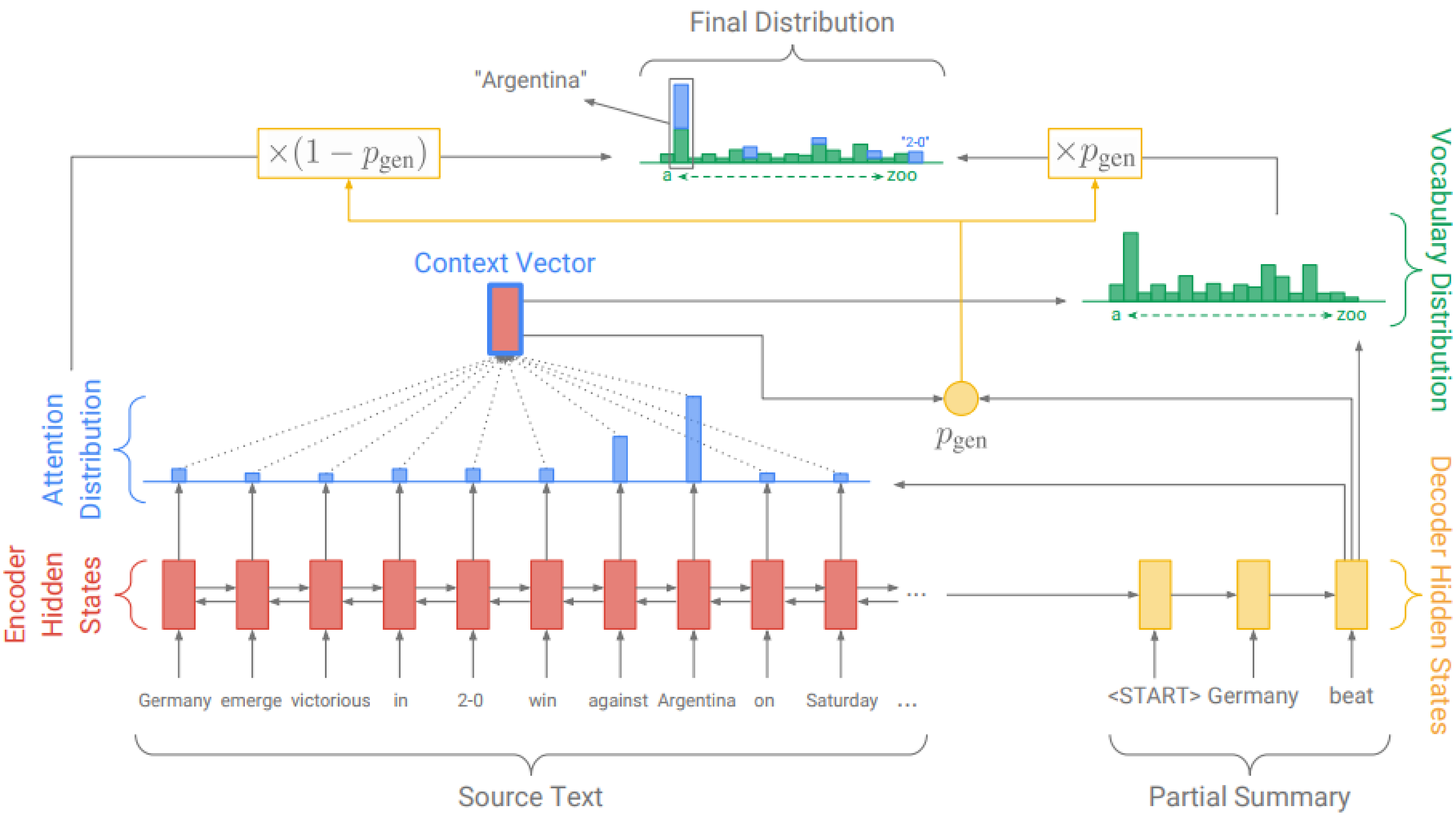
What decoder can output depends on the input.



The process stops when "END" has the largest attention weights.

https://arxiv.org/abs/1704.04368

Applications - Summarization



More Applications

Machine Translation

French: Guillaume et Cesar ont une voiture bleue a Lausanne.

Copy Copy Copy

English: Guillaume and Cesar have a blue car in Lausanne.

