Understanding Neural Networks

C.-C. Jay Kuo
University of Southern California

Three Viewpoints

- Approximation Theory Viewpoint
- Optimization Theory Viewpoint
- Signal Analysis Viewpoint

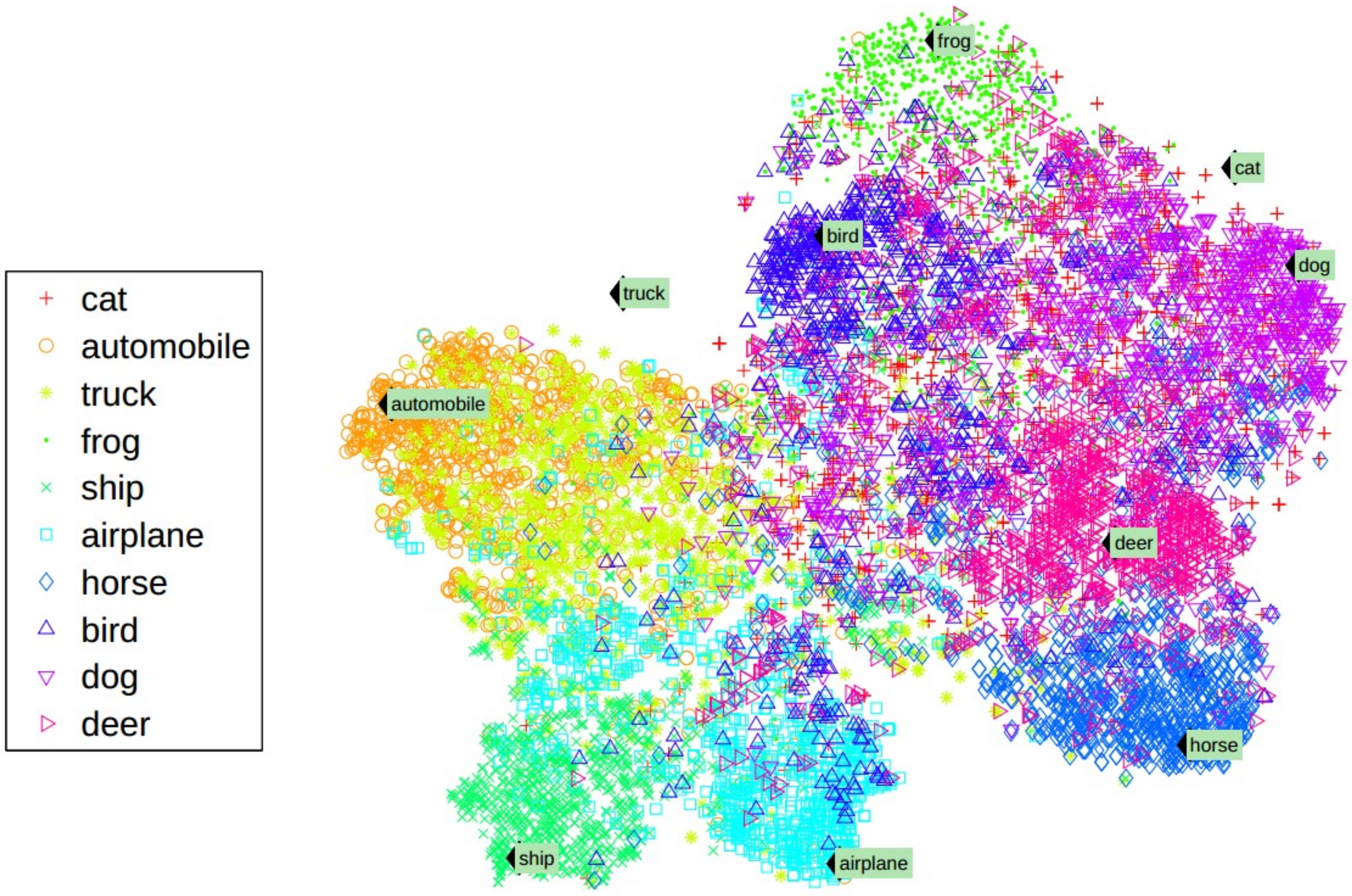
Approximation Theory

Approximation Theory: How?

