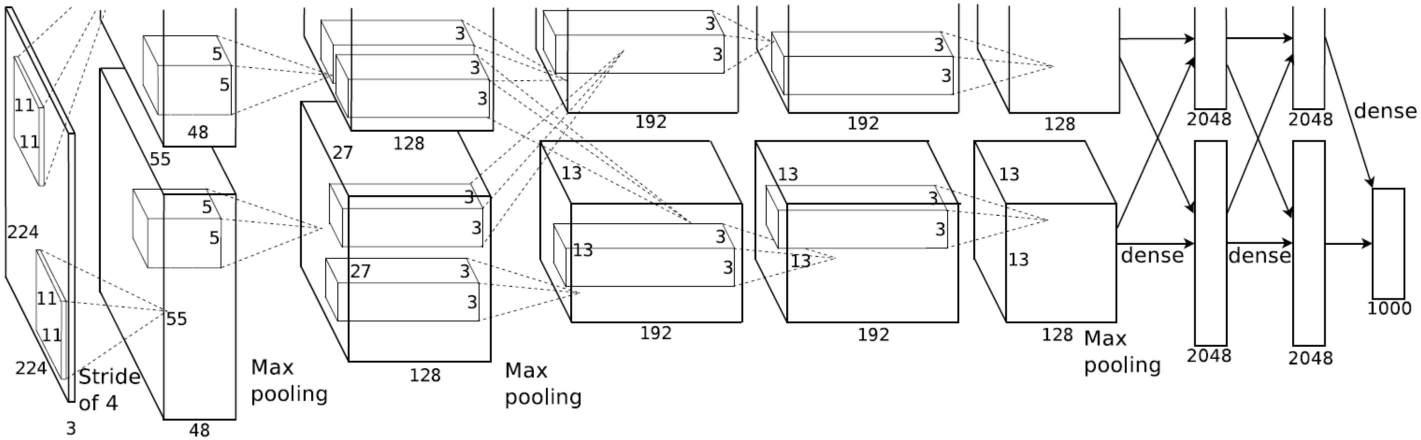
ECS289 VISUAL RECOGNITION

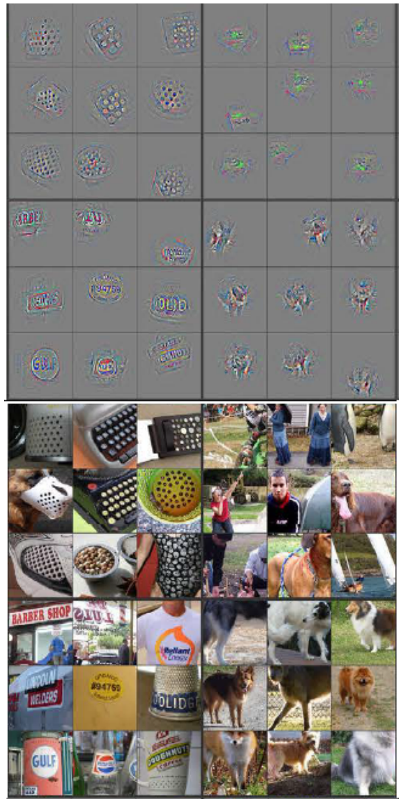
Intriguing properties of neural networks

Wei-Chih Chen(Michael)

2015/10/22

Introduction





- We know already:
  (1) Math: loss function, backpropagation
  (2) how to train: by mini batch, stochastic gradient descent
  (3) Implement details: Max pooling, Relu, dropout, architecture
  (4) Visualize features from each layers

UCDAVIS

Introduction

- We do NOT know:
  Relation between each layers.

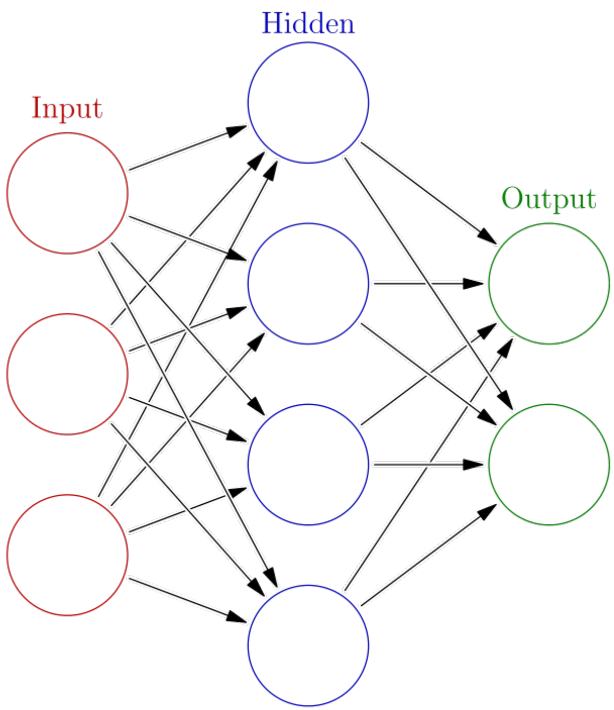
- This paper proposed:

