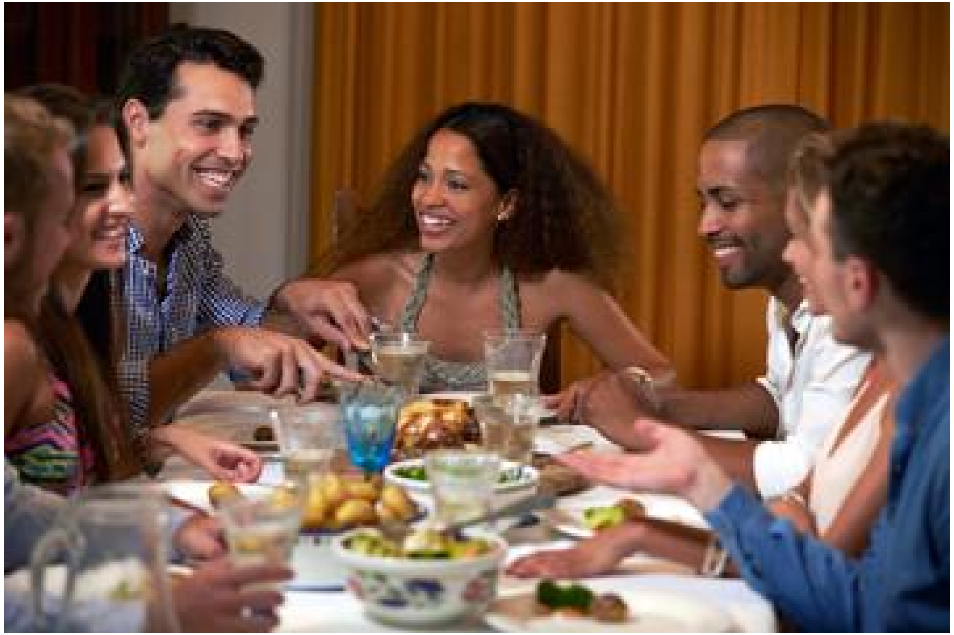
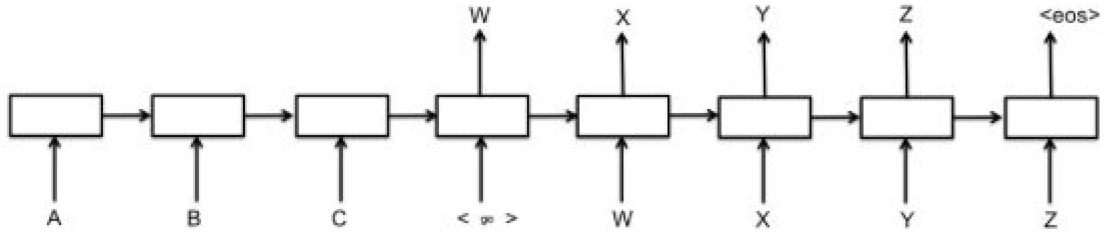
Show and Tell

1. Intro/Demo
2. Architecture
3. Show
4. Tell

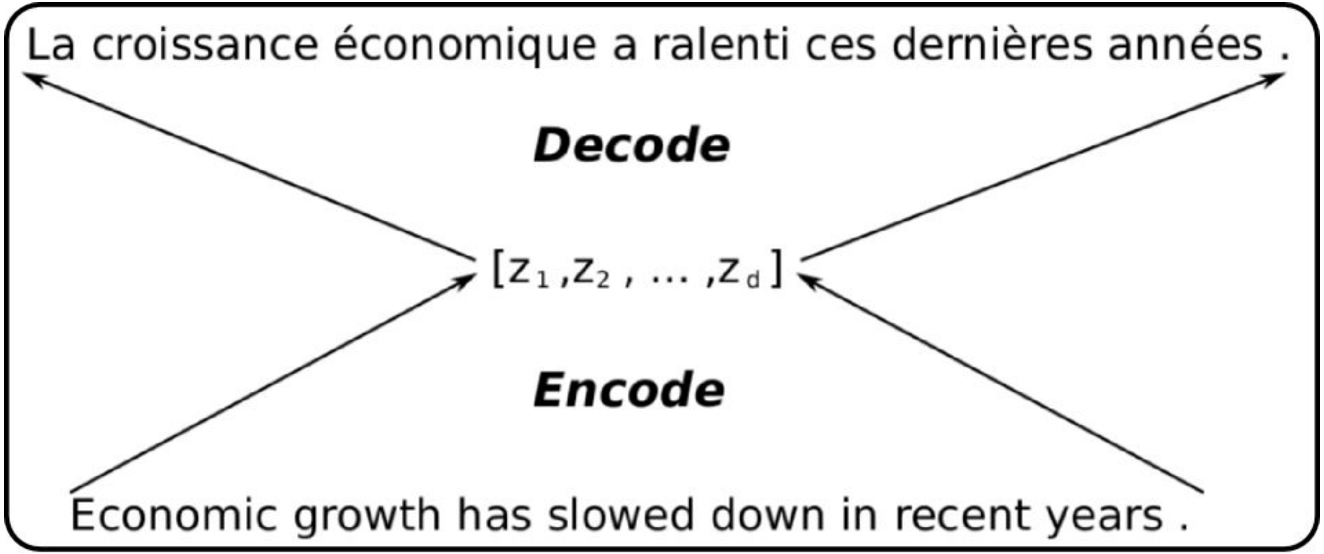


```
Captions for image dinner.jpg:
  0) a group of people sitting around a dinner table . (p=0.014804)
  1) a group of people sitting around a table with food . (p=0.003402)
  2) a group of people sitting at a table with plates of food . (p=0.001879)
```

