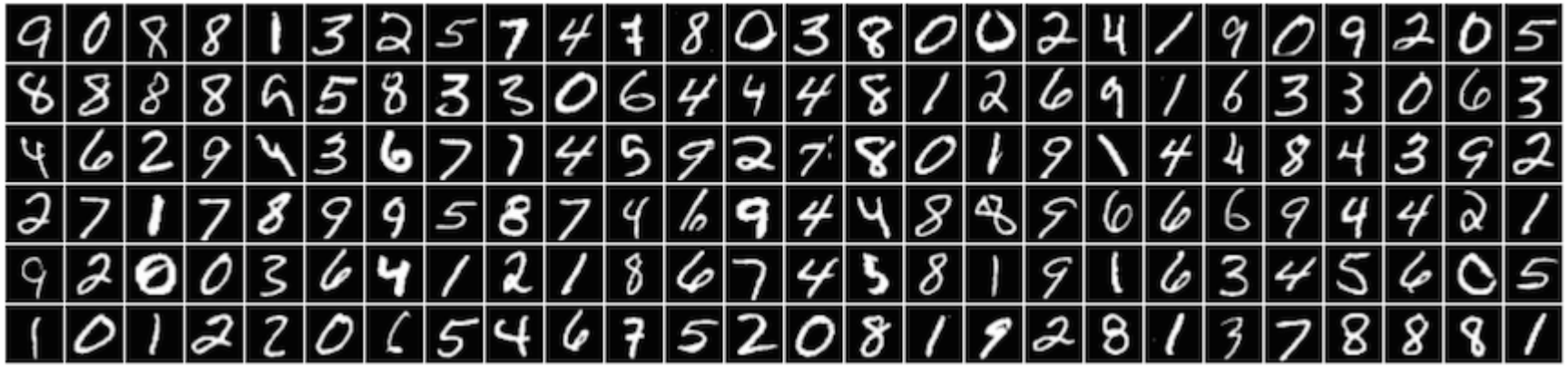
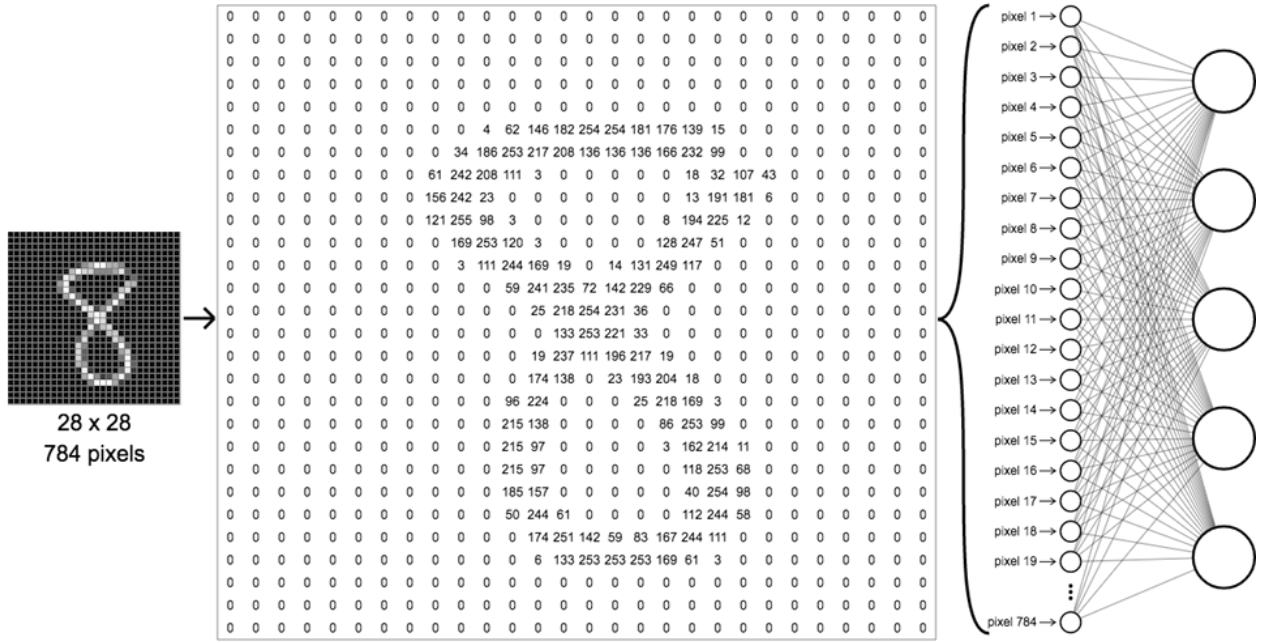
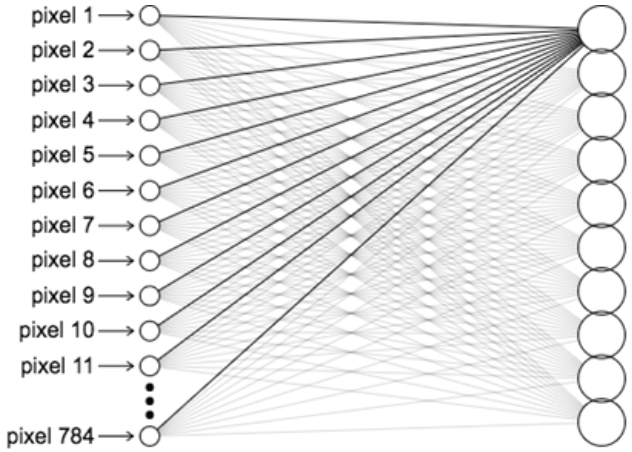
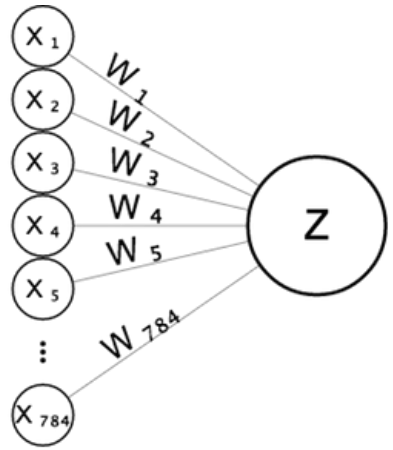
MNIST
dataset



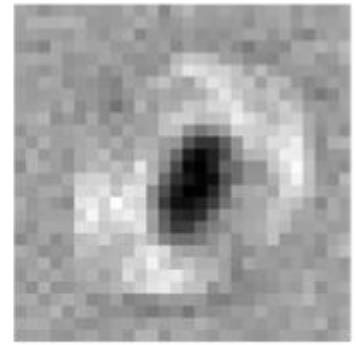


https://ml4a.github.io/ml4a/neural_networks/

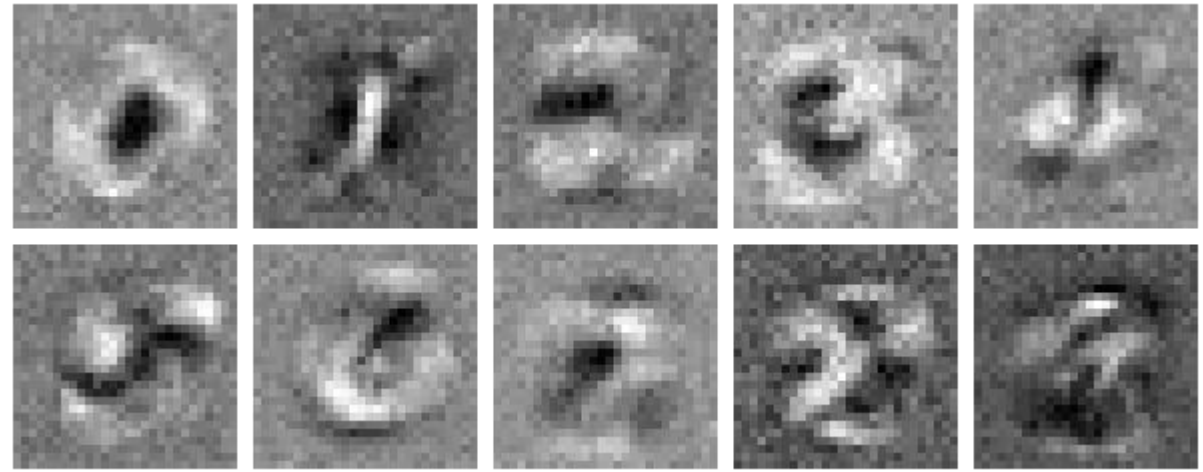




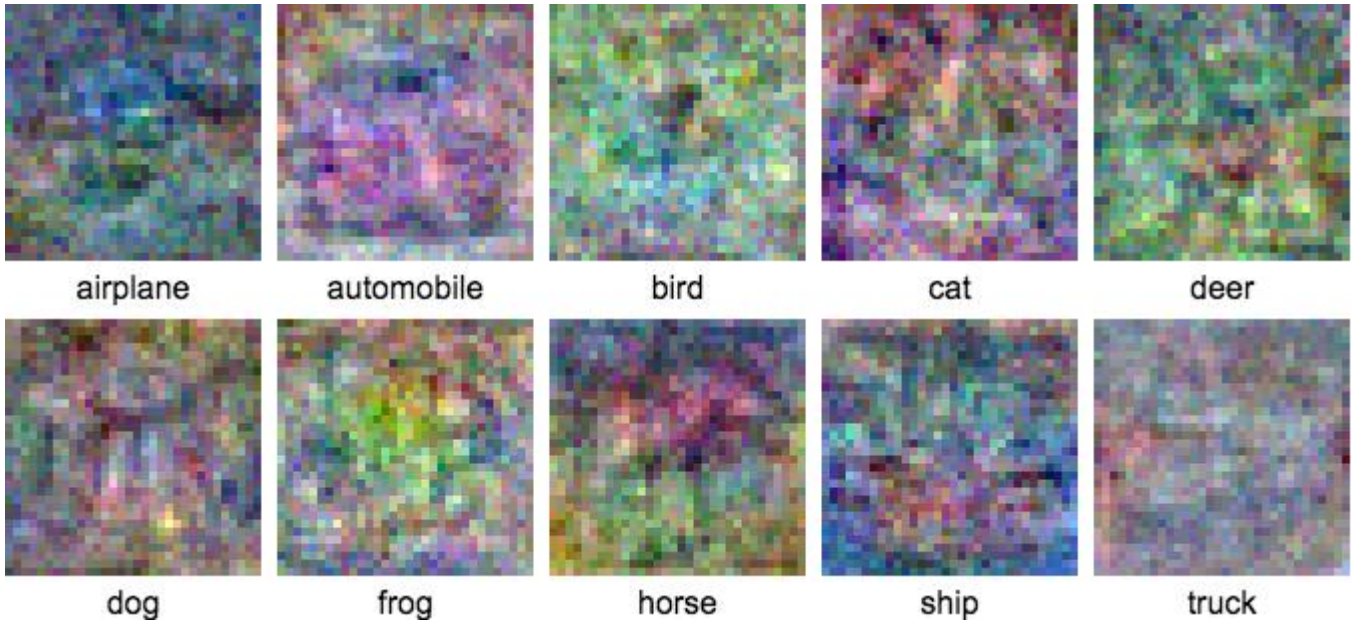
$$\begin{bmatrix} W_1 & W_2 & W_3 & \cdots & W_{28} \\ W_{29} & W_{30} & W_{31} & & W_{56} \\ W_{57} & W_{58} & W_{59} & & W_{84} \\ \vdots & & & \ddots & \\ W_{757} & W_{758} & W_{759} & & W_{784} \end{bmatrix}$$



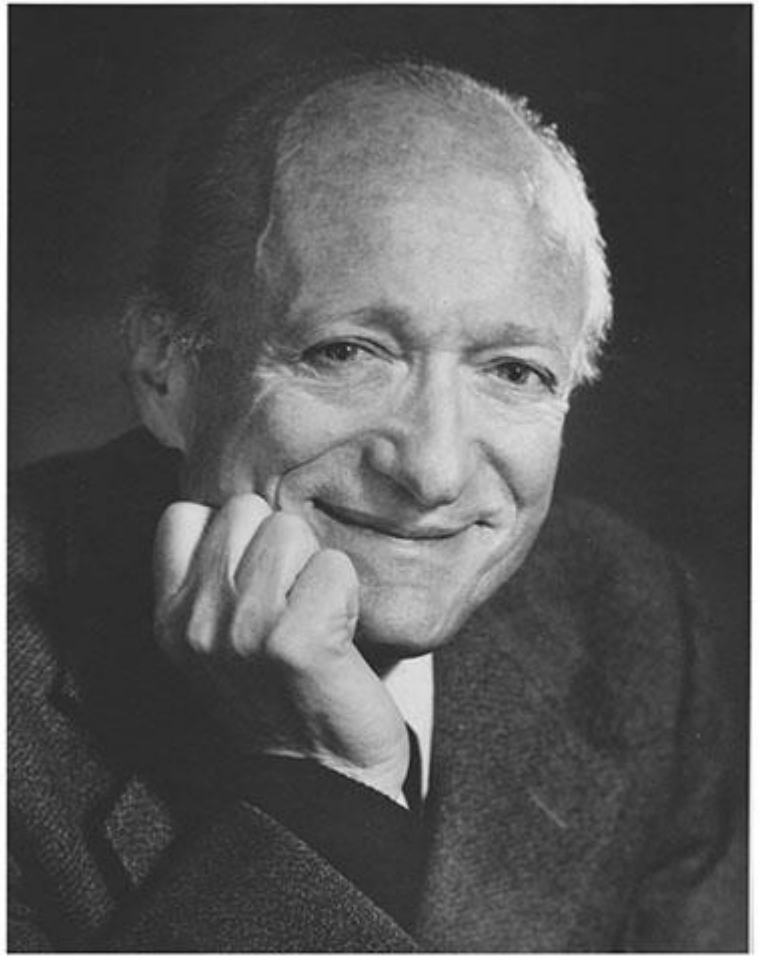
$$\begin{bmatrix} X_1 & X_2 & X_3 & \cdots & X_{28} \\ X_{29} & X_{30} & X_{31} & & X_{56} \\ X_{57} & X_{58} & X_{59} & & X_{84} \\ \vdots & & & \ddots & \\ X_{757} & X_{758} & X_{759} & & X_{784} \end{bmatrix} \circ \begin{bmatrix} W_1 & W_2 & W_3 & \cdots & W_{28} \\ W_{29} & W_{30} & W_{31} & & W_{56} \\ W_{57} & W_{58} & W_{59} & & W_{84} \\ \vdots & & & \ddots & \\ W_{757} & W_{758} & W_{759} & & W_{784} \end{bmatrix} = Z$$



CIFAR-10, a labeled set of 60,000 32x32 color images belonging to ten classes: airplanes, automobiles, birds, cats, deer, dogs, frogs, horses, ships, and trucks.