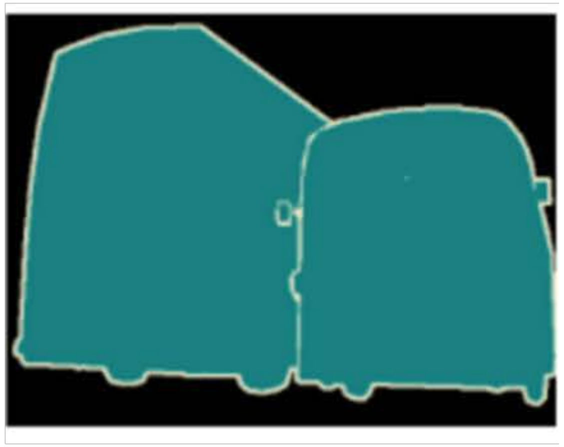
f () = ``Cat”

Embedding Words into High Dimensional Vectors

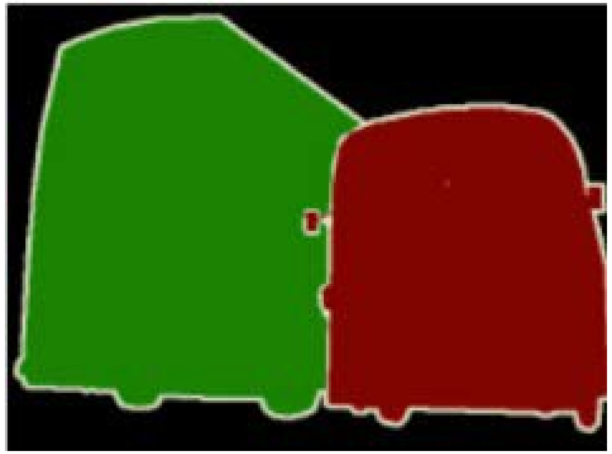
- Input variables:
  - Width
  - Height
  - Bits per pixel
  - Channel numbers
- Output variables:
  - Word Embedding
  - Map objects into a high-dimensional vector



Approximation Theory: How?

f () = 

Class-Based Segmentation

g () = 

Instance-Based Segmentation

From Images to Images

- Output function is now defined at each pixel (rather than the whole images)
- It is important to differentiate instances from classes

$f\Big($  $\Big)$ = ``4 Dogs”

Approximation Theory

- For vision problems
  - Input is either image/video
  - Output can be object classes, scene classes, localizations and even sentence descriptions

Sentence Description Example (from Microsoft CoCo Dataset)

Important: motorbike, person
Unimportant: car

Object tags: car, person, motorbike

- A ***man*** sitting on a porch with two motor ***scooters*** parked outside.
- A ***man*** with his cheeks pushed out and two ***scooters*** to the left.
- A young ***man*** holding his breath.
- A young ***man*** puffs out his cheeks in an outdoor cafe.
- A young ***man*** with a silly look on his face.

Visual Annotation

- Visual annotation by humans is extremely expensive
- If image & video can be annotated by machines, this technique will have a great impact on video files indexing and retrieval
- Usually, image/video annotation is tackled by CNN+RNN
  - CNN processes visual inputs
  - RNN processes sentence (text) outputs

