CNN as Guided Multilayer RECOS Transform

C.-C. Jay Kuo
University of Southern California

Deep Learning Networks

- Focus on one particular type commonly used for pattern recognition and computer vision:
  - M-P neuron model
  - Multi-Llayer perceptron (MLP)
  - Convolution neural network (CNN)

- Another type
  - Recurrent neural network (RNN)

Part I: Architectural Evolution

Evolution of CNNs

- Computational neuron and logic networks
  - McClulloch and Pitts (1943)
- Multi-Layer Perceptron (MLP)
  - Rosenblatt (1957)
  - Used as “decision networks”
- Convolutional Neural Networks (CNN)
  - Fukushima (1980) and LeCun et al. (1998)
  - AlexNet (2012)
  - Used as “ feature extraction & decision networks”

Artificial Neuron Model (M-P Model)

- McClulloch and Pitts (M-P) neuron model (1943)
  - “All-or-none” characteristics (logic unit)

$$y = \operatorname{sgn}(wx - \varphi)$$

$x = (x_1, x_2, \cdots, x_n)^T$—an input vector

$w = (w_1, w_2, \cdots, w_n)$—a weight vector

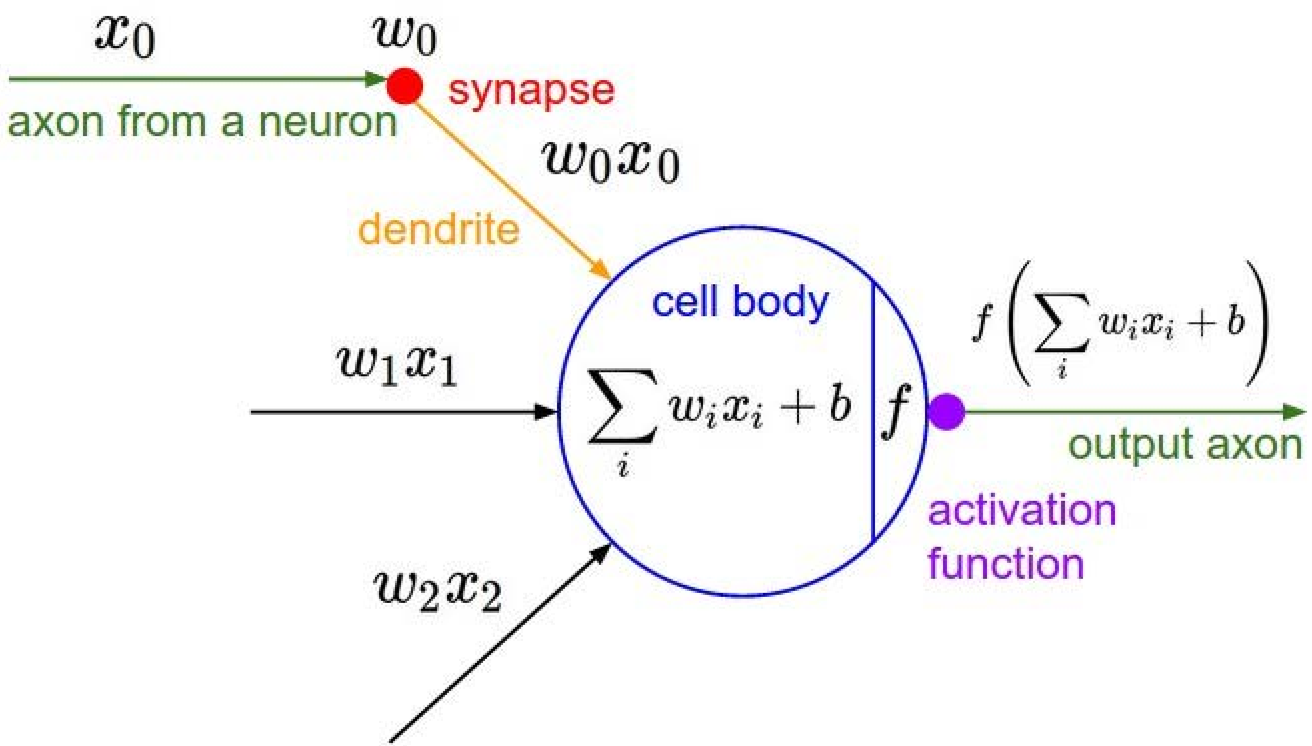
φ—a threshold

$$\operatorname{sgn}(v) = \begin{cases} 1, & v > 0 \\ 0, & v \le 0 \end{cases}$$

Step Activation Function

McCulloch WS, Pitts W. A logical calculus of the ideas immanent in nervous activity. The bulletin of mathematical biophysics. 1943 Dec 1;5(4):115-33.

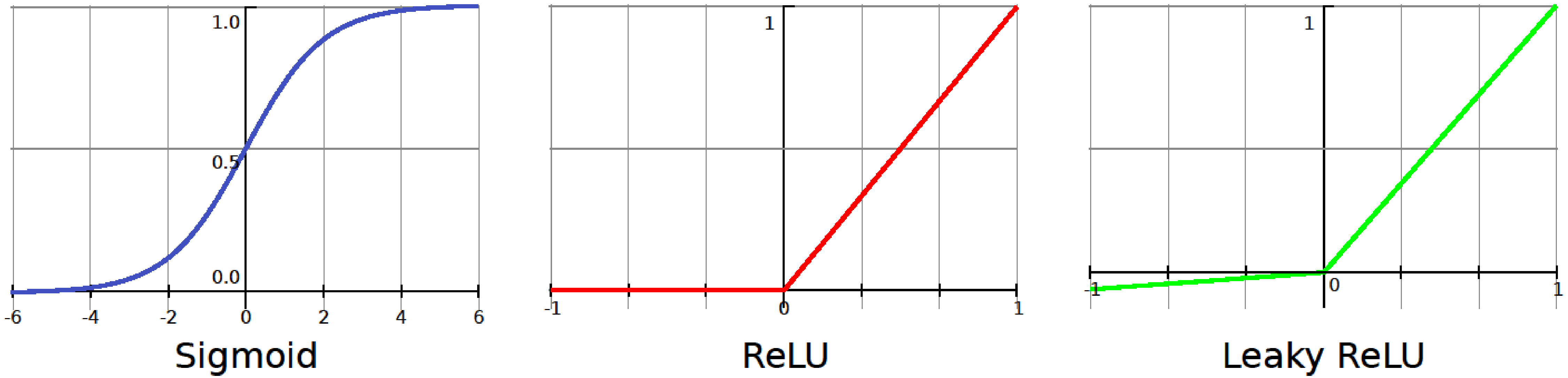
Comparison with Today's Model (Convolution + Nonlinear Activation)



The only difference is the nonlinear activation function – from a step function to other forms (sigmoid, ReLU, Leaky ReLU, etc.)