Weight matrix



Stephen Kuffler (1913 – 1980)

Often referred to as the "Father of Modern Neuroscience"

David Hubel and Torsten Wiesel

1981 Nobel Prize in Physiology or Medicine for their discoveries concerning information processing in the visual system

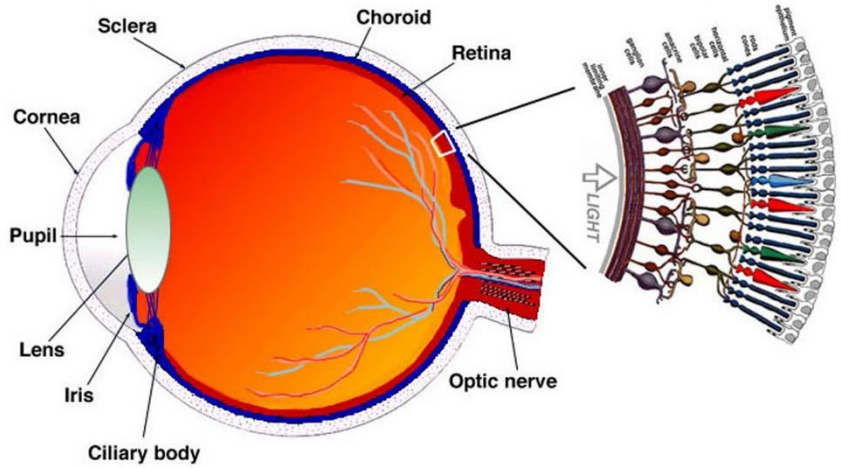
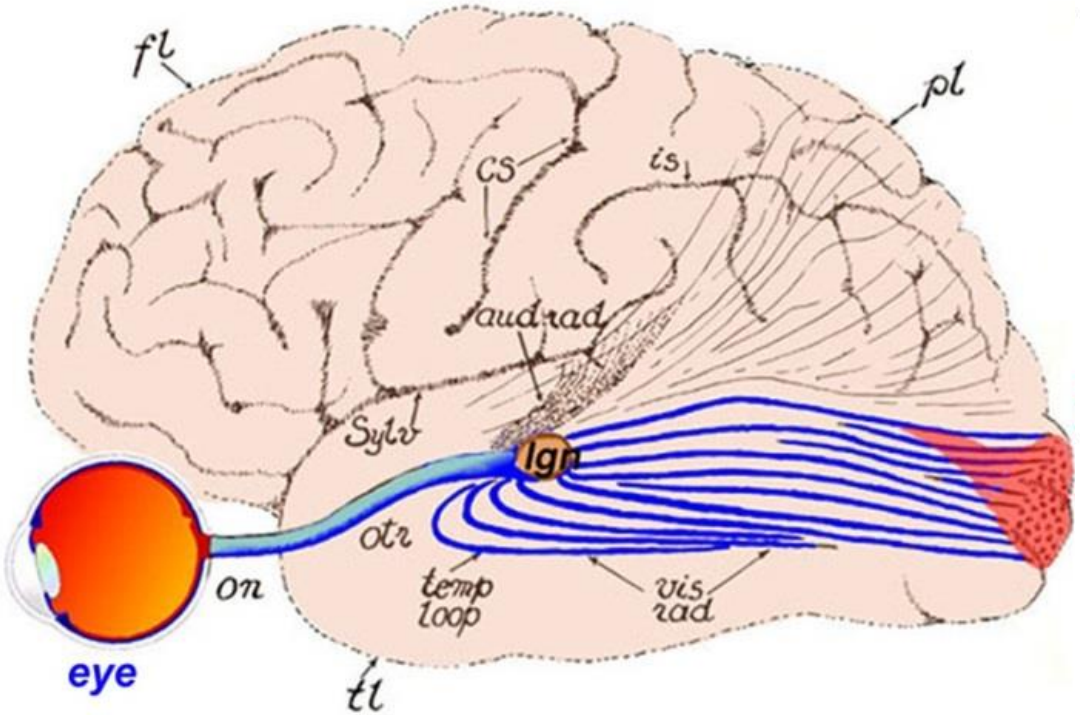
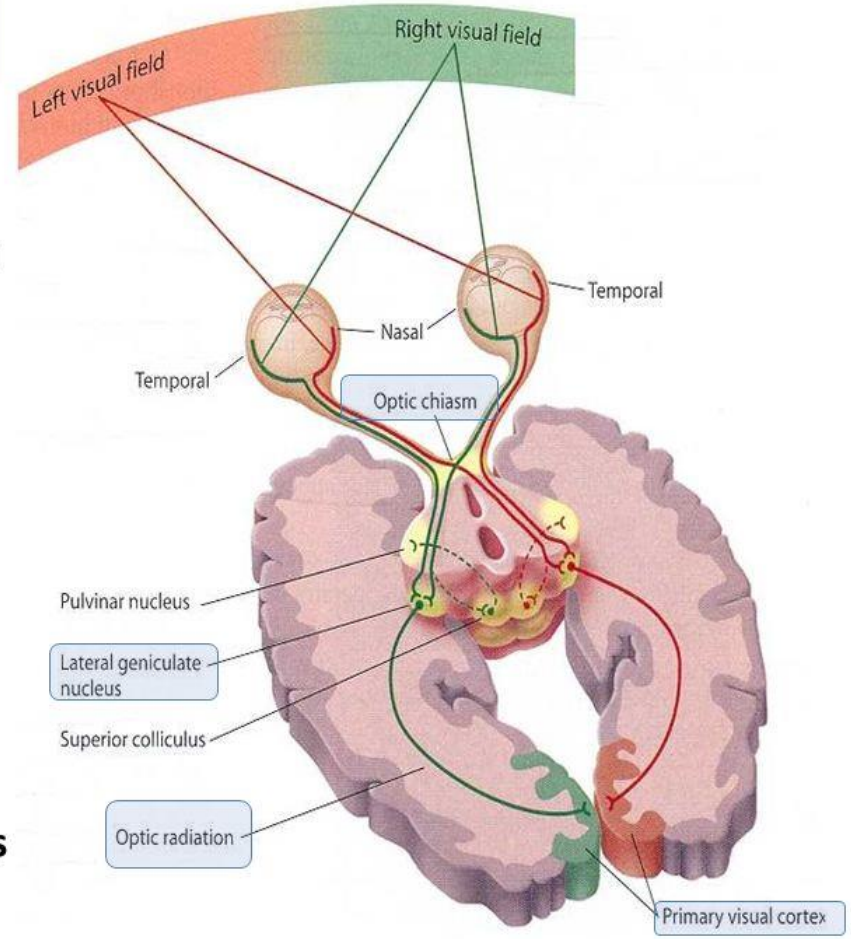


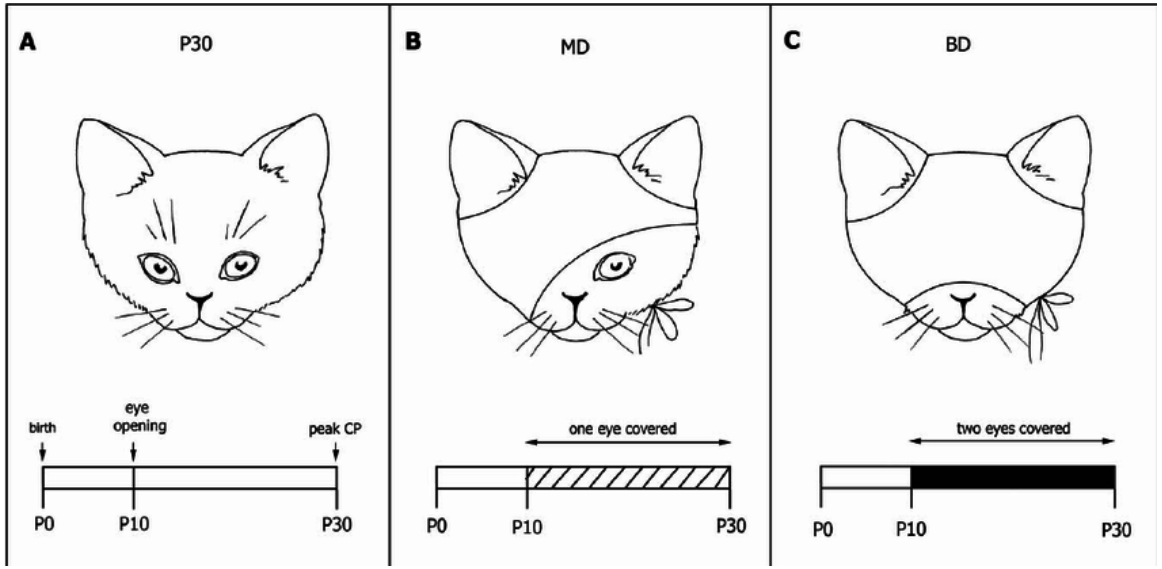
Fig. 1.1. A drawing of a section through the human eye with a schematic enlargement of the retina.

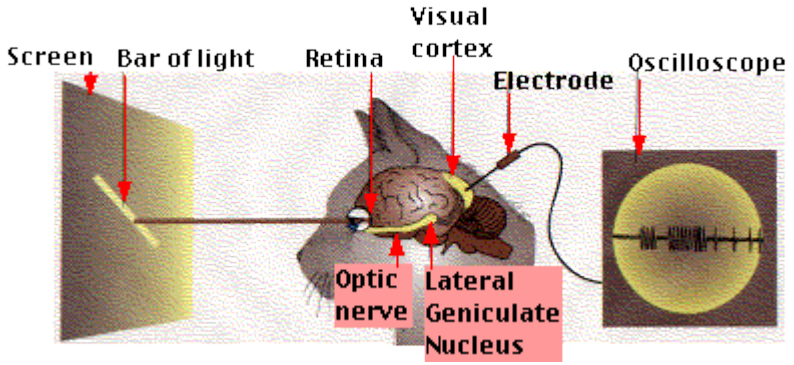


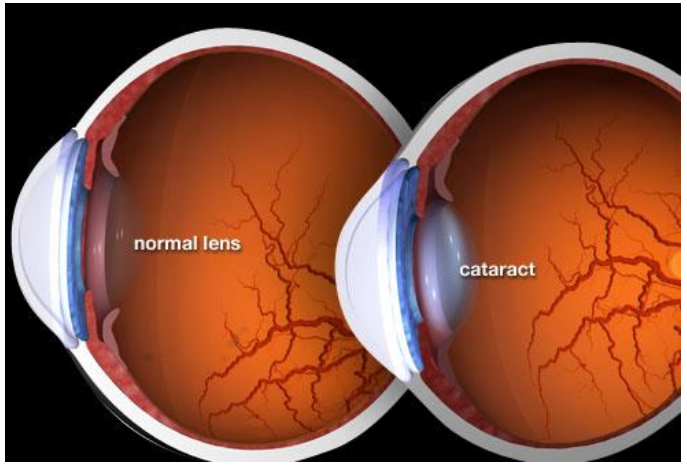
Visual Pathway

1. Cones
2. Bipolar neurons
3. Ganglion cell's axon forms the optic nerve
4. Optic nerve to the Optic Chiasm
5. Optic tract
6. Lateral geniculate nuclei of the thalamus
7. Optic Radiations
8. Primary visual areas of the occipital lobes

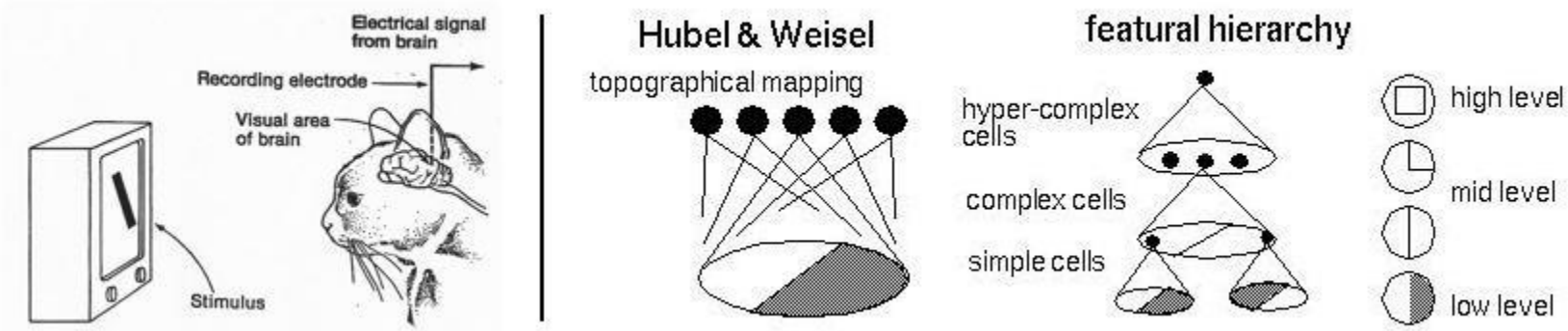






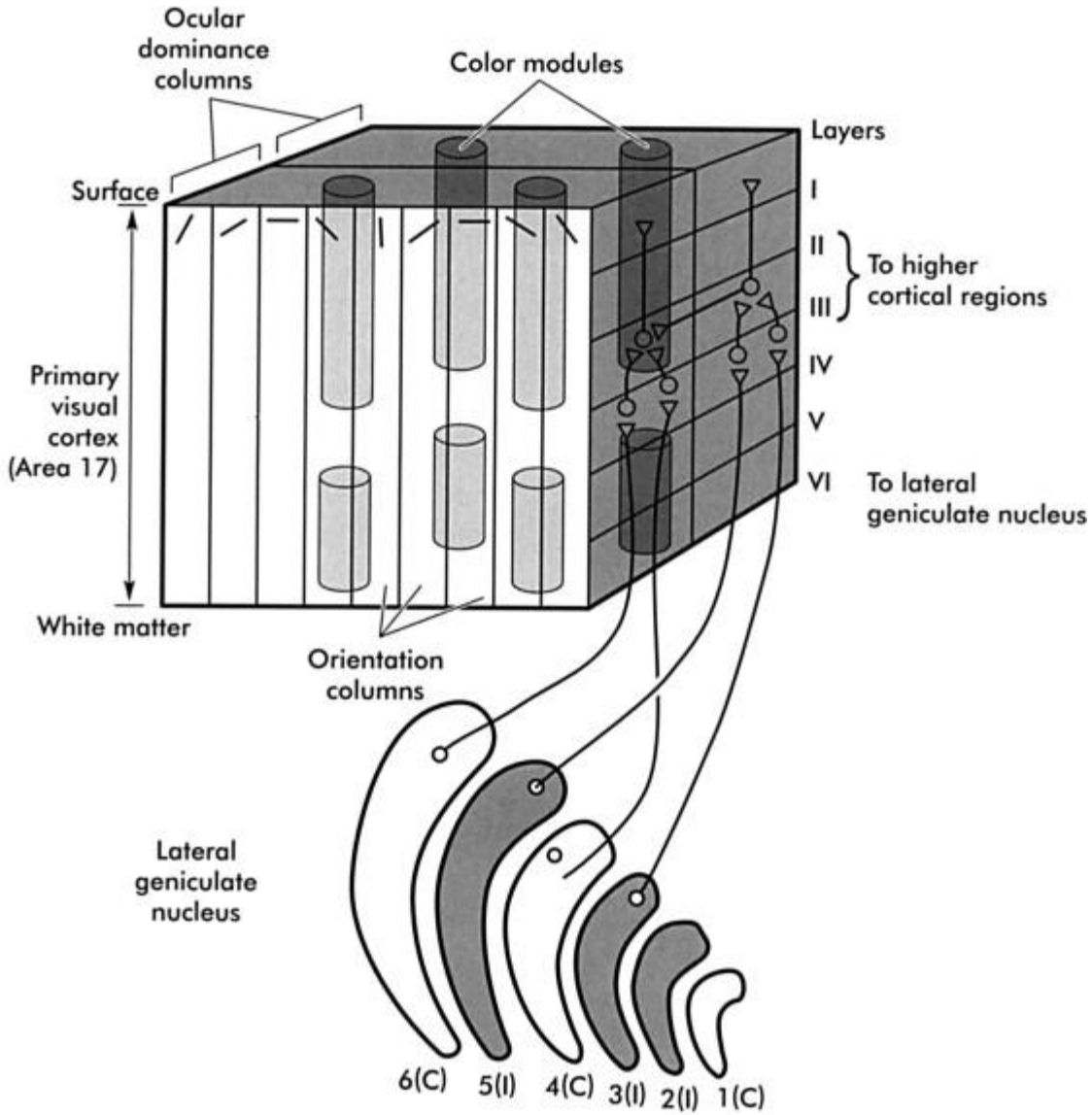


https://knowingneurons.com/2014/10/29/hubel-and-wiesel-the-neural-basis-of-visual-perception/
http://centennial.rucares.org/index.php?page=Neural_Basis_Visual_Perception

