CIT
互联网教育智能技术及应用
国 家 工 程 实 验 室

人工智能与教育学术沙龙

Academic Salon:
Artificial Intelligence and Education

北京師範大學 BEIJING NORMAL UNIVERSITY

清華大學 Tsinghua University

中国移动 China Mobile

网龙华渔教育

科大讯飞 iFLYTEK

An Introduction to AI and Deep Learning

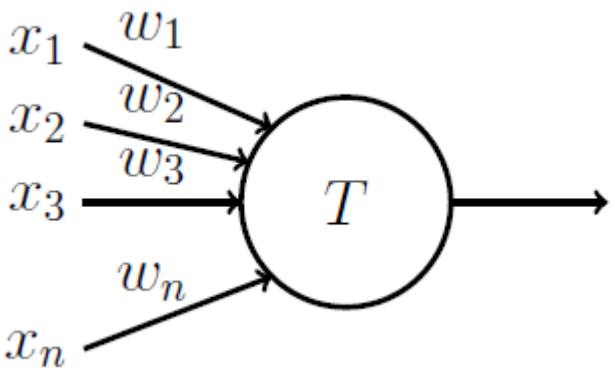
John Hopcroft

Cornell University

The information revolution

- Machine learning is an important driver of the information revolution that is changing our world.
- Deep learning is an important aspect.
- This talk will give a short introduction to machine learning and then explore some interesting research in deep learning.

Threshold logic unit



$$z = \begin{cases} 0 & \sum x_i w_i < T \\ 1 & \sum x_i w_i \geq T \end{cases}$$

data $a_1, a_2, \ldots, a_n$ $\qquad\qquad$ $l_i = \pm 1$

Algorithm

Set $w = 0$

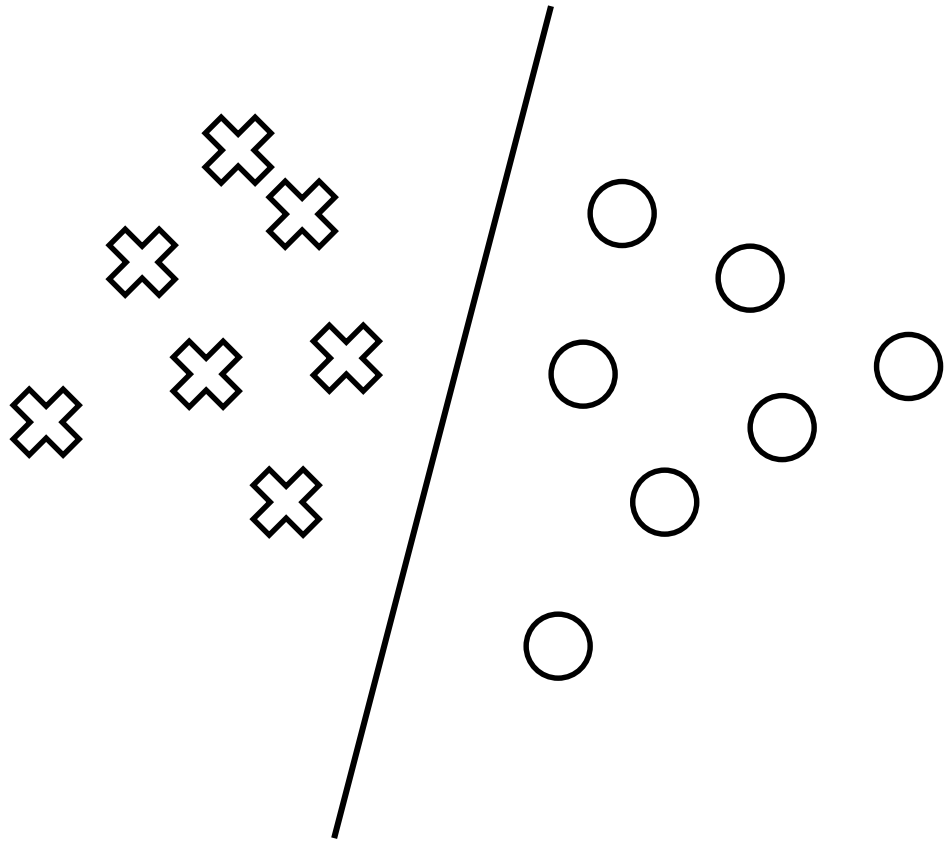
Until all patterns classified correctly

If a_i **not classified correctly** $w = w + a_i l_i$

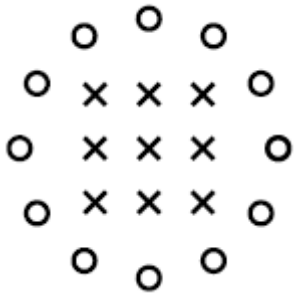
End

Note: Weight vector is a linear combination of the patterns

Linearly separated data

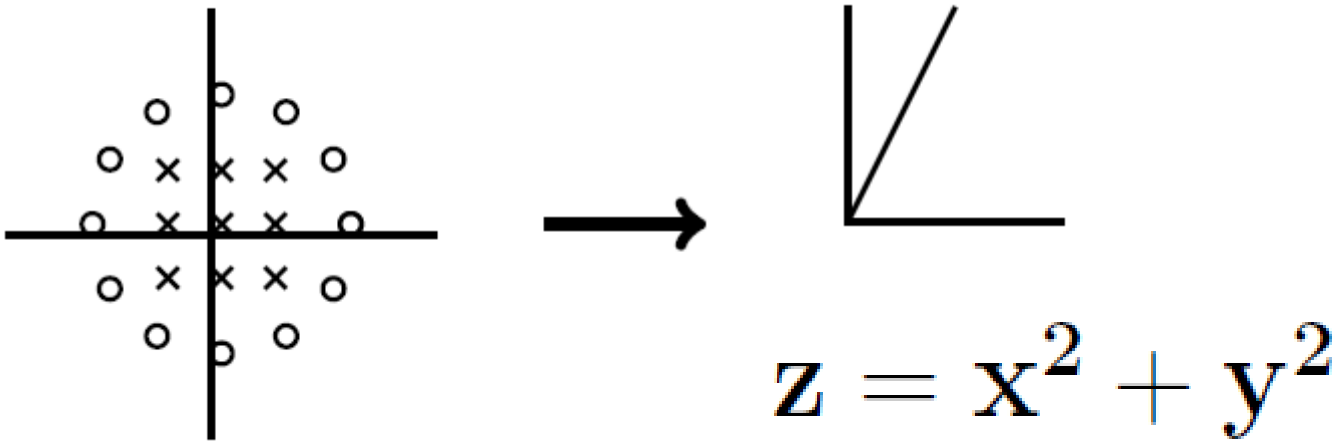


What if data is not linearly separable?



Solve problem in higher dimensional space

Map problem to higher dimensional space where the data is linearly separable



$$z = x^2 + y^2$$

$$a_i \xrightarrow{f} f(a_i)$$

Mapping may be to an infinite dimensional space.

Do not need to compute images of patterns in the higher dimensional space.

Only need the products of the images of the patterns.

$$a_i \xrightarrow{f} f(a_i)$$

$$w = \sum_{i=0}^{n} c_i f(a_i)$$

$$wf(a_j) = \sum_{i=1}^{n} c_i \underbrace{f(a_i) f(a_j)}_{\text{Products}}$$

If one knows the products of images they do not need to know the images.

$$w \longleftarrow w + f(a_j)$$

Just increase the coefficient c_j of $f(a_j)$.

Kernels

$$K = \begin{pmatrix} & \\ & \end{pmatrix} \qquad \text{where } k_{ij} = f(a_i)f(a_j).$$

Given a matrix K does there exist a function f such that $k_{ij} = f(a_i)f(a_j)$?

Given a matrix K does there exist a function f such that $k_{ij} = f(a_i)f(a_j)$?

There exists a function f if and only if K is a positive semi definite matrix.

For all x, $x^T K x \geq 0$.

Kernel matrix

Gaussian kernel $k_{ij} = e^{\frac{1}{2}\frac{(a_i - a_j)^2}{\sigma^2}}$

Support vector machine

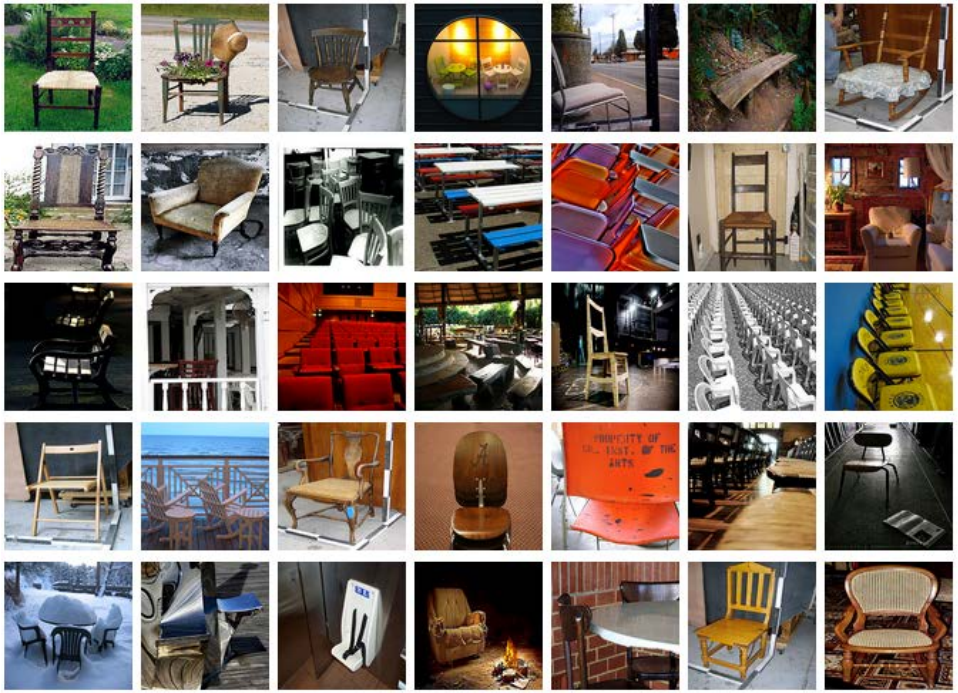
Kernels and mapping to higher dimensional space is the essence of support vector machines.