Figure Credit: Stanford CS231n Course Instruction

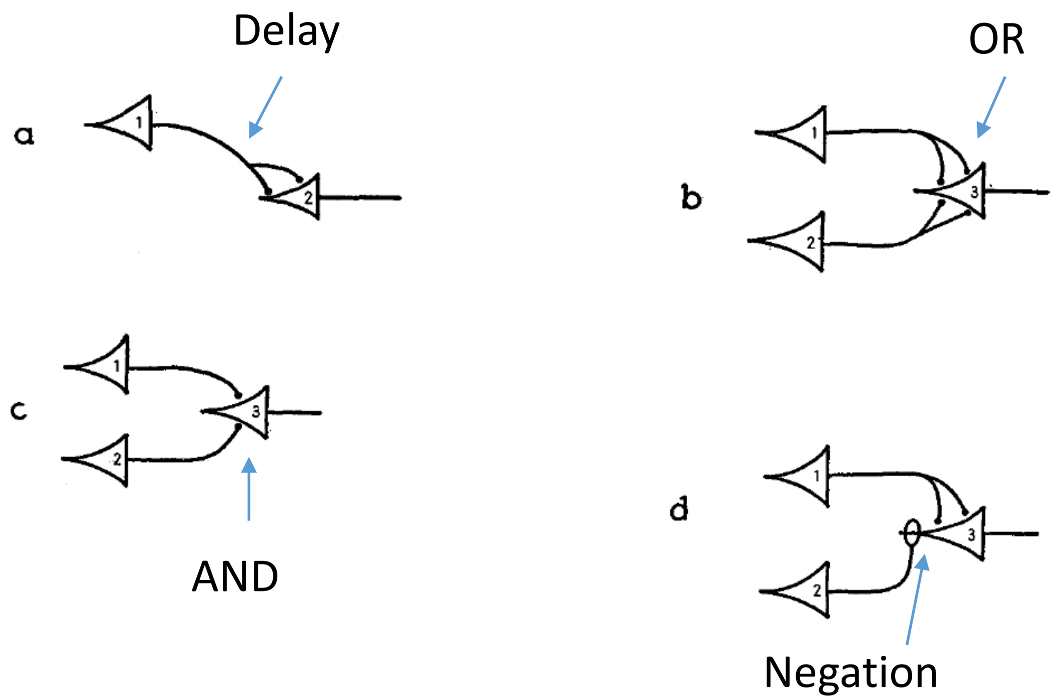
Modern Nonlinear Activation Function



M-P Model and Networks

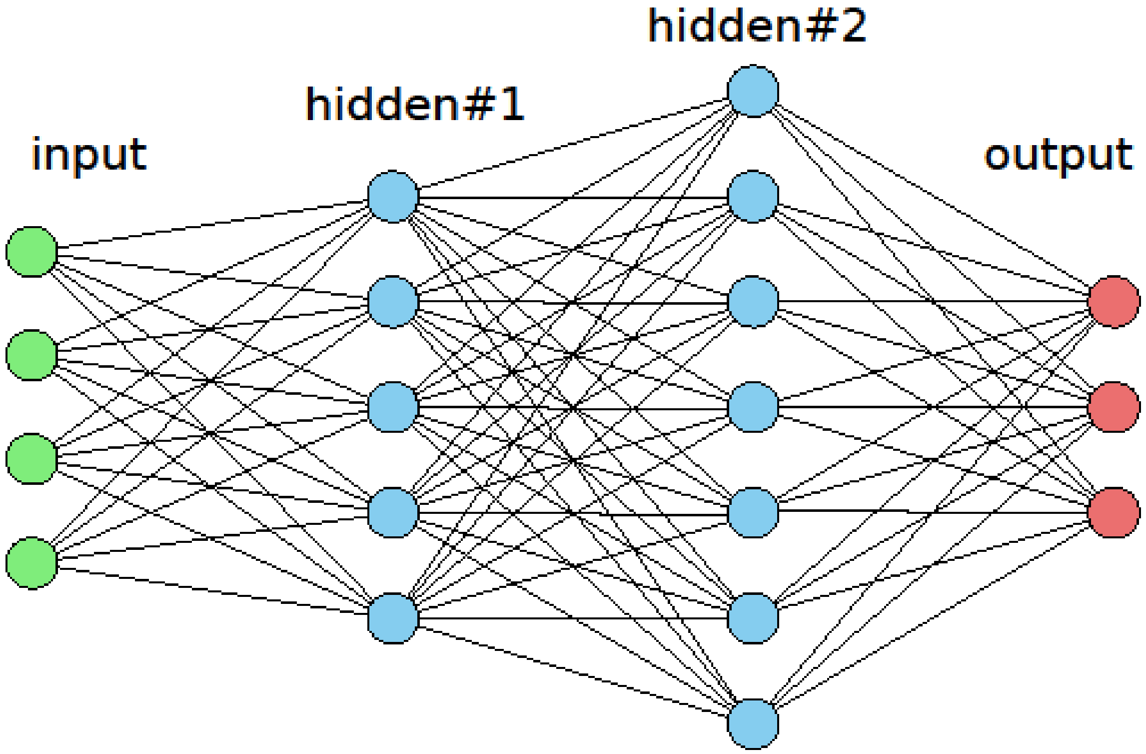
- It contains some basic elements
  - Convolution operation
  - Bias term
  - Nonlinear activation

- The M-P network is a logical circuit
  - A mathematical model used to model nervous system
  - No modern neural network architecture
  - No training considered



Multilayer Perceptron (MLP)

- Supervised learning by backpropagation (BP)
- Highly parallelism
- Fully connection between every two adjacent layers
- No connection between neurons at the same layer



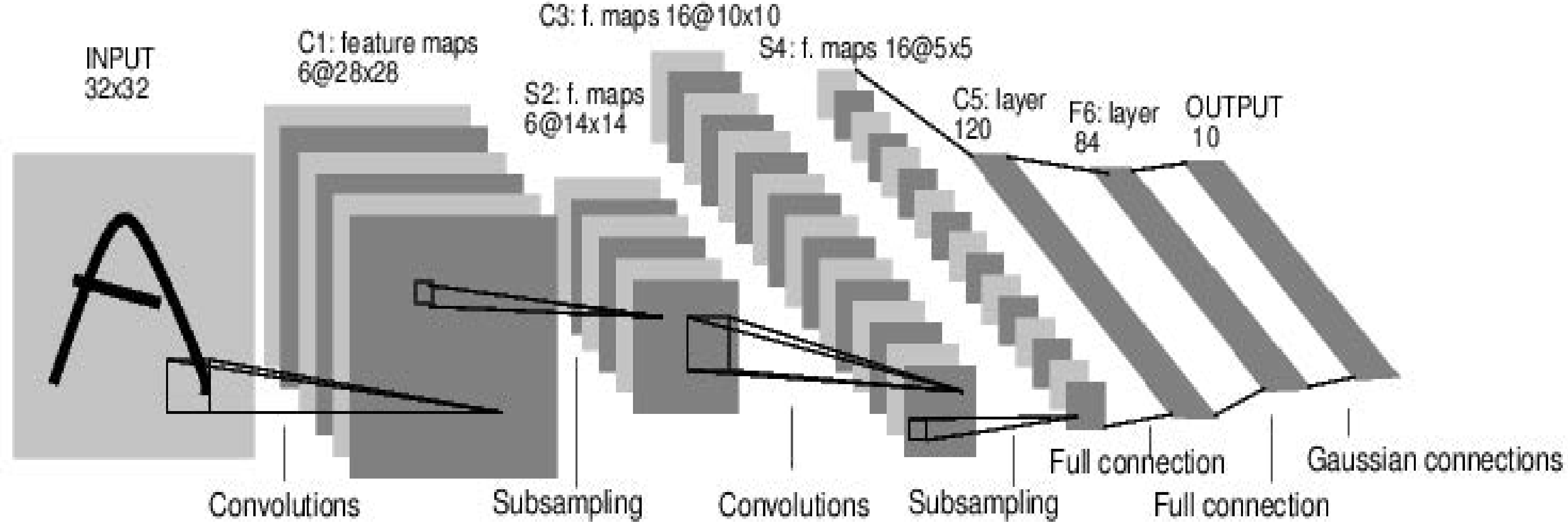
Classic 2-Hidden Layer MLP

Competitions and Limitations

- **MLPs were hot in 80's and early 90's**
  - **Use the n-D feature vector as the input**
  - **One feature per input node (n nodes in total)**
- **Competitive solutions exist**
  - **SVM**
  - **Random Forest**
- **What happens if the input is the source data?**
  (e.g. an image of size 32x32 = 1024)

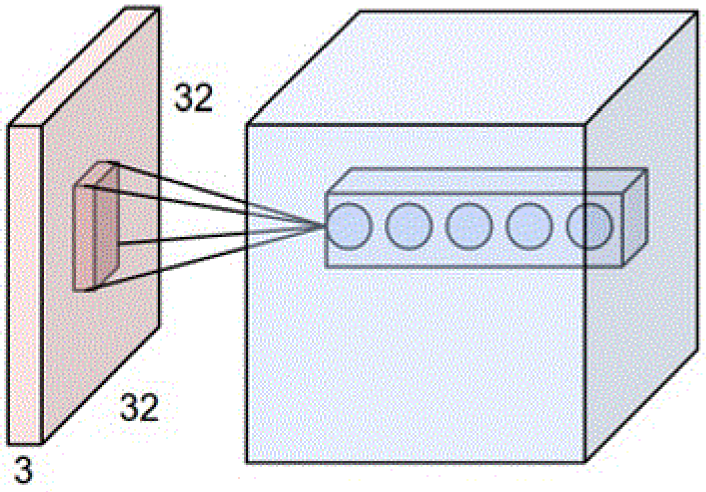
Modern Convolutional Neural Network (CNN)

- LeNet-5



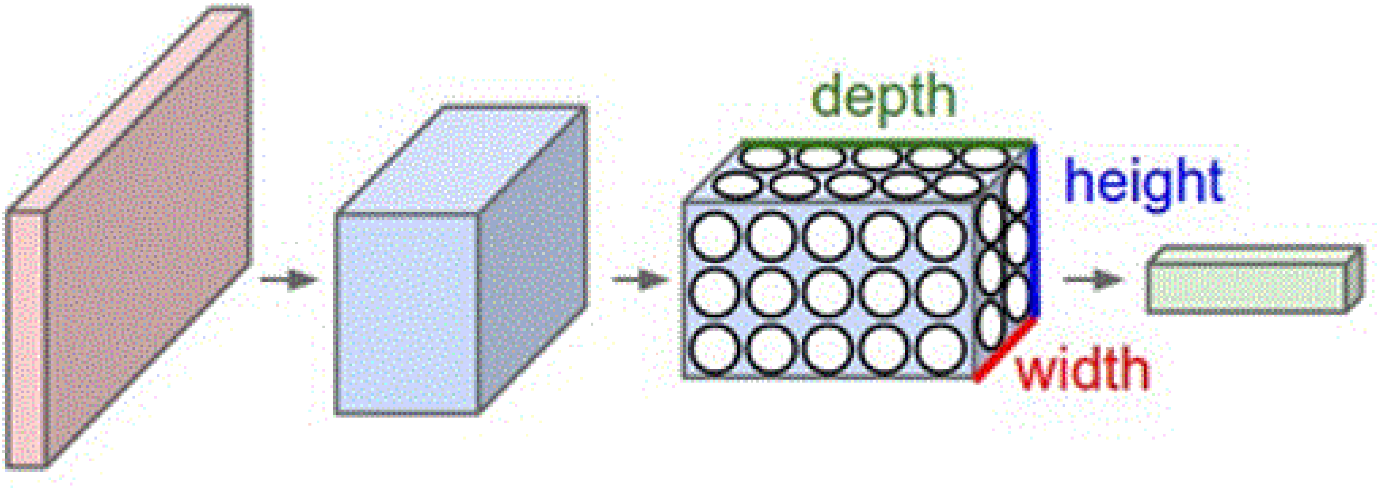
- Can handle a large image by partitioning it into small blocks
- Convolutional layers -> feature extraction module
- Fully connected layers -> decision module
- Two modules are back-to-back connected