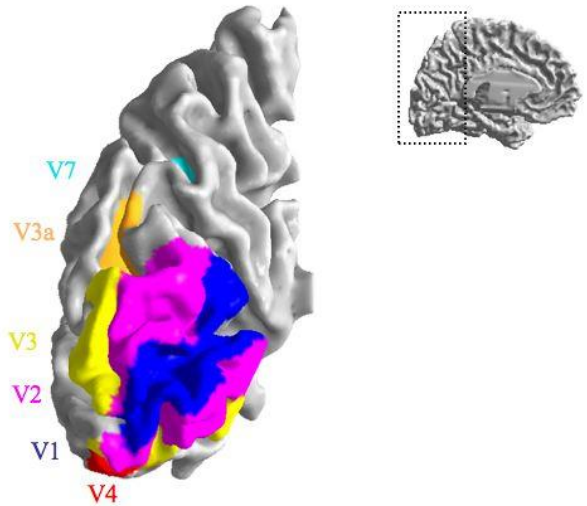


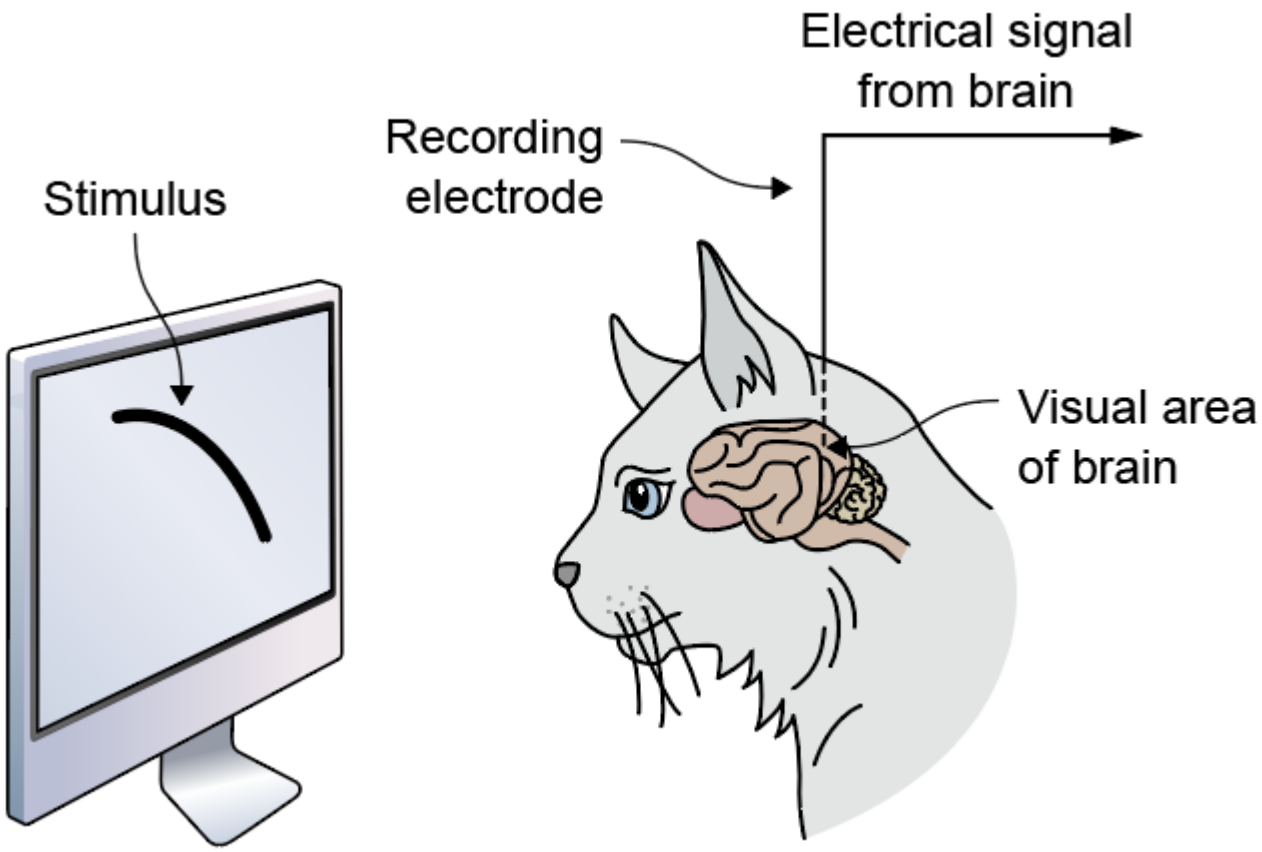
- Hubel and Weisel measured the electrical responses from a cat's brain while stimulating it with simple patterns on a television screen.
- They found was that neurons in the early visual cortex are organized in a hierarchical fashion, where the first cells connected to the cat's retinas are responsible for detecting simple patterns like edges and bars, followed by later layers responding to more complex patterns by combining the earlier neuronal activities.
- Their experiments would provide an early inspiration to artificial intelligence researchers seeking to construct well-defined computational frameworks for computer vision.

Ref : https://ml4a.github.io/ml4a/convnets/



Primary visual cortex (V1)





J. Physiol. (1962), **160**, *pp.* 106–154
With 2 plates and 20 text-figures
Printed in Great Britain

RECEPTIVE FIELDS, BINOCULAR INTERACTION AND FUNCTIONAL ARCHITECTURE IN THE CAT'S VISUAL CORTEX

BY D. H. HUBEL AND T. N. WIESEL

From the Neurophysiology Laboratory, Department of Pharmacology Harvard Medical School, Boston, Massachusetts, U.S.A.

(*Received* 31 *July* 1961)

What chiefly distinguishes cerebral cortex from other parts of the central nervous system is the great diversity of its cell types and interconnexions. It would be astonishing if such a structure did not profoundly modify the response patterns of fibres coming into it. In the cat's visual cortex, the receptive field arrangements of single cells suggest that there is indeed a degree of complexity far exceeding anything yet seen at lower levels in the visual system.

In a previous paper we described receptive fields of single cortical cells, observing responses to spots of light shone on one or both retinas (Hubel & Wiesel, 1959). In the present work this method is used to examine receptive fields of a more complex type (Part I) and to make additional observations on binocular interaction (Part II).

This approach is necessary in order to understand the behaviour of individual cells, but it fails to deal with the problem of the relationship of one cell to its neighbours. In the past, the technique of recording evoked slow waves has been used with great success in studies of functional anatomy. It was employed by Talbot & Marshall (1941) and by Thompson, Woolsey & Talbot (1950) for mapping out the visual cortex in the rabbit, cat, and monkey. Daniel & Whitteridge (1959) have recently extended this work in the primate. Most of our present knowledge of retinotopic projections, binocular overlap, and the second visual area is based on these investigations. Yet the method of evoked potentials is valuable mainly for detecting behaviour common to large populations of neighbouring cells; it cannot differentiate functionally between areas of cortex smaller than about 1 mm^2. To overcome this difficulty a method has in recent years been developed for studying cells separately or in small groups during long micro-electrode penetrations through nervous tissue. Responses are correlated with cell location by reconstructing the electrode

Kunihiko Fukushima

Hubel and Wiesel's experiments inspired Kunihiko Fukushima in devising the Neocognitron, a neural network which attempted to mimic these hierarchical and compositional properties of the visual cortex.

The neocognitron was the first neural network architecture to use hierarchical layers where each layer is responsible for detecting a pattern from the output of the previous one, using a sliding filter to locate it anywhere in the image.

Biol. Cybernetics 36, 193–202 (1980)

Biological Cybernetics

Neocognitron: A Self-organizing Neural Network Model for a Mechanism of Pattern Recognition Unaffected by Shift in Position

Kunihiko Fukushima

NHK Broadcasting Science Research Laboratories, Kinuta, Setagaya, Tokyo, Japan

Abstract. A neural network model for a mechanism of visual pattern recognition is proposed in this paper. The network is self-organized by "learning without a teacher", and acquires an ability to recognize stimulus patterns based on the geometrical similarity (Gestalt) of their shapes without affected by their positions. This network is given a nickname "neocognitron". After completion of self-organization, the network has a structure similar to the hierarchy model of the visual nervous system proposed by Hubel and Wiesel. The network consists of an input layer (photoreceptor array) followed by a cascade connection of a number of modular structures, each of which is composed of two layers of cells connected in a cascade. The first layer of each module consists of "S-cells", which show characteristics similar to simple cells or lower order hypercomplex cells, and the second layer consists of "C-cells" similar to complex cells or higher order hypercomplex cells. The afferent synapses to each S-cell have plasticity and are modifiable. The network has an ability of unsupervised learning: We do not need any "teacher" during the process of self-organization, and it is only needed to present a set of stimulus patterns repeatedly to the input layer of the network. The network has been simulated on a digital computer. After repetitive presentation of a set of stimulus patterns, each stimulus pattern has become to elicit an output only from one of the C-cells of the last layer, and conversely, this C-cell has become selectively responsive only to that stimulus pattern. That is, none of the C-cells of the last layer responds to more than one stimulus pattern. The response of the C-cells of the last layer is not affected by the pattern's position at all. Neither is it affected by a small change in shape nor in size of the stimulus pattern.

1. Introduction

The mechanism of pattern recognition in the brain is little known, and it seems to be almost impossible to reveal it only by conventional physiological experiments. So, we take a slightly different approach to this problem. If we could make a neural network model which has the same capability for pattern recognition as a human being, it would give us a powerful clue to the understanding of the neural mechanism in the brain. In this paper, we discuss how to synthesize a neural network model in order to endow it an ability of pattern recognition like a human being.

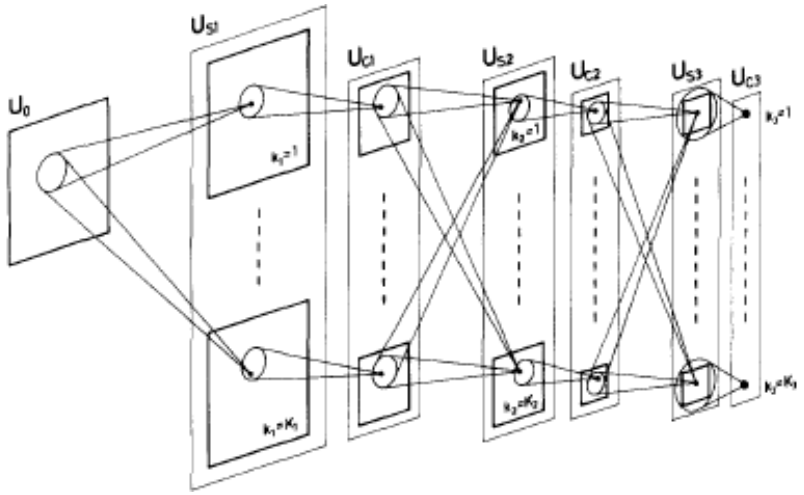
Several models were proposed with this intention (Rosenblatt, 1962; Kabrisky, 1966; Giebel, 1971; Fukushima, 1975). The response of most of these models, however, was severely affected by the shift in position and/or by the distortion in shape of the input patterns. Hence, their ability for pattern recognition was not so high.

In this paper, we propose an improved neural network model. The structure of this network has been suggested by that of the visual nervous system of the vertebrate. This network is self-organized by "learning without a teacher", and acquires an ability to recognize stimulus patterns based on the geometrical similarity (Gestalt) of their shapes without affected by their position nor by small distortion of their shapes.

This network is given a nickname "neocognitron"[1], because it is a further extention of the "cognitron", which also is a self-organizing multilayered neural network model proposed by the author before (Fukushima, 1975). Incidentally, the conventional cognitron also had an ability to recognize patterns, but its response was dependent upon the position of the stimulus patterns. That is, the same patterns which were presented at different positions were taken as different patterns by the conventional cognitron. In the neocognitron proposed here, however, the response of the network is little affected by the position of the stimulus patterns.

1 Preliminary report of the neocognitron already appeared elsewhere (Fukushima, 1979a, b)

0340-1200/80/0036/0193/$02.00



Schematic diagram illustrating the interconnections between layers in the neocognitron

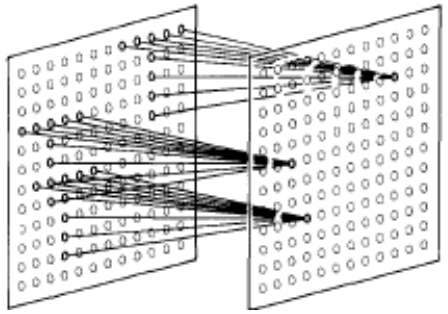


Illustration showing the input interconnections to the cells within a single cell-plane

- Although the neocognitron achieved some success in pattern recognition tasks and introduced convolutional filters to neural networks, it was limited by its lack of a training algorithm to learn the filters.
- The pattern detectors were manually engineered for the specific task, using a variety of heuristics and techniques from computer vision
- Backpropagation had not yet been applied to train neural nets, and thus there was no easy way to optimize neocognitrons or reuse them on different vision tasks.

Correlation and Convolution

❖Correlation and Convolution are basic operations that we will perform to extract information from images.

❖They are simple, they can be analyzed and understood very well, and they are also easy to implement and can be computed very efficiently.

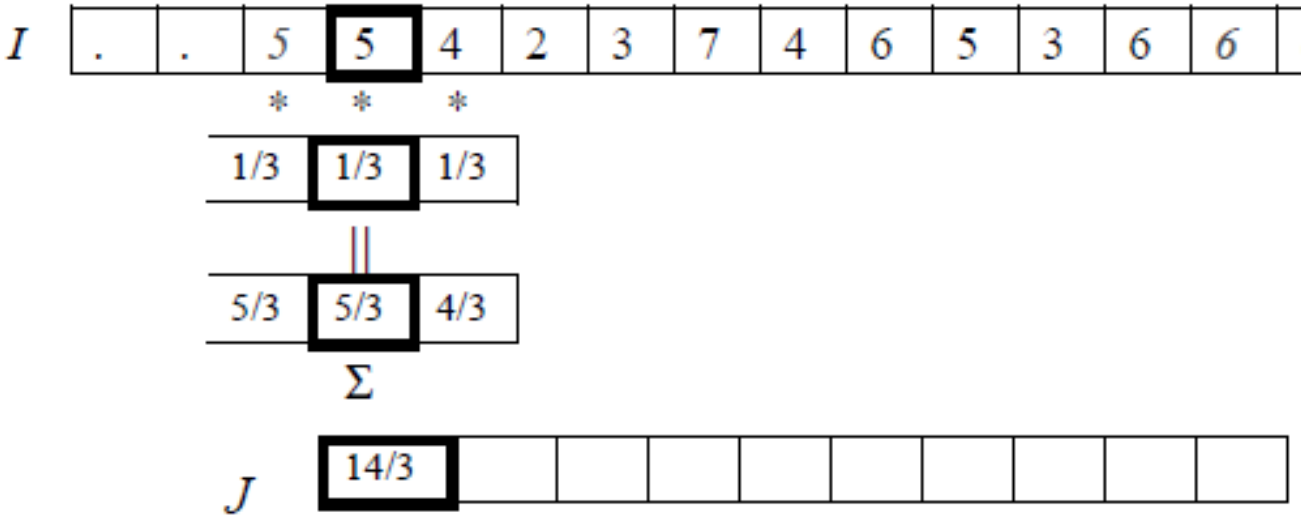
❖These operations have two key features: they are *shift-invariant, and they are linear.*

❖Shift-invariant means that we perform the same operation at every point in the image.

❖Linear means that this operation is linear, that is, we replace every pixel with a linear combination of its neighbors

One of the simplest operations that we can perform with correlation is local averaging.

5	4	2	3	7	4	6	5	3	6



$$F \circ I(x) = \sum_{i=-N}^{N} F(i) I(x+i)$$

$$F \circ I(x,y) = \sum_{j=-N}^{N} \sum_{i=-N}^{N} F(i,j) I(x+i, y+j)$$

Source : David Jacobs, Correlation and Convolution, http://www.cs.umd.edu/~djacobs/CMSC426/Convolution.pdf

Convolution

$$F * I(x) = \sum_{N} F(i) I(x-i)$$

$$F * I(x,y) = \sum_{j=-N}^{N} \sum_{i=-N}^{N} F(i,j) I(x-i, y-j)$$

Convolution is just like correlation, except that we flip over the filter before correlating.

Correlation and convolution are identical when the filter is symmetric.

The key difference between the two is that convolution is associative.