There exist many kernels such as the Gaussian kernel.

The next advance is in deep learning.

Advent of deep learning

Image net competition



1.2 million images, 1,000 categories

Something new is happening

ImageNet competition 1.2 million images, 1,000 categories

2011 error rate 25%

2012 Alexnet error rate 15%

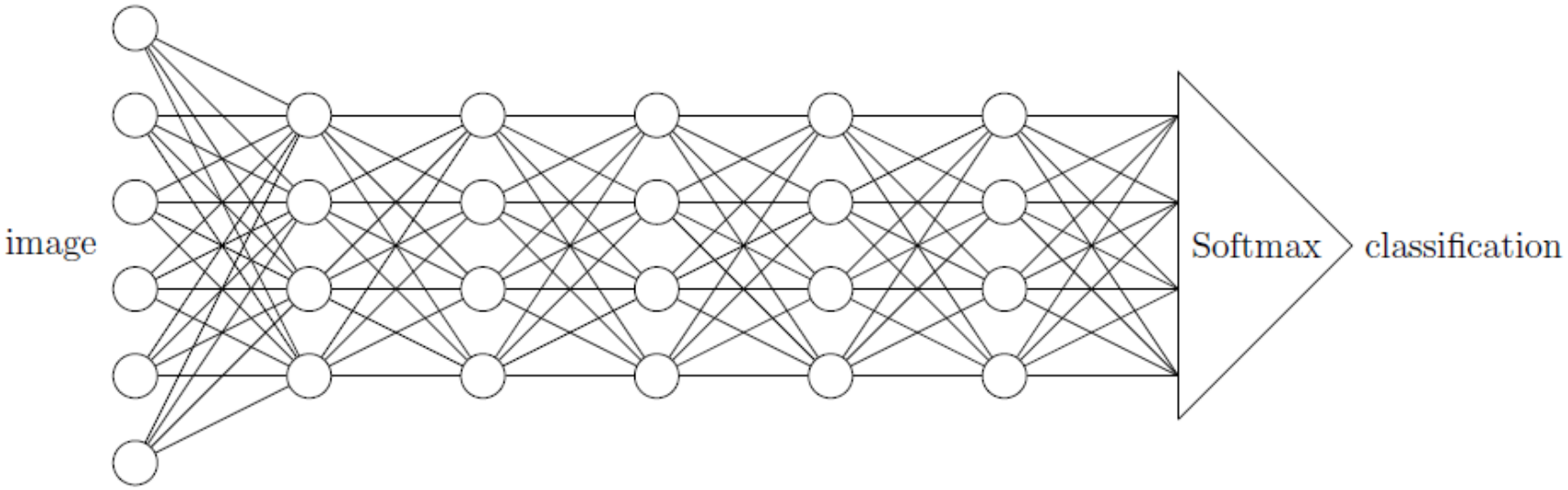
2013 Zfnet error rate 14.8%

2014 GoogleNet error rate 6.67