Chat-bot

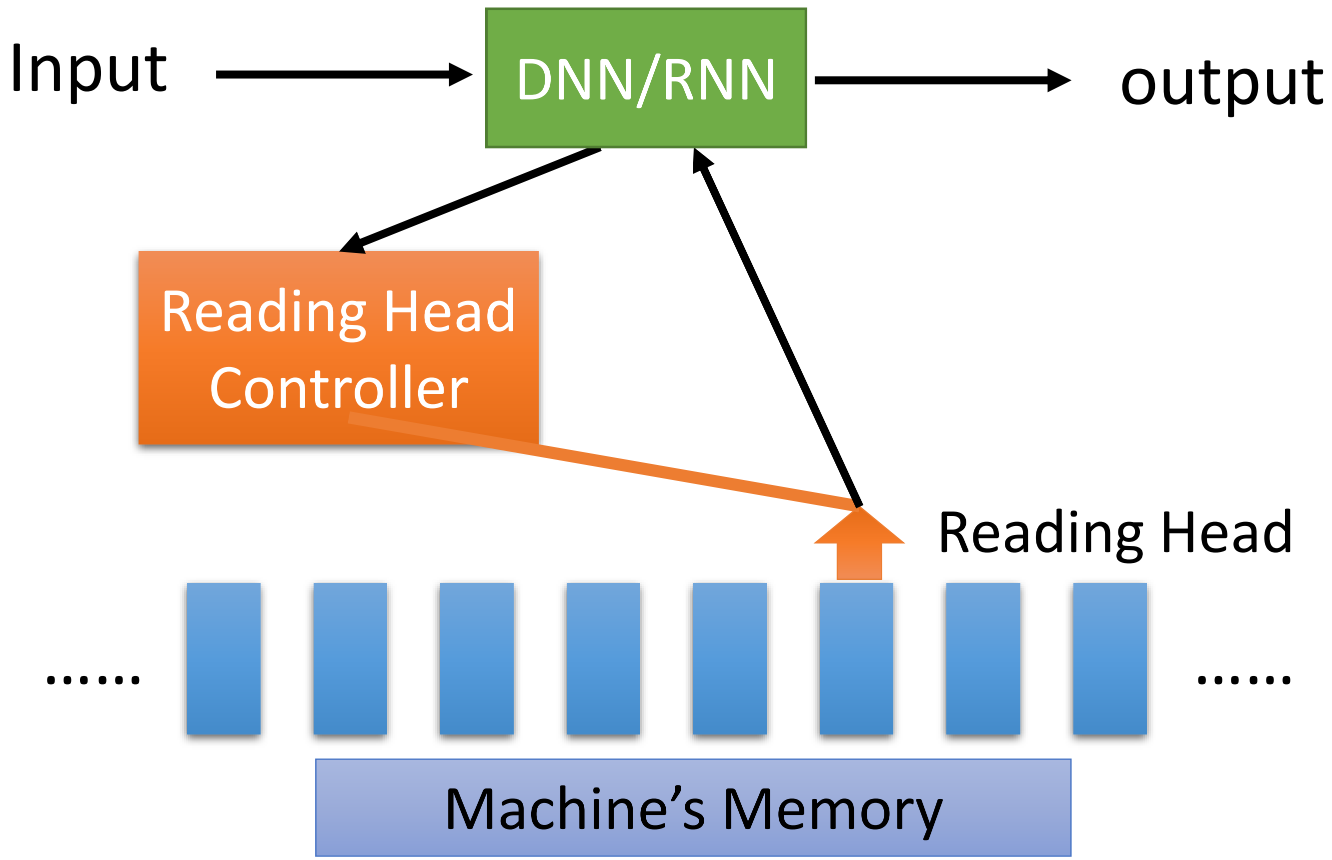
User: X寶你好，我是庫洛洛

Machine:庫洛洛你好，很高興認識你

Outline

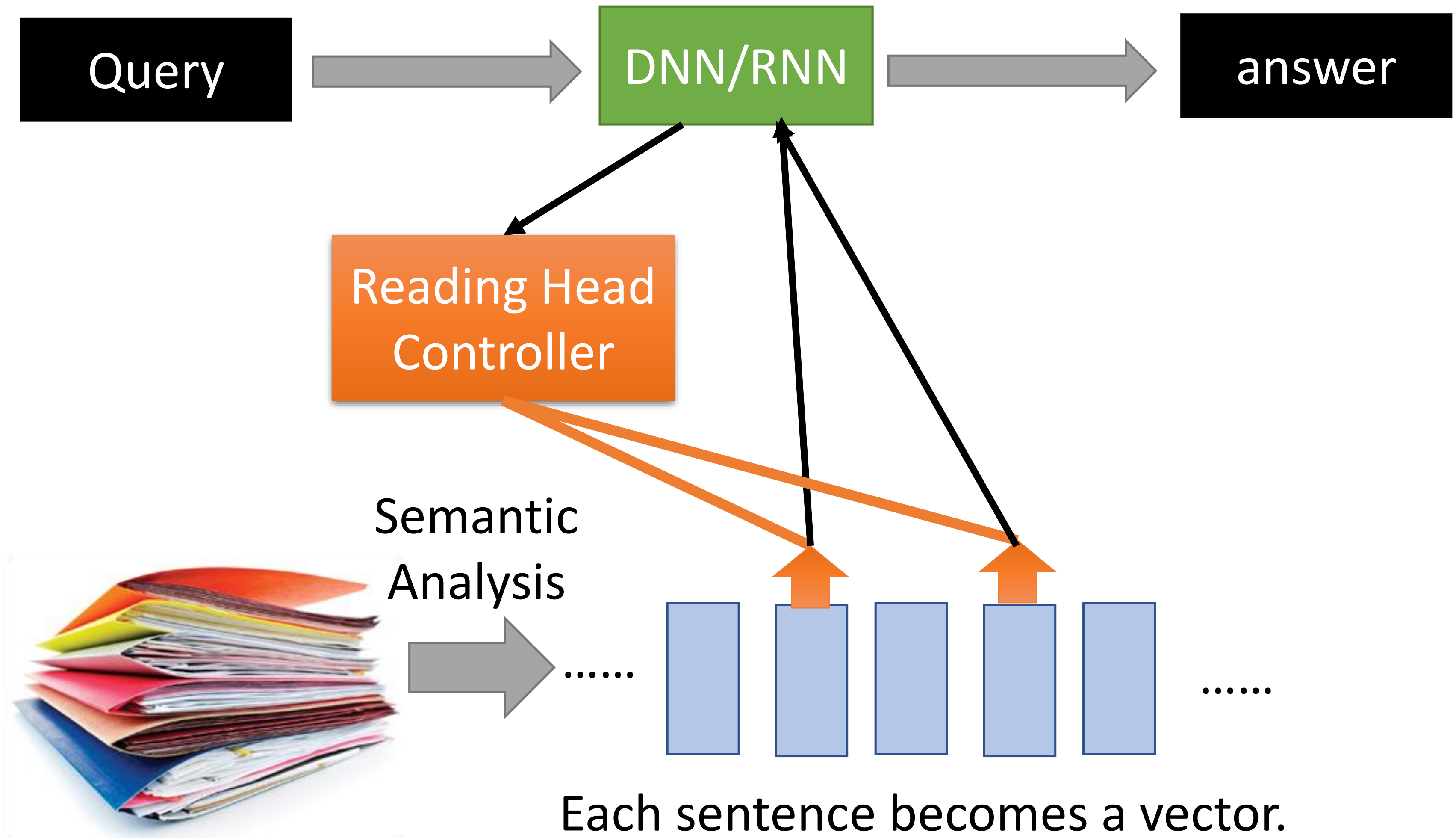
- Convolutional Neural Network (Review)
- Spatial Transformer
- Highway Network & Grid LSTM
- Pointer Network
- External Memory

External Memory

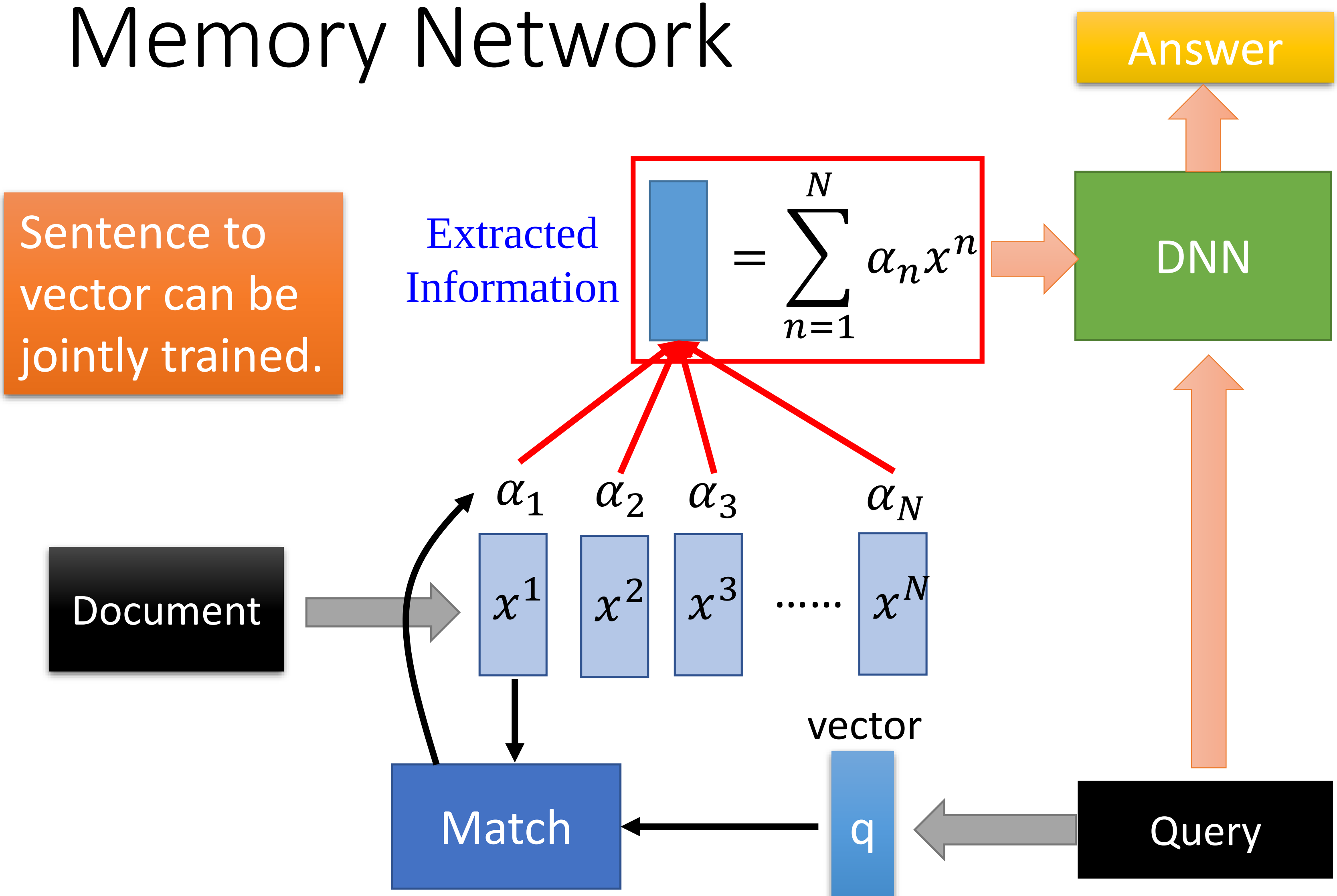


Ref: http://speech.ee.ntu.edu.tw/~tlkagk/courses/MLDS_2015_2/Lecture/Attain%20(v3).ecm.mp4/index.html