F(G*I) = (F*G)*I.*

The Filter Matrix

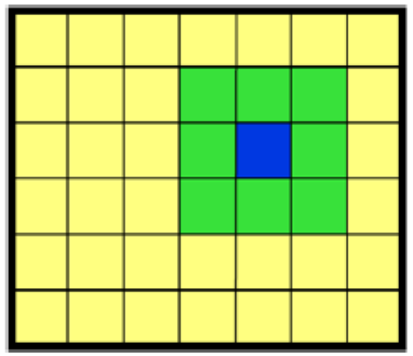
$$I'(u,v) \leftarrow \frac{p_0 + p_1 + p_2 + p_3 + p_4 + p_5 + p_6 + p_7 + p_8}{9}$$

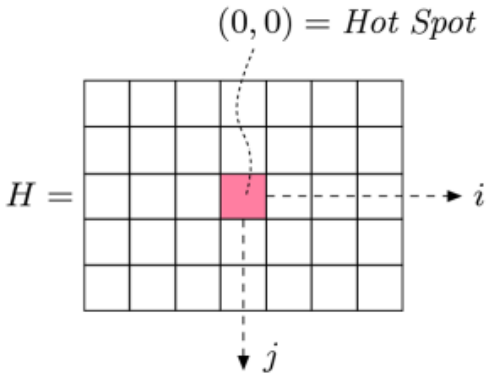
$$I'(u,v) \leftarrow \tfrac{1}{9} \cdot [\, I(u-1,v-1) + I(u,v-1) + I(u+1,v-1) + I(u-1,v) + I(u,v) + I(u+1,v) + I(u-1,v+1) + I(u,v+1) + I(u+1,v+1) \,]$$

$$H(i,j) = \begin{bmatrix} 1/9 & 1/9 & 1/9 \\ 1/9 & 1/9 & 1/9 \\ 1/9 & 1/9 & 1/9 \end{bmatrix} = \frac{1}{9}\begin{bmatrix} 1 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \end{bmatrix}$$

Filter operation can be expressed as a matrix
Example: averaging filter

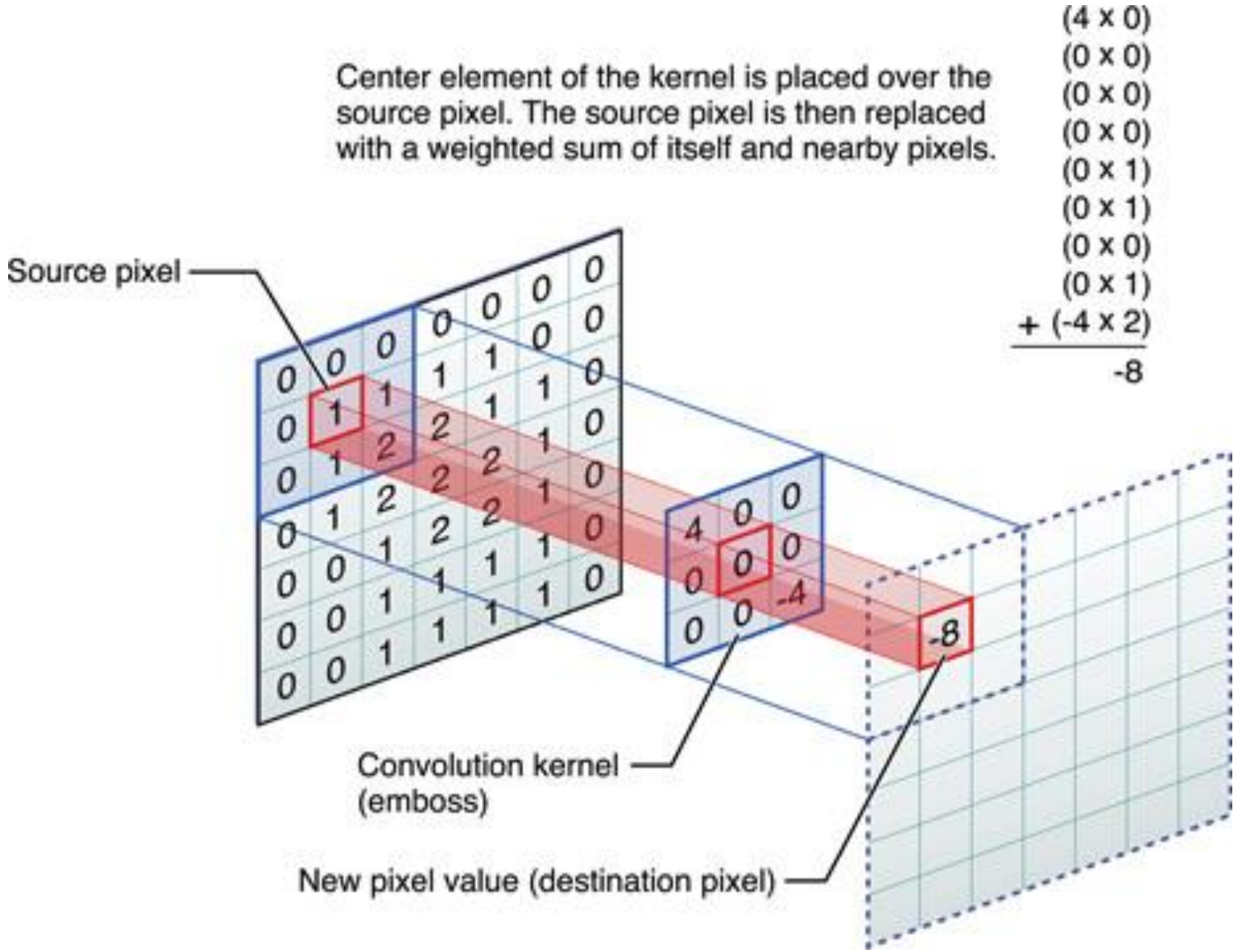
Filter matrix also called filter mask ***H(i,j)***

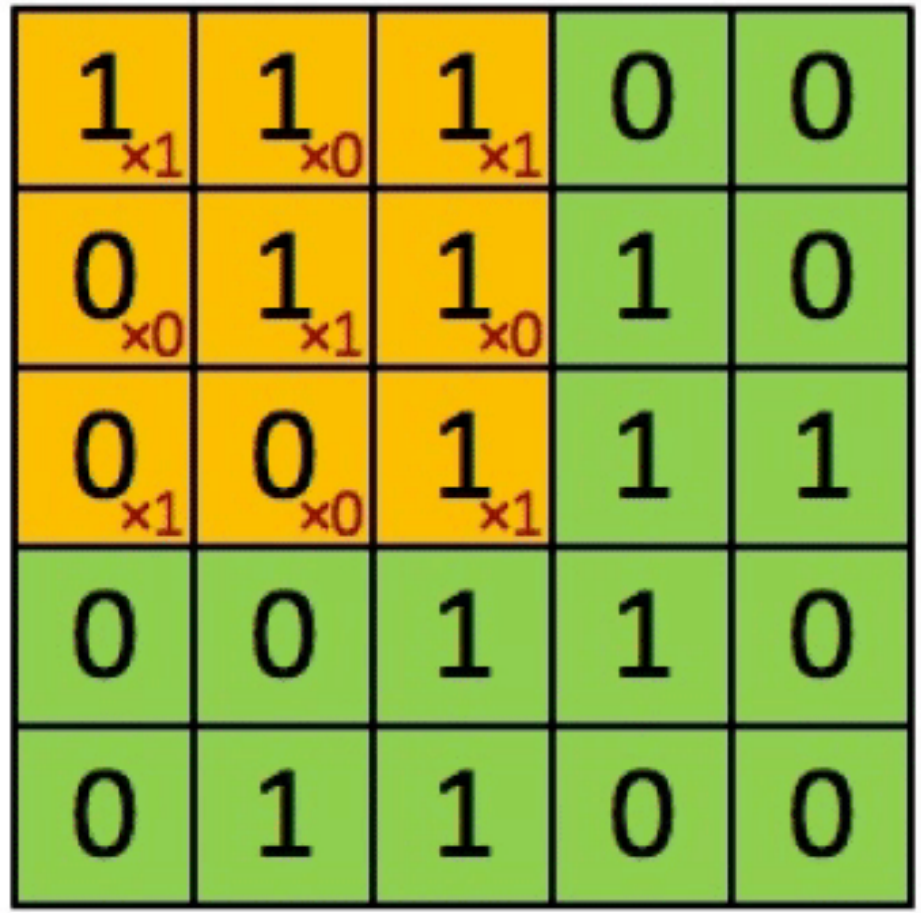




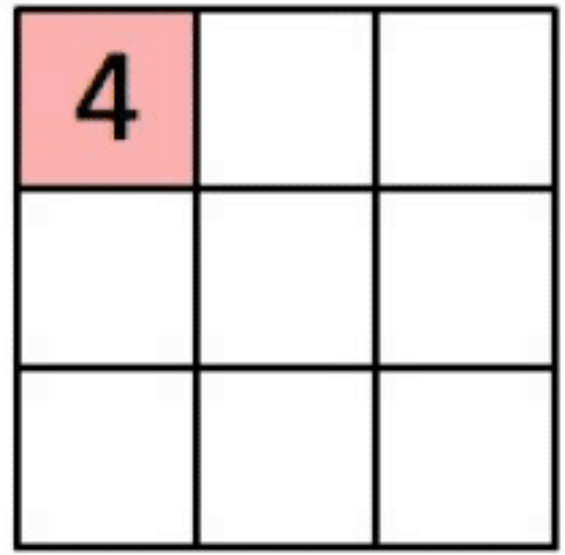
Source : Emmanuel Agu, http://web.cs.wpi.edu/~emmanuel/courses/cs545/S14/

Center element of the kernel is placed over the source pixel. The source pixel is then replaced with a weighted sum of itself and nearby pixels.





Image



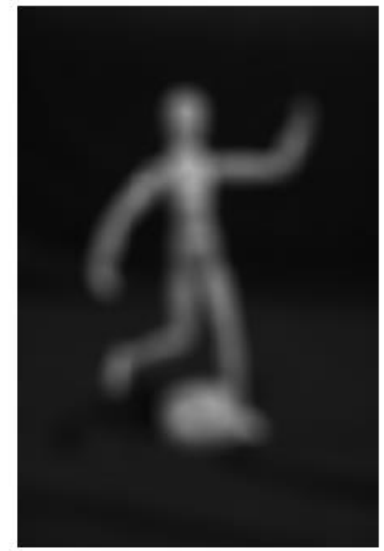
Convolved Feature

Mean Filters: Effect of Filter Size

Original

7×7

15×15



41×41

Kernel Size: The kernel size defines the field of view of the convolution. A common choice for 2D is 3 — that is 3x3 pixels.

Stride: The stride defines the step size of the kernel when traversing the image.

Padding: The padding defines how the border of a sample is handled.

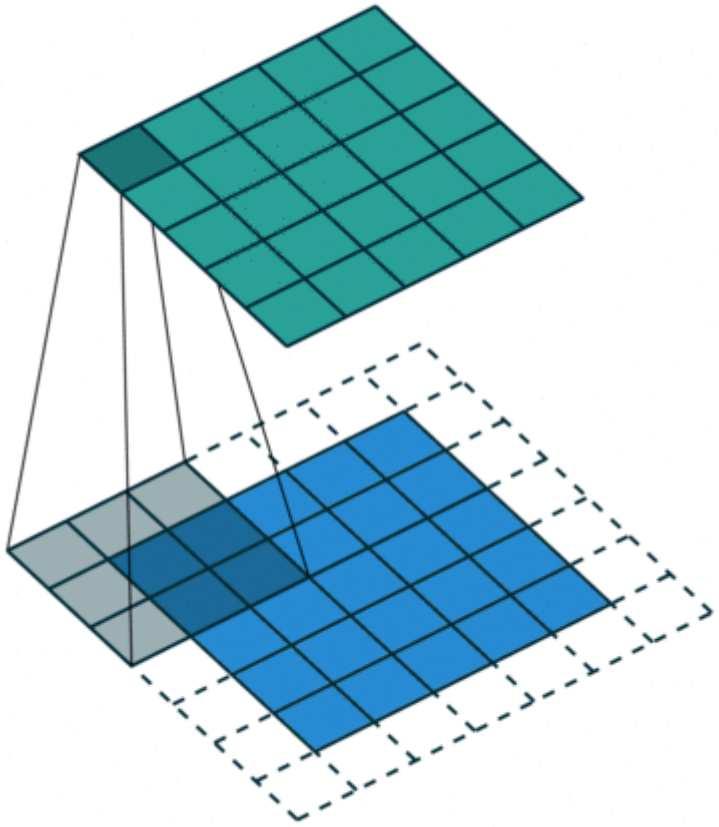


Image Size after convolution

Image size : n x n , filter size : f x f

Image size after convolution : $(n - f + 1) \times (n - f + 1)$

With zero padding: $(n + 2p - f + 1) \times (n + 2p - f + 1)$

Strided convolution: $\left(\left\lfloor \frac{n + 2p - f}{s} + 1 \right\rfloor \times \left\lfloor \frac{n + 2p - f}{s} + 1 \right\rfloor\right)$

❑ Conv 3x3 with stride=1,padding=0

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

6x6 Image

21	28	36	31
33	42	51	41
38	47	43	34
34	37	30	22

4x4

❑ Conv 3x3 with **stride=2**,padding=0

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

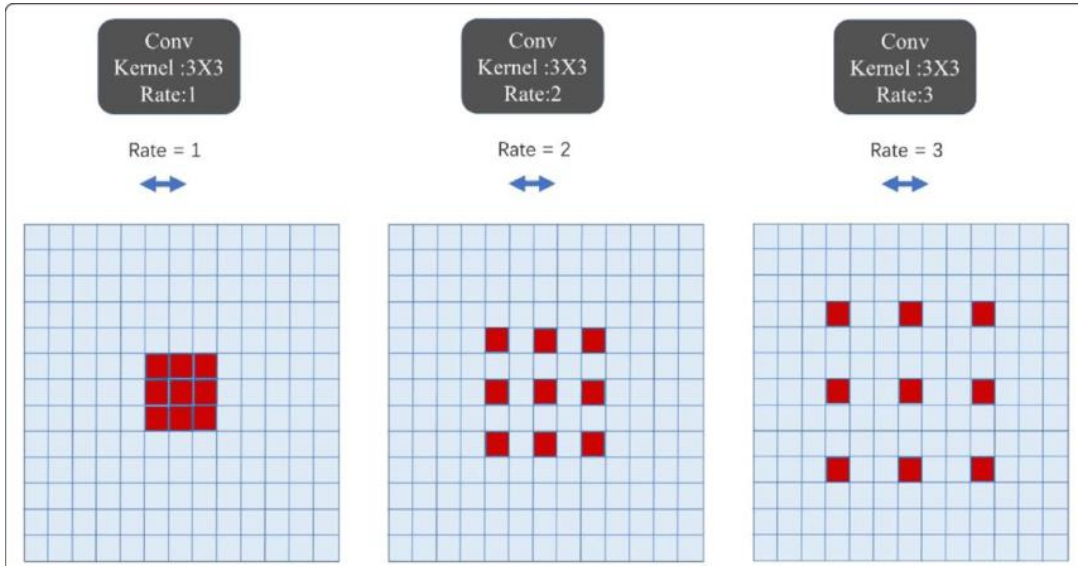
7x7 Image

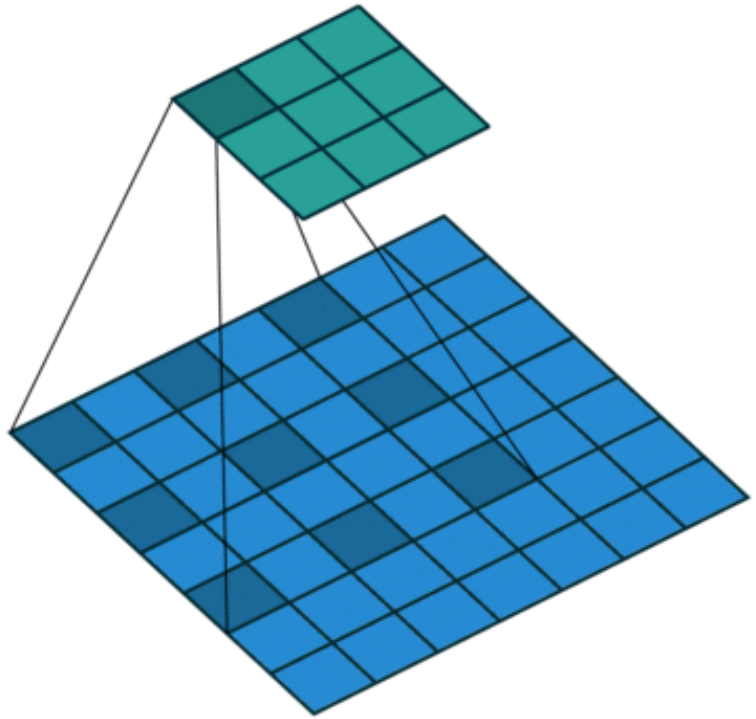
14	36	23
38	43	23
21	18	12

3x3

Dilated Convolutions (Atrous convolution)

- Dilated convolutions introduce another parameter to convolutional layers called the **dilation rate**.
- This defines a spacing between the values in a kernel.
- A 3x3 kernel with a dilation rate of 2 will have the same field of view as a 5x5 kernel, while only using 9 parameters.
- This delivers a wider field of view at the same computational cost.