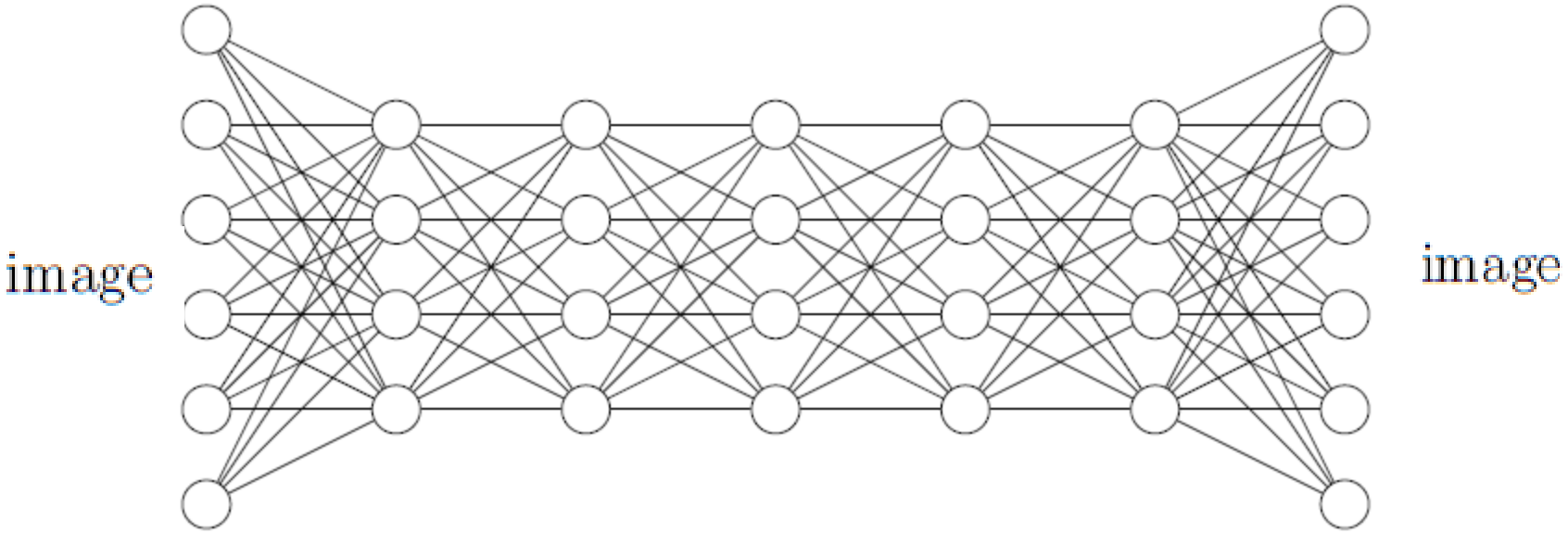
2015 ResNet error rate 3.57%

Human error rate 5% with training

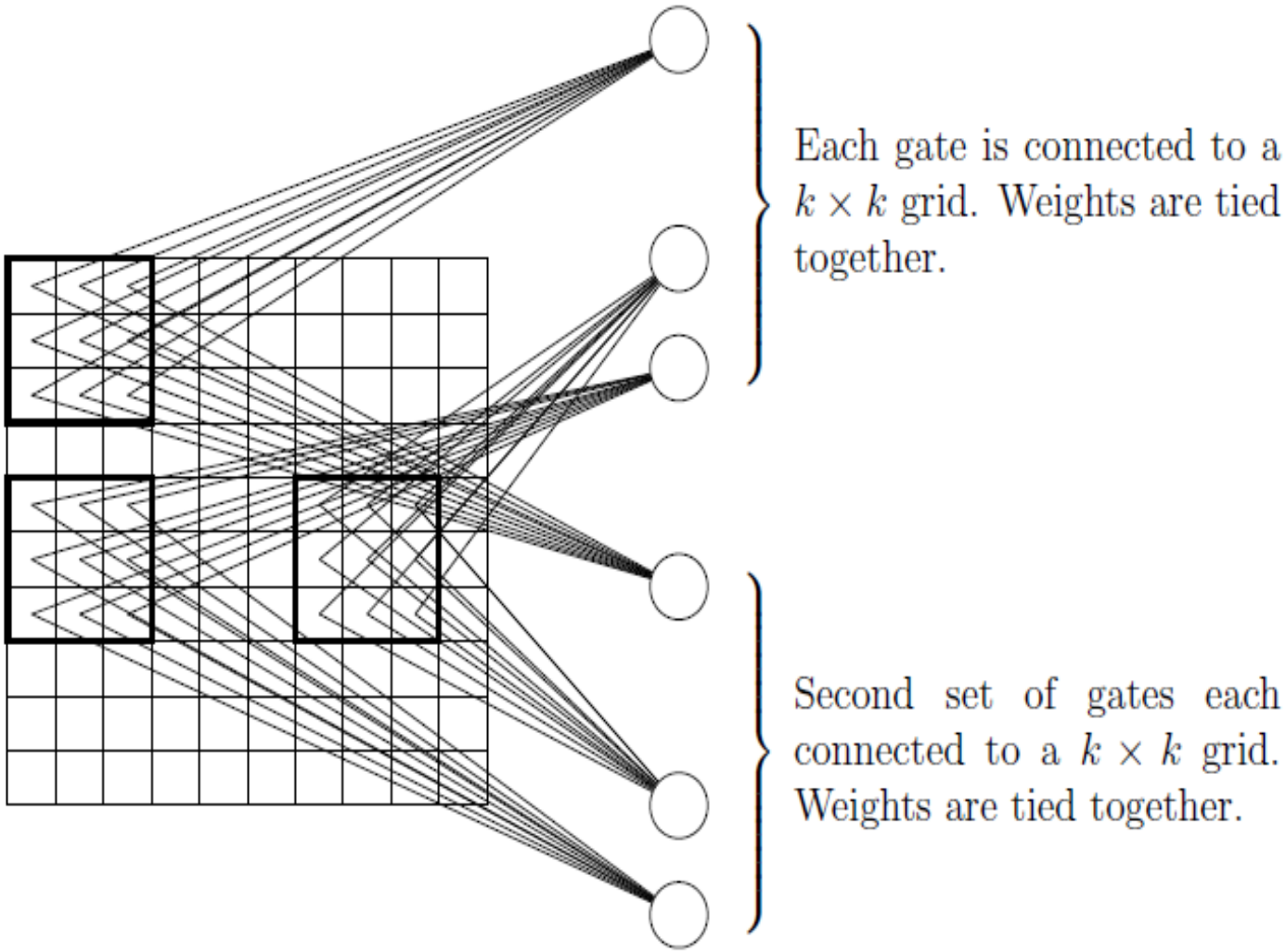
Supervised learning

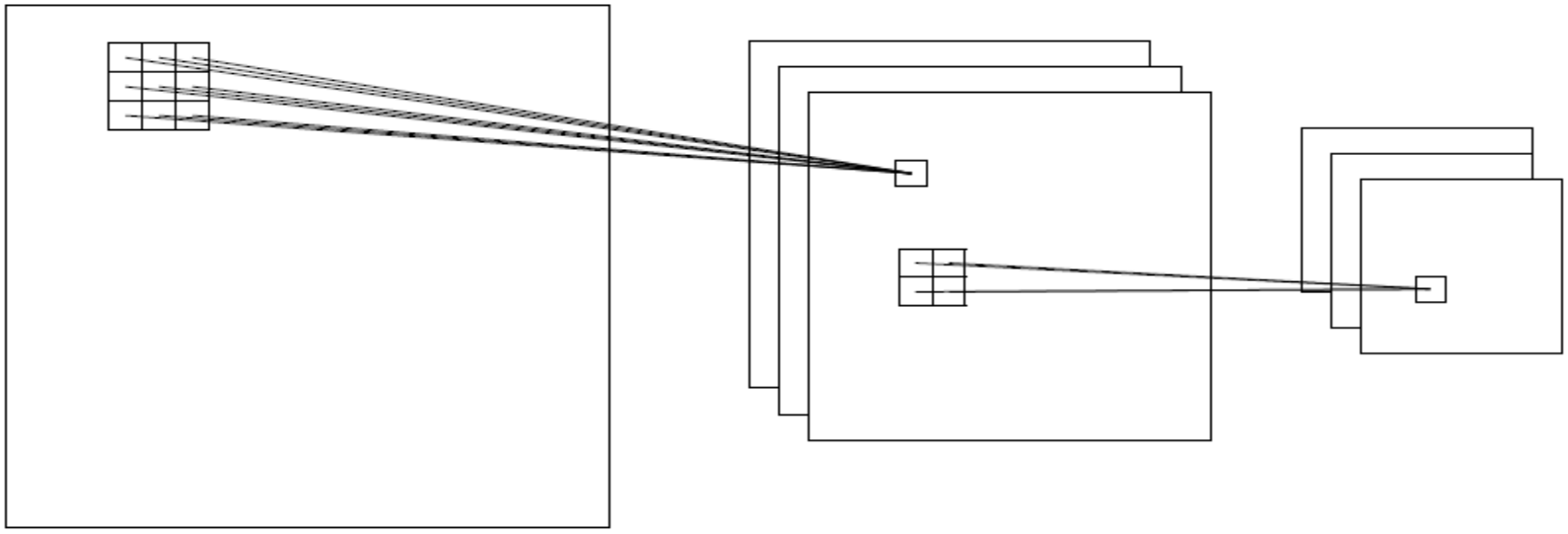


Unsupervised learning



Convolution level




convolution
pooling

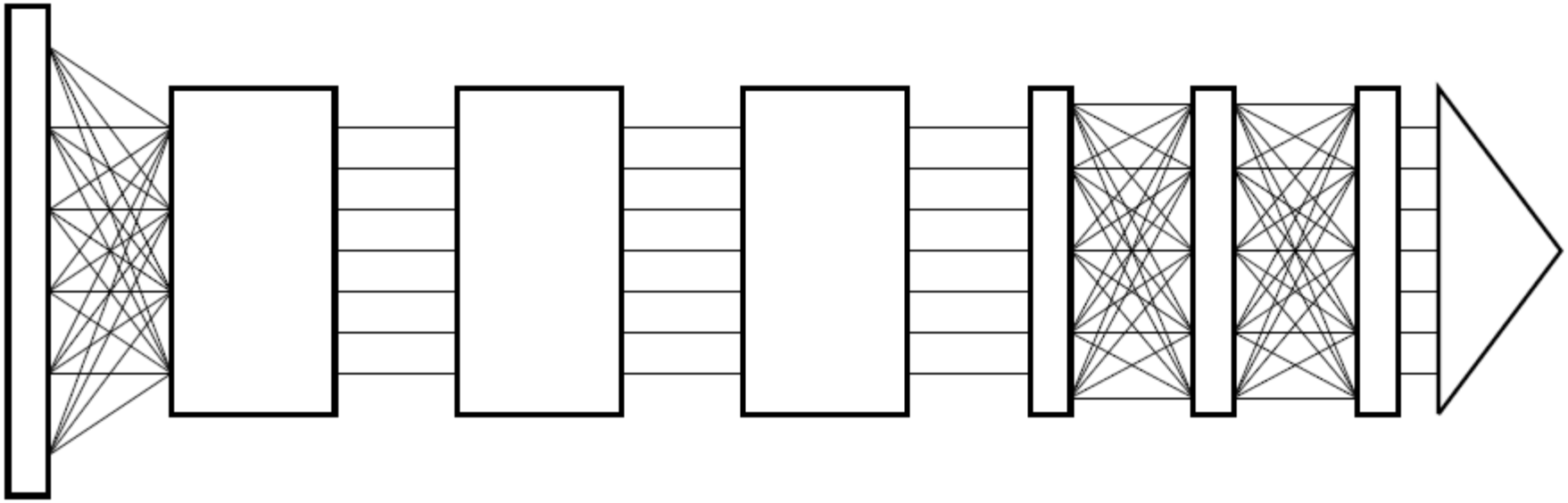
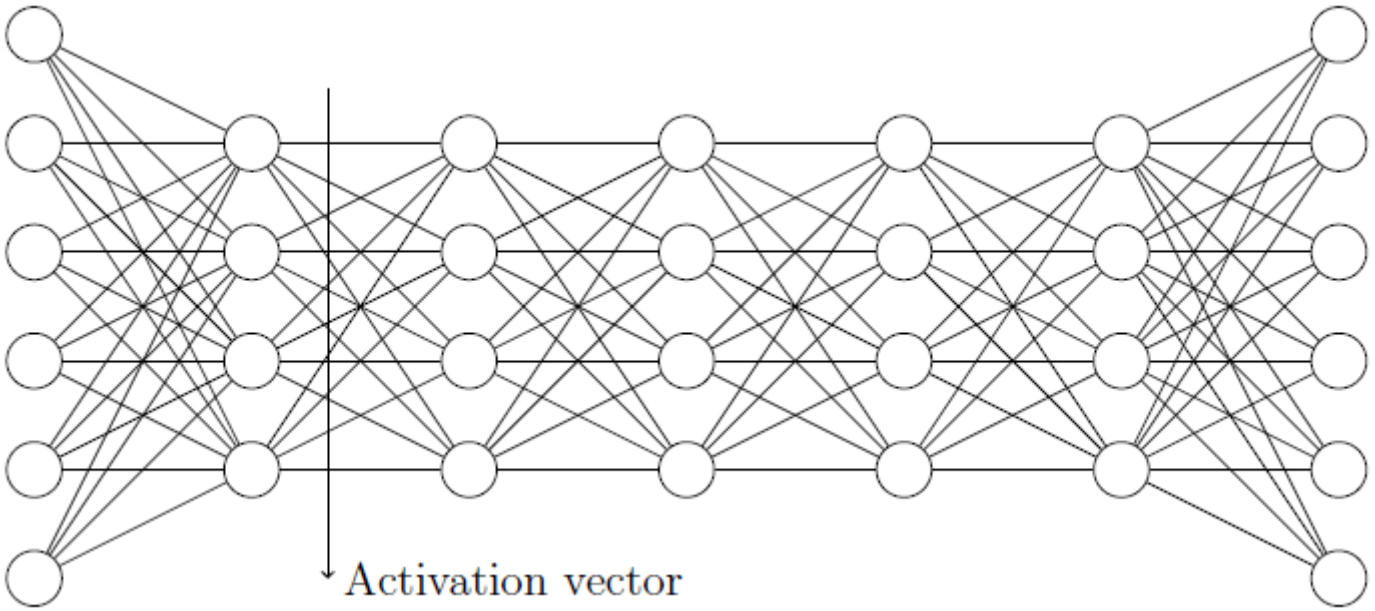
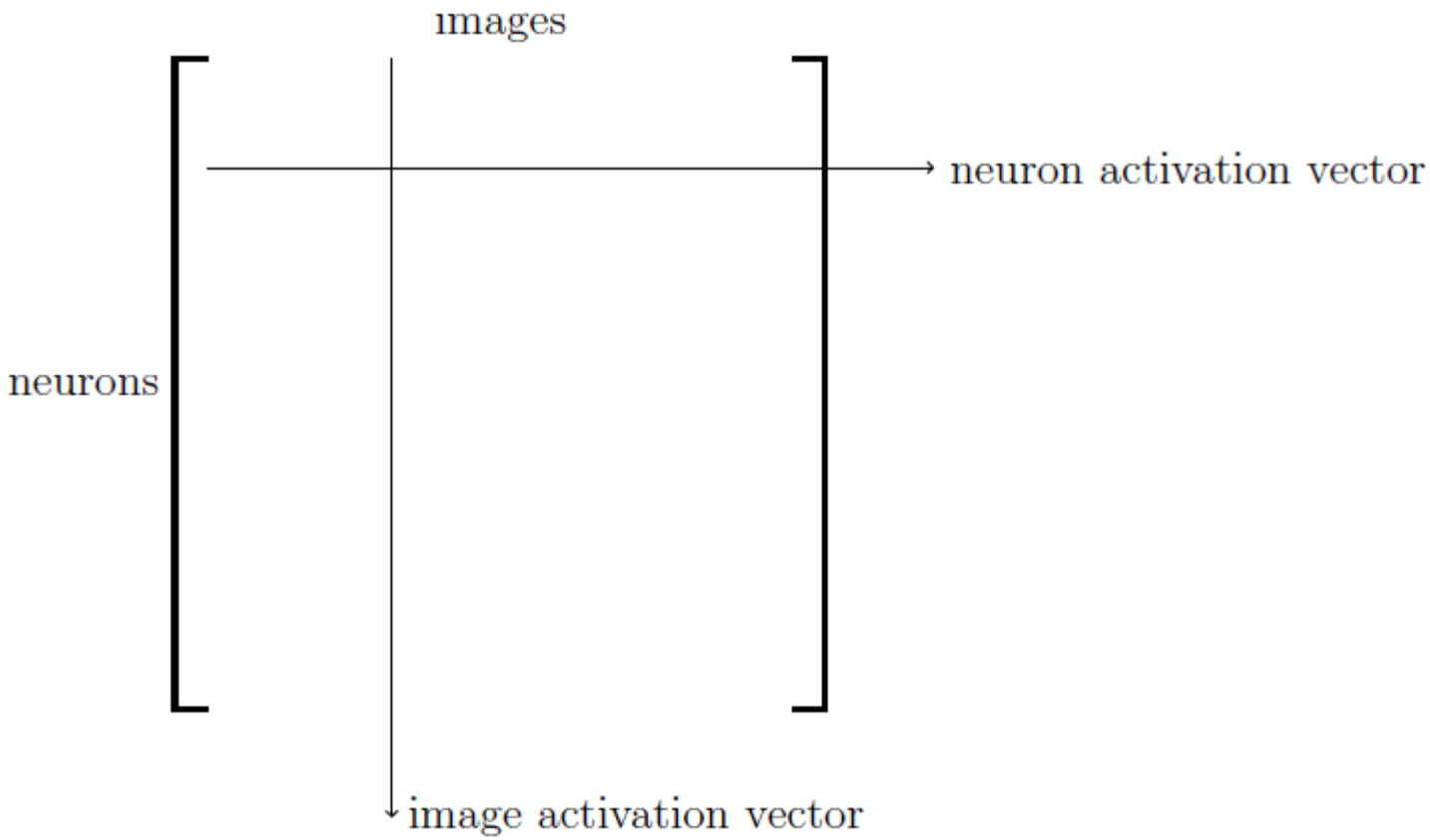