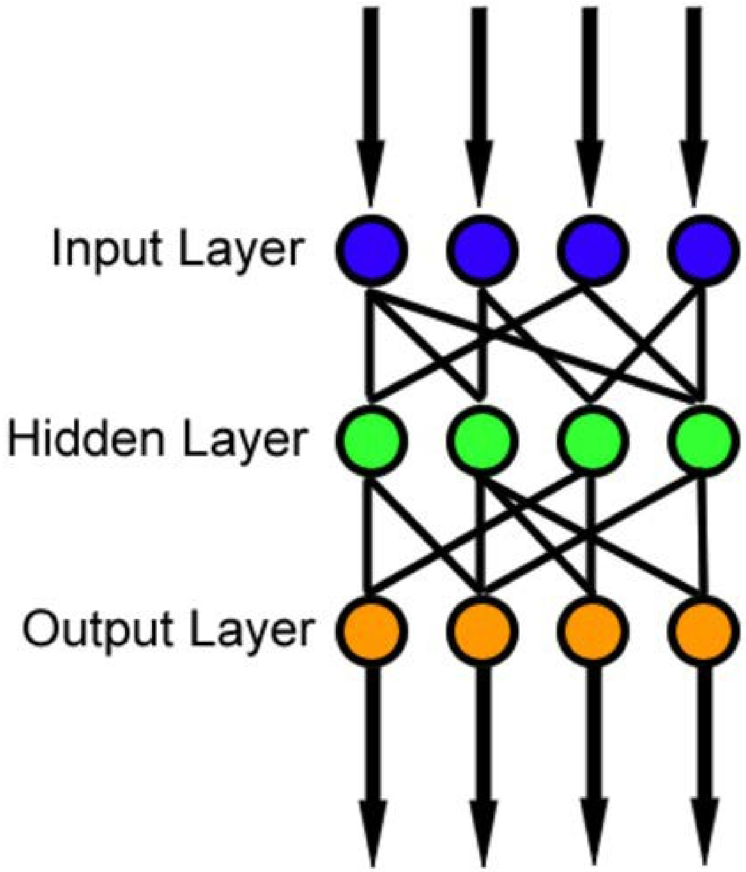
Main Result in Approximation Theory

- Multilayer feedforward networks (i.e. CNNs) with as few as one hidden layer using arbitrary squashing function are capable of approximating any measurable function to any desired degree of accuracy
  - These networks are universal approximators
  - Any lack of success in applications must arise from inadequate learning, insufficient numbers of hidden units, or the lack of a deterministic relationship between input and target
  - Squashing function means the “S” shape function (e.g. sigmoid or logistic function)

Hornik, Stinchcombe and White “Multilayer feedforward networks are universal approximators,” *Neural Networks*. 1989 Dec.

Single-Layer Neural Network

- All missing links can be viewed as a link with weight value “0”
- If there is a single output node, it is a scalar function; otherwise, it is a vector function
- Why squashing function?
  - If there is no squashing function, the hidden layer can be removed by the product of two matrices (necessary but not sufficient)
- Demand a geometrical interpretation



Too Abstract to Comprehend?

- We will revisit it using a more geometric explanation

Optimization Theory

Problem in Loss Function Optimization (Backpropagation)

- The loss function is highly non-convex in a high-dimensional space (the filter weight space)
- Backpropagation is a stochastic descent algorithm used to search optimal weights over the loss surface
- There are many local minima
  - Most solutions are trapped by local minima

Spherical Spin Glass (SSG) Model

- Spin glass is a “disordered magnet”, where the magnetic spin of component atoms are not aligned in a regular pattern

- The spherical model describes a set of particles on a lattice containing *N* sites.
  - For each site *j*, a spin σ_j interacts only with its nearest neighbors and an external field *H*
  - Ising model: $\sigma_j \in \{1, -1\}$
  - Spherical model: $\sum_{j=1}^{N} \sigma_j^2 = N$

Energy Surface of SSG

- What is the distribution of critical points (maxima, minima, and saddle points) of the loss function?
  - There are results from random matrix theory applied to spherical spin glasses
  - These functions have a large number of saddle points
  - While local minima are numerous, they are relatively easy to find and they are all more or less equivalent in terms of performance on the test set

Main Result

- For a N-dimensional spherical spin glass (SSG) model
  - Its minimum energy values depend on the initial state yet form a layered band structure
  - These bands are lower bounded by the global minimum
  - The probability of finding them outside the band diminishes exponentially with *N*
- A link between the above model and the CNN parameter optimization can be established

- A. Choromanska, M. Henaff, M. Mathieu, G. B. Arous, and Y. LeCun. "The Loss Surfaces of Multilayer Networks." In AISTATS. 2015