2D convolution using a 3 kernel with a dilation rate of 2 and no padding

$$y[i] = \sum_{k=1}^{K} x[i + r \cdot k] w[k]$$

Introduce zeros between filter values

Separable Filters

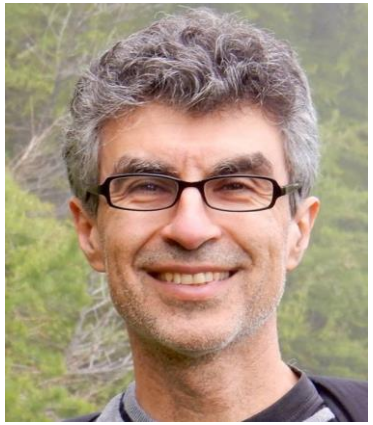
Generally, 2D correlation is more expensive than 1D correlation because the sizes of the filters we use are larger.

If our filter is NxN in size, and our image contains MxM pixels, then the total number of multiplications we must perform is N^2M^2.

With an important class of filters, we can save on this computation.

$$\begin{bmatrix} \frac{1}{9} & \frac{1}{9} & \frac{1}{9} \\ \frac{1}{9} & \frac{1}{9} & \frac{1}{9} \\ \frac{1}{9} & \frac{1}{9} & \frac{1}{9} \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix} \begin{bmatrix} \frac{1}{3} & 0 \\ \frac{1}{3} & 0 \\ \frac{1}{3} & 0 \end{bmatrix} \circ \begin{bmatrix} 0 & 0 & 0 \\ \frac{1}{3} & \frac{1}{3} & \frac{1}{3} \\ 0 & 0 & 0 \end{bmatrix}$$

Yann LeCun



Yoshua Bengio

Gradient-Based Learning Applied to Document Recognition

Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner

Abstract—

Multilayer Neural Networks trained with the backpropagation algorithm constitute the best example of a successful Gradient-Based Learning technique. Given an appropriate network architecture, Gradient-Based Learning algorithms can be used to synthesize a complex decision surface that can classify high-dimensional patterns such as handwritten characters, with minimal preprocessing. This paper reviews various methods applied to handwritten character recognition and compares them on a standard handwritten digit recognition task. Convolutional Neural Networks, that are specifically designed to deal with the variability of 2D shapes, are shown to outperform all other techniques.

Real-life document recognition systems are composed of multiple modules including field extraction, segmentation, recognition, and language modeling. A new learning paradigm, called Graph Transformer Networks (GTN), allows such multi-module systems to be trained globally using Gradient-Based methods so as to minimize an overall performance measure.

Two systems for on-line handwriting recognition are described. Experiments demonstrate the advantage of global training, and the flexibility of Graph Transformer Networks.

A Graph Transformer Network for reading bank check is also described. It uses Convolutional Neural Network character recognizers combined with global training techniques to provides record accuracy on business and personal checks. It is deployed commercially and reads several million checks per day.

Keywords— **Neural Networks, OCR, Document Recognition, Machine Learning, Gradient-Based Learning, Convolutional Neural Networks, Graph Transformer Networks, Finite State Transducers.**

NOMENCLATURE

- GT Graph transformer.
- GTN Graph transformer network.
- HMM Hidden Markov model.
- HOS Heuristic oversegmentation.
- K-NN K-nearest neighbor.
- NN Neural network.
- OCR Optical character recognition.
- PCA Principal component analysis.
- RBF Radial basis function.
- RS-SVM Reduced-set support vector method.
- SDNN Space displacement neural network.
- SVM Support vector method.
- TDNN Time delay neural network.
- V-SVM Virtual support vector method.

The authors are with the Speech and Image Processing Services Research Laboratory, AT&T Labs-Research, 100 Schulz Drive Red Bank, NJ 07701. E-mail: {yann,leonb,yoshua,haffner}@research.att.com. Yoshua Bengio is also with the Département d'Informatique et de Recherche Opérationelle, Université de Montréal, C.P. 6128 Succ. Centre-Ville, 2920 Chemin de la Tour, Montréal, Québec, Canada H3C 3J7.

I. INTRODUCTION

Over the last several years, machine learning techniques, particularly when applied to neural networks, have played an increasingly important role in the design of pattern recognition systems. In fact, it could be argued that the availability of learning techniques has been a crucial factor in the recent success of pattern recognition applications such as continuous speech recognition and handwriting recognition.

The main message of this paper is that better pattern recognition systems can be built by relying more on automatic learning, and less on hand-designed heuristics. This is made possible by recent progress in machine learning and computer technology. Using character recognition as a case study, we show that hand-crafted feature extraction can be advantageously replaced by carefully designed learning machines that operate directly on pixel images. Using document understanding as a case study, we show that the traditional way of building recognition systems by manually integrating individually designed modules can be replaced by a unified and well-principled design paradigm, called *Graph Transformer Networks*, that allows training all the modules to optimize a global performance criterion.

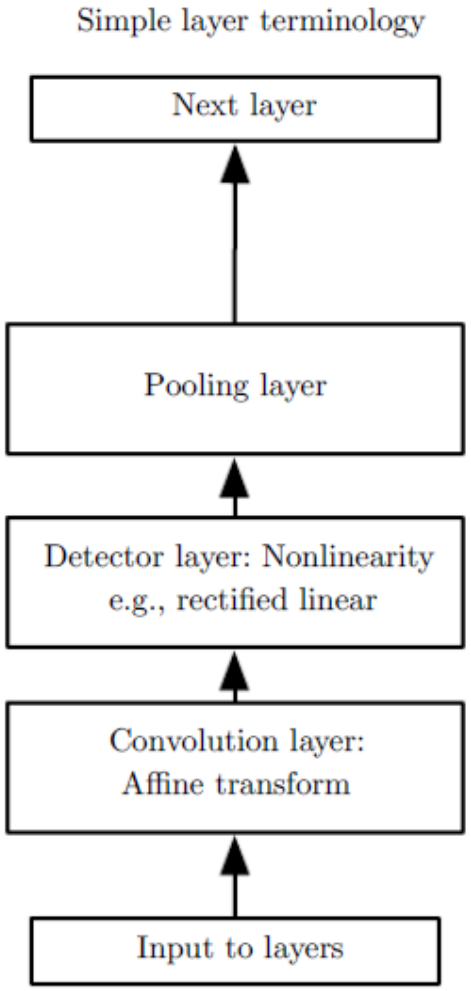
Since the early days of pattern recognition it has been known that the variability and richness of natural data, be it speech, glyphs, or other types of patterns, make it almost impossible to build an accurate recognition system entirely by hand. Consequently, most pattern recognition systems are built using a combination of automatic learning techniques and hand-crafted algorithms. The usual method of recognizing individual patterns consists in dividing the system into two main modules shown in figure 1. The first module, called the feature extractor, transforms the input patterns so that they can be represented by low-dimensional vectors or short strings of symbols that (a) can be easily matched or compared, and (b) are relatively invariant with respect to transformations and distortions of the input patterns that do not change their nature. The feature extractor contains most of the prior knowledge and is rather specific to the task. It is also the focus of most of the design effort, because it is often entirely hand-crafted. The classifier, on the other hand, is often general-purpose and trainable. One of the main problems with this approach is that the recognition accuracy is largely determined by the ability of the designer to come up with an appropriate set of features. This turns out to be a daunting task which, unfortunately, must be redone for each new problem. A large amount of the pattern recognition literature is devoted to describing and comparing the relative

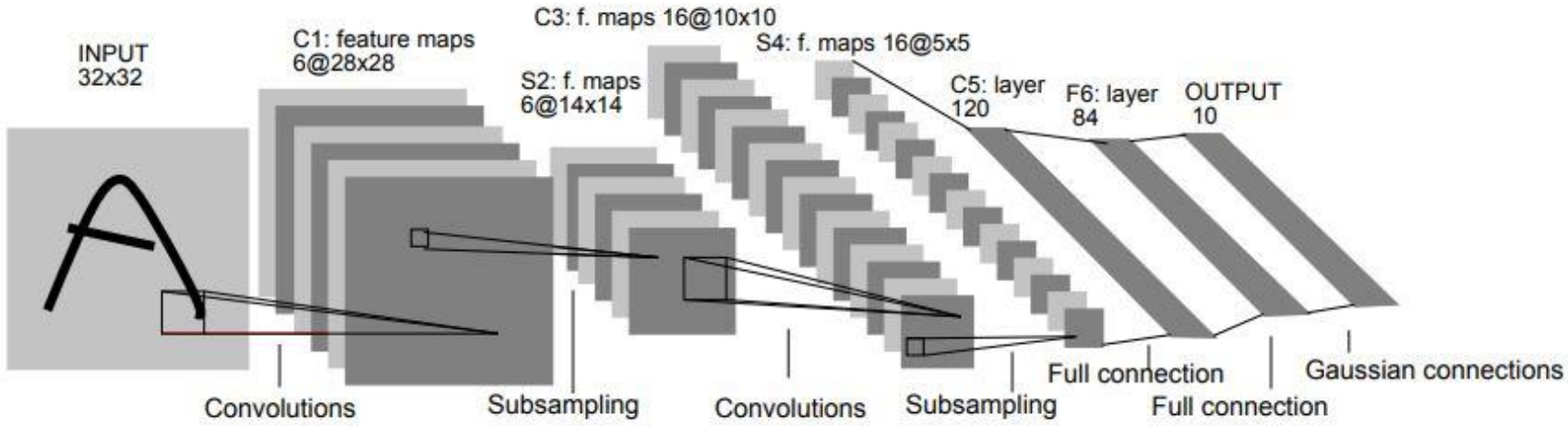
- The main message of this paper is that better pattern recognition systems can be built by relying more on **automatic learning**, and less on **hand-designed heuristics**.

$$E_{test} - E_{train} = k(h/P)^{\alpha}$$

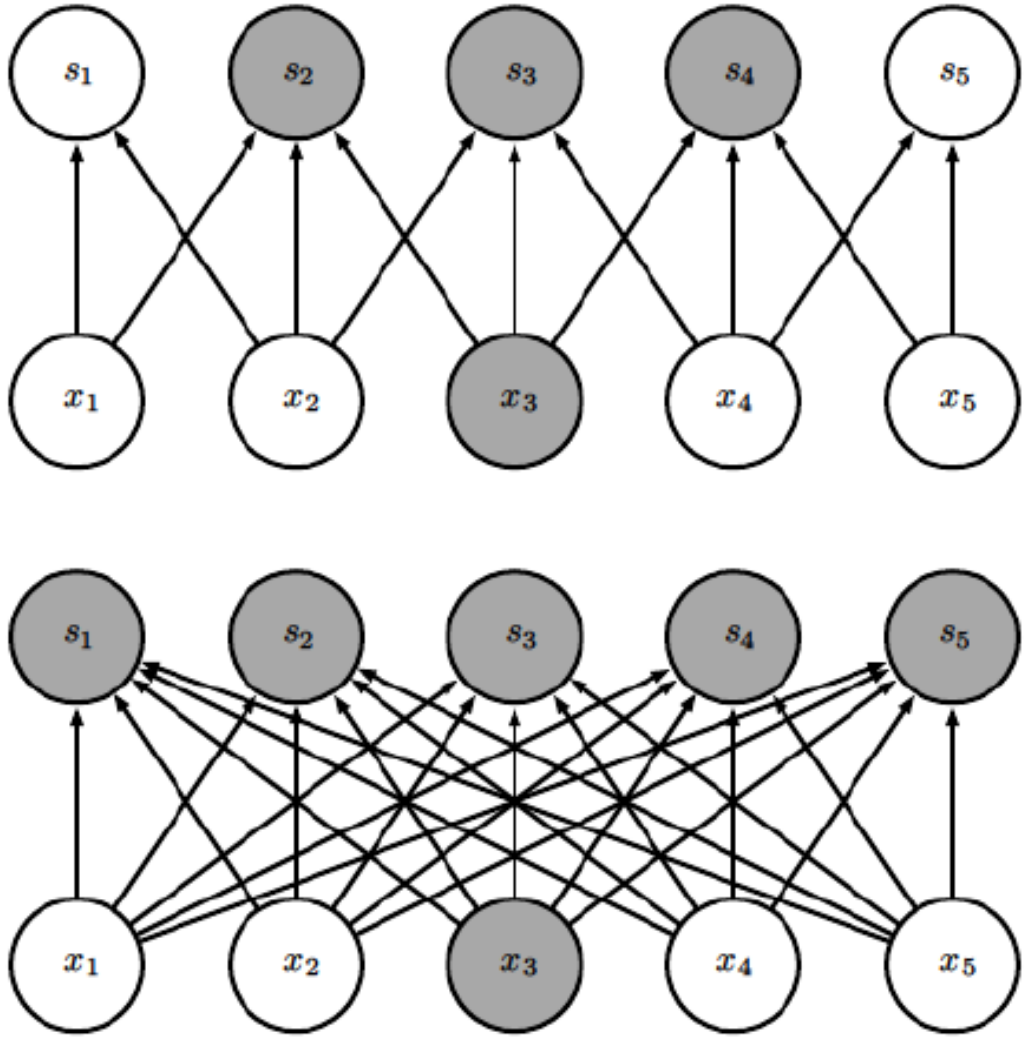
- P is the number of training samples, h is the measure of effective capacity (complexity of the machine), α is a number between 0.5 and 1, and k is a constant.
- Use backpropagation for training the proposed model.
- Experiments were conducted on MNIST dataset.

LeNet 5 - 1998





- Convolutional networks are simply neural networks that use convolution in place of general matrix multiplication in at least one of their layers.
- Convolution leverages three important ideas that can improve a machine learning system : **sparse interactions**, **parameter sharing**, **equivariant representations**.
- The spatial relationship between the pixels is not considered in ANNS.
- Convolution provides a means for working with inputs of variable size.



Number of parameters will be very less compared to conventional ANN

Figure 9.2: *Sparse connectivity, viewed from below:* We highlight one input unit, x_3, and also highlight the output units in $\boldsymbol{s}$ that are affected by this unit. *(Top)* When $\boldsymbol{s}$ is formed by convolution with a kernel of width 3, only three outputs are affected by $\boldsymbol{x}$. *(Bottom)* When $\boldsymbol{s}$ is formed by matrix multiplication, connectivity is no longer sparse, so all of the outputs are affected by x_3.