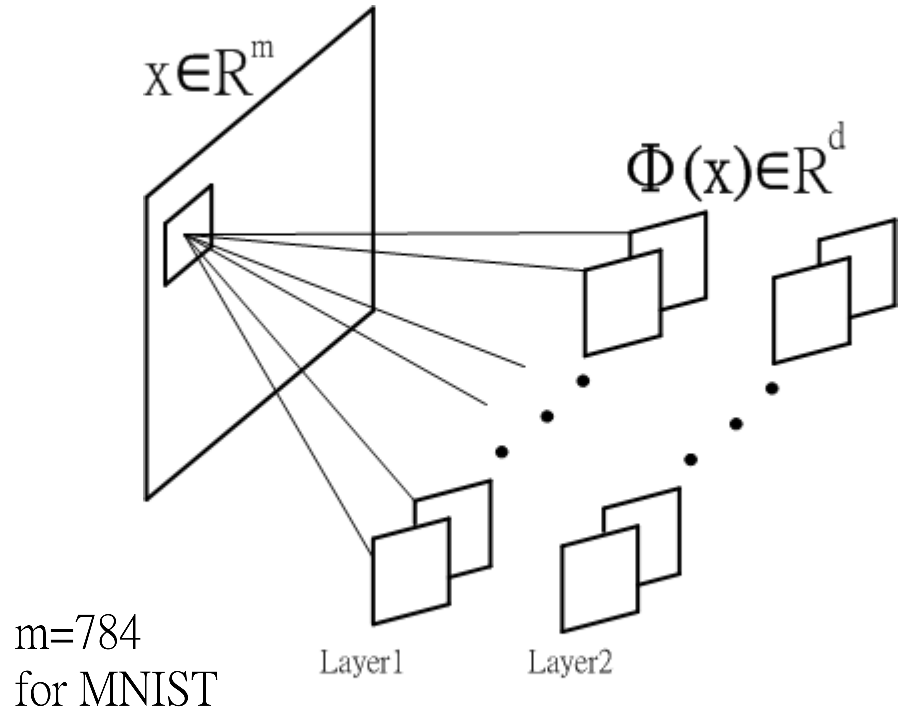
  -Space rather than individual units contain semantic info.

  -misclassify an image by applying imperceptible perturbation

Space?

- Space rather than individual units contain semantic info
- Representation Φ as an function mapping an image x to feature space.

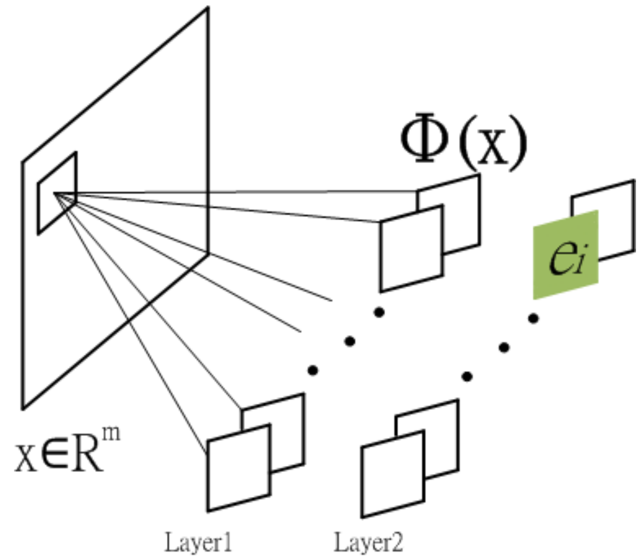




Space?