Encoder-Decoder



“Sequence to Sequence Learning with Neural Networks” (Sutskever, 2014)



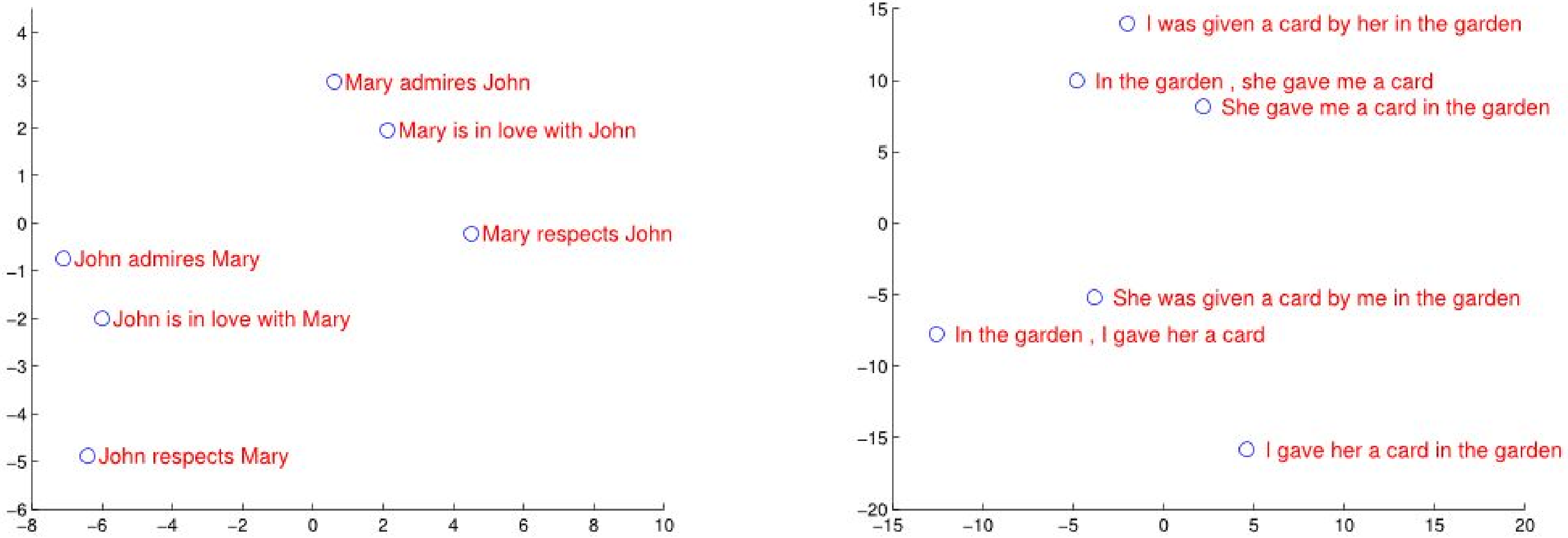
La croissance économique a ralenti ces dernières années .

Decode

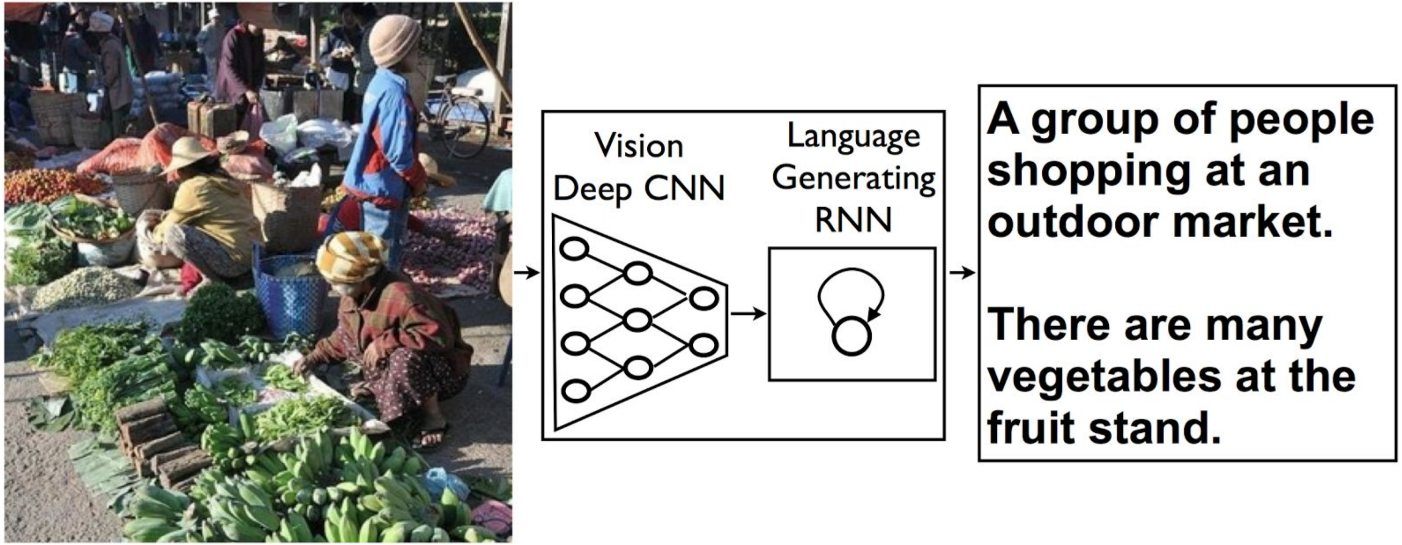
$[z_1, z_2, \ldots, z_d]$

Encode

Economic growth has slowed down in recent years .