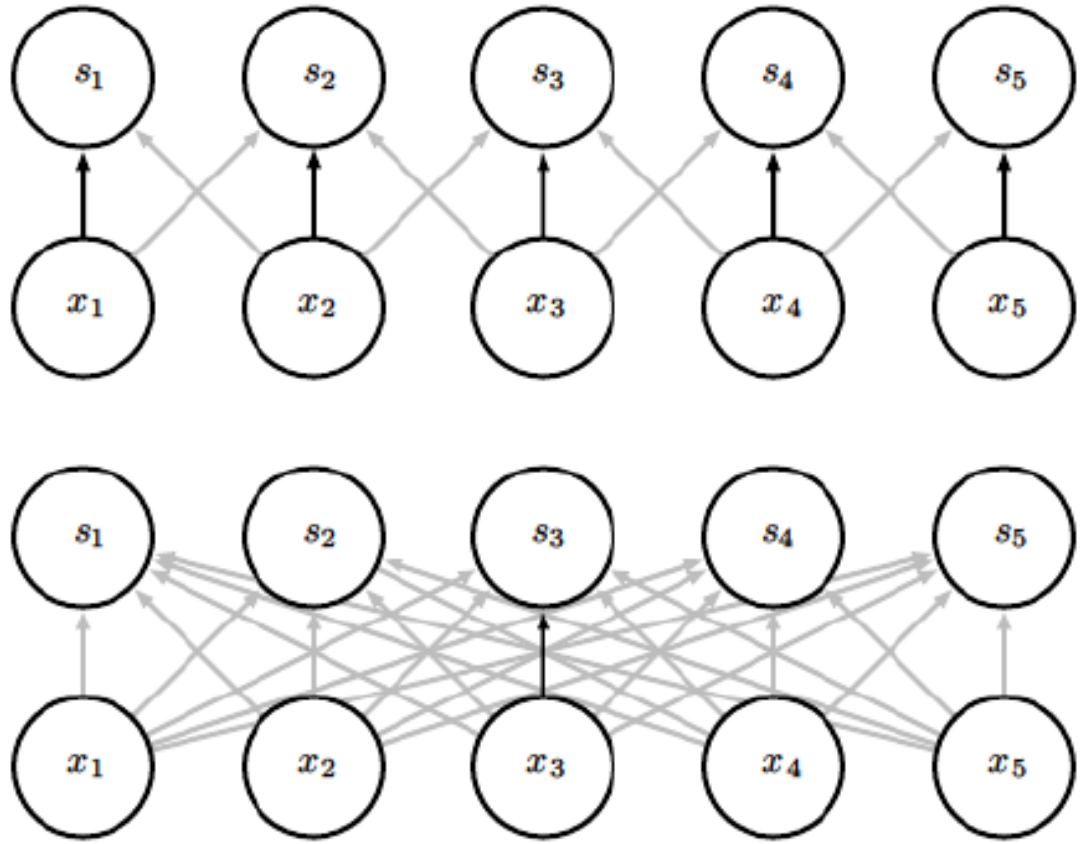
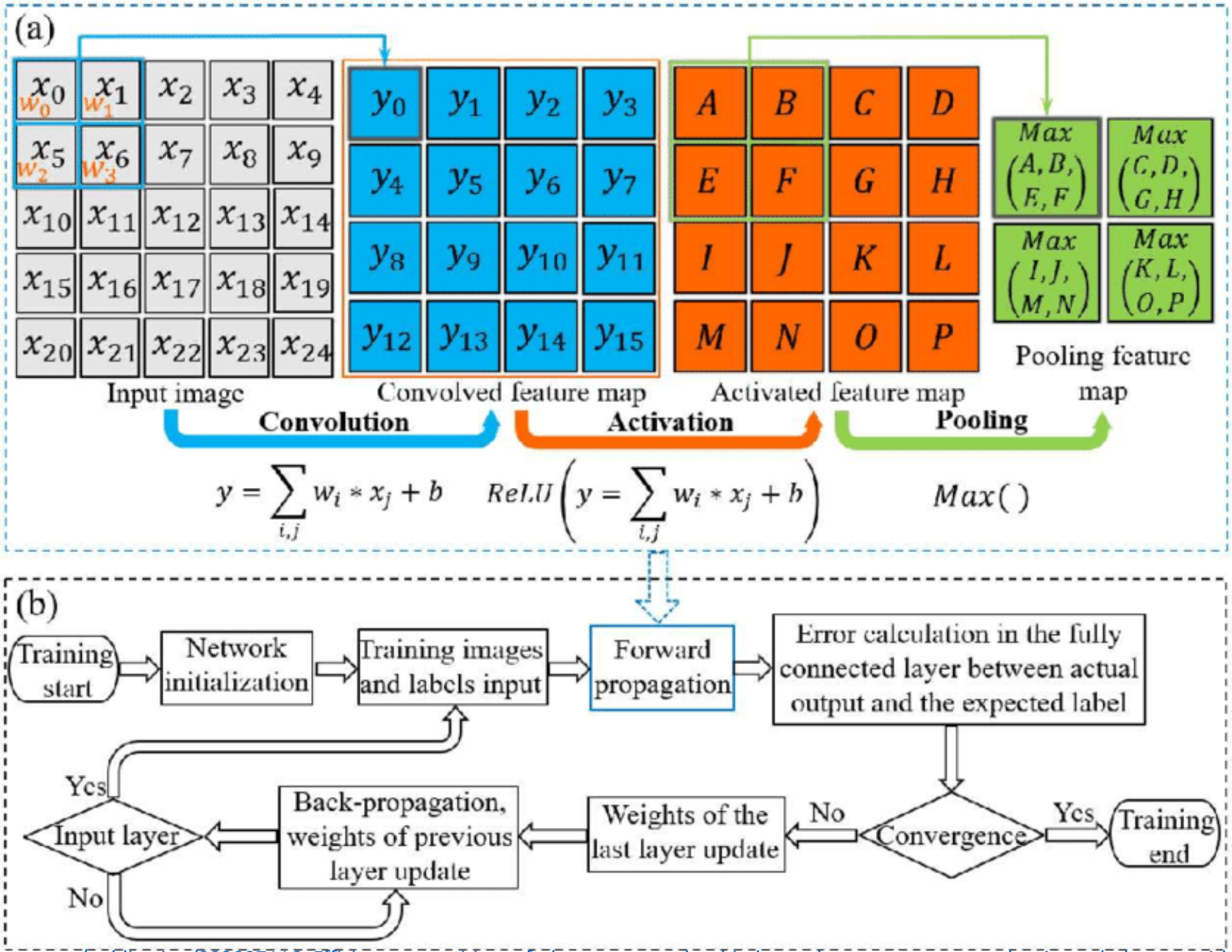


Figure 9.5: *Parameter sharing:* Black arrows indicate the connections that use a particular parameter in two different models. *(Top)* The black arrows indicate uses of the central element of a 3-element kernel in a convolutional model. Due to parameter sharing, this single parameter is used at all input locations. *(Bottom)* The single black arrow indicates the use of the central element of the weight matrix in a fully connected model. This model has no parameter sharing so the parameter is used only once.



https://www.researchgate.net/publication/333007116_Efficient_recognition_of_the_propagated_orbital_angular_momentum_modes_in_turbulences_with_the_convolutional_neural_network

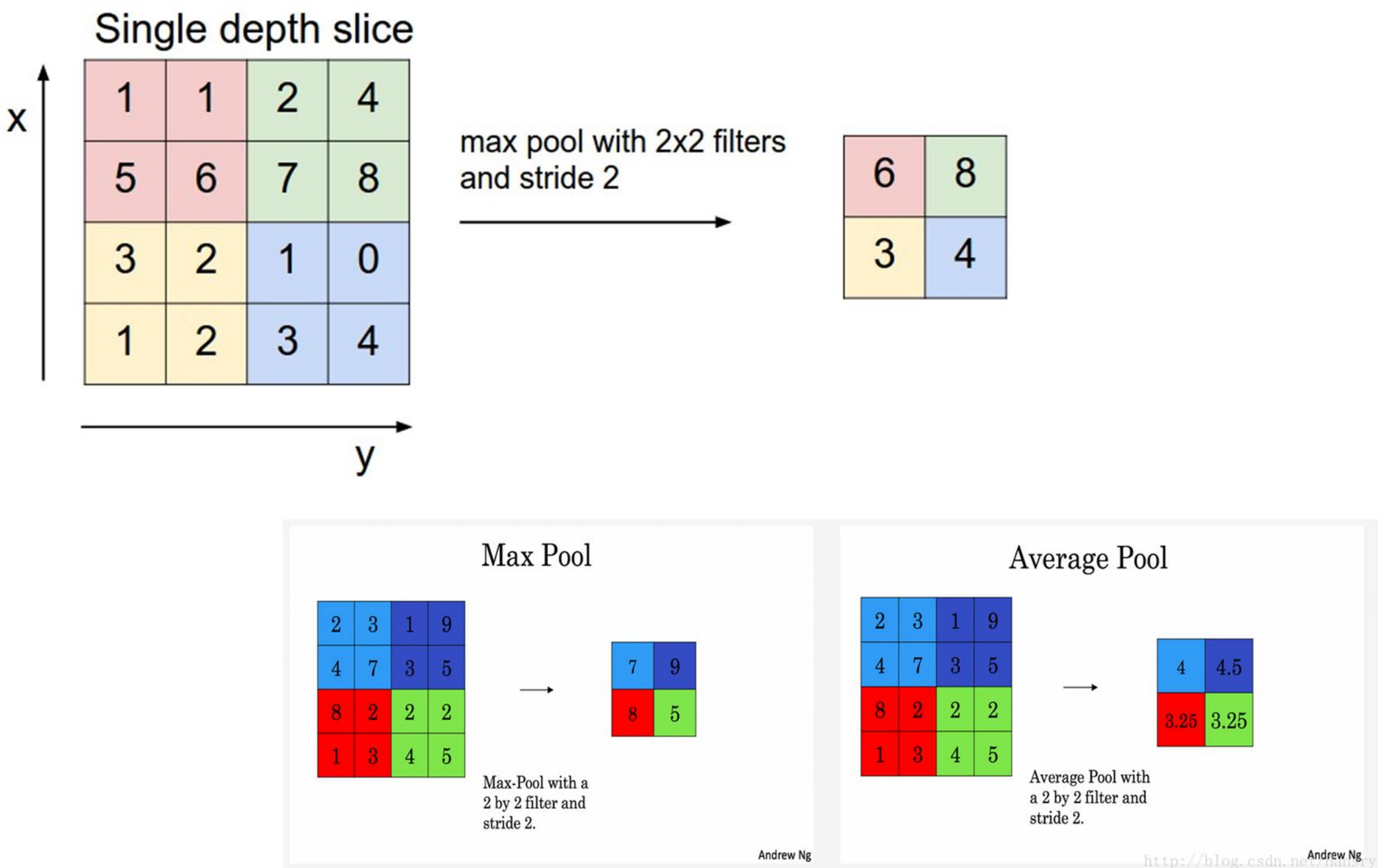
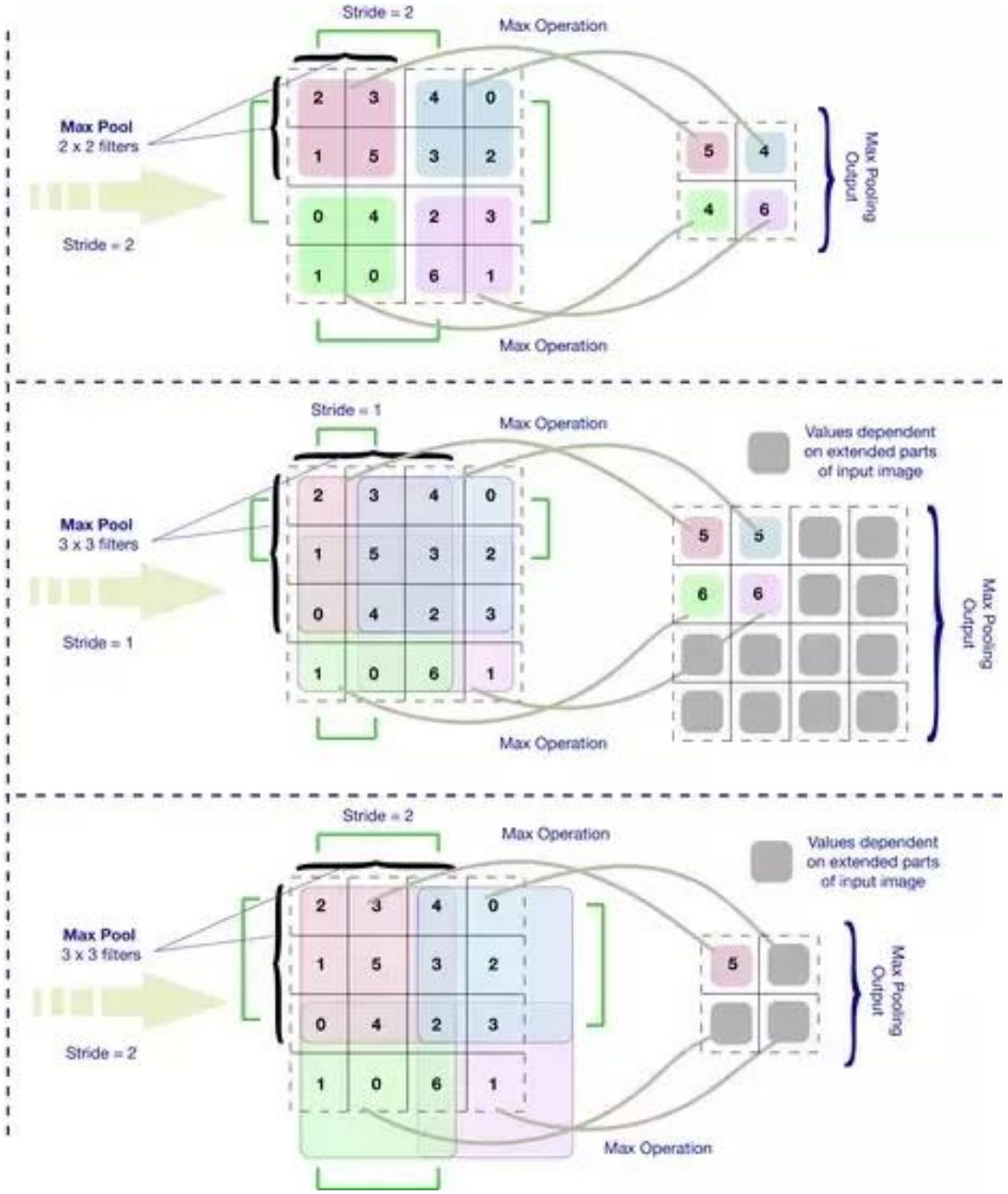


Image size : n x n , filter size : f x f

Image size after pooling : $\left(\frac{n-f}{s}+1 \times \frac{n-f}{s}+1\right)$

- The purpose of pooling layers is to perform dimensionality reduction to widen subsequent convolutional layers' receptive fields.
- The same effect can be achieved by using a convolutional layer: using a stride of 2 also reduces the dimensionality of the output and widens the receptive field of higher layers.
- The resulting operation differs from a max-pooling layer in that
  - –it cannot perform a true max operation
  - –it allows pooling across input channels



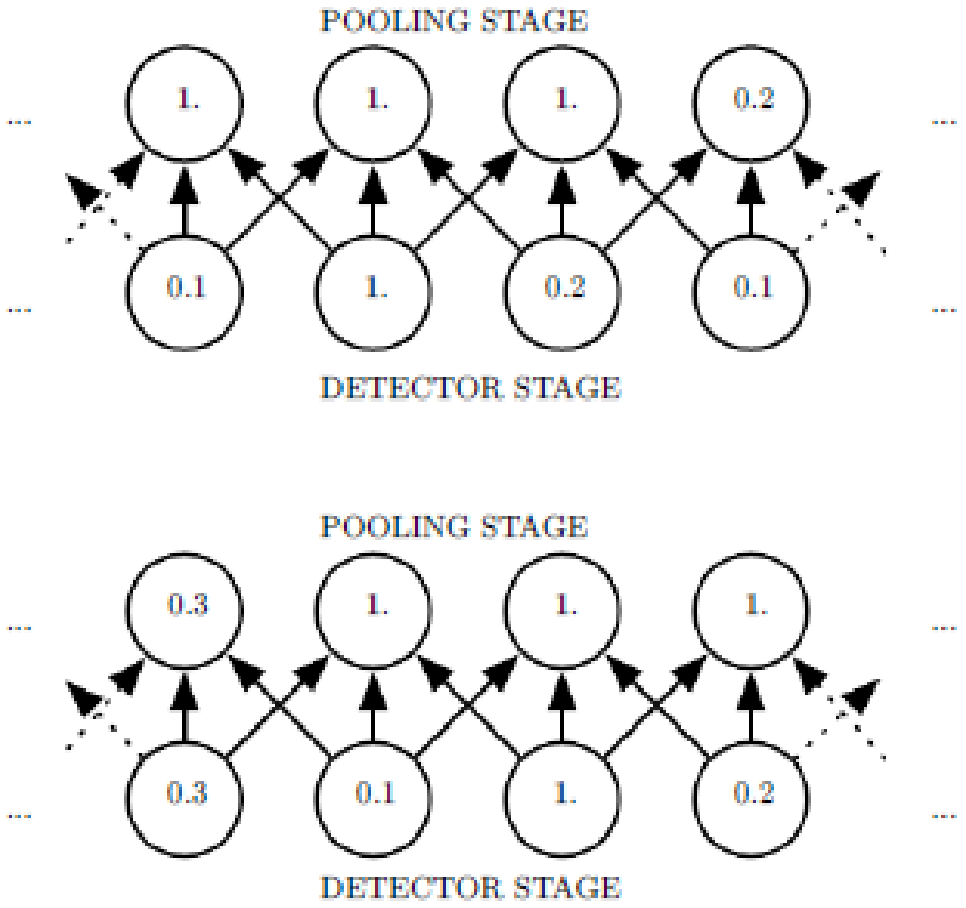


Figure 9.8: Max pooling introduces invariance. *(Top)* A view of the middle of the output of a convolutional layer. The bottom row shows outputs of the nonlinearity. The top row shows the outputs of max pooling, with a stride of one pixel between pooling regions and a pooling region width of three pixels. *(Bottom)* A view of the same network, after the input has been shifted to the right by one pixel. Every value in the bottom row has changed, but only half of the values in the top row have changed, because the max pooling units are only sensitive to the maximum value in the neighborhood, not its exact location.