Convolutional Layer

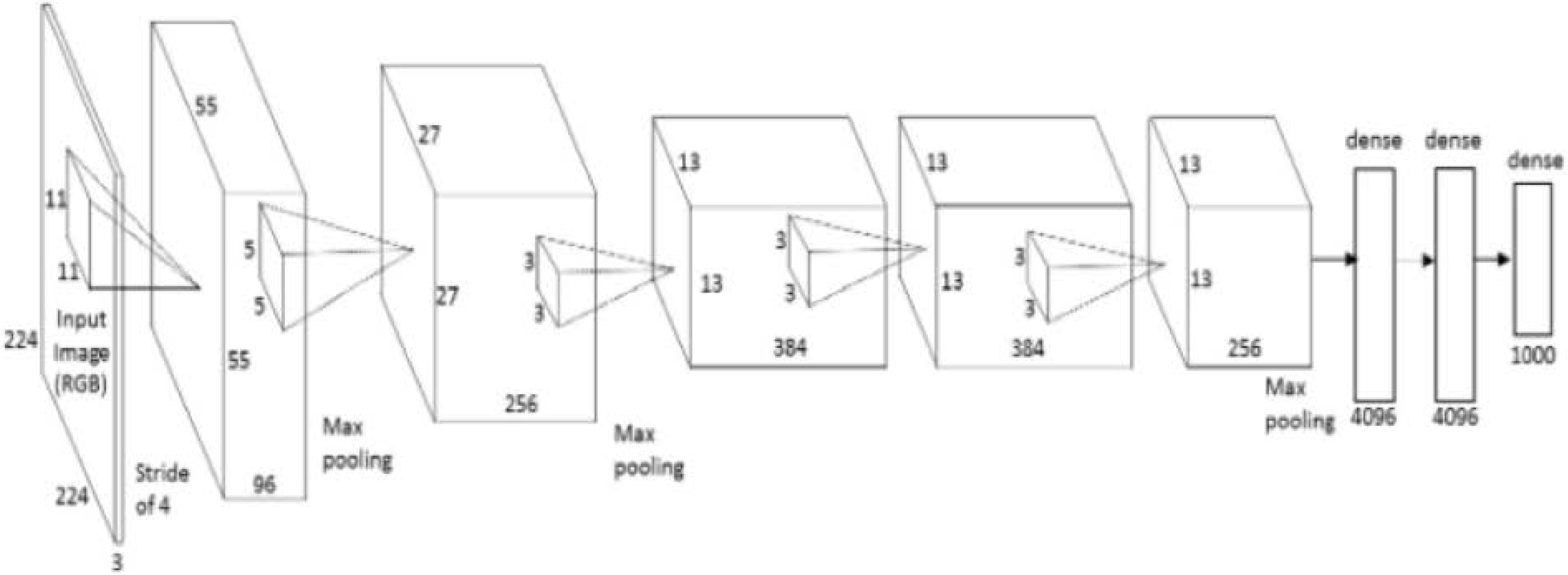


Operations in one convolutional layer

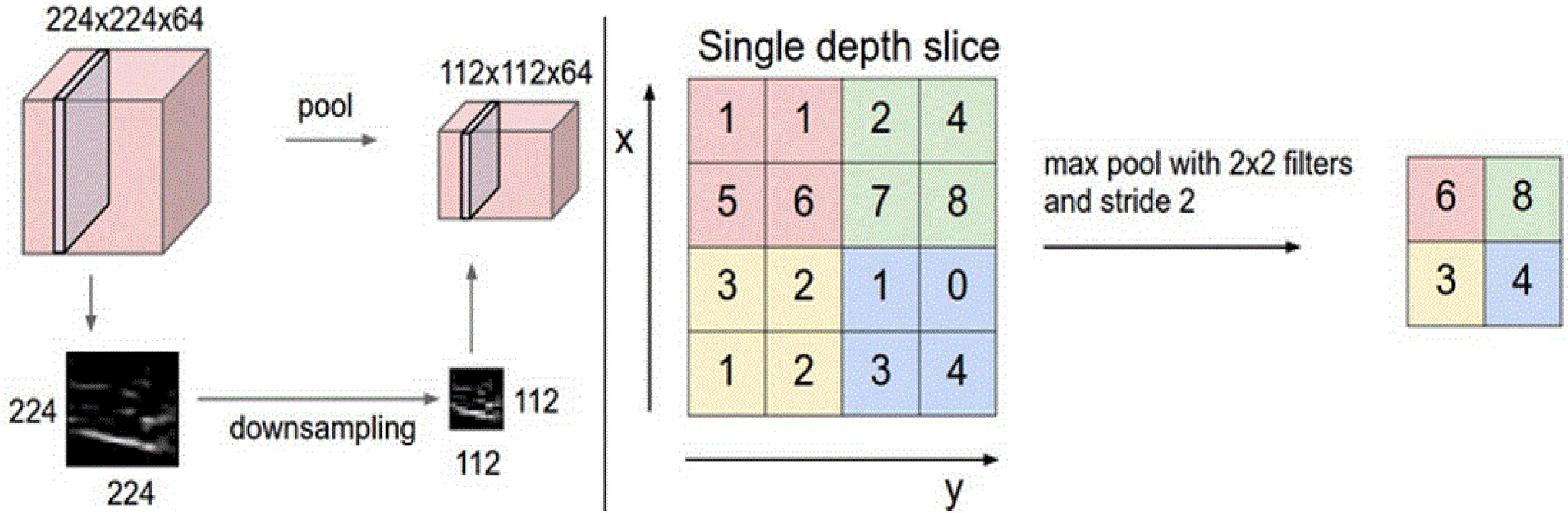
Multiple Convolutional Layers in Cascade



Alex Net


55
27
13
13
13
dense
dense
dense
11
11
5
5
3
3
3
3
3
3
27
13
13
13
224
Input
Image
(RGB)
55
384
384
256
1000
256
Max
pooling
4096
4096
Max
pooling
Max
pooling
224
Stride
of 4
96
3

Max Pooling



224x224x64
pool
112x112x64
224
downsampling
112
224
112
Single depth slice
x
y
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

Connection between LeNet-5 and MLP

- The Input to layer C5 in LeNet-5 is the output from S5 (a hybrid spatial-spectral feature space)
- “S4->C5->F6->Output” can be viewed as a 2-hidden layer MLP
- Generally, a CNN consists of two sub-networks
  - Feature extraction sub-network (a feature vector extractor)
  - Decision sub-network (a classifier)
  - They are inter-connected
    - By conducting BP up to S4, we train the decision module only (learning a new decision network)
    - By conducting BP up to input, we train both the feature and decision modules (“learned features + learned decisions” versus “handcrafted features + learned decisions”)

Part II: Theoretical Foundation

Three Viewpoints

- Signal Processing Viewpoint
- Approximation Theory Viewpoint
- Optimization Theory Viewpoint

Single Layer Signal Analysis (1)

- Signal Modeling

$$\mathbf{x} = \mathbf{A}\mathbf{c},$$

$$\mathbf{A} \in R^{N \times M}$$

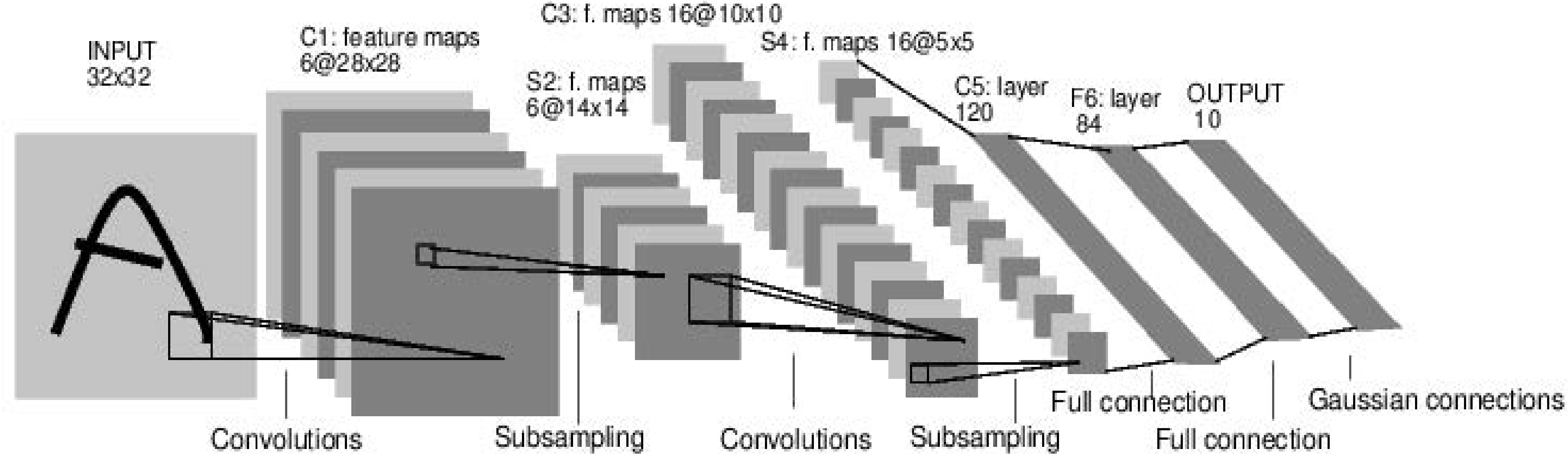
- **x** are a class of observed signals
- **A** and **c** are to be determined

Single Layer Signal Analysis (2)

- Signal Transform (M=N)
  - Fourier transform: sinusoid components in **x**
  - Wavelet transform: multi-scale components in **x**
- Sparse Coding (M>N)
  - Find the most suitable dictionary **A** for **x** under constraints on **c** (e.g. sparsity)
  - Dictionary learning
- Feature extraction
  - Coefficient **c** for an observed instance, **x,** can be used as its features

Where CNN Stores "Learned Knowledge"?