PCA projection of LSTM hidden states



Vision
Deep CNN

Language
Generating
RNN

A group of people shopping at an outdoor market.

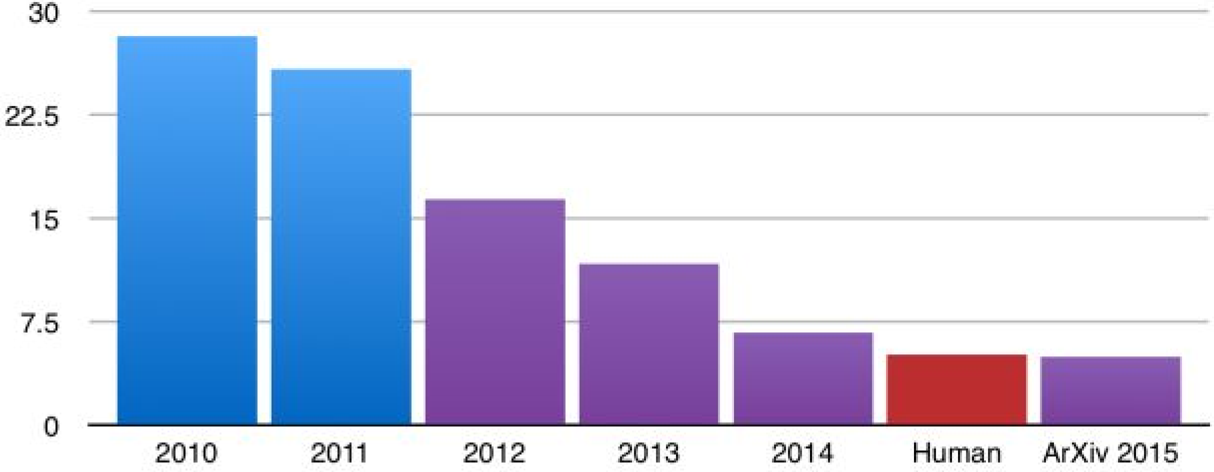
There are many vegetables at the fruit stand.

ImageNet/Large Scale Visual Recognition Challenge

Classification

- 1.2 million images in training set
- 100,000 test set
- 1000 categories
- Predict 5 out of 1000