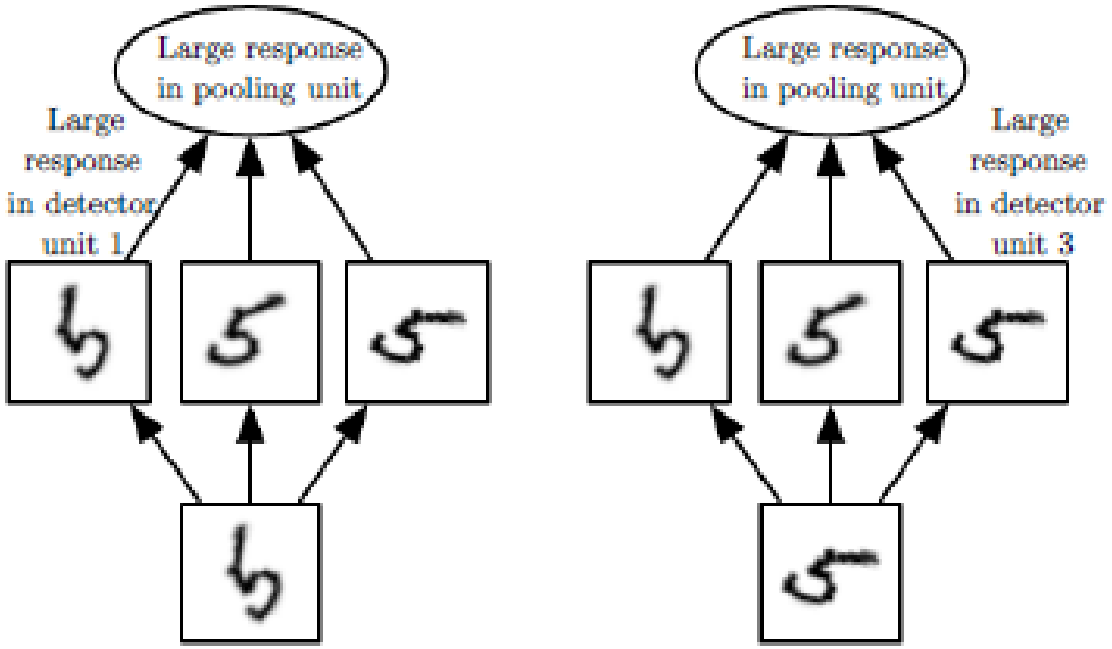
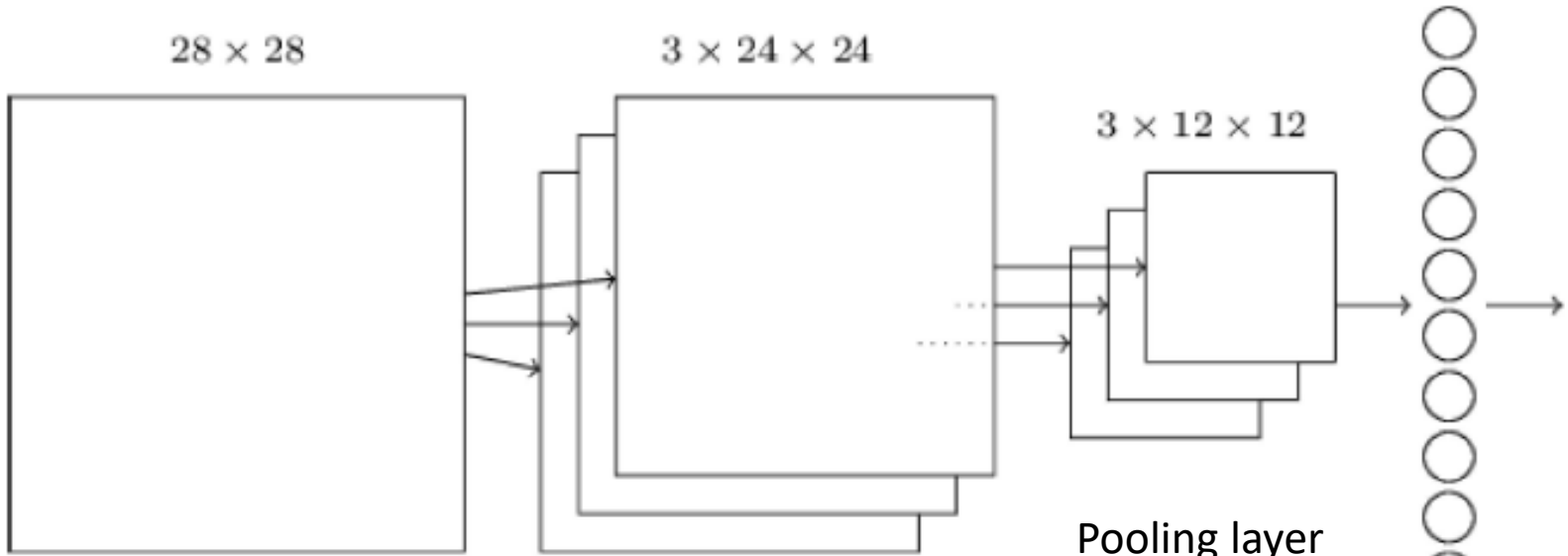
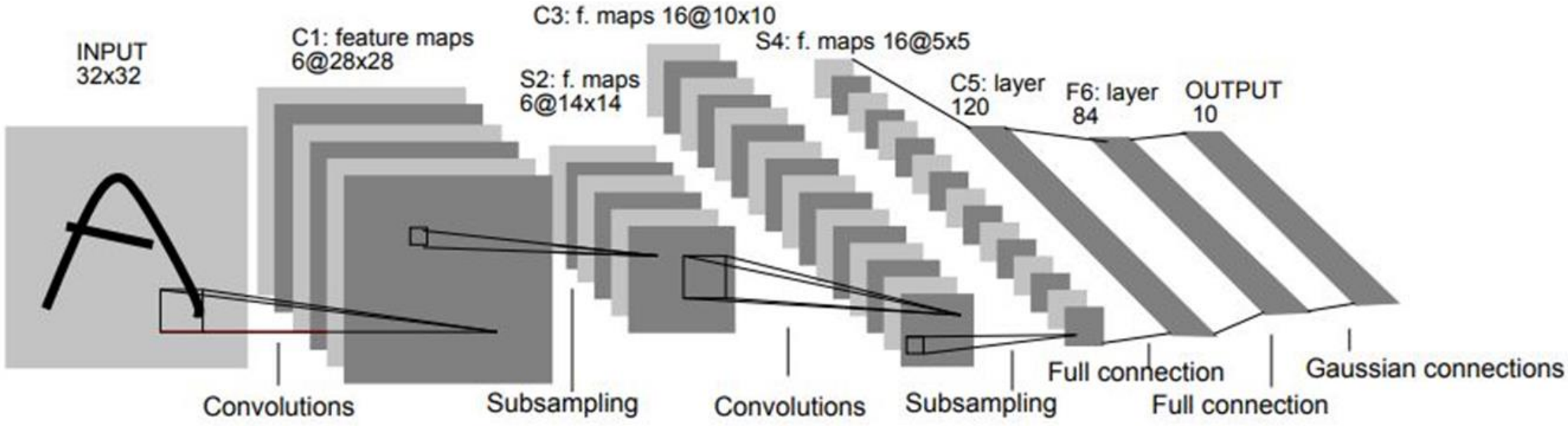
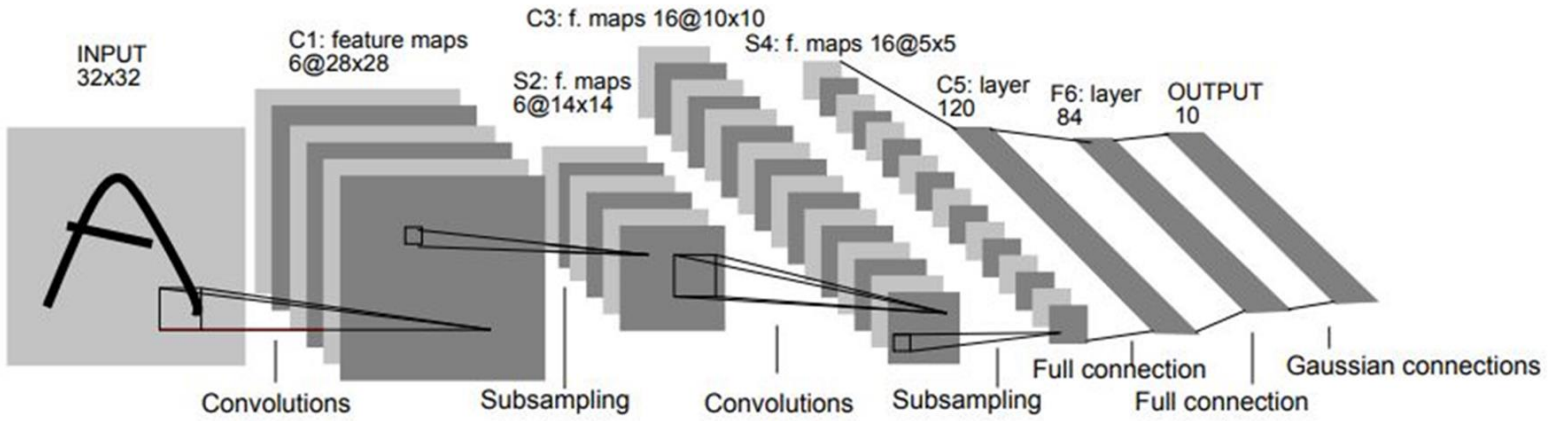


Figure 9.9: *Example of learned invariances:* A pooling unit that pools over multiple features that are learned with separate parameters can learn to be invariant to transformations of the input. Here we show how a set of three learned filters and a max pooling unit can learn to become invariant to rotation. All three filters are intended to detect a hand-written 5. Each filter attempts to match a slightly different orientation of the 5. When a 5 appears in the input, the corresponding filter will match it and cause a large activation in a detector unit. The max pooling unit then has a large activation regardless of which pooling unit was activated. We show here how the network processes two different inputs, resulting in two different detector units being activated. The effect on the pooling unit is roughly the same either way. This principle is leveraged by maxout networks (Goodfellow *et al.*, 2013a) and other convolutional networks. Max pooling over spatial positions is naturally invariant to translation; this multi-channel approach is only necessary for learning other transformations.


28 × 28
3 × 24 × 24
3 × 12 × 12
Input layer
Convolutional layer
Pooling layer
Fully connected layer

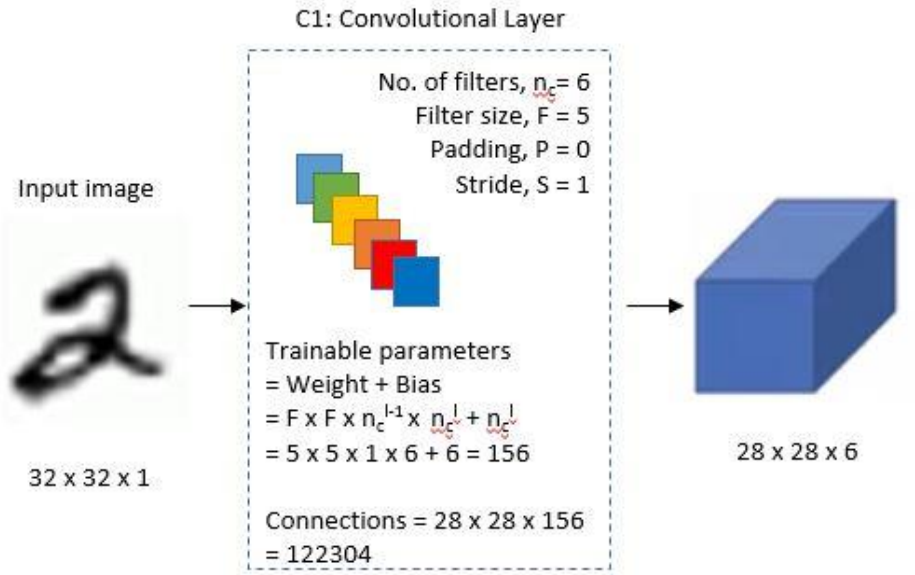


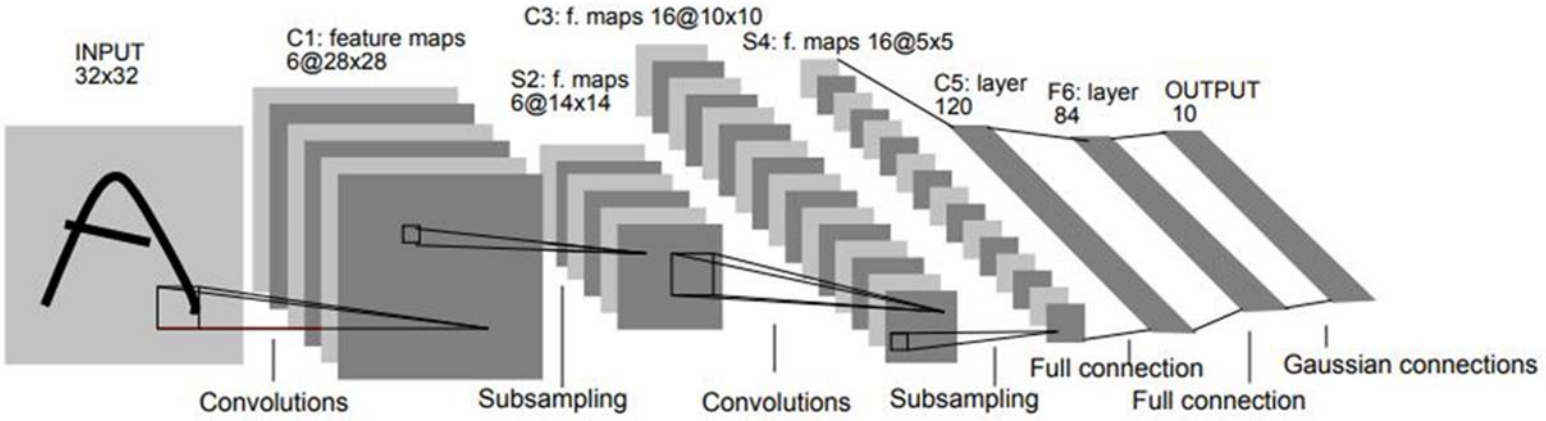
- LeNet-5 comprises 5 layers (excluding input layer), all of which contains trainable parameters.
- The input is a 32 x 32 pixel image. The values of the input pixels are normalized so that the background level (white) corresponds to a value of -0.1 and the foreground (black) corresponds to 1.175.
- Convolutional layers are labeled Cx, sub sampling layers are labeled Sx and fully connected layers are labeled Fx, where x is the layer index.