Image
Convolution levels
Fully connected levels
Softmax

Activation space



Activation space



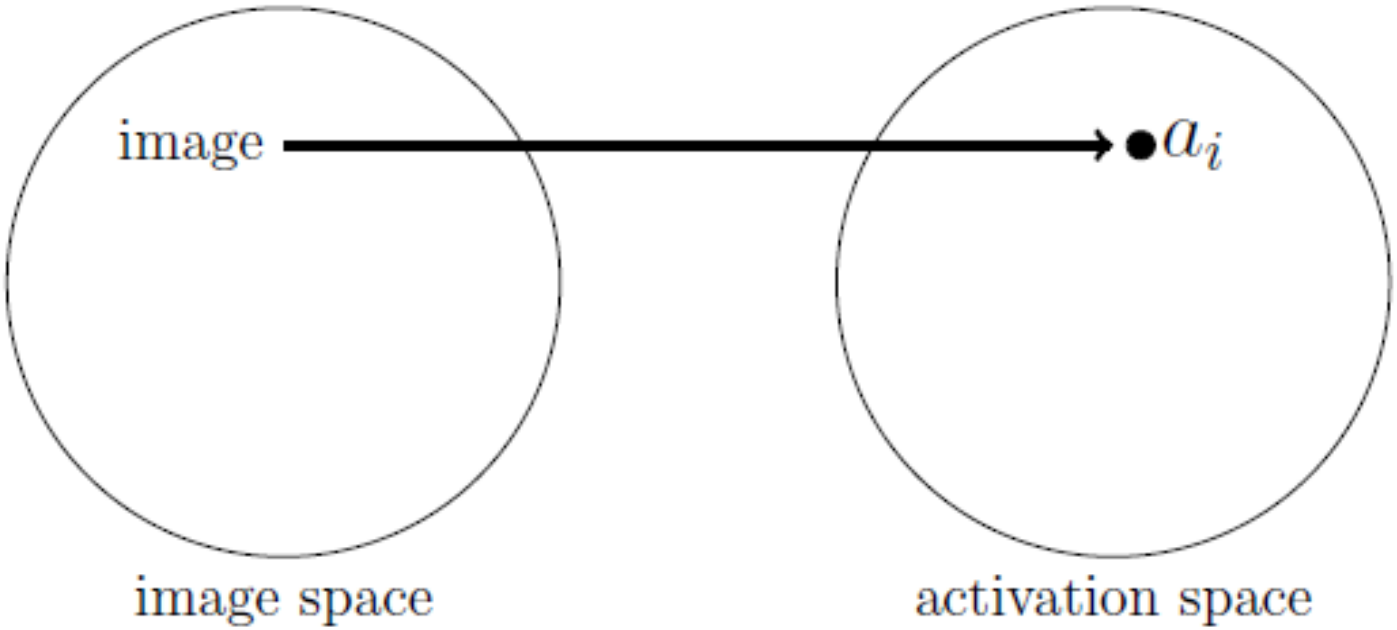
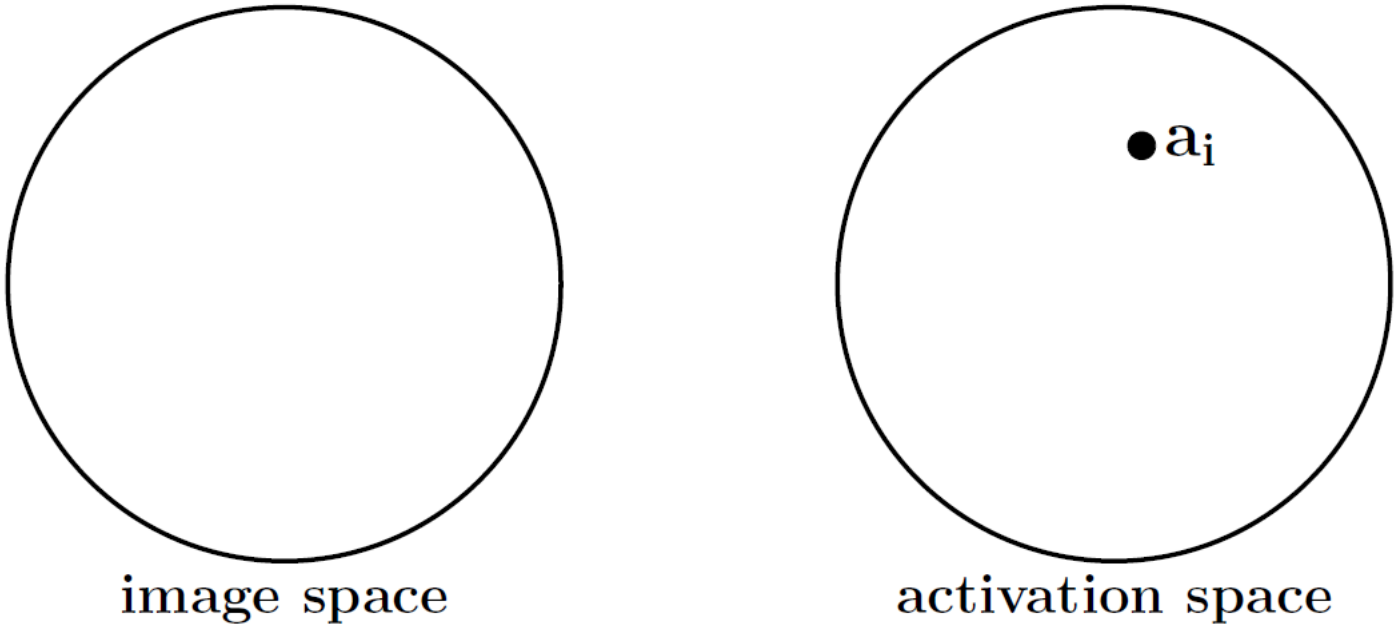