ILSVRC top-5 error on ImageNet

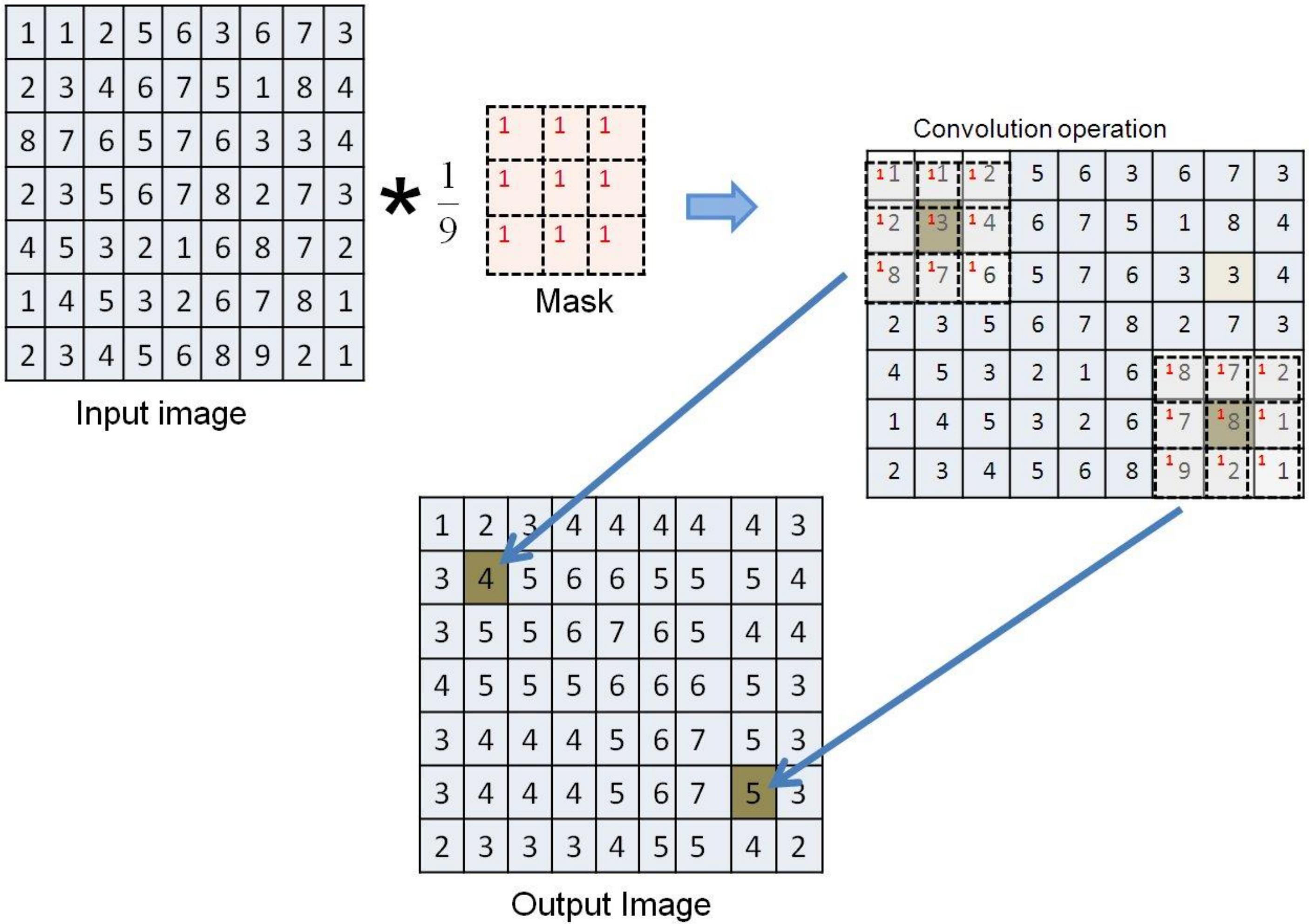


Convolution

I_{11}	I_{12}	I_{13}	I_{14}	I_{15}	I_{16}	I_{17}	I_{18}	I_{19}
I_{21}	I_{22}	I_{23}	I_{24}	I_{25}	I_{26}	I_{27}	I_{28}	I_{29}
I_{31}	I_{32}	I_{33}	I_{34}	I_{35}	I_{36}	I_{37}	I_{38}	I_{39}
I_{41}	I_{42}	I_{43}	I_{44}	I_{45}	I_{46}	I_{47}	I_{48}	I_{49}
I_{51}	I_{52}	I_{53}	I_{54}	I_{55}	I_{56}	I_{57}	I_{58}	I_{59}
I_{61}	I_{62}	I_{63}	I_{64}	I_{65}	I_{66}	I_{67}	I_{68}	I_{69}

K_{11}	K_{12}	K_{13}
K_{21}	K_{22}	K_{23}

$$O(i,j)=\sum_{k=1}^{m}\sum_{l=1}^{n} I(i+k-1,j+l-1)K(k,l)$$



Input image

1	1	2	5	6	3	6	7	3
2	3	4	6	7	5	1	8	4
8	7	6	5	7	6	3	3	4
2	3	5	6	7	8	2	7	3
4	5	3	2	1	6	8	7	2
1	4	5	3	2	6	7	8	1
2	3	4	5	6	8	9	2	1

$* \frac{1}{9}$

Mask

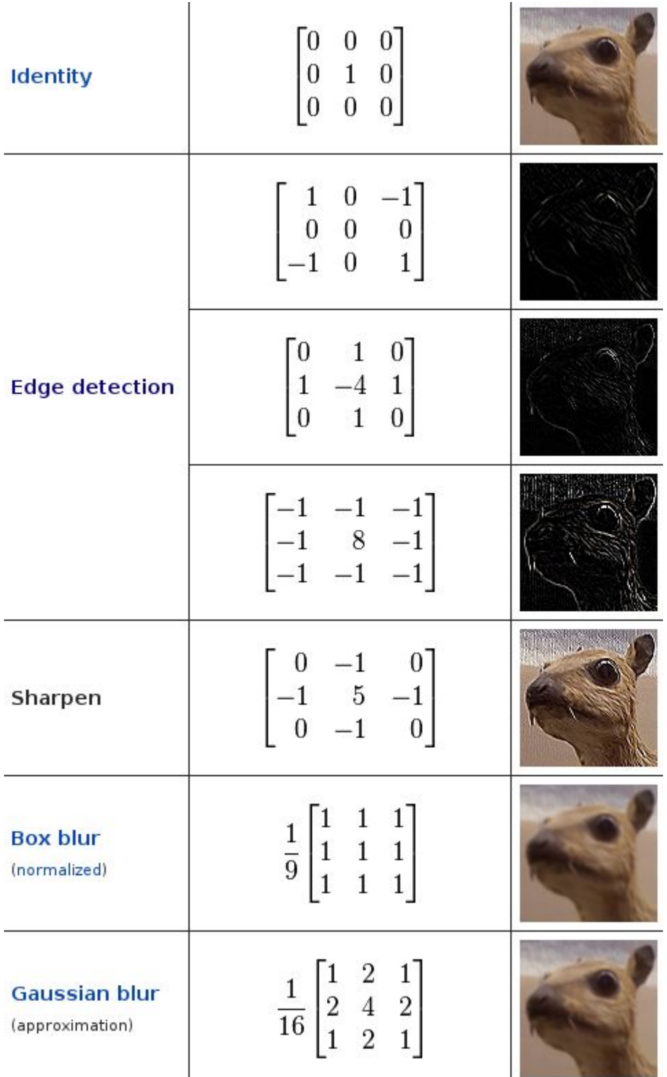
1	1	1
1	1	1
1	1	1

Convolution operation

$^{1}1$	$^{1}1$	$^{1}2$	5	6	3	6	7	3
$^{1}2$	$^{1}3$	$^{1}4$	6	7	5	1	8	4
$^{1}8$	$^{1}7$	$^{1}6$	5	7	6	3	3	4
2	3	5	6	7	8	2	7	3
4	5	3	2	1	6	$^{1}8$	$^{1}7$	$^{1}2$
1	4	5	3	2	6	$^{1}7$	$^{1}8$	$^{1}1$
2	3	4	5	6	8	$^{1}9$	$^{1}2$	$^{1}1$