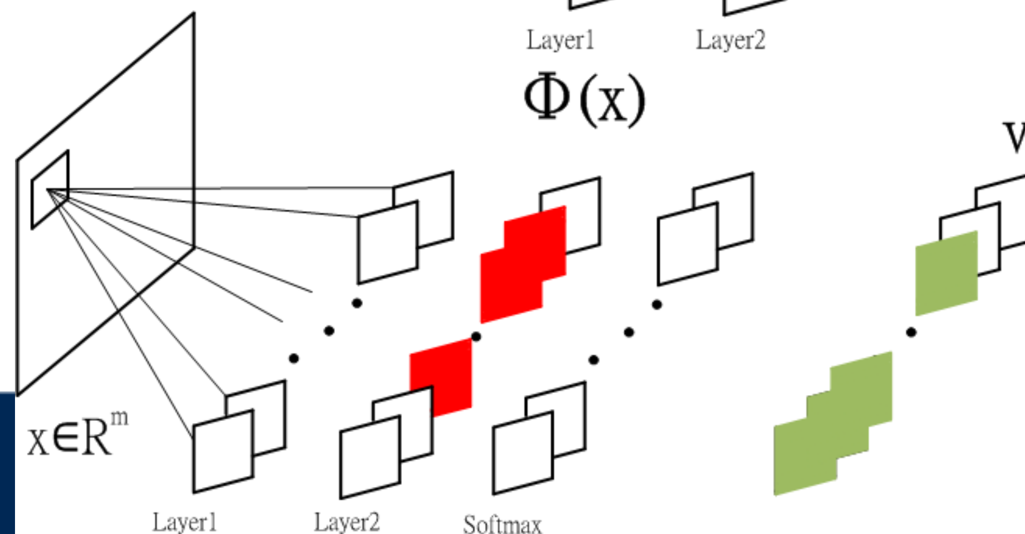
- Using the natural basis of the i-th hidden unit:

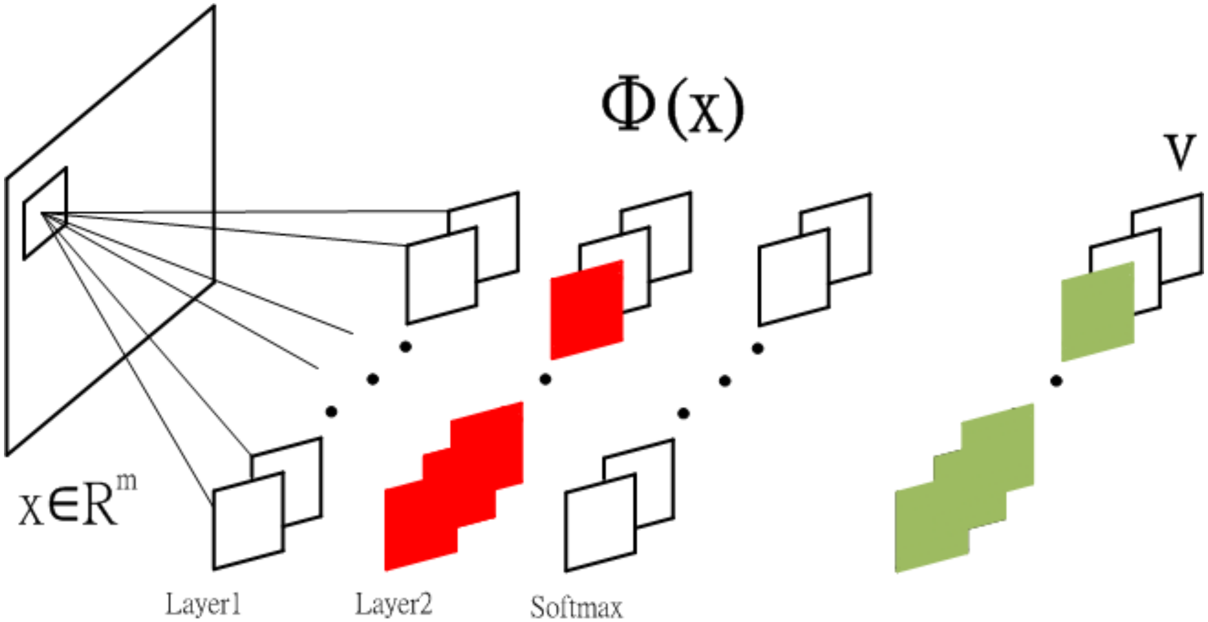
$$x' = \arg\max_{x \in \mathcal{I}} \langle \phi(x), e_i \rangle$$