image
a_i
image space
activation space


a_i
image space
activation space

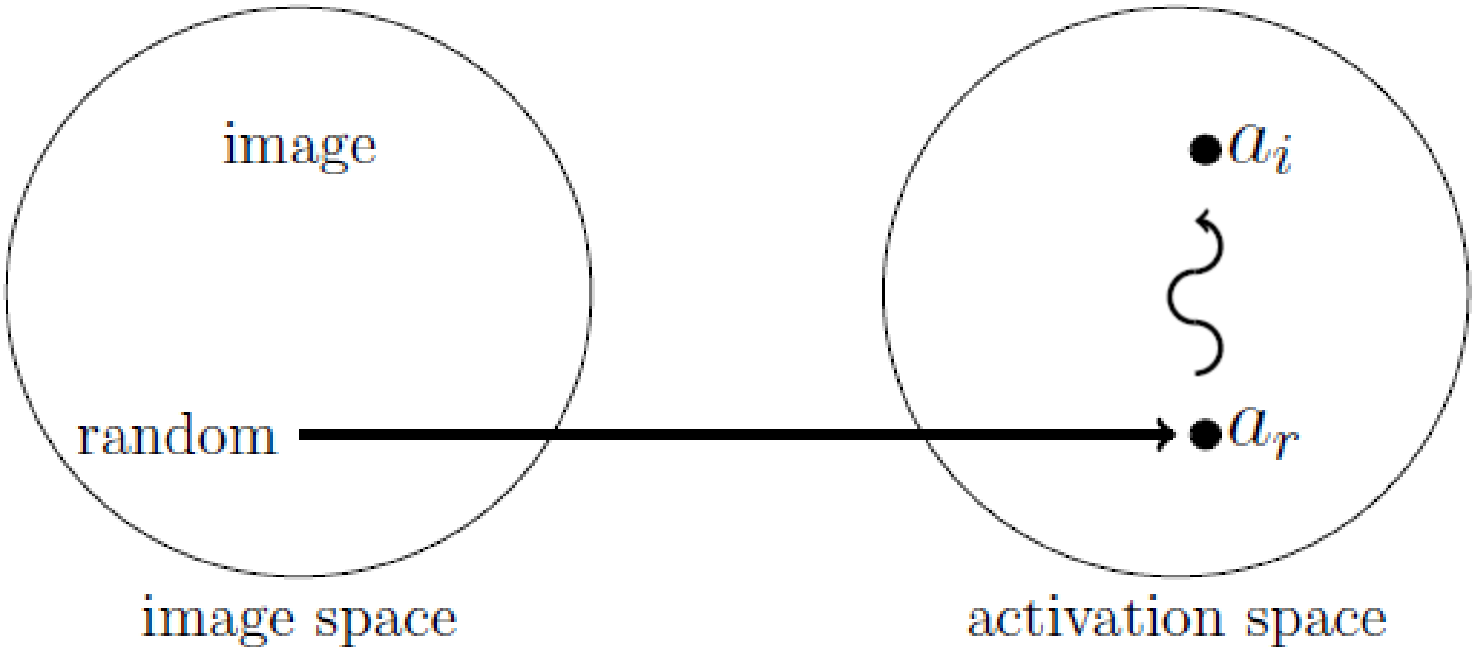
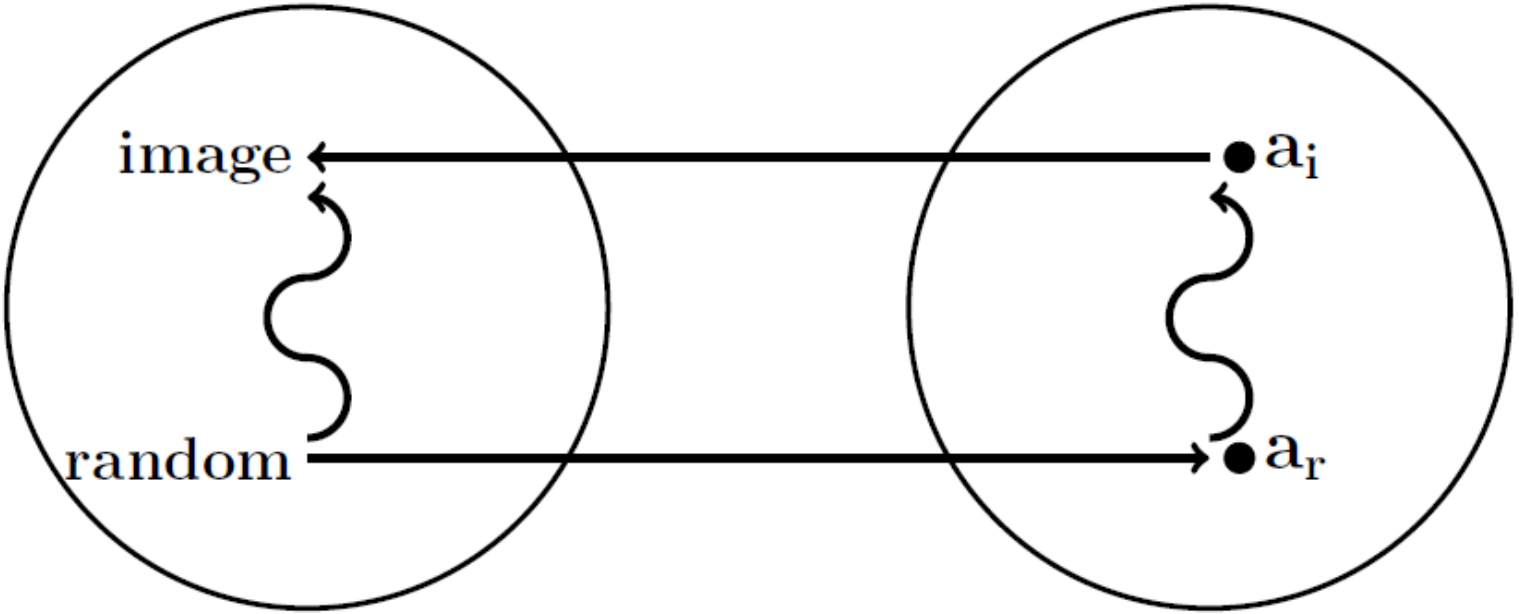
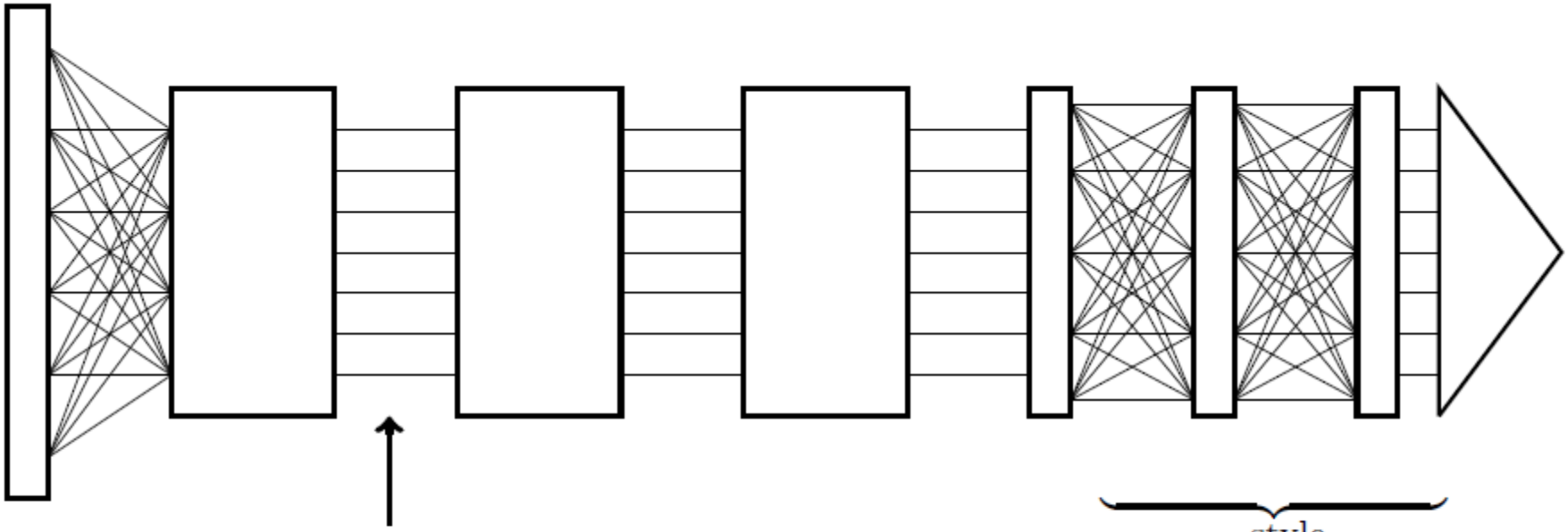