- All training/learning results are summarized in filter weights
  - Filter weights play a critical role in understanding CNN



Each convolutional or fully connected layer defines a transform matrix

CNN as Multi-Layer Signal Transform

- Comparison of single- and multi-layer methods

Single-layer Approach

- There is only one transform matrix
- Learning **A** from a class of signals
- Determine **c** from an instance of **x**
- Use **c** as the features for decision

Multi-layer Approach

- There are multiple transform matrices
- Learning **A's** from a class of signals and their decision labels (**d**)
- Feed an instance of **x** into the network for its decision **d**
- Need a nonlinear activation between layers

Road Map

- Explain the operation of “one perceptron layer” as “clustering”
- Why nonlinear activation?
- Benefits of cascaded layers?
- Explain the multi-layer signal transform
- What is the self-organization property?
- What is the role of supervised learning?

Operation in one Perceptron Layer

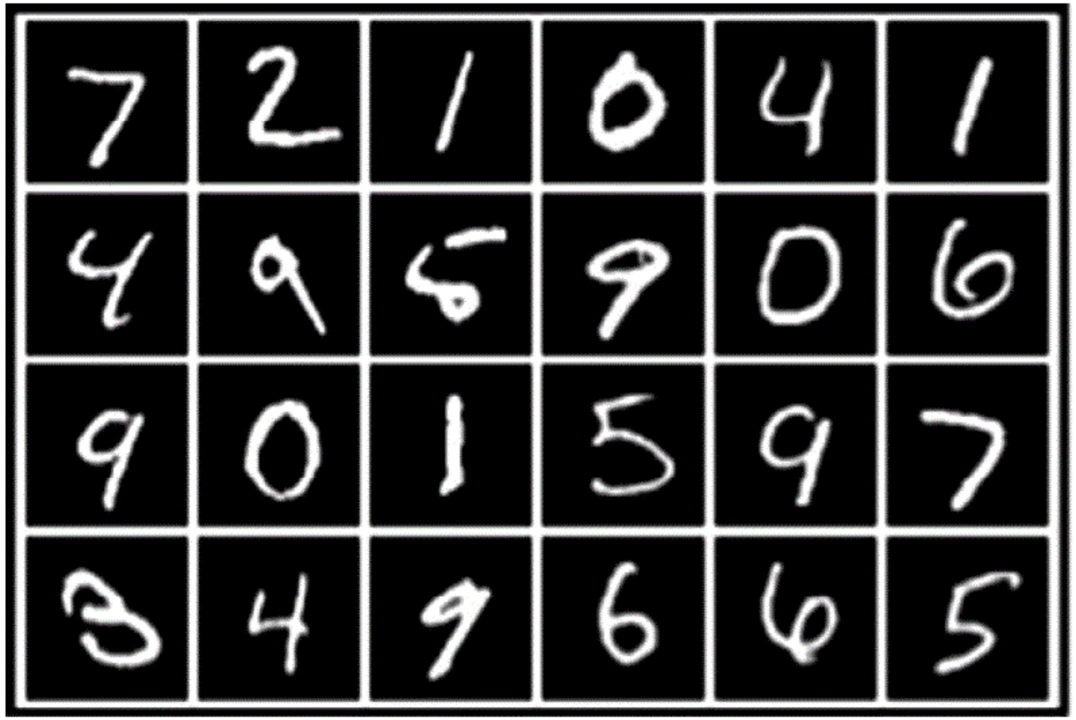
$$\mathbf{y} = \mathbf{A}\mathbf{x}, \quad \mathbf{A}^T = [\mathbf{a}_1 \cdots \mathbf{a}_k \cdots \mathbf{a}_K]$$

$$y_k = \mathbf{a}_k^T \mathbf{x} \text{ and } \mathbf{A} \in R^{K \times N}$$

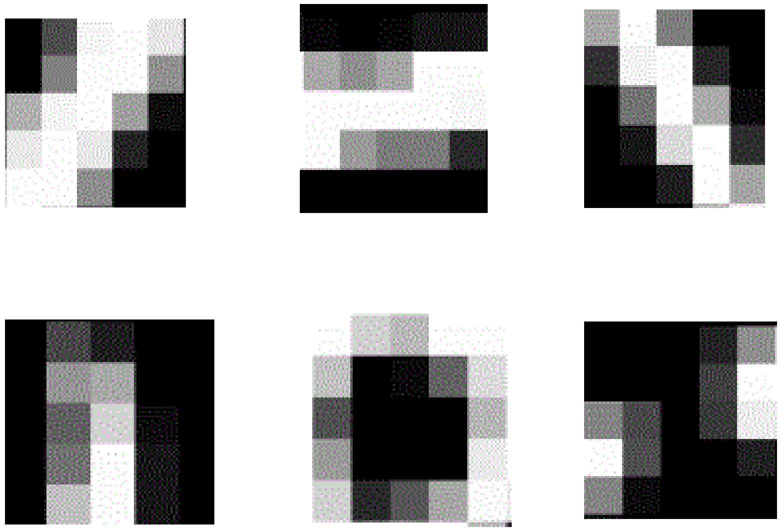
$$\mathbf{y} = (y_1, \cdots, y_k, \cdots y_K)^T \in R^K$$

We view $\mathbf{a}_k$ as a visual pattern

MNIST Dataset



6 Representative Patterns



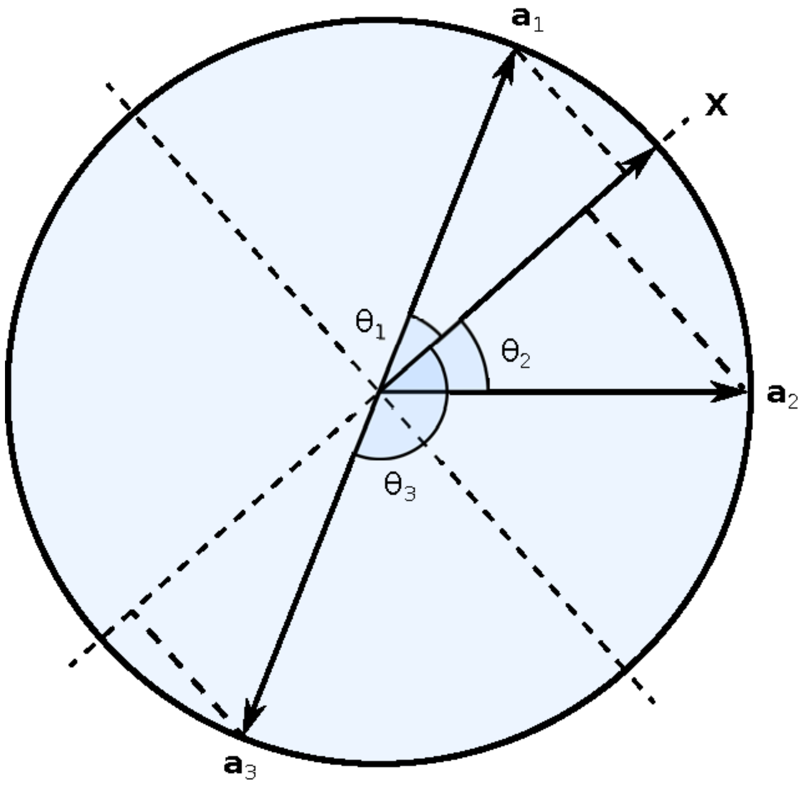
Pattern Matching by Correlation $y_k = \mathbf{a}_k^T \mathbf{x}$

Convolution is “Vector Inner Product” or “Projection”

- All intermediate layers contain convolutional operations:
  - Convolutional layers
  - Fully connected layers
- A convolution operation can be viewed as the inner product to two vectors
- Filter Weights are fixed in the test stage
  - Called anchor vectors
- Why rectification is essential?

REctified COrrelation on a Sphere (RECOS) Model

- Consider clustering in the unit sphere
- The distance is measured by the geodesic distance
- A shorter geodesic distance implies a small intersection angle between two vectors
- What happens to negative correlation (or projection)?

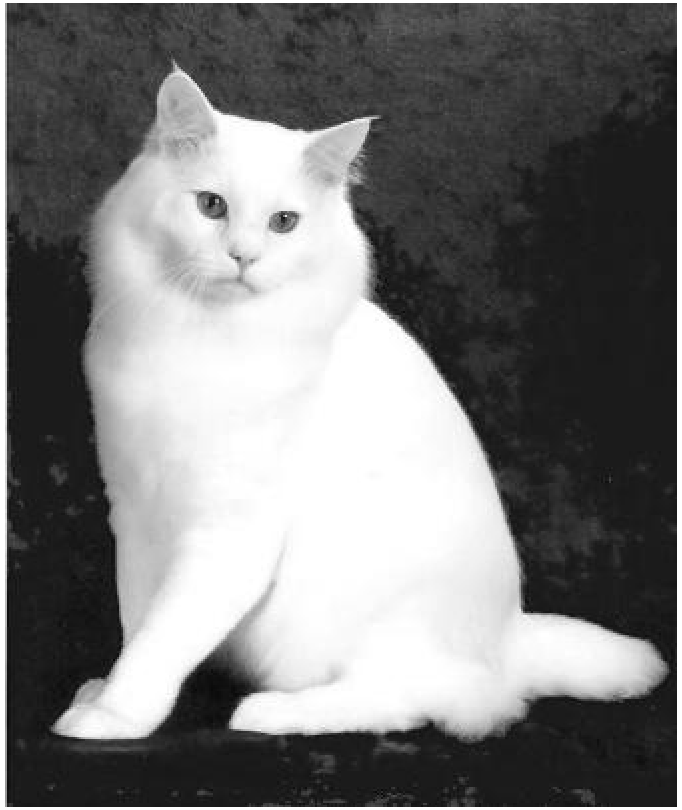


C.-C. Jay Kuo, "Understanding Convolutional Neural Networks with A Mathematical Model", the Journal of Visual Communication and Image Representation, Vol. 41, pp. 406-413, 2016