Reading Comprehension

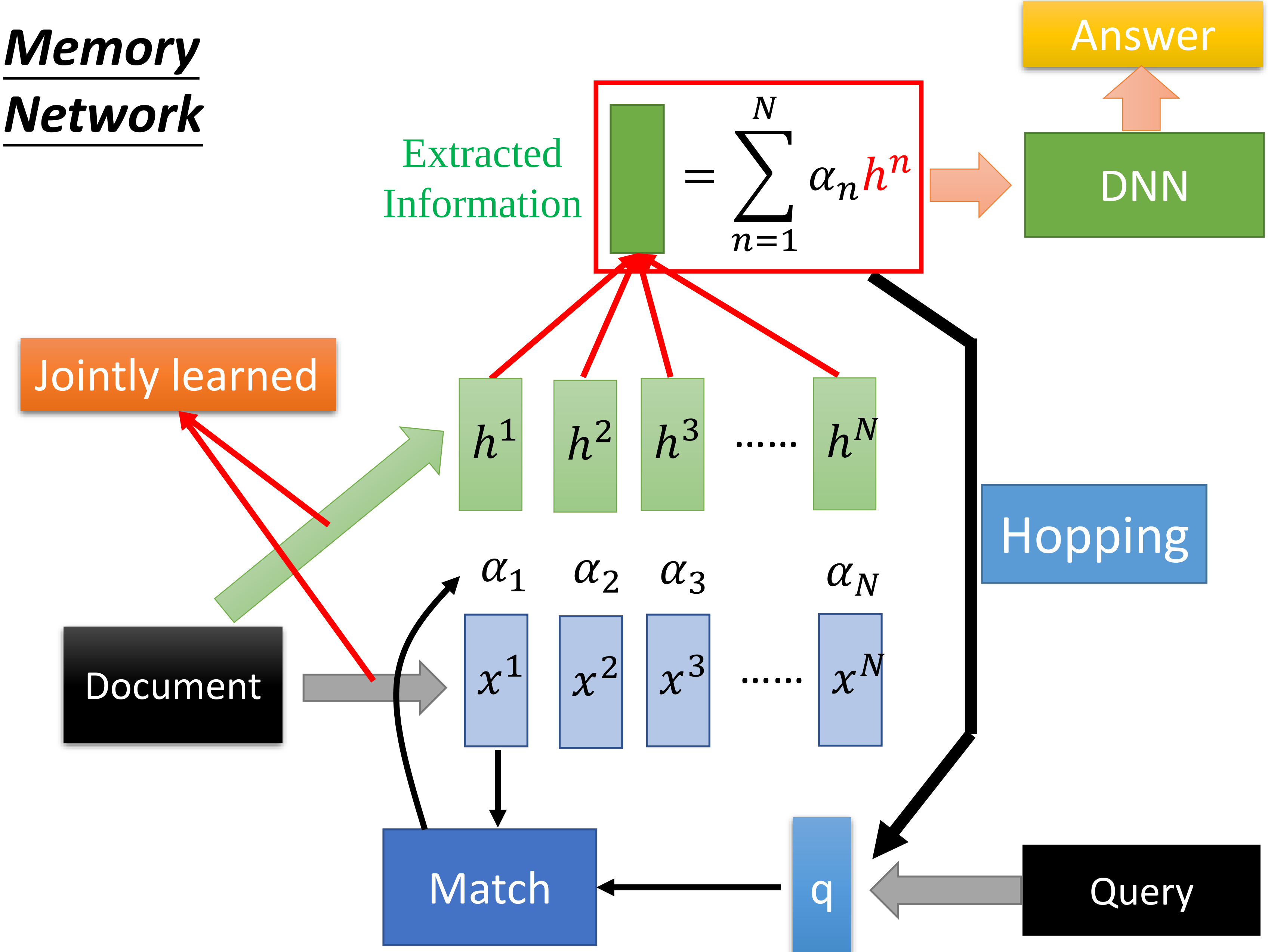


Memory Network

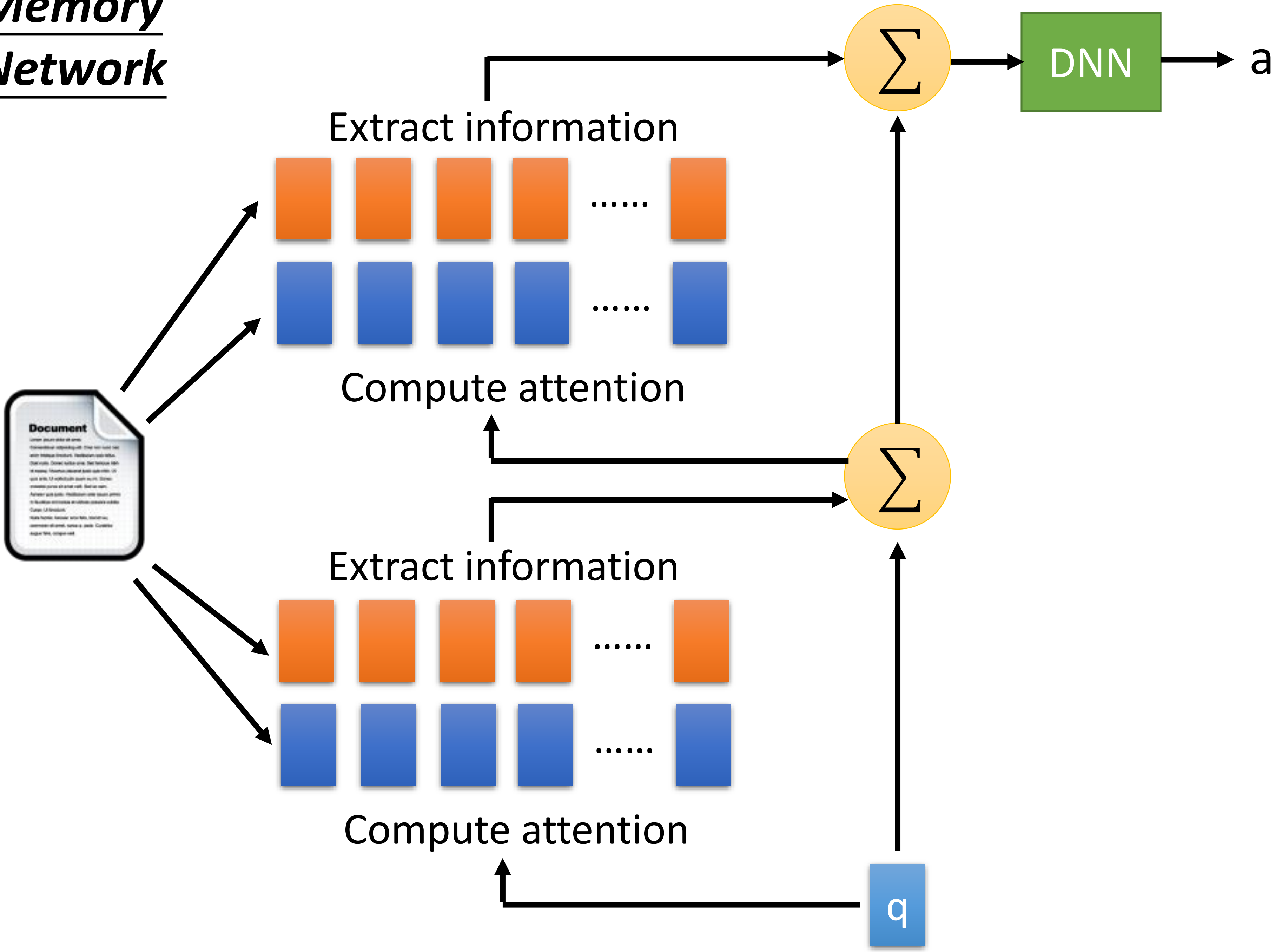


Sainbayar Sukhbaatar, Arthur Szlam, Jason Weston, Rob Fergus, “End-To-End Memory Networks”, NIPS, 2015