image
random
a_i
a_r
image space
activation space

Image from activation vector





$\lambda_1 \times$ content difference $+ \lambda_2 \times$ style difference

Changing young to old

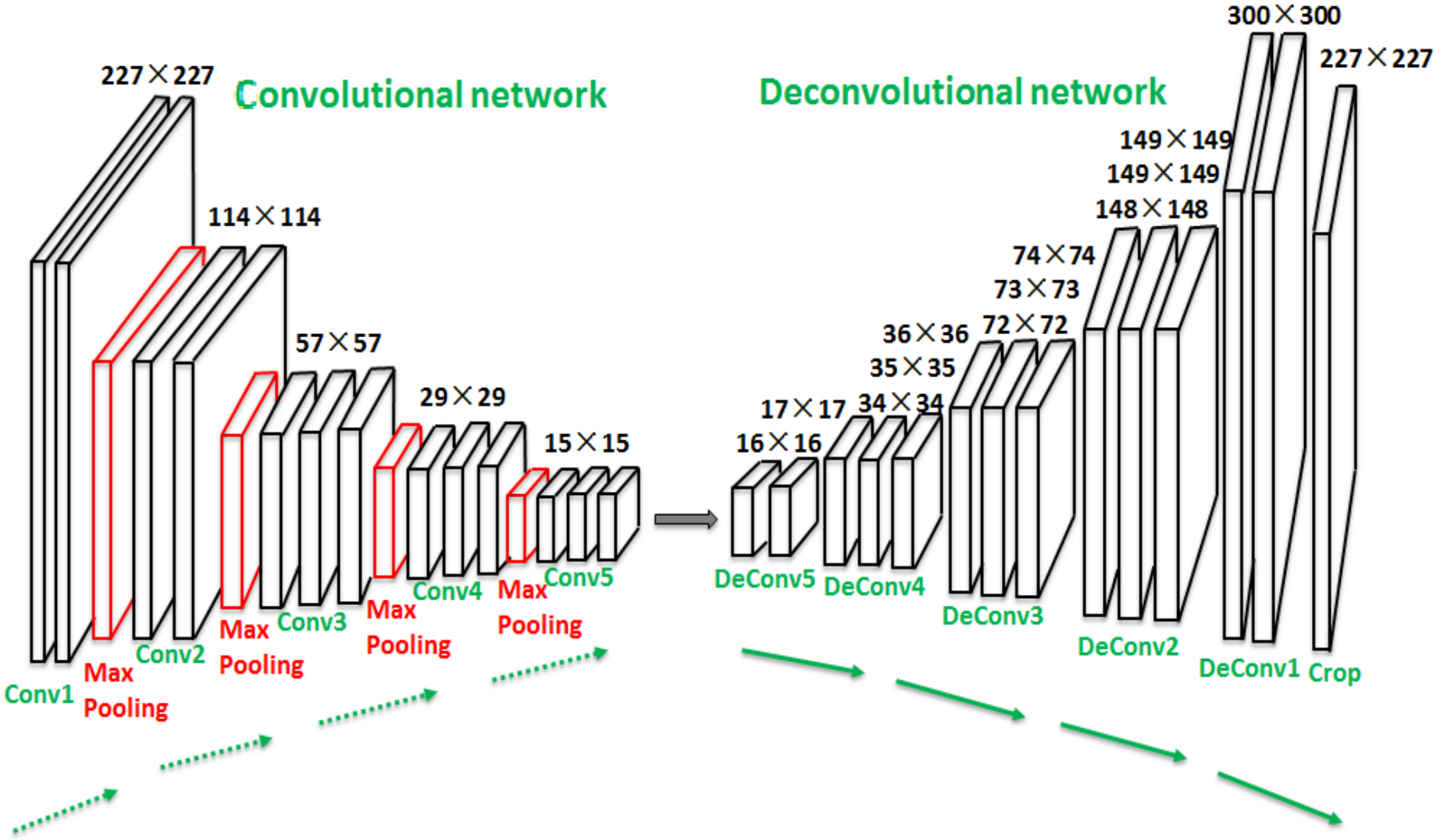
Jacob R. Gardner*, Paul Upchurch*, Matt J. Kusner, Yixuan Li, Kilian Q. Weinberger, Kavita Bala, John E. Hopcroft

Chinese Painting Meets Cornell Campus

* Top left: Elegant photo taken by Parvez Sukheswalla.

* Bottom left: Guanzhong Wu, *Bridges*, 1985.

Convolution Deconvolution



Conv1 Conv2 Conv3 Conv4 Conv5

rwAlexNet

Style transfer on untrained deep networks

[Kun He, Yang Wang, John Hopcroft, NIPS 2016] A Powerful Generative Model Using Random Weights for the Deep Image Representation.

Some research questions

- What do individual gates learn?
- How does what the second level gates learn differ from what the first level gates learn?
- How does what a gate learns evolve over time?
- Train with two different sets of starting weights. Do gates learn the same things?
- Train two networks with different sets of photographs. Do early gates learn the same things?

Do two gates learn the same thing?

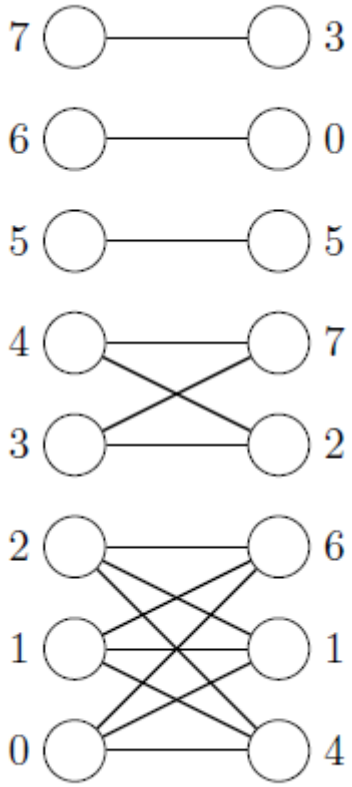
Consider two gates a_i and a_j

$$\text{covariance}(a_i, a_j) = \sum_{\text{images}} \frac{(a_{i,\text{image}} - u_i)(a_{j,\text{image}} - u_j)}{\sigma_i \sigma_j}$$

Do two gates learn the same thing?

Gates in network 2	Gates in network 1: 1	2	3	4	5	6	7	8
1	0.0304	-0.0108	-0.0296	-0.0214	-0.0576	0.0072	-0.0493	-0.0162
2	0.0036	0.0069	0.0227	0.0011	-0.0061	-0.0130	0.0097	-0.0220
3	0.0227	-0.0072	0.1090	-0.0110	0.0415	-0.0135	0.0136	0.2901
4	-0.0062	0.0032	-0.0007	-0.0362	0.0604	0.0037	0.0536	0.1028
5	-0.0177	0.0748	-0.0332	-0.0533	-0.0397	0.0179	-0.0204	0.0610
6	-0.0124	0.0076	0.0332	0.0314	0.0348	-0.0305	0.0296	-0.0003
7	0.0046	0.0034	-0.0301	0.0390	-0.0298	0.0137	0.0103	-0.0365
8	-0.0537	-0.0636	-0.0759	0.0419	-0.0746	0.0019	0.0518	-0.0629

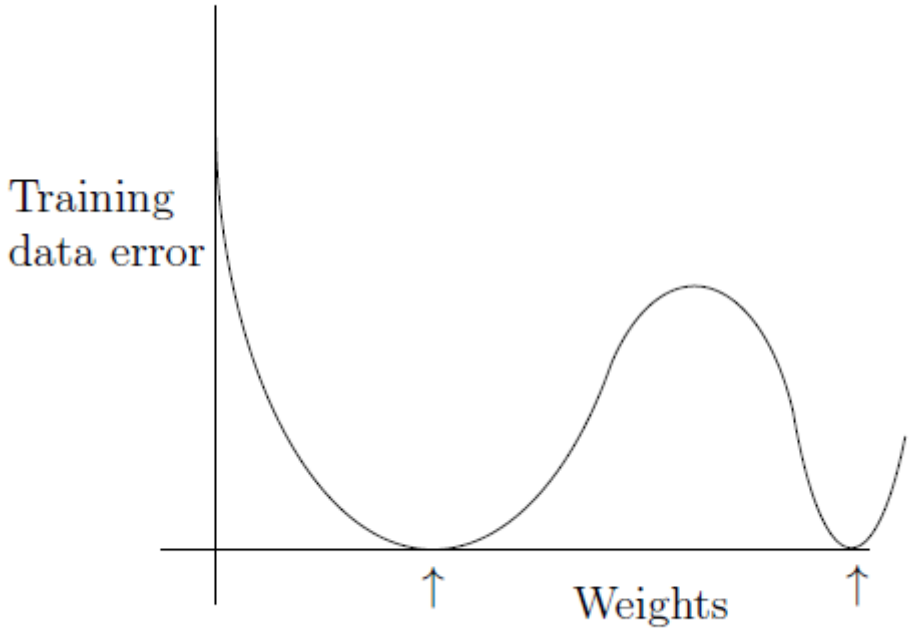
Matching gates in two networks



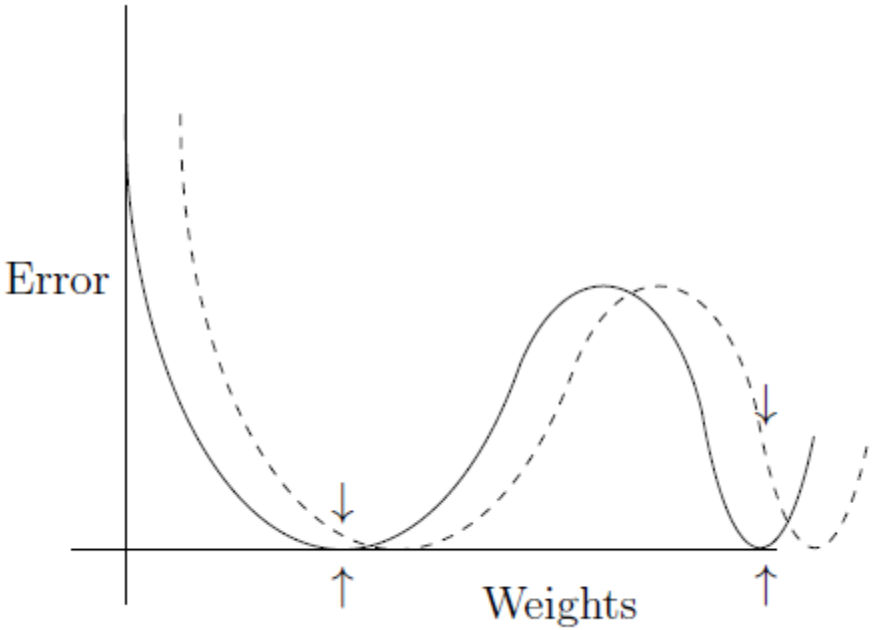
Training a deep network

- Many local minima
- Some are better than others.
- Training takes a long time, can we possibly speed it up?

Multiple local minima



Broad minima seem to be better.