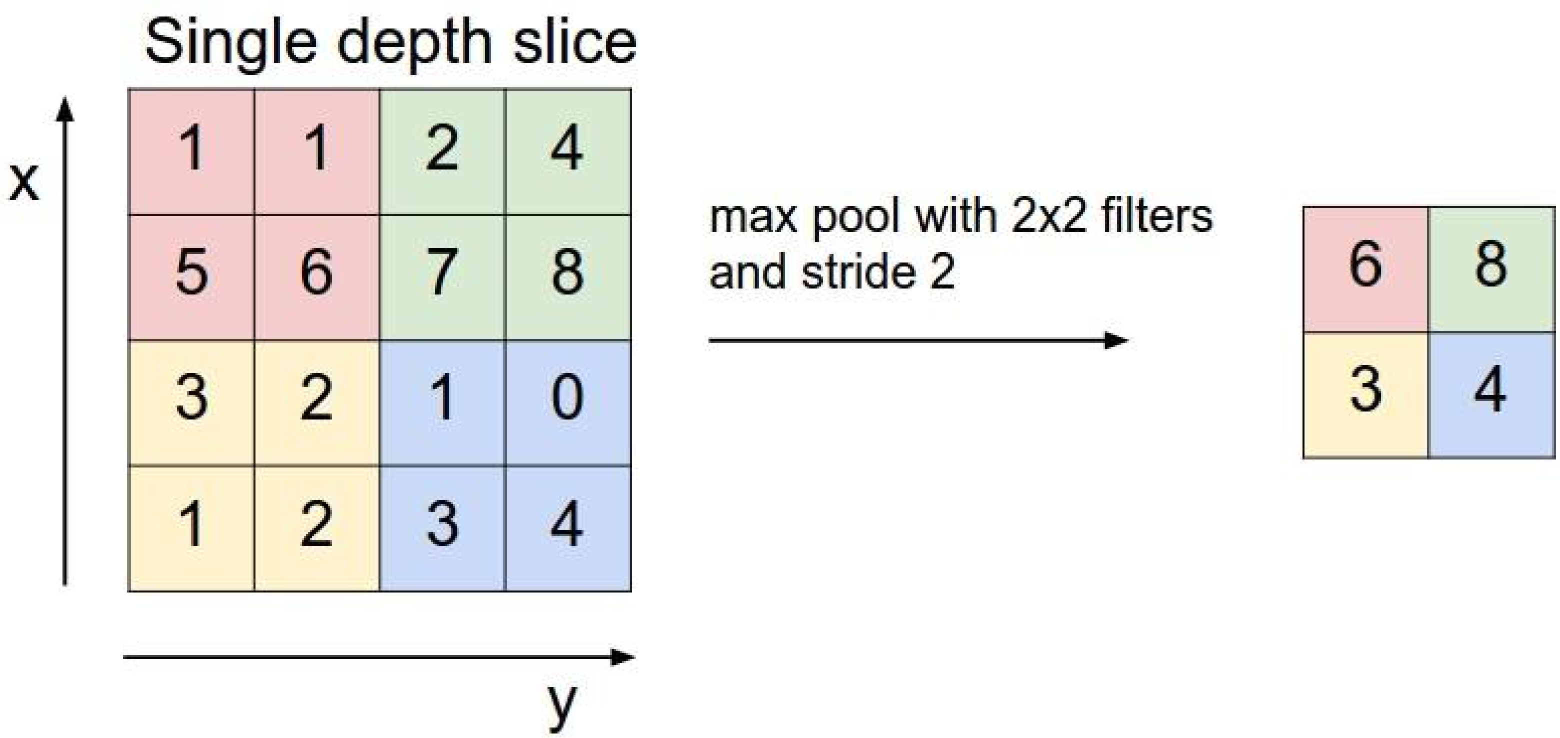
1	2	3	4	4	4	4	4	3
3	4	5	6	6	5	5	5	4
3	5	5	6	7	6	5	4	4
4	5	5	5	6	6	6	5	3
3	4	4	4	5	6	7	5	3
3	4	4	4	5	6	7	5	3
2	3	3	3	4	5	5	4	2

Output Image



Identity	$\begin{bmatrix} 0 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{bmatrix}$	
Edge detection	$\begin{bmatrix} 1 & 0 & -1 \\ 0 & 0 & 0 \\ -1 & 0 & 1 \end{bmatrix}$	
	$\begin{bmatrix} 0 & 1 & 0 \\ 1 & -4 & 1 \\ 0 & 1 & 0 \end{bmatrix}$	
	$\begin{bmatrix} -1 & -1 & -1 \\ -1 & 8 & -1 \\ -1 & -1 & -1 \end{bmatrix}$	
Sharpen	$\begin{bmatrix} 0 & -1 & 0 \\ -1 & 5 & -1 \\ 0 & -1 & 0 \end{bmatrix}$	
Box blur (normalized)	$\frac{1}{9}\begin{bmatrix} 1 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \end{bmatrix}$	
Gaussian blur (approximation)	$\frac{1}{16}\begin{bmatrix} 1 & 2 & 1 \\ 2 & 4 & 2 \\ 1 & 2 & 1 \end{bmatrix}$	