Memory Network



Memory Network



Multiple-hop

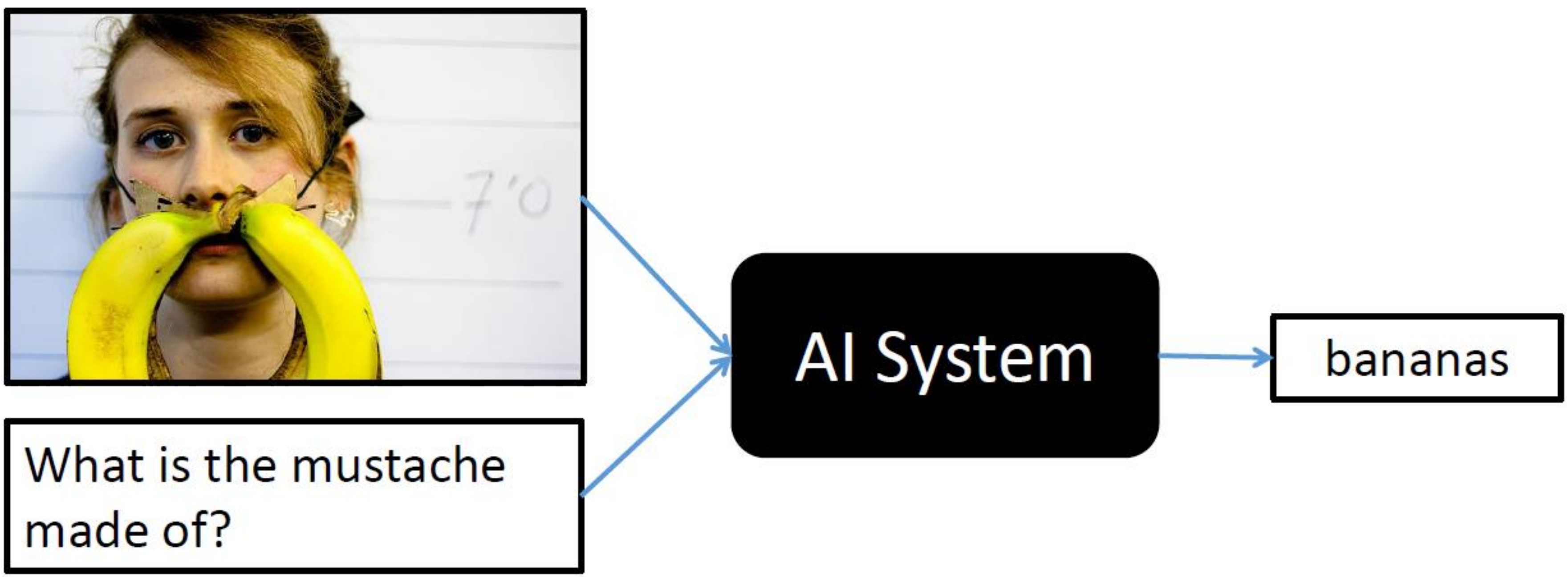
- End-To-End Memory Networks. S. Sukhbaatar, A. Szlam, J. Weston, R. Fergus. NIPS, 2015.

The position of reading head:

Story (16: basic induction)	Support	Hop 1	Hop 2	Hop 3
Brian is a frog.	yes	0.00	0.98	0.00
Lily is gray.		0.07	0.00	0.00
Brian is yellow.	yes	0.07	0.00	1.00
Julius is green.		0.06	0.00	0.00
Greg is a frog.	yes	0.76	0.02	0.00
What color is Greg? Answer: yellow Prediction: yellow				

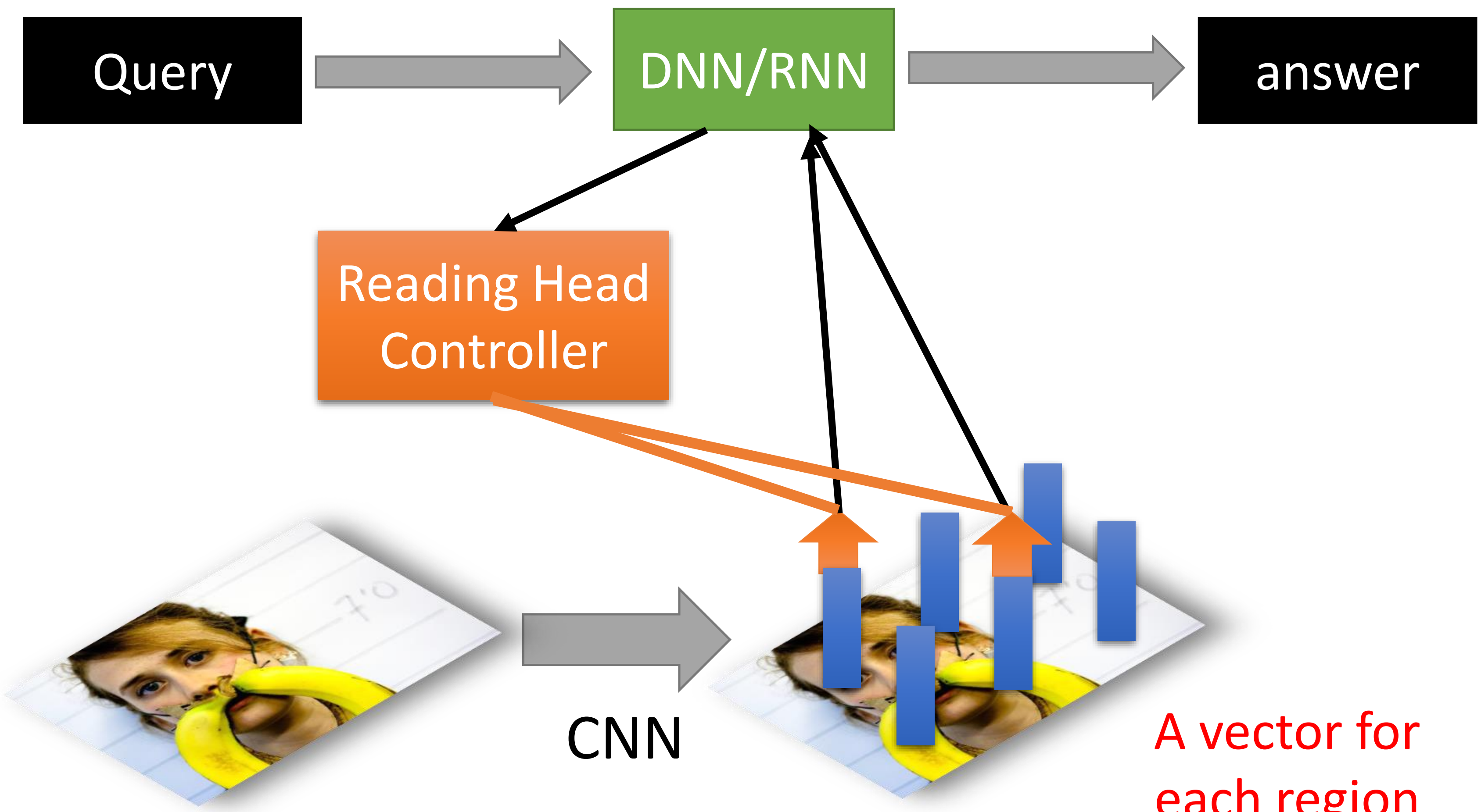
Keras has example:
https://github.com/fchollet/keras/blob/master/examples/babi_memnn.py