Physical Meaning of “Unit Sphere”

- Local mean removal before the inner product
  - A constant does not carry visual pattern information
  - The constant effect can be added at the output of the inner product
- Normalized magnitude
  - The magnitude of an input patch is its contrast
  - Low contrast -> weak visual pattern information -> treated as zero vector
  - Other cases -> contrast adjustment has little impact on the visual pattern

Comparison of Positive & Negative Correlations



Confusion Caused by Negative Correlations

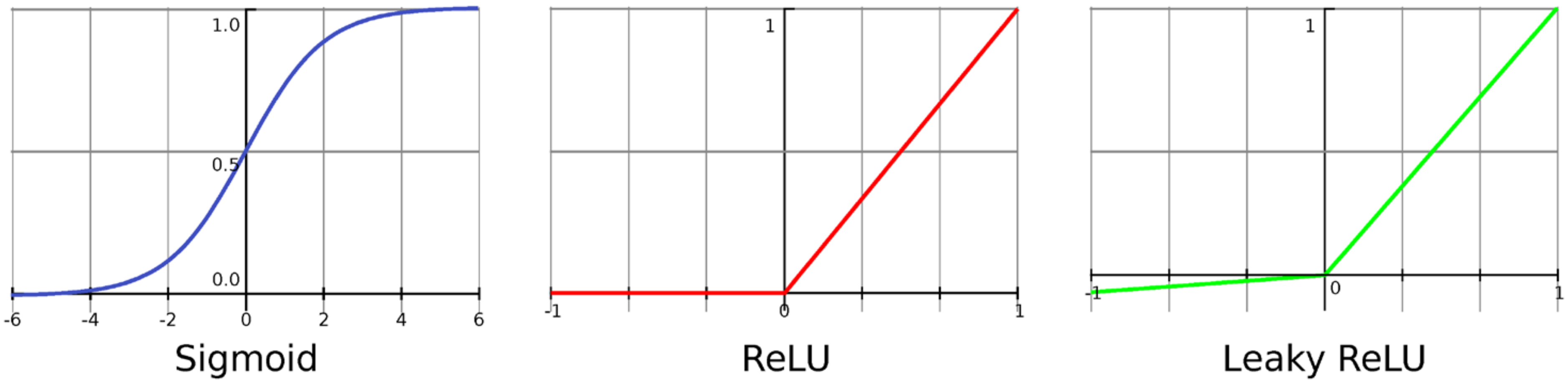
- When two convolutional filters are in cascade, the cascaded system cannot differentiate the following scenarios:

- Confusing Case #1
  - A positive correlation in stage 1 and a positive filter coefficient in stage 2
  - A negative correlation in stage 1 and a negative filter coefficient in stage 2

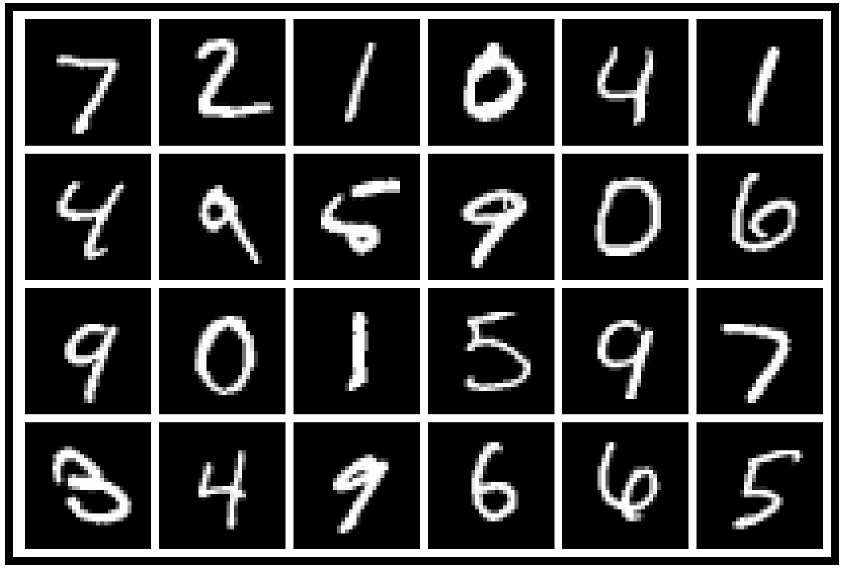
- Confusing Case #2
  - A positive correlation in stage 1 and a negative filter coefficient in stage 2
  - A negative correlation in stage 1 and a positive filter coefficient in stage 2

Nonlinear Activation Functions:

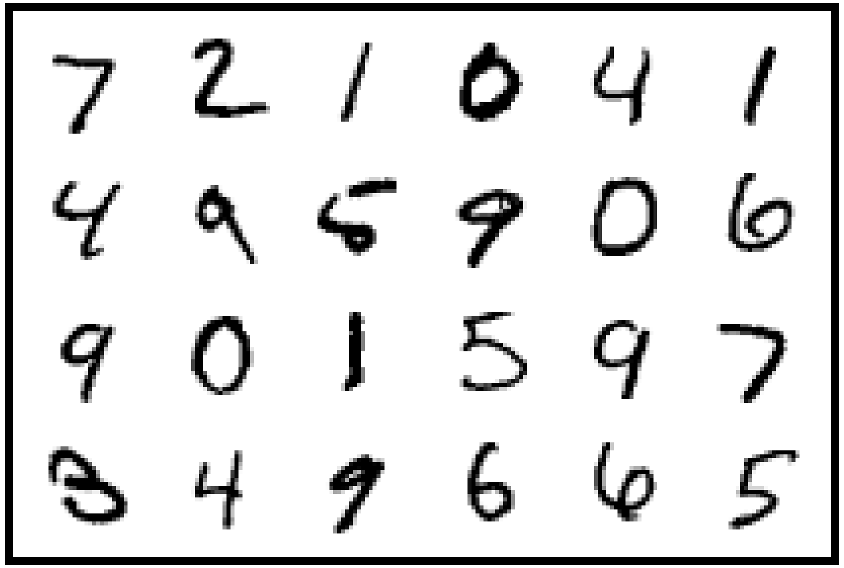
- When two convolutional filters are in cascade, nonlinear activation is used to clip negative correlations