Max Pooling



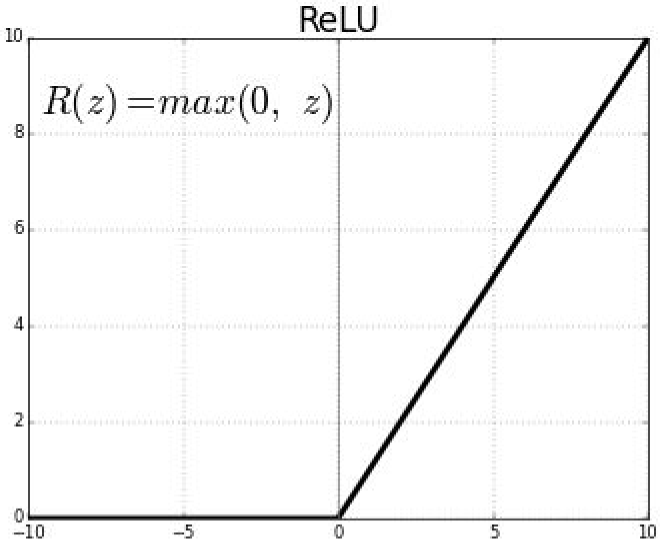
Single depth slice

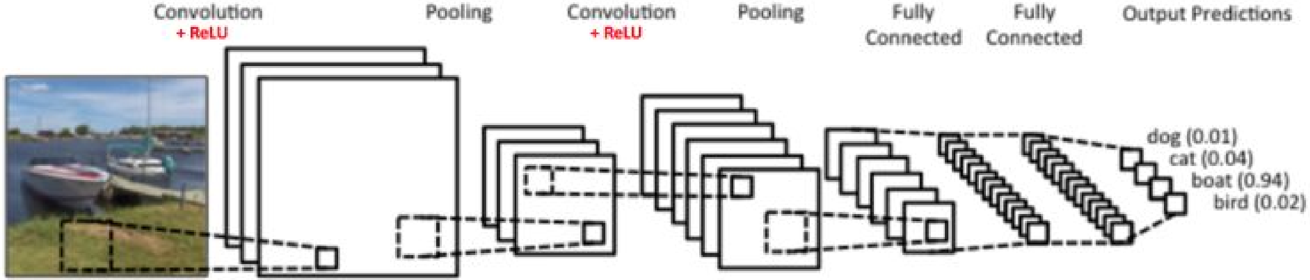
1	1	2	4
5	6	7	8
3	2	1	0
1	2	3	4

max pool with 2x2 filters and stride 2

6	8
3	4

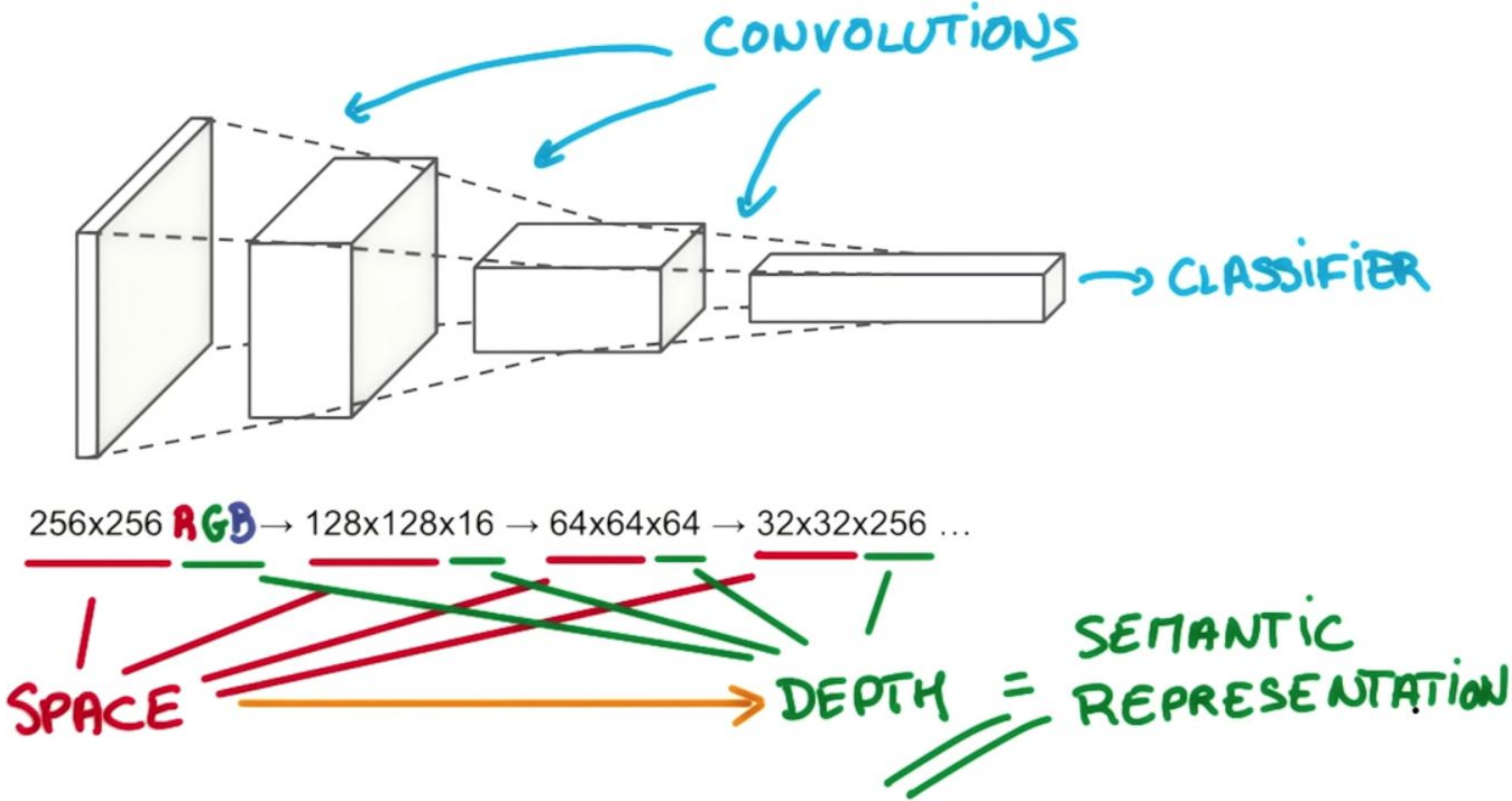
ReLu





Convolution
+ ReLU
Pooling
Convolution
+ ReLU
Pooling
Fully Connected