Visual Question Answering



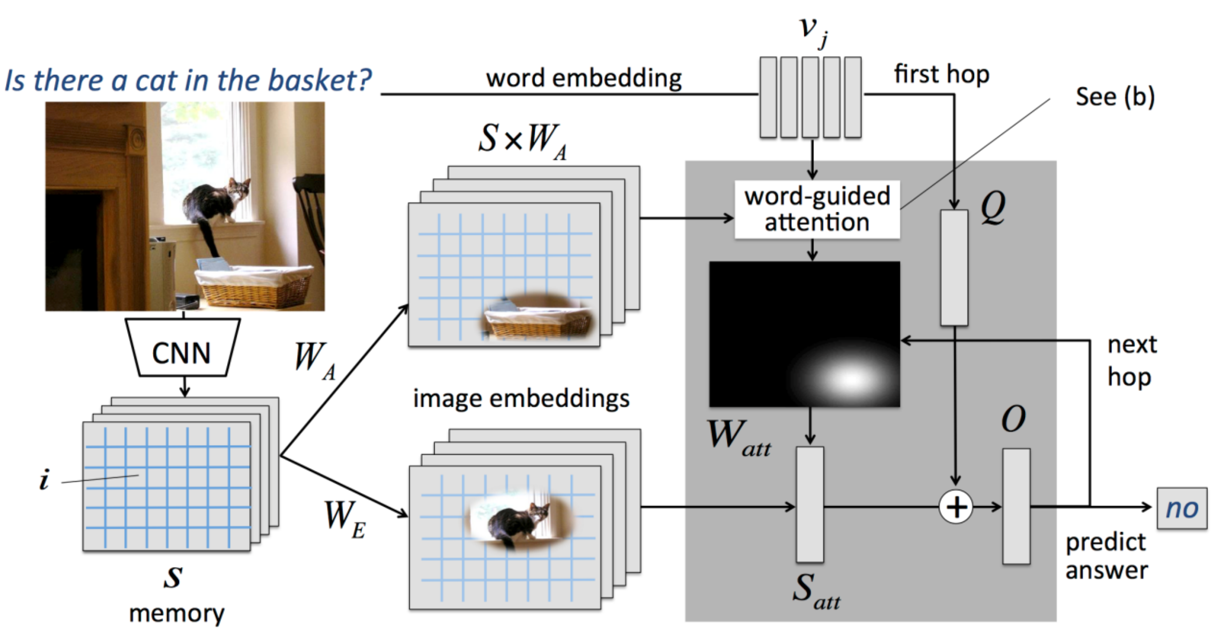
source: http://visualqa.org/

Visual Question Answering

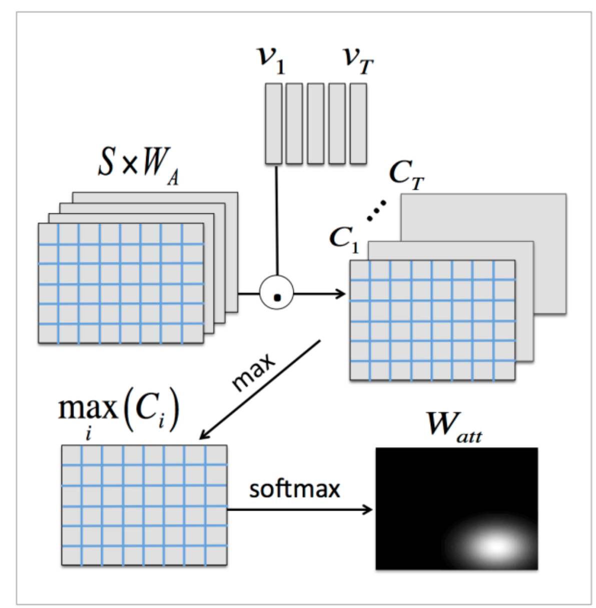


Visual Question Answering

- Huijuan Xu, Kate Saenko. Ask, Attend and Answer: Exploring Question-Guided Spatial Attention for Visual Question Answering. arXiv Pre-Print, 2015



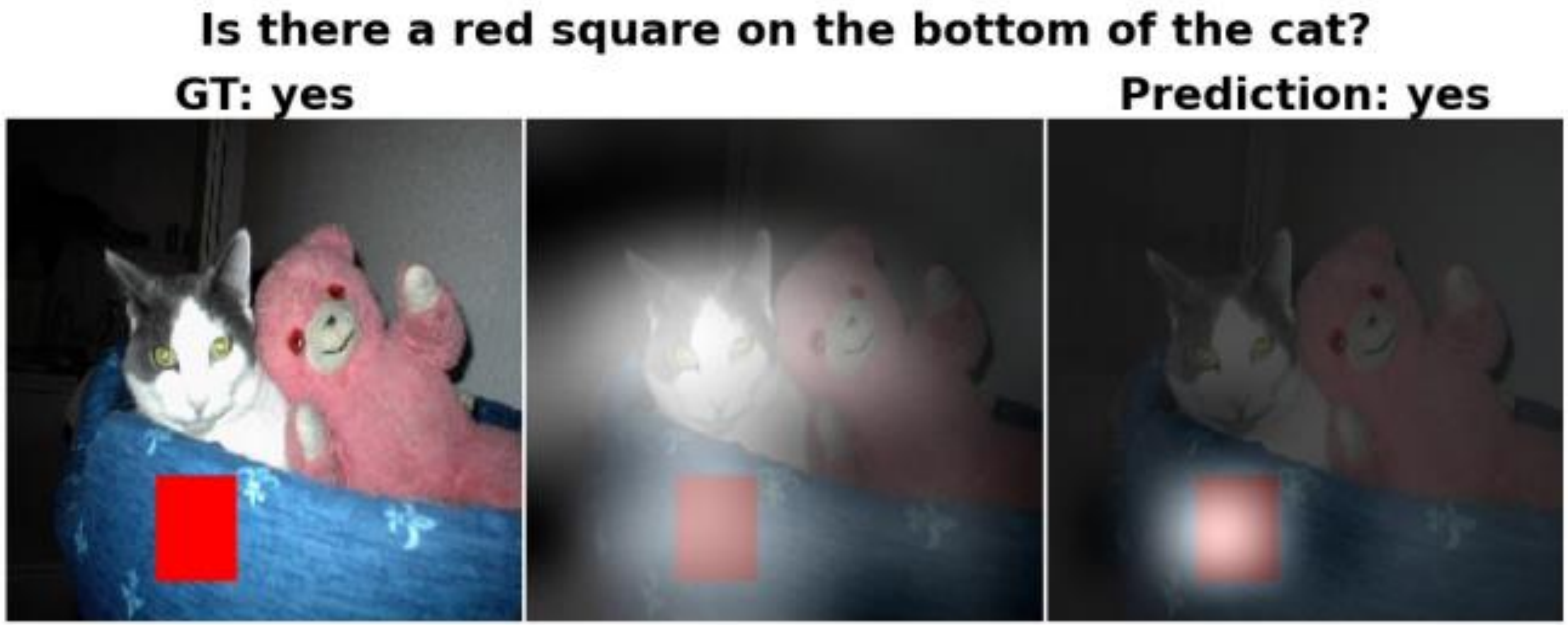
(a) Overview



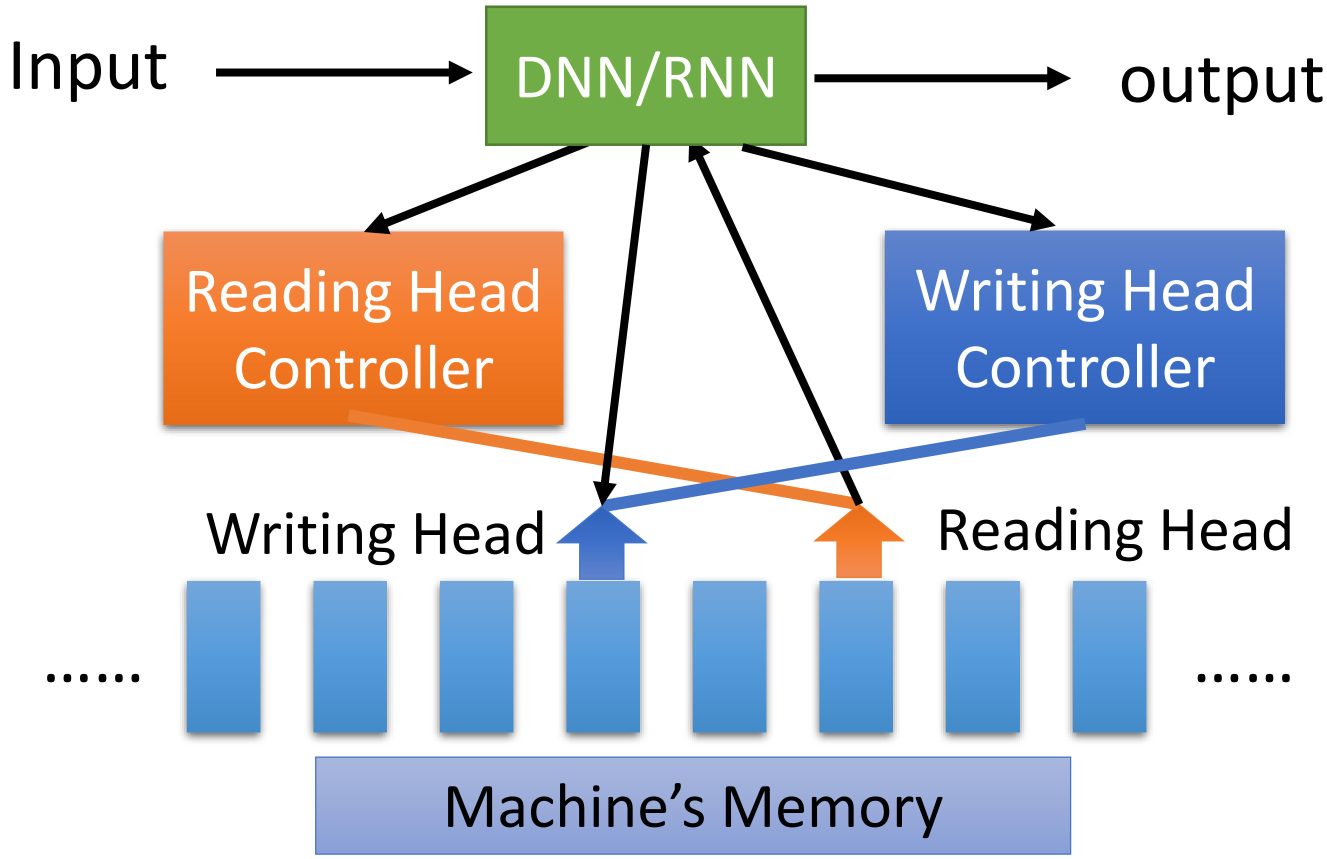
(b) Word-guided attention

Visual Question Answering

- Huijuan Xu, Kate Saenko. Ask, Attend and Answer: Exploring Question-Guided Spatial Attention for Visual Question Answering. arXiv Pre-Print, 2015