Versions of gradient descent

$$J(w) = \min_{w} \left(\sum_{images} error(w, image) \right)$$

Gradient descent

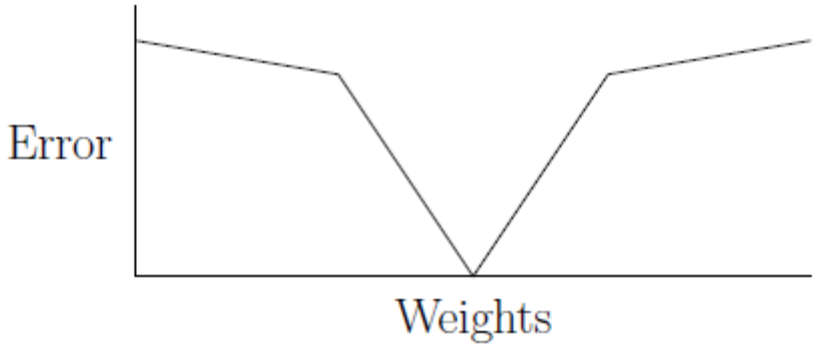
$$w \leftarrow w - \nabla J(w)$$

Stochastic gradient descent

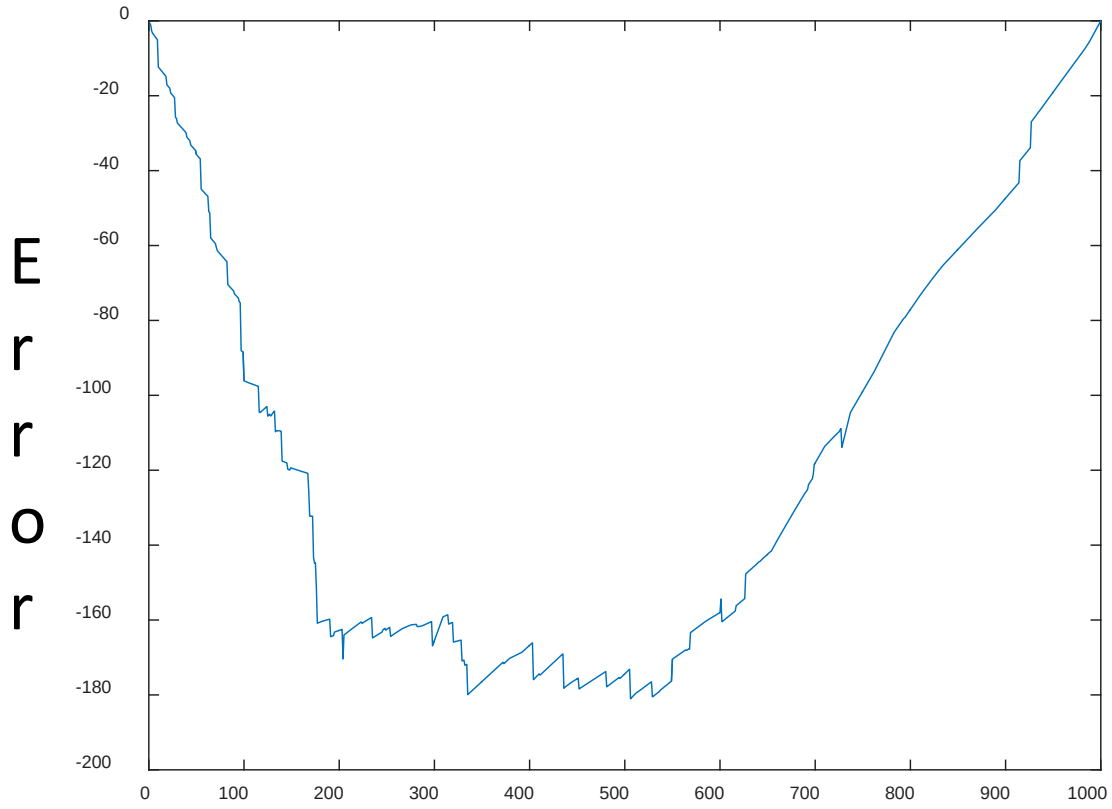
One image at a time

Small batch of images

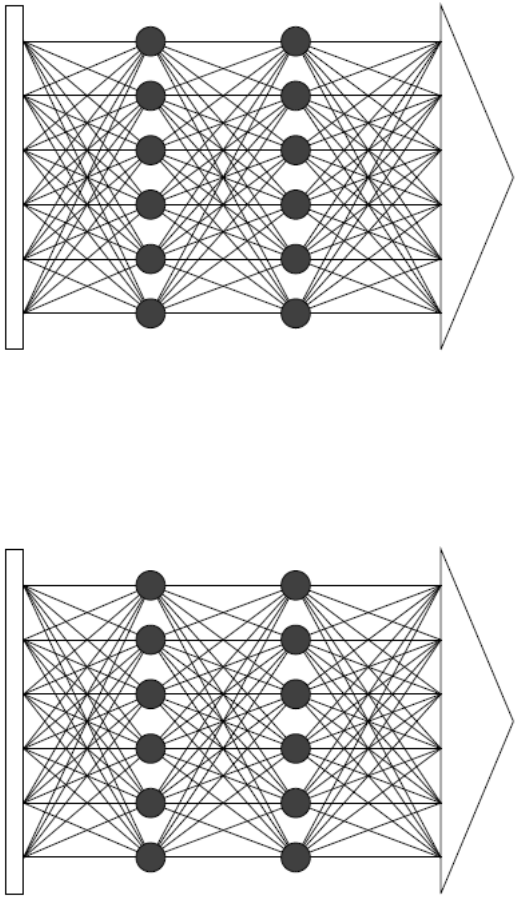
Gradient descent



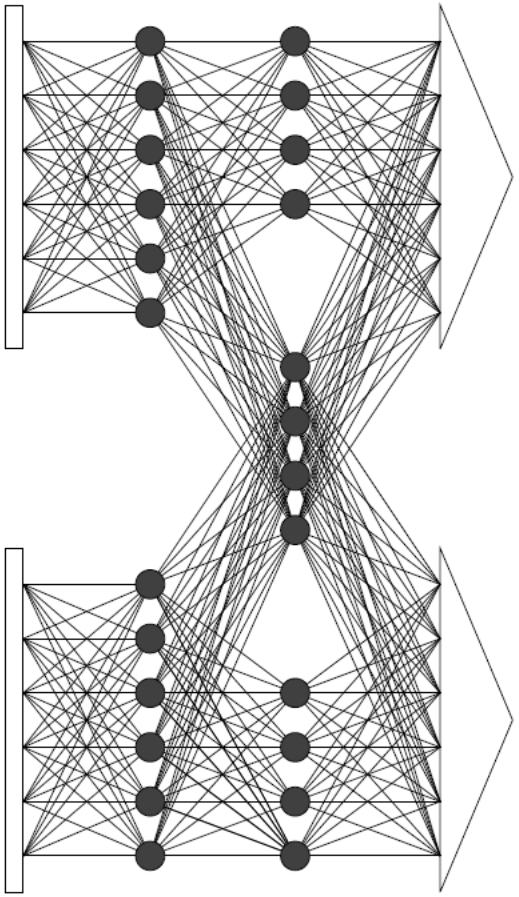
Error function for a single image. Actual error function is the sum over thousands of single image error functions.


Error
0
-20
-40
-60
-80
-100
-120
-140
-160
-180
-200
0
100
200
300
400
500
600
700
800
900
1000

Learning two tasks separately



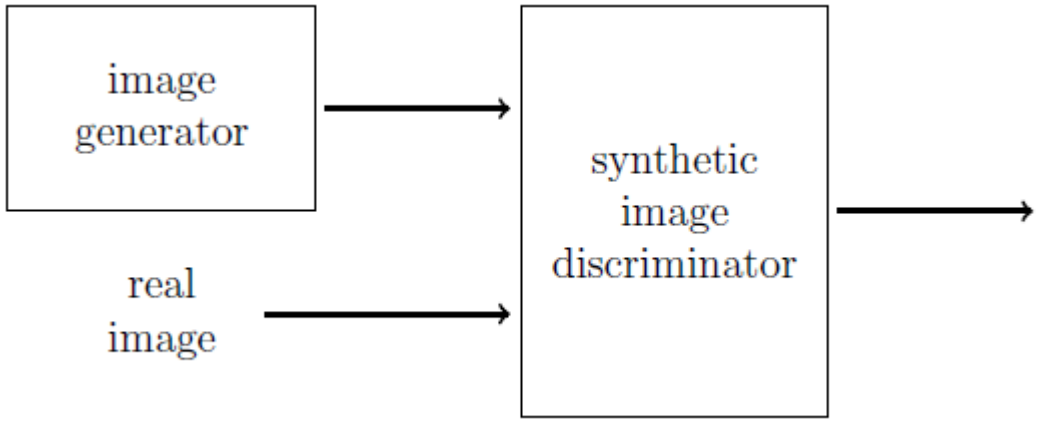
What is common to the two tasks?



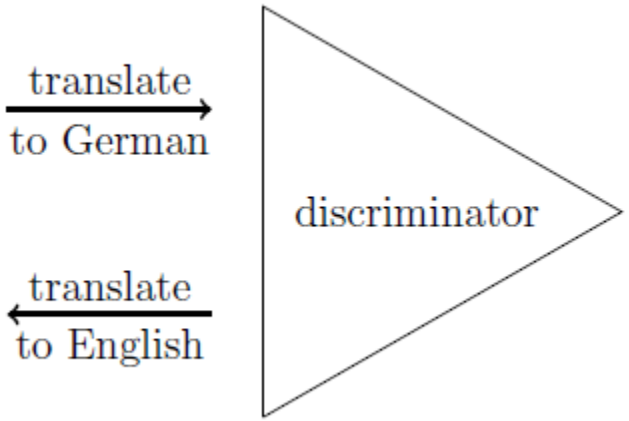
Learn one task first and then another or learn both together

- If we learn one task first and then add an additional task are the two tasks in different regions of activation space?
- If we learn them together do they share the same region in activation space?