Fully Connected
Output Predictions
dog (0.01)
cat (0.04)
boat (0.94)
bird (0.02)

CONVOLUTIONAL PYRAMID



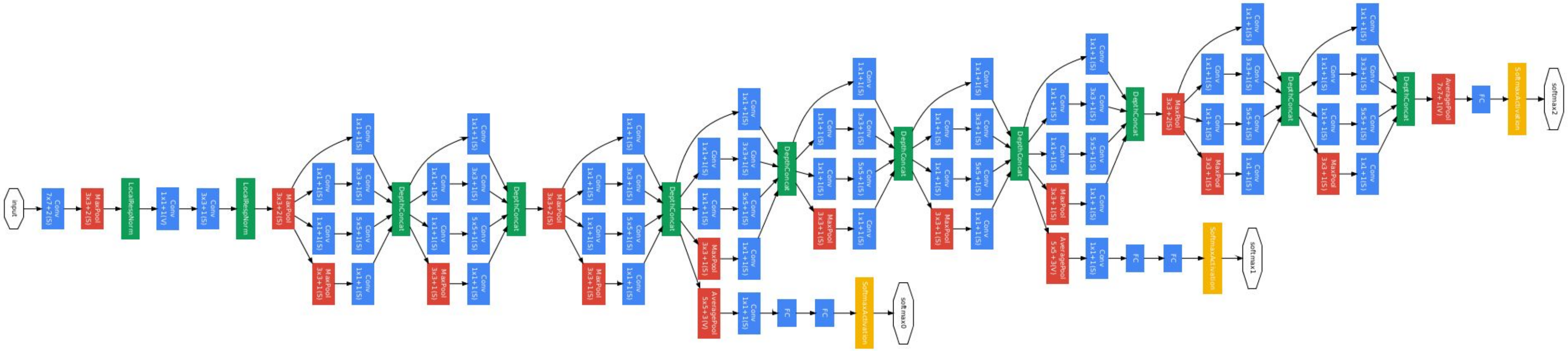
256x256 RGB → 128x128x16 → 64x64x64 → 32x32x256 ...

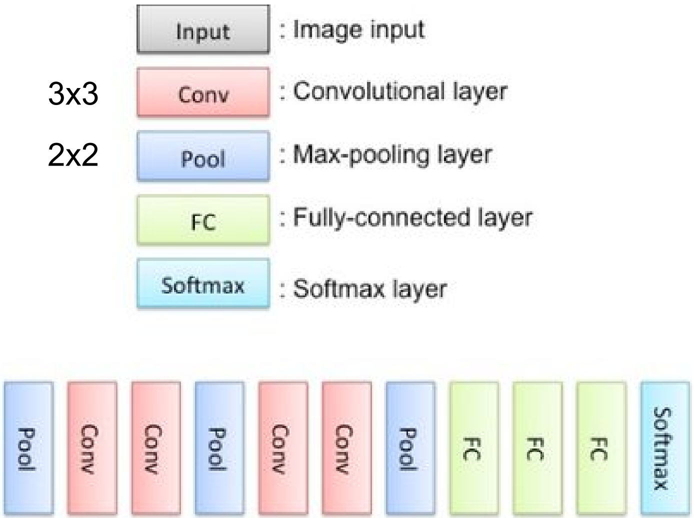
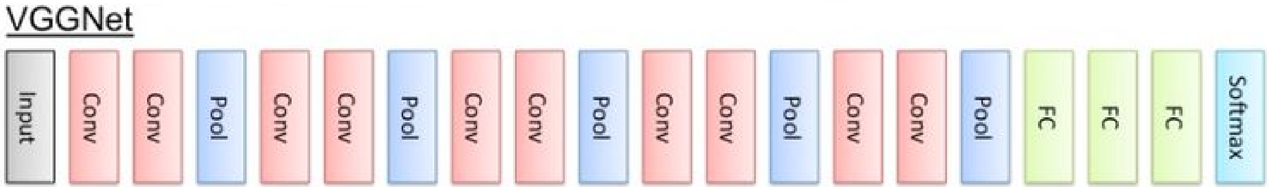
Going deeper with convolutions, 2014 (GoogLenet)