Experiments on MNIST



Original



Negative

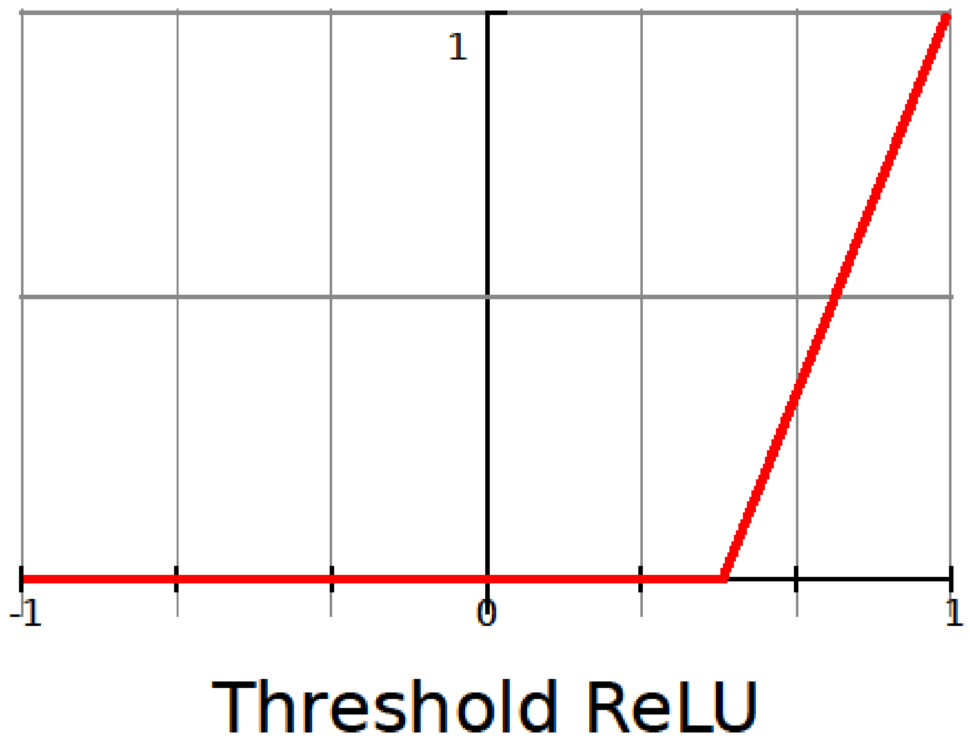
Test Performance of LeNet-5

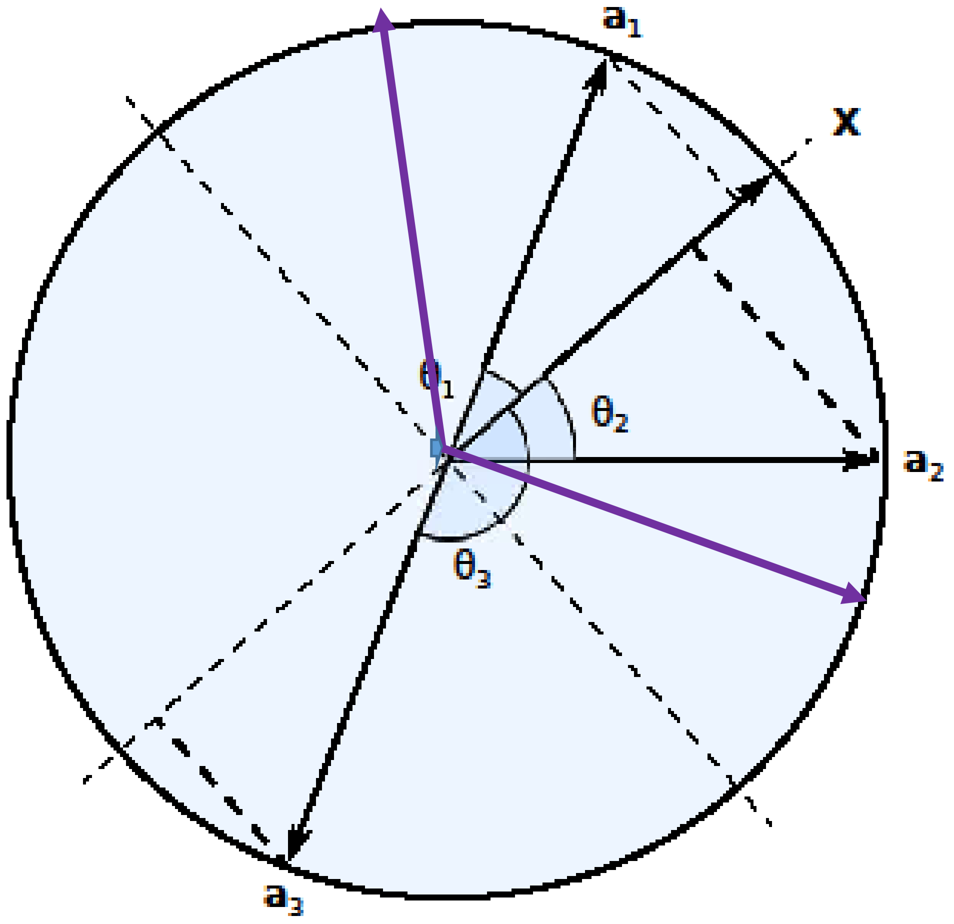
- Original: 98.94% (trained by original)
- Negative: 37.36% (trained by original)

Test Performance of LeNet-5

- Original: 37.36% (trained by negative)
- Negative: 98.94% (trained by negative)

More about Rectification: Threshold ReLU

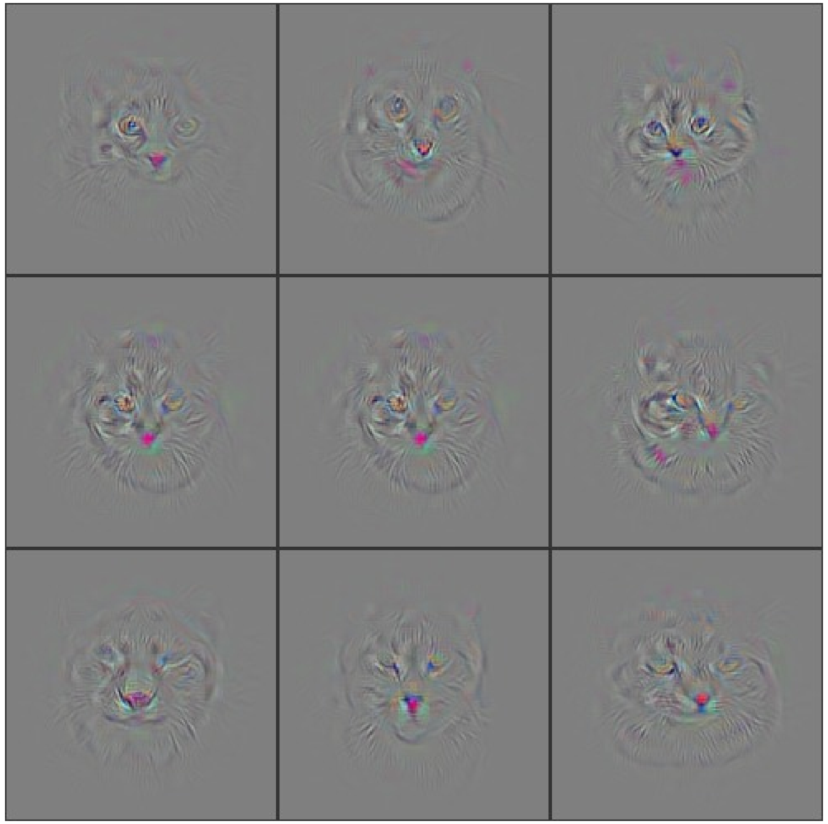




Benefit of Cascaded RECOS Model

What are the common salient regions of all 9 cat Images ?

Top 9 Input Activation Images



Deconv Image

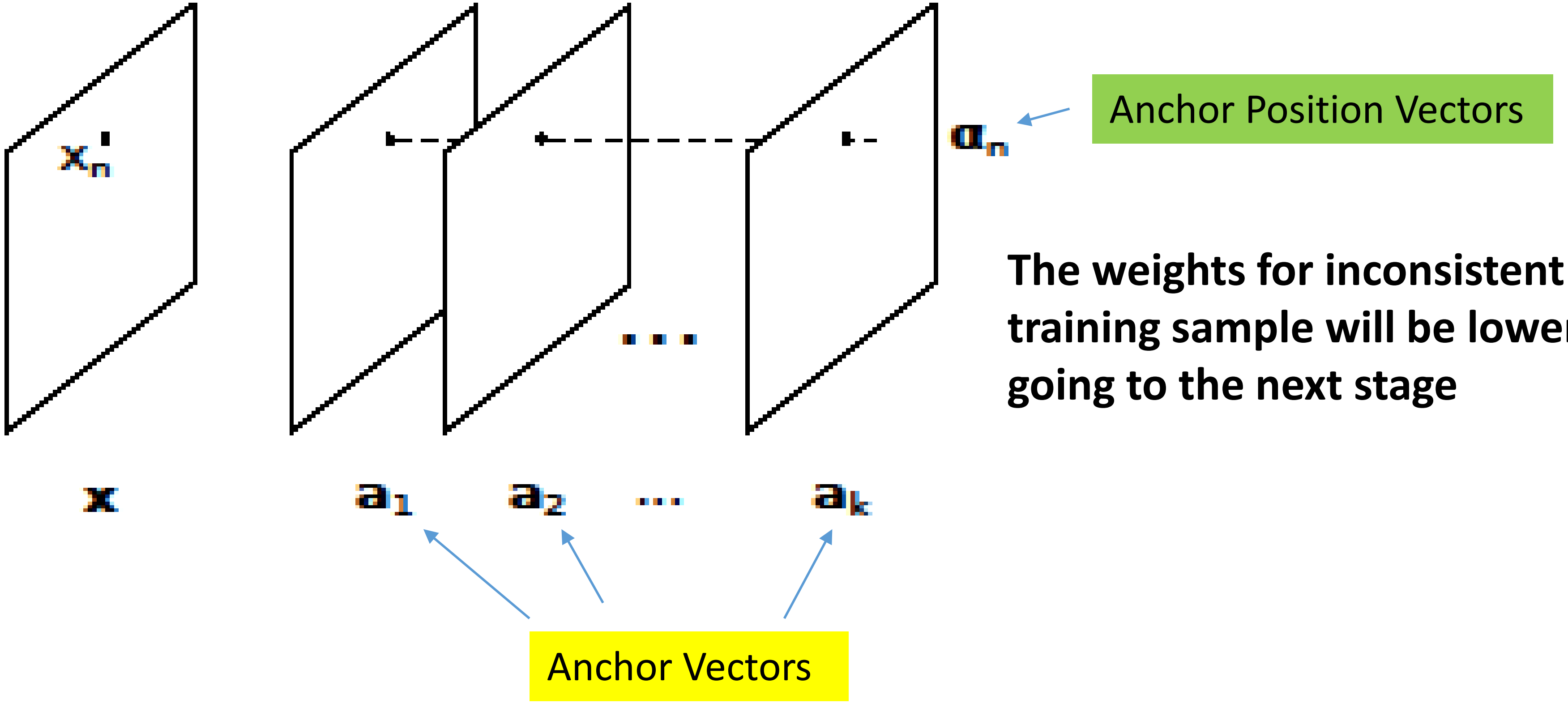
Can CNN extract them automatically?

Consistency Across Multiple Samples of the Same Class

Foreground is consistent while background is not

Why Background Being Removed?