Generative Adversarial Networks



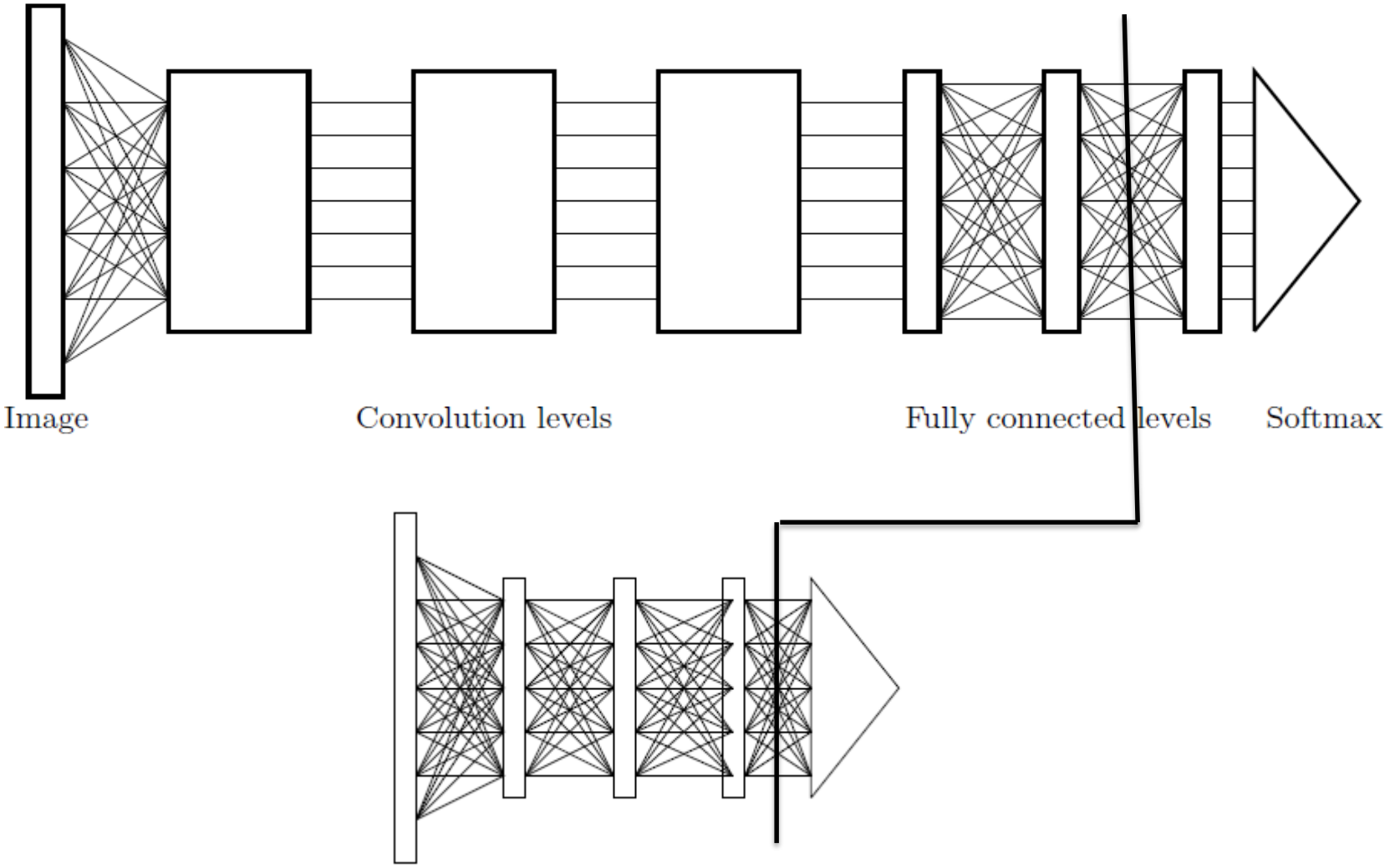
Automatic language translator



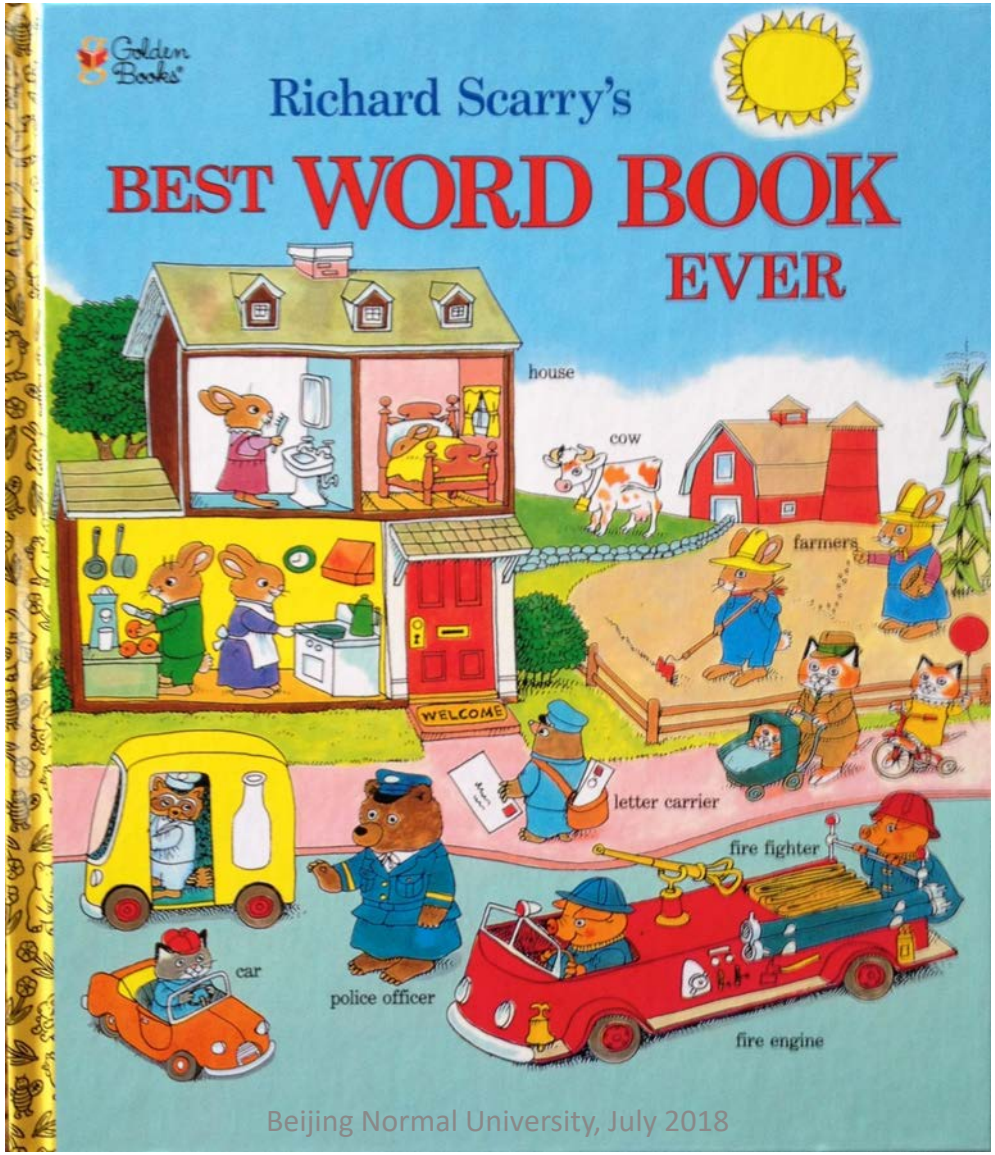
First train the discriminator to determine if a sentence is a real sentence in German as opposed to a synthetic sentence.

Then train a translator for English to German and a translator from German to English.

Compression



Learning from a single image




CLAY TWP
FIRE DEPT
ENGINE
241

Fooling deep learning

Cat

Automobile

Is Artificial Intelligence Real?

No!

At the current state artificial intelligence is pattern recognition in high dimensional space.

AI programs do not extract the essence of an object and understanding its function or other important aspects.

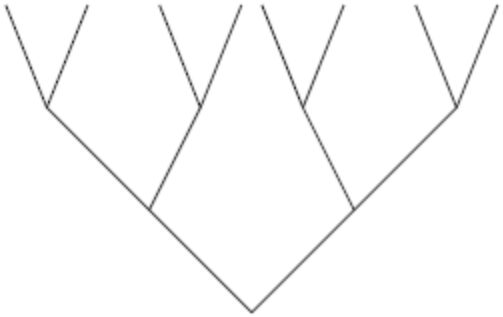
Another revolution in 40 years may accomplish that.

Is artificial intelligence real?

Computing power

Not all intelligent-like tasks need AI.
Some just need computing power and access to large data.

Chess



Tree of board positions given move in chess

Computers are doing more and more things that one thought required intelligence.

Thank you

CIT 互联网教育智能技术及应用国家工程实验室

http://cit.bnu.edu.cn cit@bnu.edu.cn 010-58807205

北京市海淀区学院南路12号 北京师范大学南院 京师科技大厦A座3层和12层

扫描二维码 关注公众号

THANKS