Distribution of Minima of Loss Function

Spin-Glass

CNN

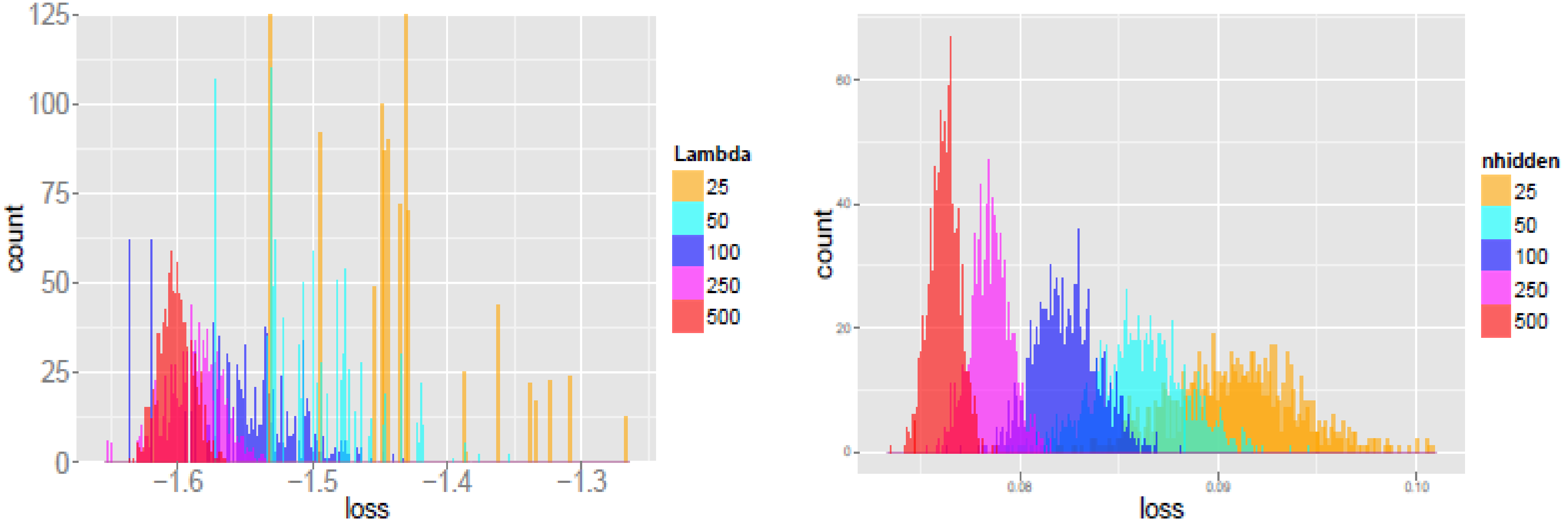


Figure 3: Distributions of the scaled test losses for the spin-glass (left) and the neural network (right) experiments.

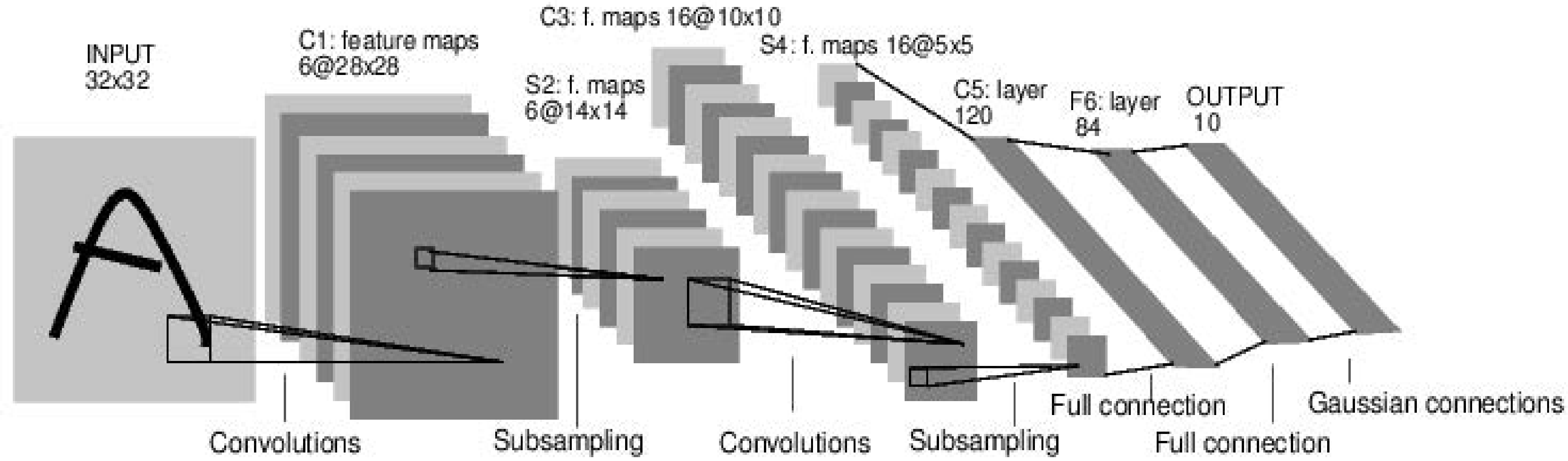
Engineering Practices

- Environments: Linux/Windows
- Software Platform: Caffe, Torch and Tensor Flow, etc.
- Network filter weights initialization
- Training parameters: learning rates, momentum, mini-batch & epoch
- Training techniques: Dropout

Signal Analysis Viewpoint

Where CNN Stores “Learned Knowledge”?