- Inconsistent background can be removed since its variance is higher

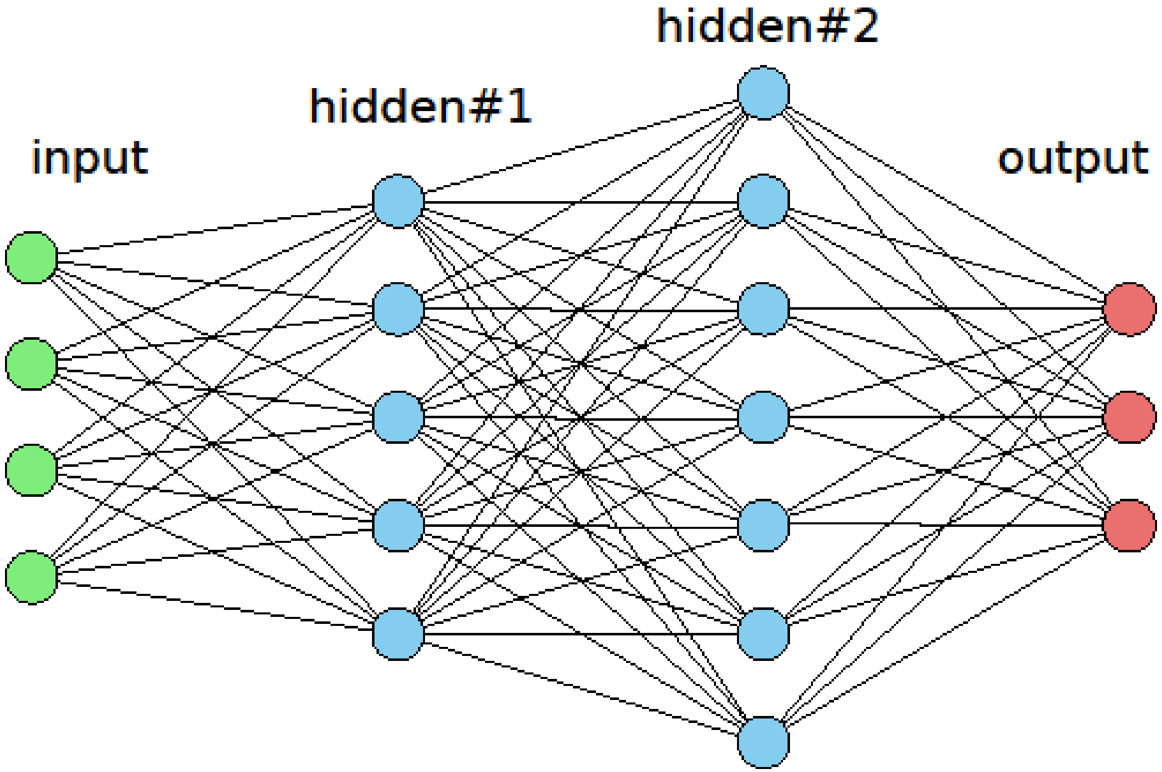


The weights for inconsistent APV across training sample will be lowered when going to the next stage

Handwritten Digits with Color Background

Can CNN recognize these digits with background?
If there is no correlation between the background and digits, it is feasible

Guided Multi-Layer RECOS Transform (1)

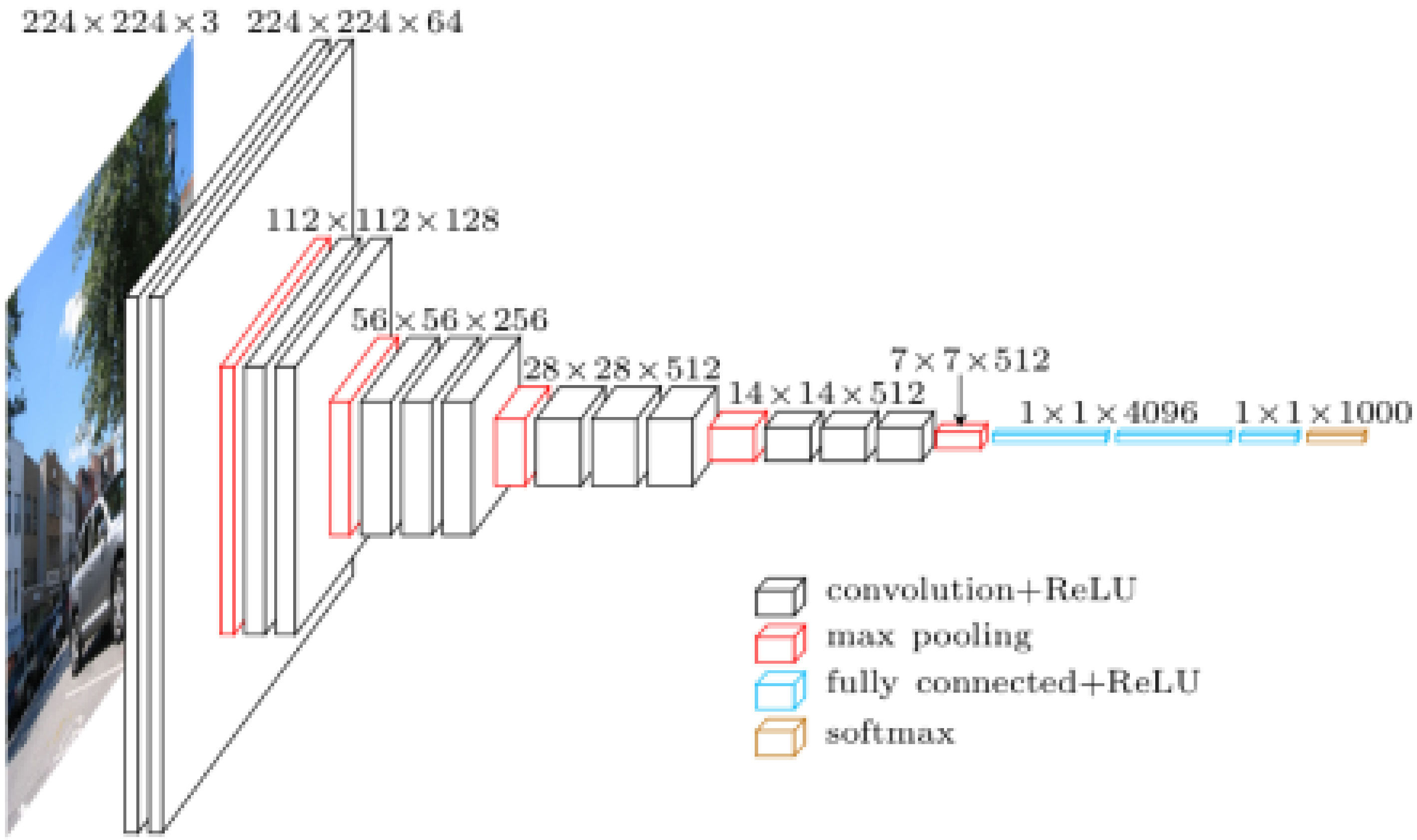


$$\mathbf{d} = \mathbf{B}_L \cdots \mathbf{B}_l \cdots \mathbf{B}_1 \mathbf{x},$$

$$\mathbf{B}_l = \mathbf{R} \circ \mathbf{A}_l$$

$$\mathbf{x} = \mathbf{x}_0 \xrightarrow{\mathbf{B}_1^F} \mathbf{x}_1 \xrightarrow{\mathbf{B}_2^F} \cdots \xrightarrow{\mathbf{B}_{L-1}^F} \mathbf{x}_{L-1} \xrightarrow{\mathbf{B}_L^F} \mathbf{x}_L = \mathbf{d},$$

Guided Multi-Layer RECOS Transform (2)



$$\mathbf{B}_l^C = \mathbf{P} \bigcup_{s\in\Omega} \mathbf{R} \circ \mathbf{A}_{l,s},$$

$$\mathbf{B}_l^F = \mathbf{R} \circ \mathbf{A}_l,$$

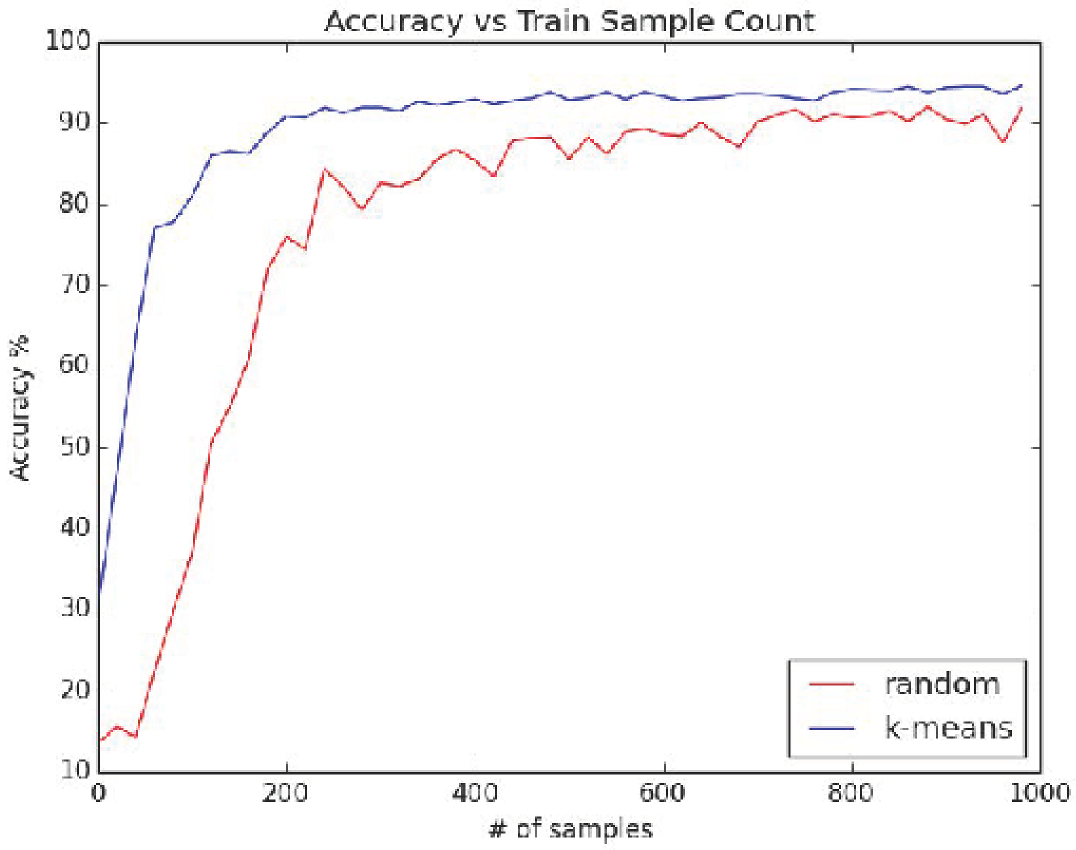
$$\mathbf{x} = \mathbf{x}_0 \xrightarrow{\mathbf{B}_1^C} \mathbf{x}_1 \xrightarrow{\mathbf{B}_2^C} \cdots \xrightarrow{\mathbf{B}_m^C} \mathbf{x}_m \xrightarrow{\mathbf{B}_{m+1}^F} \mathbf{x}_{m+1} \xrightarrow{\mathbf{B}_{m+2}^F} \cdots \xrightarrow{\mathbf{B}_{L-1}^F} \mathbf{x}_{L-1} \xrightarrow{\mathbf{B}_L^F} \mathbf{x}_L = \mathbf{d},$$