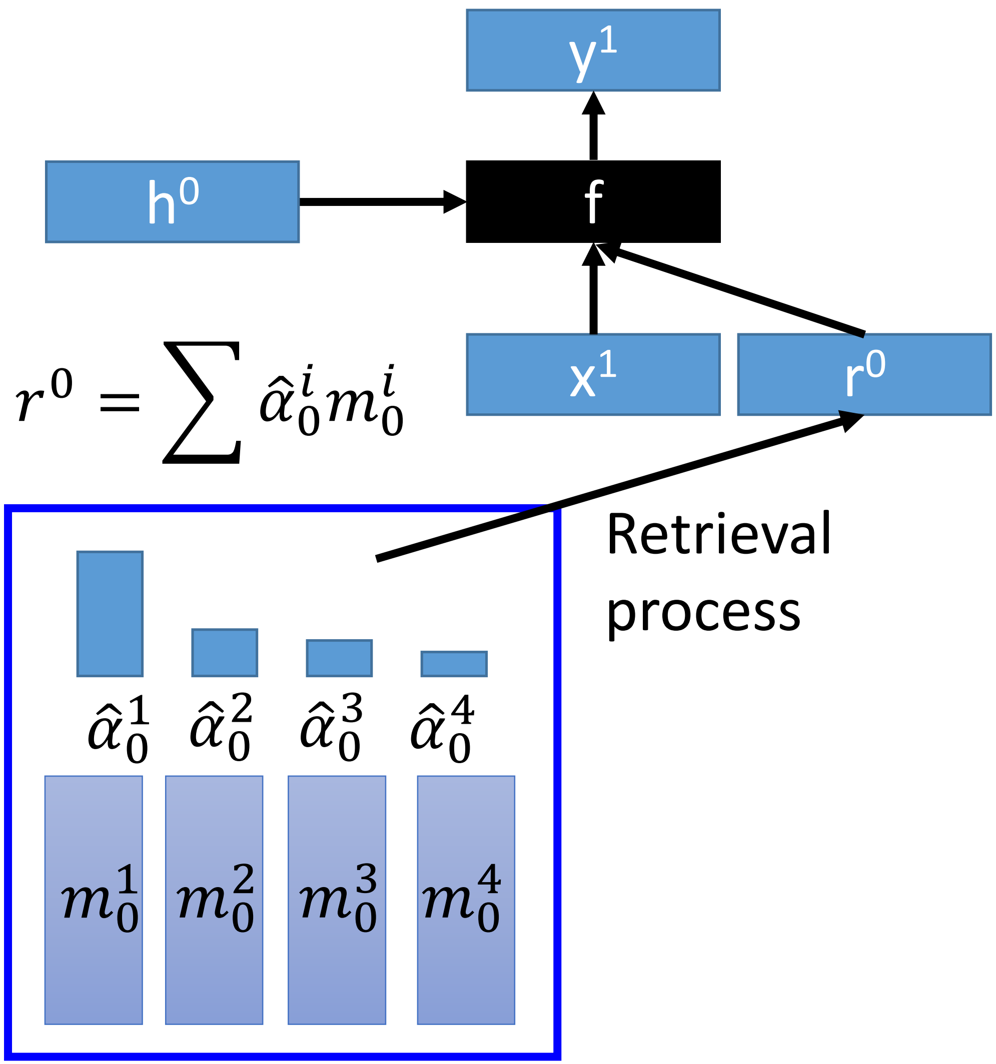


External Memory v2

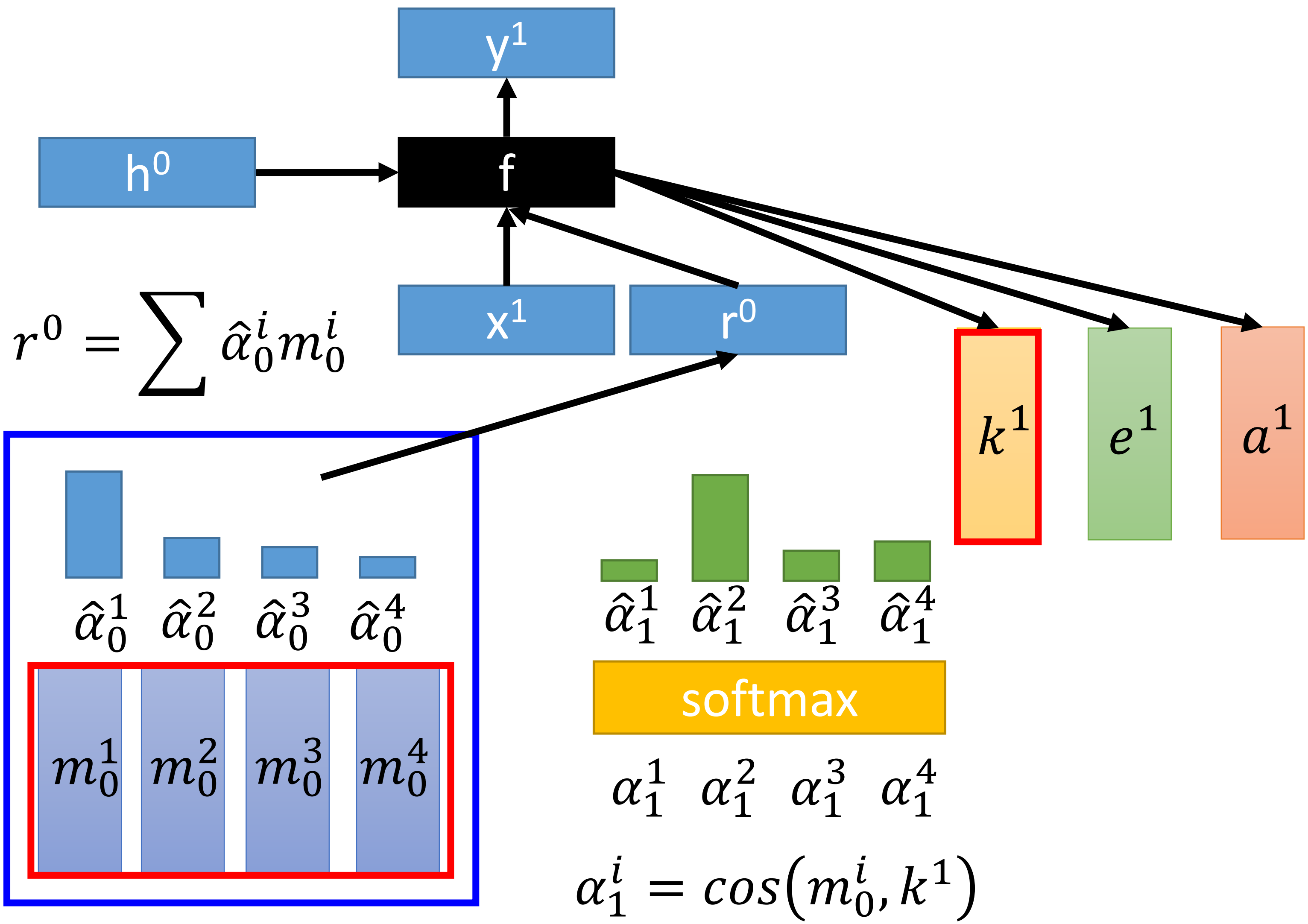


Neural Turing Machine

Neural Turing Machine



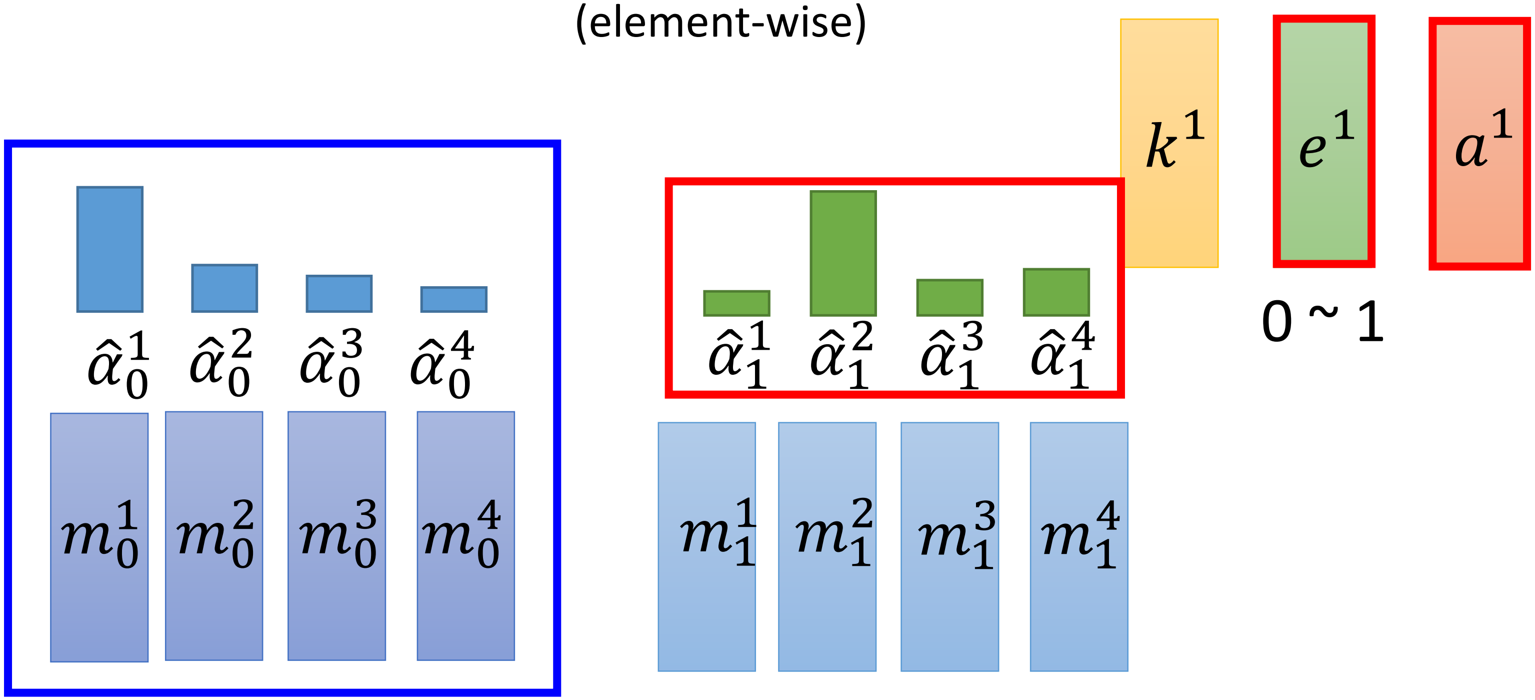
Neural Turing Machine