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

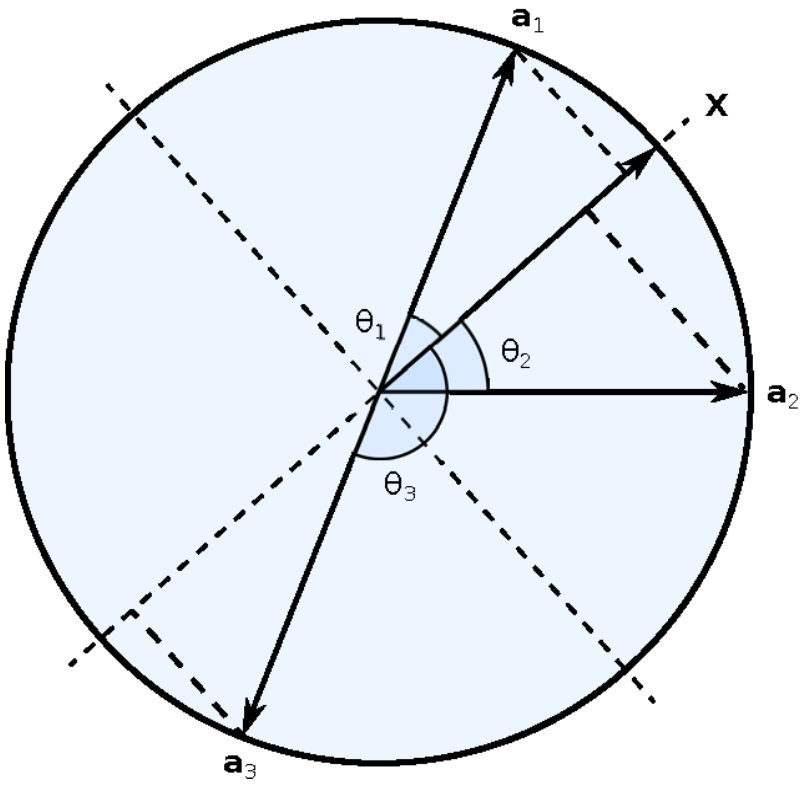


Convolution is “Vector Inner Product” or “Projection”

- 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?

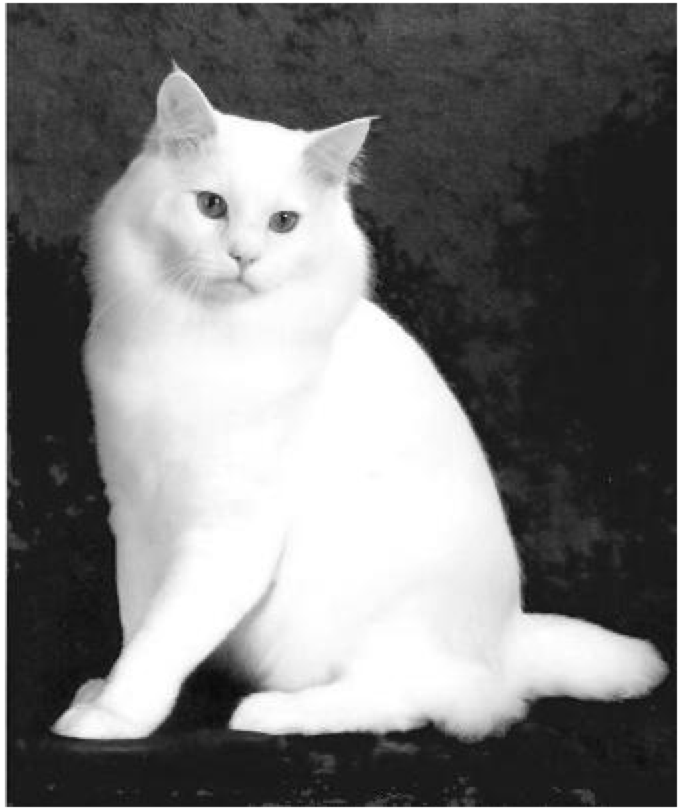
REctified COrrelation on a Sphere (RECOS) Model

- 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", arXiv:1609.04112 and to appear in the Journal of Visual Communication and Image Representation