Layer C1

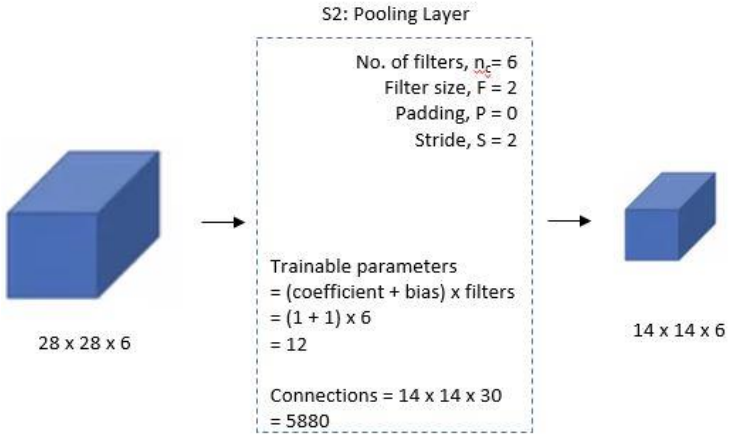
- Convolutional layer with 6 feature maps (6 filters)
- Convolutional filter size : 5 x 5
- Feature map size (image size after convolution) : 28 x 28
- Total trainable parameters : 156 (5 x 5 x 6 = 150w + 6b = 156)

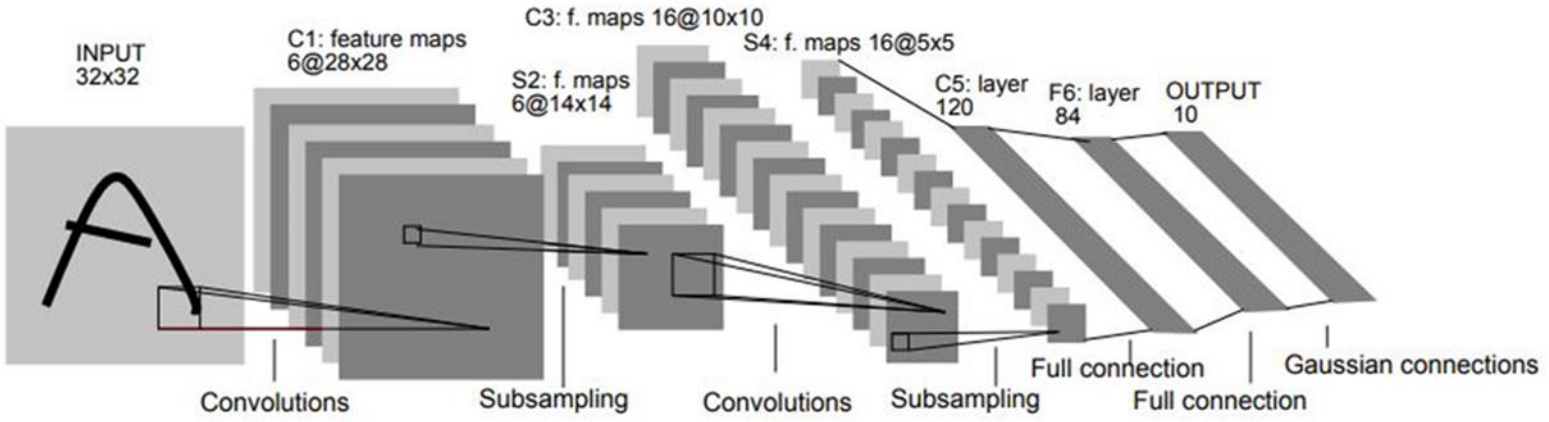




Layer S2

- Subsampling layer (pooling layer)
- Each unit in each feature map is connected to a 2 x 2 neighborhood in the corresponding feature map in C1.
- The four inputs to a unit in S2 are added, then multiplied by a trainable coefficient, and added to a trainable bias.
- The result is passed through a *tanh* function.
- The 2 x 2 receptive fields are non-overlapping, hence the output will be half the size of the input.
- Layer S2 has 12 trainable parameters (6w+6b)





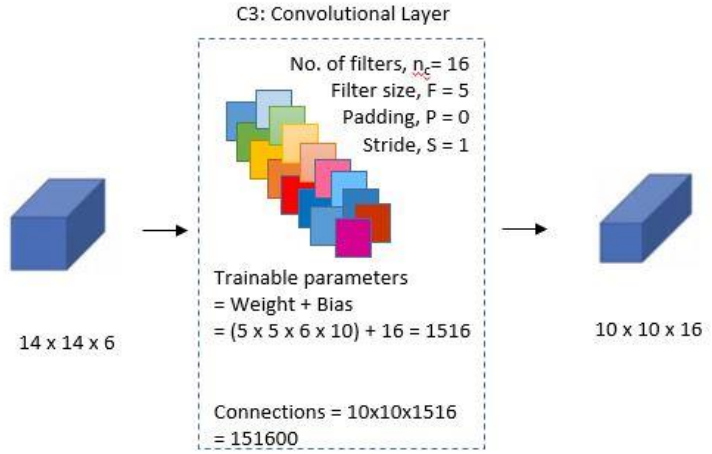
	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
0	X				X	X	X			X	X	X	X		X	X
1	X	X				X	X	X			X	X	X	X		X
2	X	X	X				X	X	X			X		X	X	X
3		X	X	X			X	X	X	X			X		X	X
4			X	X	X			X	X	X	X		X	X		X
5				X	X	X			X	X	X	X		X	X	X

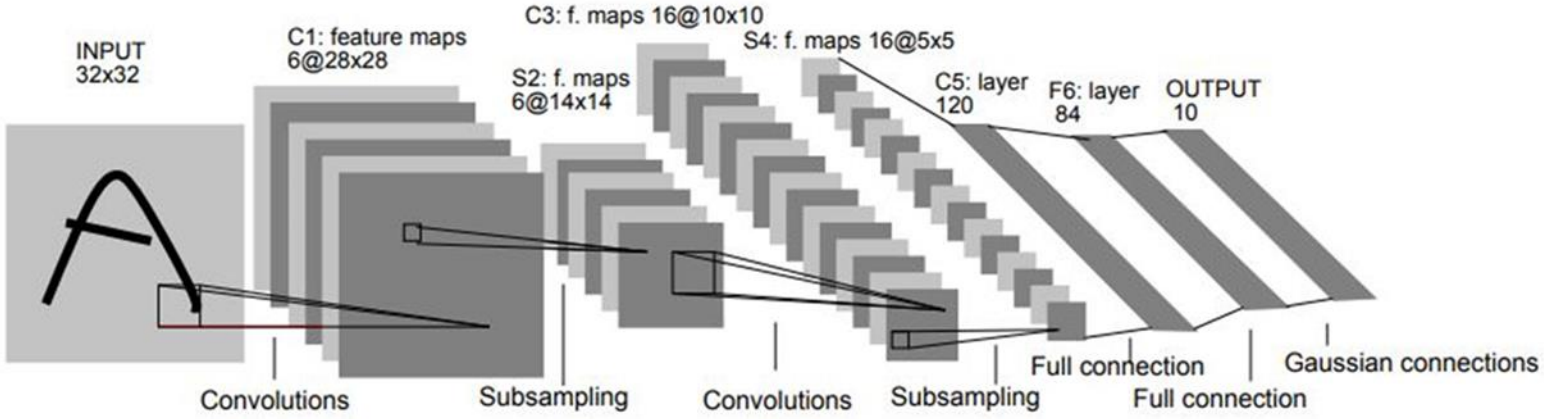
TABLE I

EACH COLUMN INDICATES WHICH FEATURE MAP IN S2 ARE COMBINED BY THE UNITS IN A PARTICULAR FEATURE MAP OF C3.

Layer C3

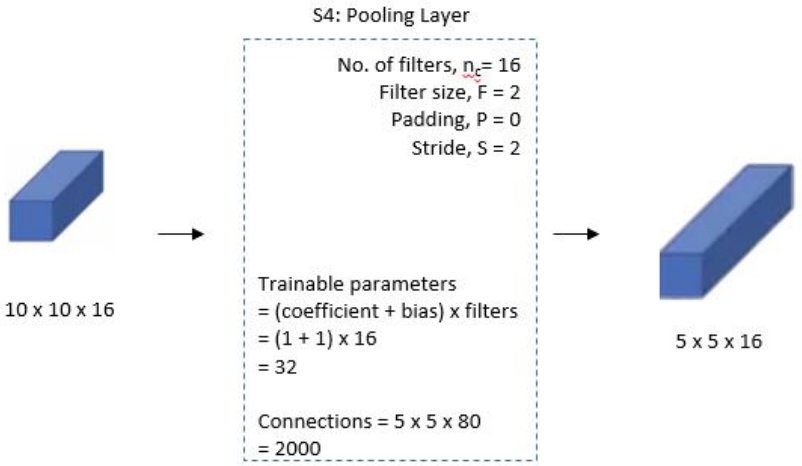
- Convolutional layer with 16 feature maps (16 filters)
- Convolutional filter size : 5 x 5 x Z
- The first six C3 feature maps take inputs from every contiguous subsets of three feature maps in S2 (filter size : 5 x 5 x 3)
- The next six take input from every contiguous subset of four (filter size : 5 x 5 x 4)
- The next three take input from some discontinuous subsets of four (filter size : 5 x 5 x 4)
- The last one takes input from all S2 feature maps (filter size : 5 x 5 x 6)
- Total trainable parameters : **1516** (450w + 600w + 300w + 150w + 16b)

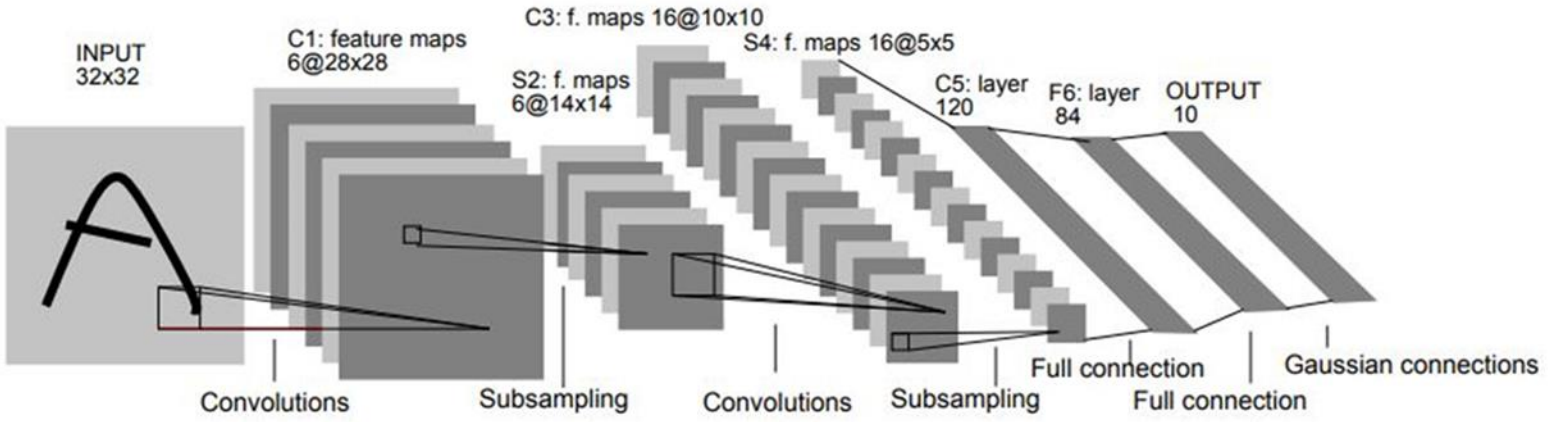


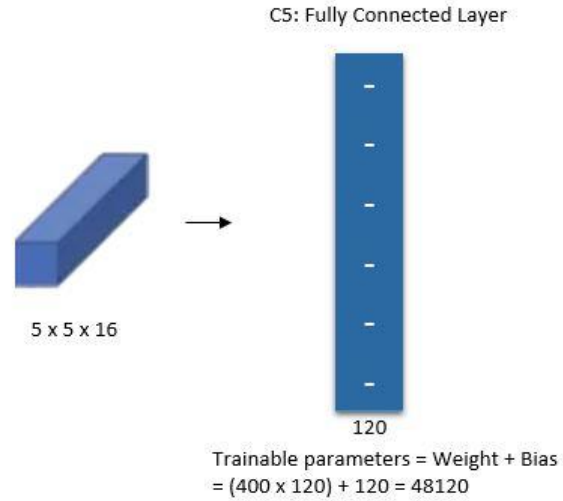


Layer S4

- Subsampling layer (pooling layer) with 16 feature maps of size 5 x 5
- Each unit in each feature map is connected to a 2 x 2 neighborhood in the corresponding feature map in C3, in a similar way as C1 and S2.
- Layer S4 has 32 trainable parameters (16w+16b)





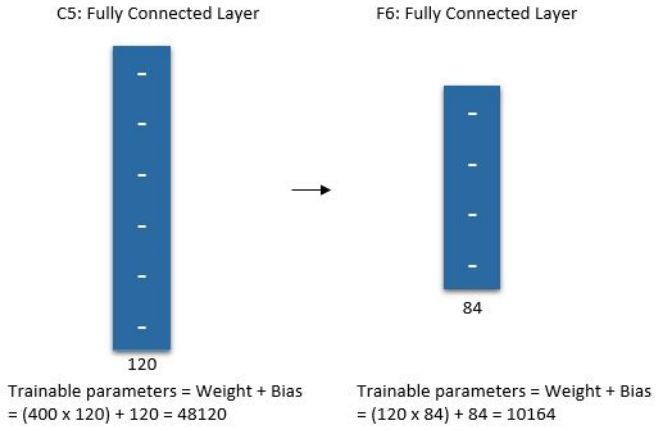


Layer C5

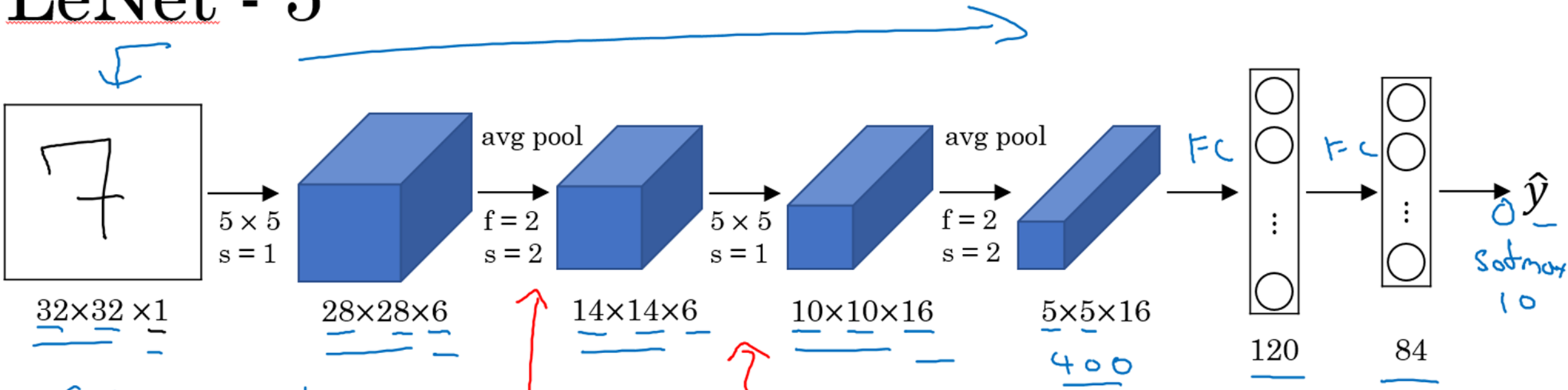
- Convolutional layer with 120 feature maps (120 filters)
- Convolutional filter size : 5 x 5 x 16
- Total trainable parameters : **48120** (48000w+120b)

Layer F6

- Contains 84 neurons and is fully connected to C5.
- Layer F6 has **10164** trainable parameters (10080 w + 84 b)



LeNet - 5



60k parameters.

$n_H, n_w \downarrow$ $n_c \uparrow$

Non-linearity after pooling

$n_H \times n_w \times n_c$ $f \times f \times n_c$

conv pool conv pool fc fc output

Advanced: sigmoid/tanh ReLU

II III

[LeCun et al., 1998. Gradient-based learning applied to document recognition]