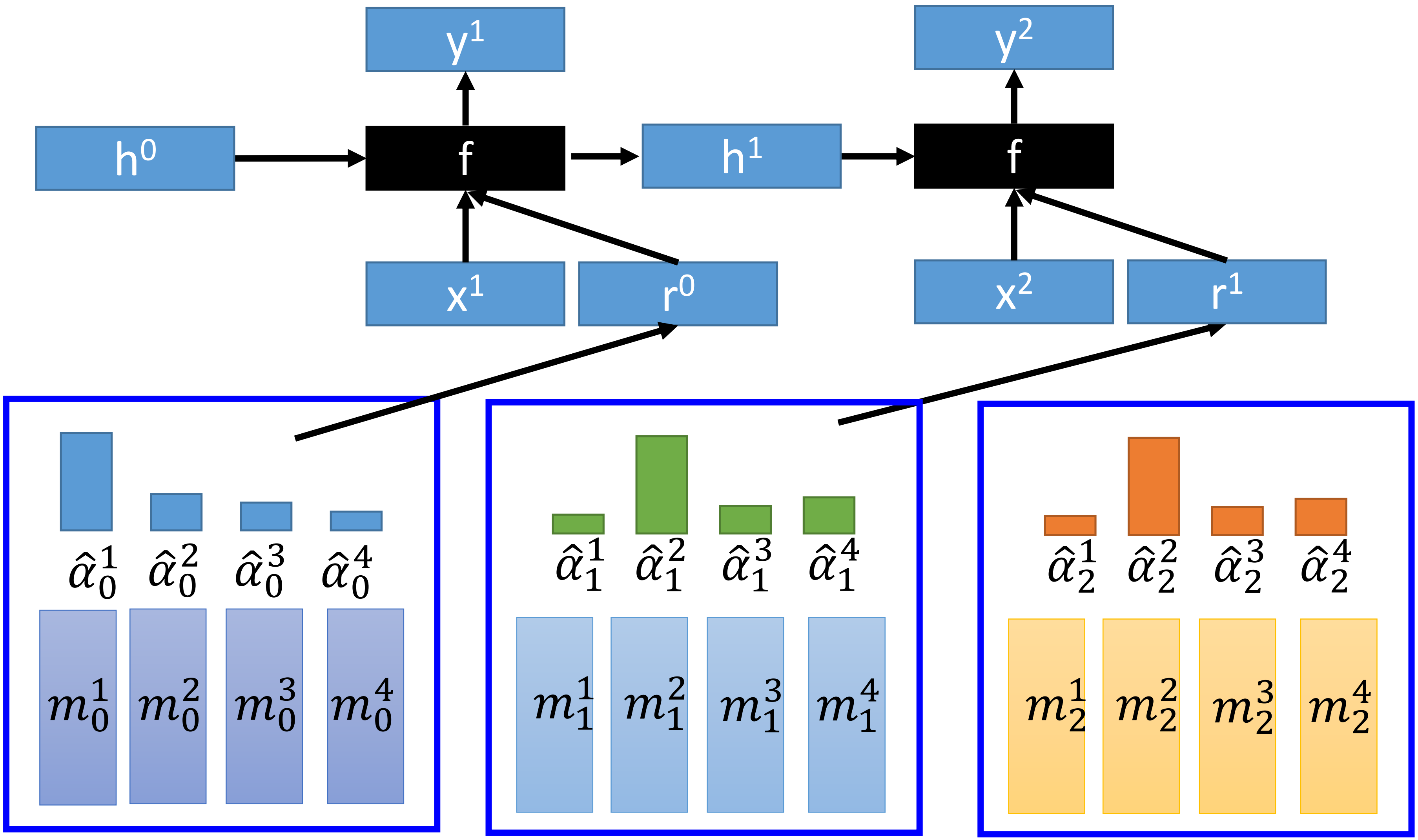
Neural Turing Machine

$$m_1^i = m_0^i - \hat{\alpha}_1^i \, e^1 \odot m_0^i + \hat{\alpha}_1^i \, a^1$$

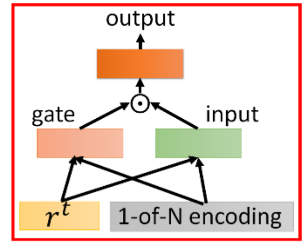
(element-wise)