Comparison of Positive & Negative Correlations



Confusion Caused by Negative Correlations

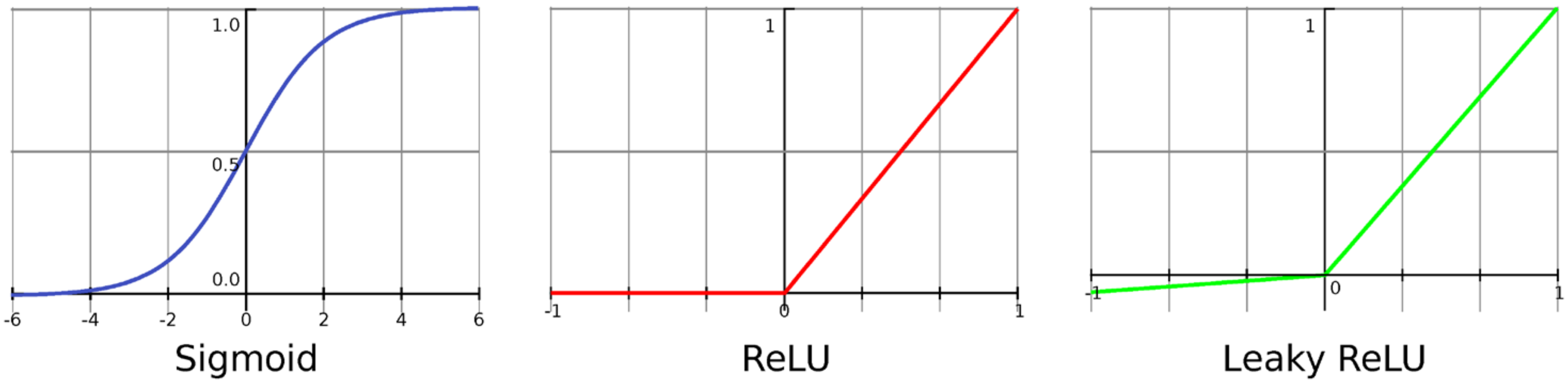
- When two convolutional filters are in cascade, the cascaded system cannot differentiate the following scenarios:

- 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

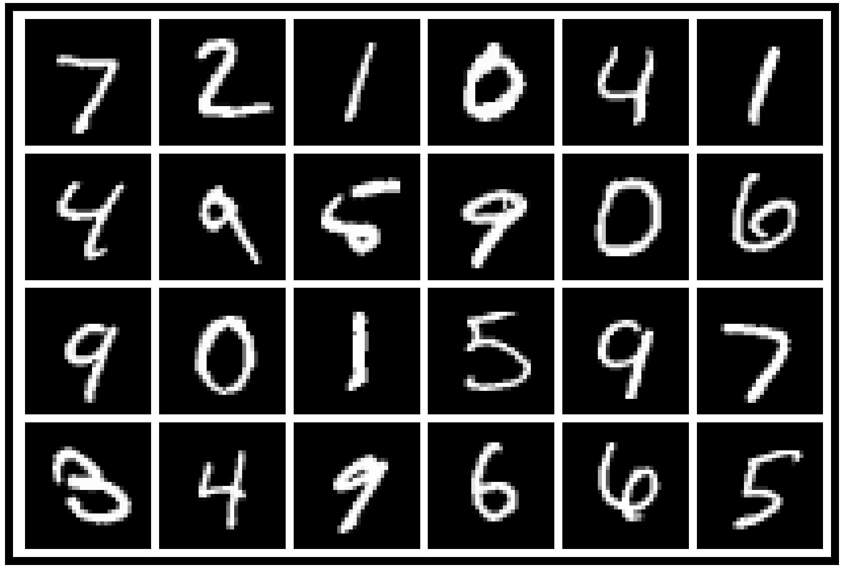
- Confusing Case #2
  - A positive correlation in stage 1 and a negative filter coefficient in stage 2
  - A negative correlation in stage 1 and a positive filter coefficient in stage 2

Nonlinear Activation Functions:

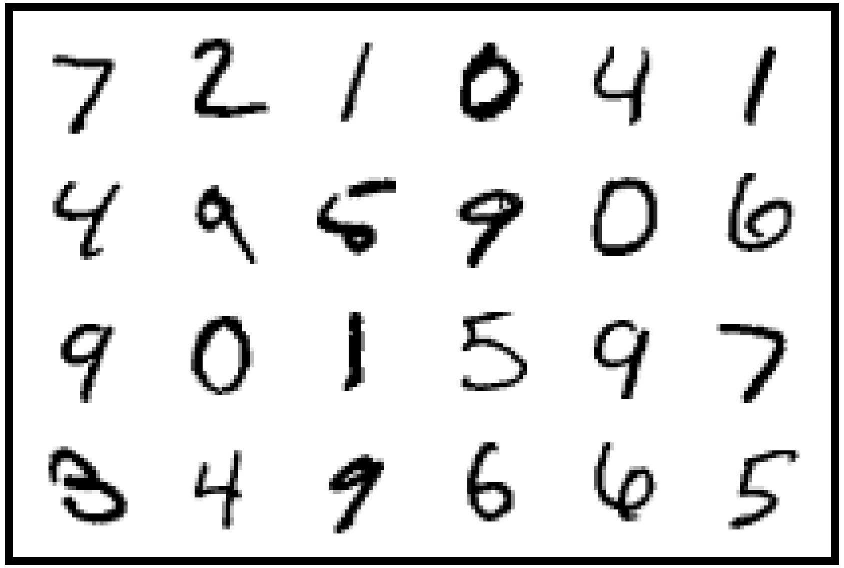
- When two convolutional filters are in cascade, nonlinear activation is used to clip negative correlations