CNN Self-Organization Property

- Self-organization property
  - Learning without a teacher [1]
    - The network is repeatedly presented with a set of stimulus patterns to the input layer, but it does not receive any label about the patterns
    - One can cluster all kinds of dogs together without knowing their names
    - Unsupervised learning
  - This property was examined in depth in 80’s and 90’s, yet its significance is dropped in recent years

- CNN provides a wide spectrum solution
  - From un-supervised to weakly and heavily supervised learning paradigms

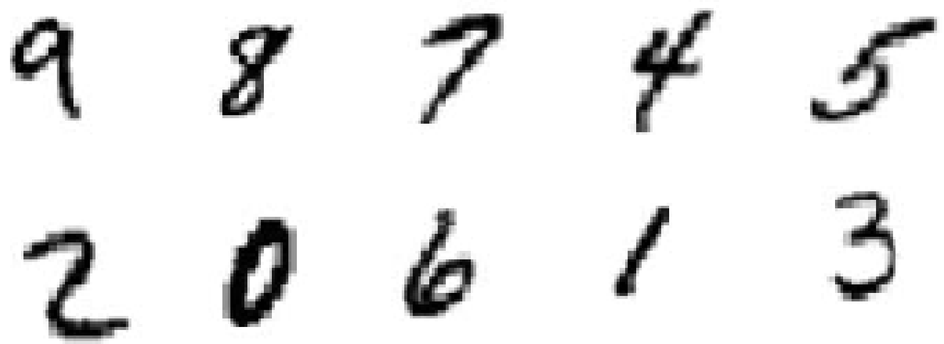
Comparison of LeNet-5 Initializations (1)



Comparison of LeNet-5 Initializations (2)

Random Initialization

K-Means Initialization



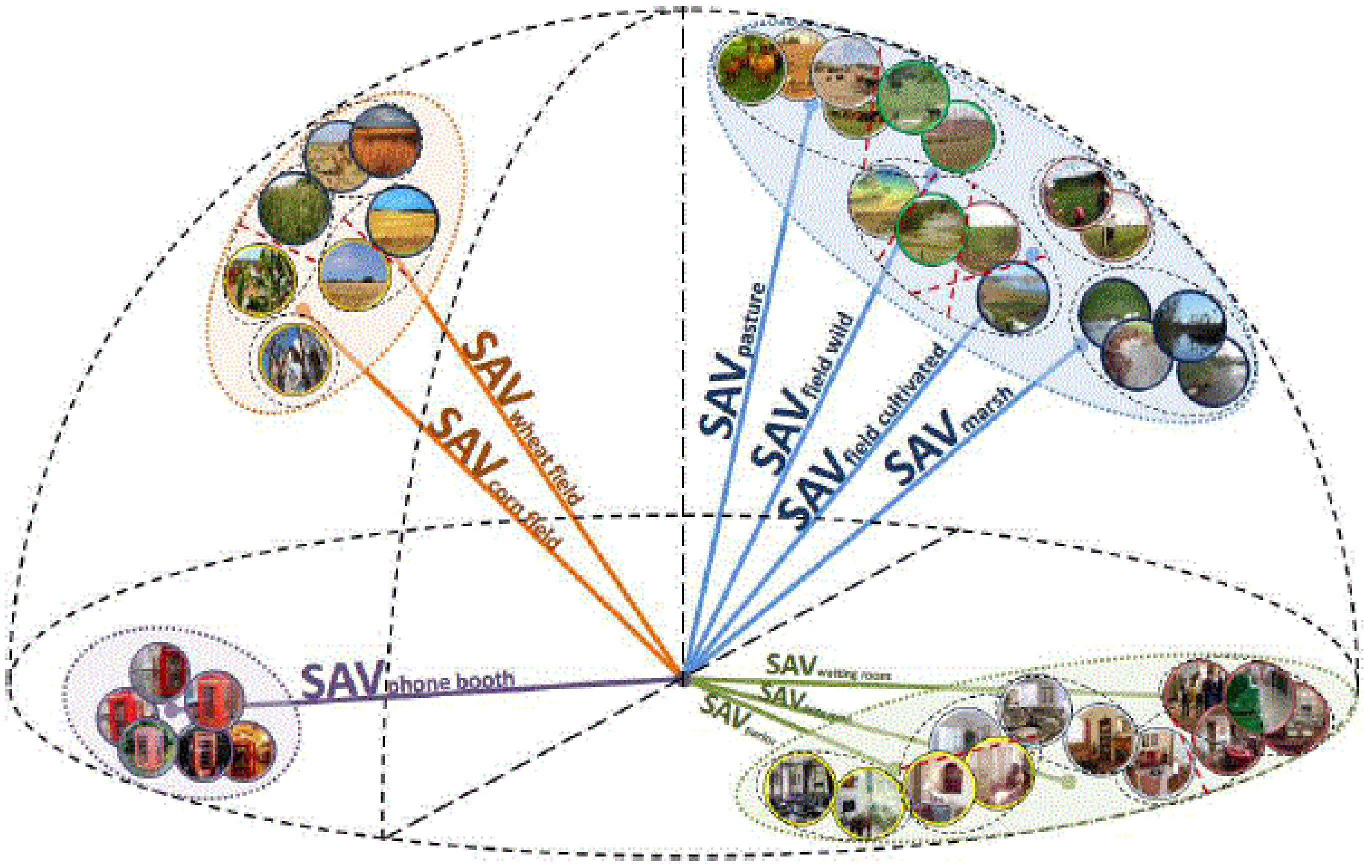
Comparison of LeNet-5 Initializations (3)

Averaged Orientation Changes of Anchor Vectors

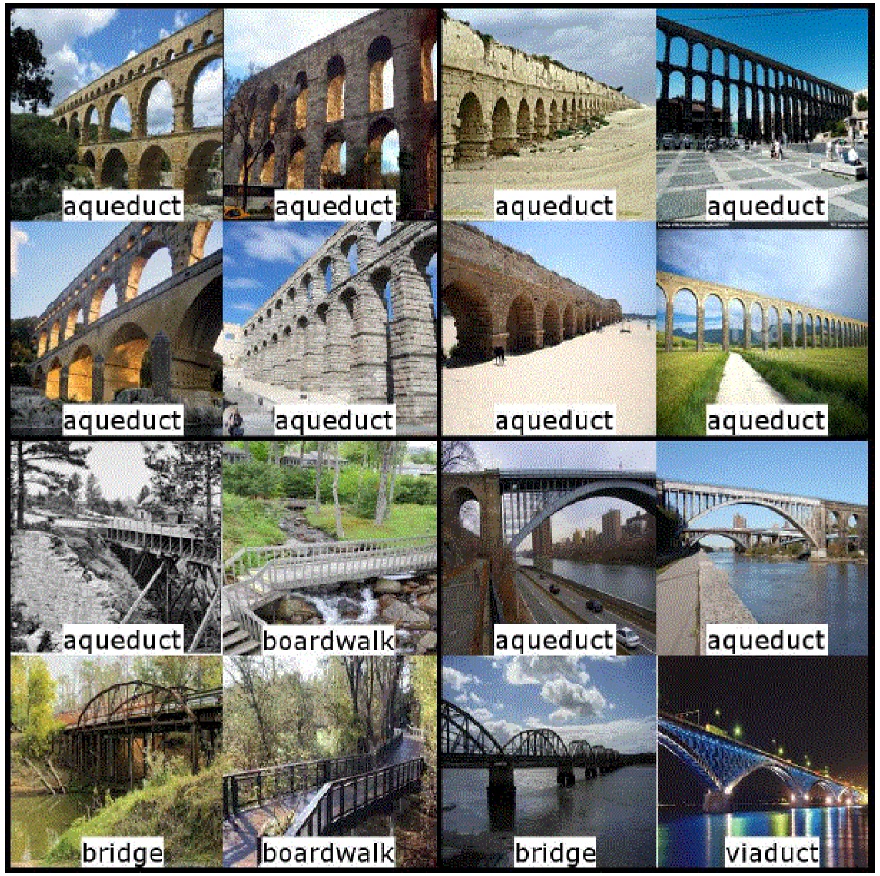
In/Out layers	k-means	random
Input/S2	0.155 (or 8.881^o)	1.715 (or 98.262^o)
S2/S4	0.169 (or 9.683^o)	1.589 (or 91.043^o)
S4/C5	0.204 (or 11.688^o)	1.567 (or 89.783^o)
C5/F6	0.099 (or 5.672^o)	1.579 (or 90.470^o)
F6/Output	0.300 (or 17.189^o)	1.591 (or 91.158^o)

Scene Anchor Vectors

Each anchor vector is associated with a scene class label



Four Sub-classes under Aqueduct Class obtained via unsupervised split



Unsupervised Split of the “Snake” Class

Road Map Revisited

- Explain the operation of “one perceptron layer” as “clustering”
- Why nonlinear activation?
- Benefits of cascaded layers?
- Explain the multi-layer signal transform
- What is the self-organization property?
- What is the role of supervised learning?

Conclusion

- Several known results can be explained using the guided multi-layer RECOS transform
  - Robustness to wrong labels
  - Overfitting
  - Data augmentation
  - Dataset bias

Future Work

- Network architecture design
- Weakly-supervised learning
- Transfer learning
- Localization and attention