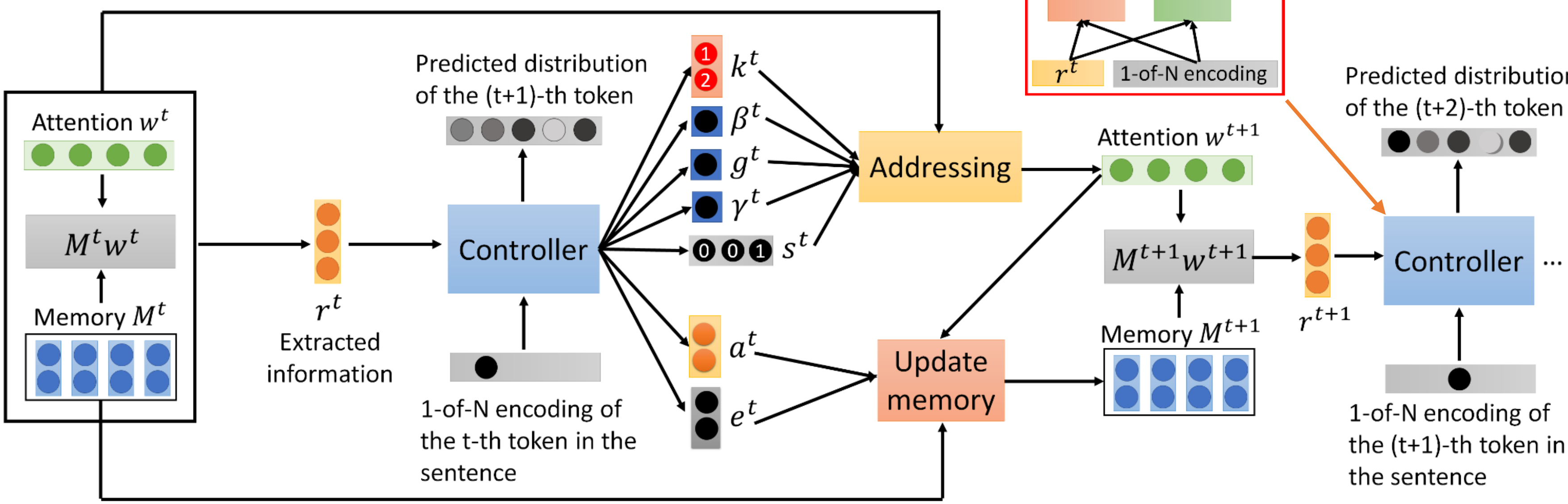


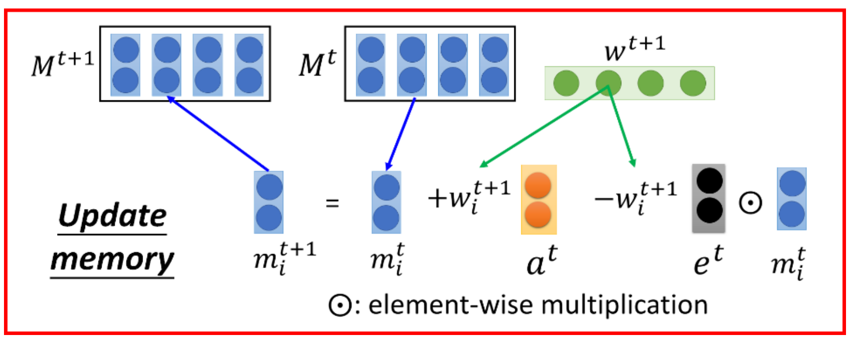
Neural Turing Machine

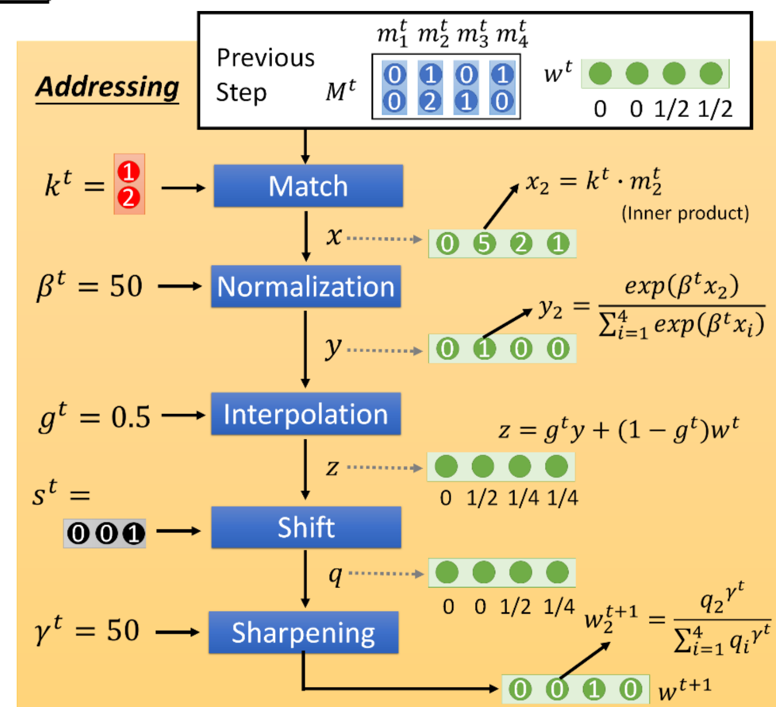


Neural Turing Machine for LM





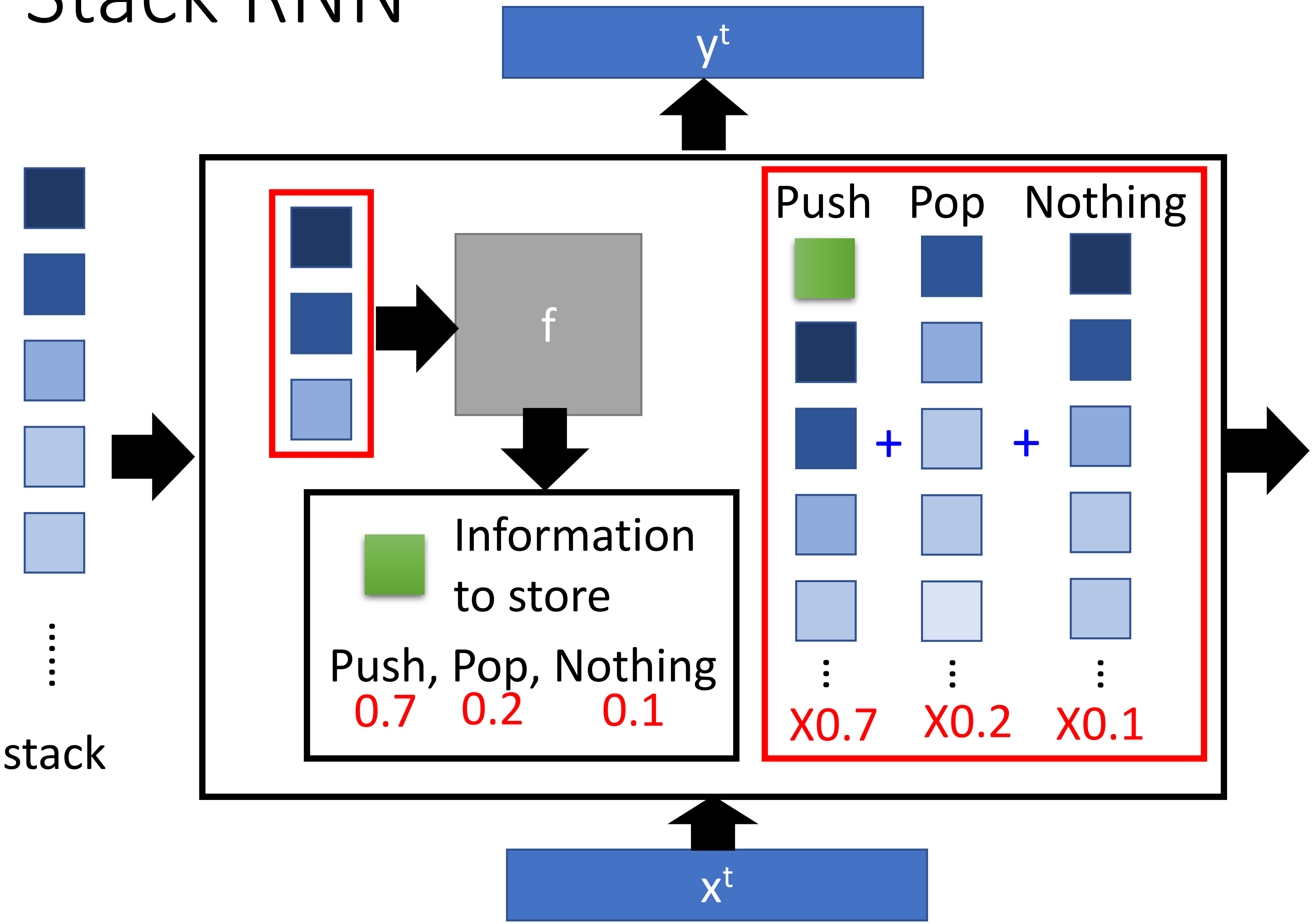




Wei-Jen Ko, Bo-Hsiang Tseng, Hung-yi Lee, “Recurrent Neural Network based Language Modeling with Controllable External Memory”, ICASSP, 2017

Armand Joulin, Tomas Mikolov, Inferring Algorithmic Patterns with Stack-Augmented Recurrent Nets, arXiv Pre-Print, 2015

Stack RNN



Concluding Remarks

- Convolutional Neural Network (Review)
- Spatial Transformer
- Highway Network & Grid LSTM
- Pointer Network
- External Memory