Experiments on MNIST



Original



Negative

Test Performance of LeNet-5

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

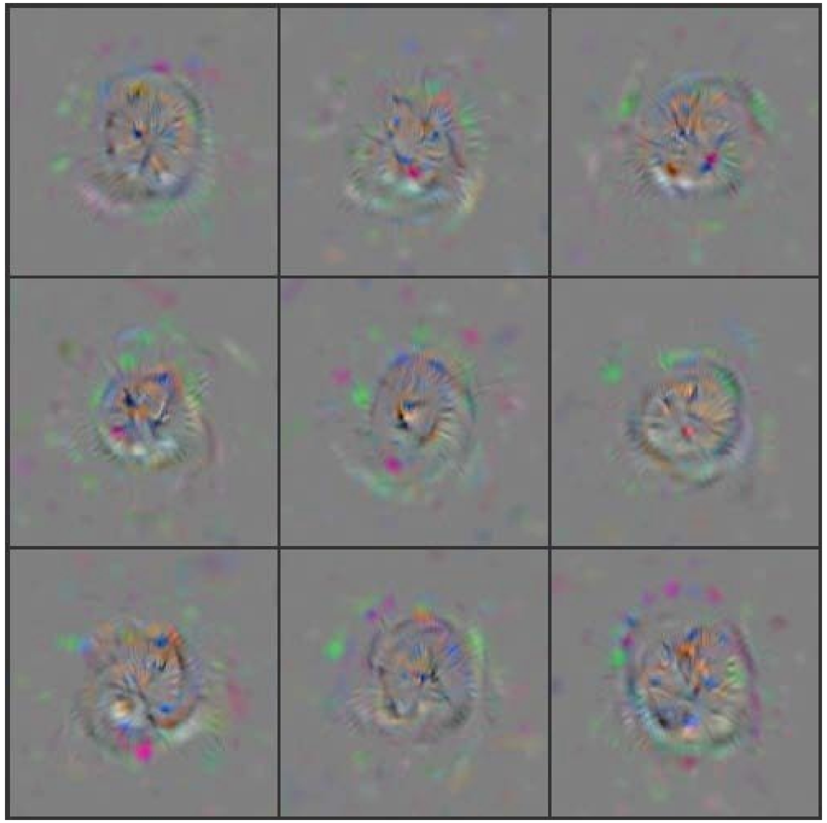
Test Performance of LeNet-5

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

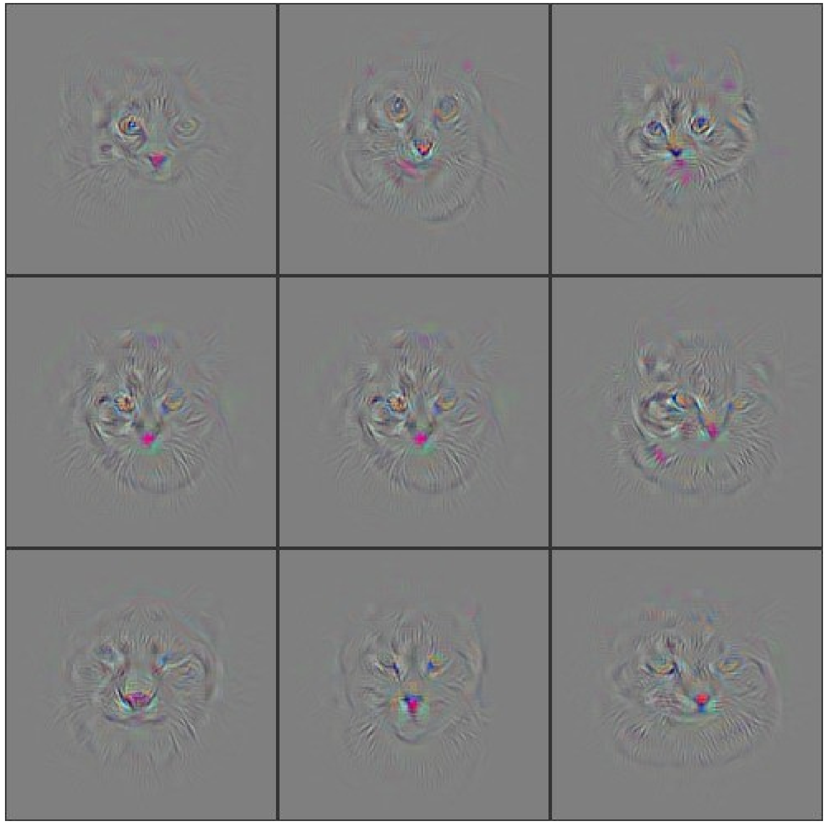
Benefit of Cascaded RECOS Model

Example 1: Cat Image

Top 9 Input Activation Images



Max Reconstructed Input Activation



Deconv Image

Example 2: Handwritten Digits

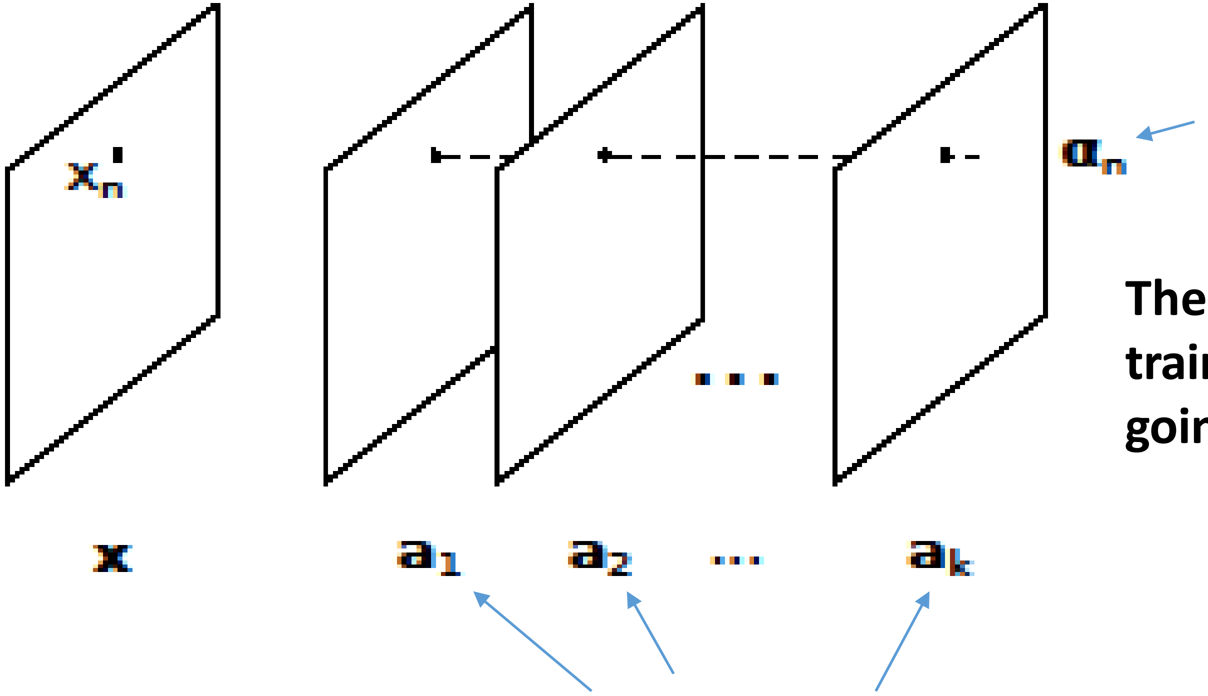
Can CNN recognize these digits with background?
If there is no correlation between the background and digits, it is feasible

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



Anchor Position Vectors

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

Anchor Vectors

What is Convergence Rate?

- An open question
  - How fast approximation errors go to zero as the no. of training samples becomes larger or the network architecture become more complex?
- Scenario #1
  - Given dataset
    - ImageNet
    - Places
  - Investigate the relationship between the numerical convergence rate and the network parameters (# of layers, # of filters per layer, filter size, etc.)
- Scenario #2
  - Given an application domain
    - How to find meaningful training data (data diversity)?
  - Consider the statistical behavior of samples

Conclusion

- The superior performance of CNNs is rooted in deep theoretical foundation
  - Approximation Theory
  - Optimization Theory
  - Signal Analysis Theory
- Today's research is too much focused on
  - Applications and performances for existing datasets
    - Being top in some datasets does not imply solving real problems
    - Difficult to have breakthrough if no new labeled datasets are built
  - Blind construction of datasets
    - We need to know what to build to improve training diversity
  - Heuristic engineering practices
    - Theoretical understanding is essential to the advancement of the field