M. Lin: Network In Network, 2013

Provable bounds for learning some deep representations, 2014

Parameters: 4M VS 140M (VGGNet)





Input : Image input

3x3 Conv : Convolutional layer

2x2 Pool : Max-pooling layer

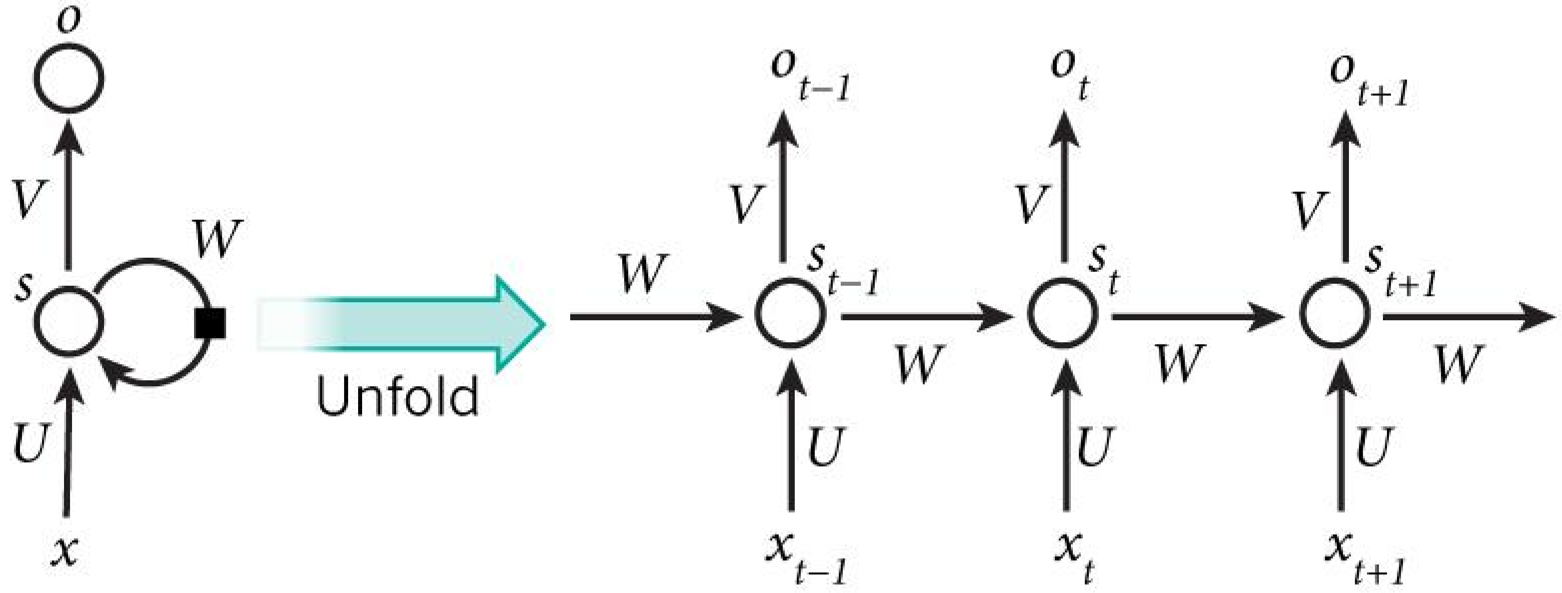
FC : Fully-connected layer

Softmax : Softmax layer

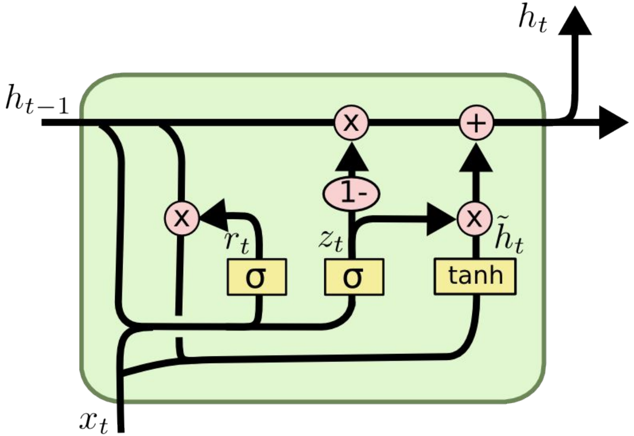
VGGNet

Input → Conv → Conv → Pool → Conv → Conv → Pool → Conv → Conv → Pool → Conv → Conv → Pool → Conv → Conv → Pool → FC → FC → FC → Softmax

RNN



LSTM



$$z_t = \sigma\left(W_z \cdot [h_{t-1}, x_t]\right)$$

$$r_t = \sigma\left(W_r \cdot [h_{t-1}, x_t]\right)$$

$$\tilde{h}_t = \tanh\left(W \cdot [r_t * h_{t-1}, x_t]\right)$$

$$h_t = (1 - z_t) * h_{t-1} + z_t * \tilde{h}_t$$