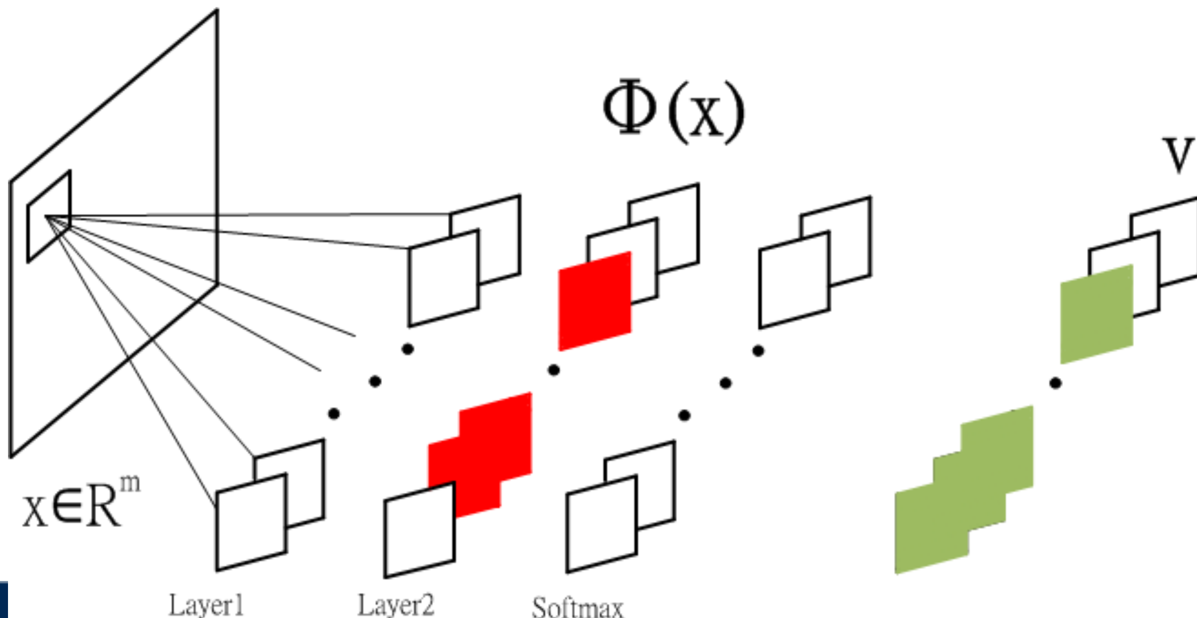
- Feature vector direction v∈R^n:

$$x' = \arg\max_{x \in \mathcal{I}} \langle \phi(x), v \rangle$$





- Natural basis direction



- Random direction

UCDAVIS

Result on MNIST

- Natural basis direction

(b) Unit sensitive to upper round stroke, or lower straight stroke.

(d) Unit senstive to diagonal straight stroke.

- Random basis

(b) Direction sensitive to lower left loop.

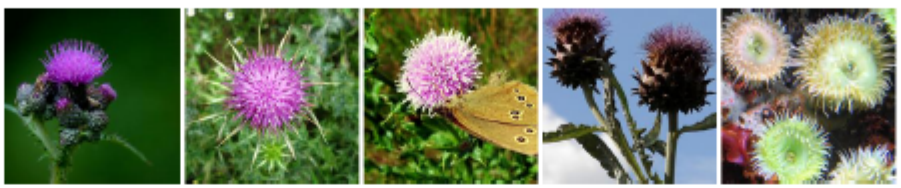
(d) Direction sensitive to right, upper round stroke.

UCDAVIS

Result on ImageNet

- Natural basis direction

(a) Unit sensitive to white flowers.



(c) Unit senstive to round, spiky flowers.

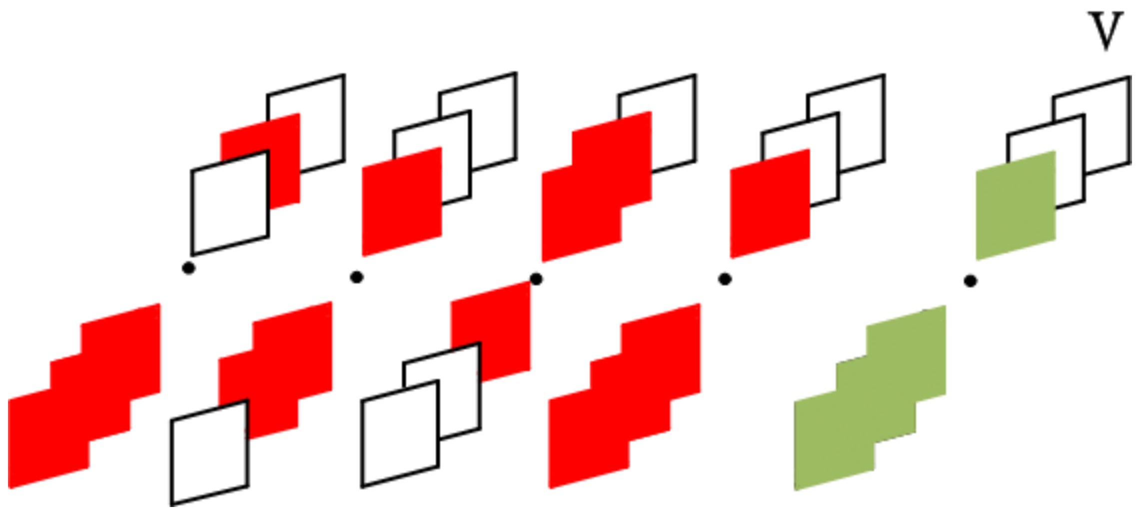
- Random basis

(a) Direction sensitive to white, spread flowers.

(c) Direction sensitive to spread shapes.

Conclude for first question

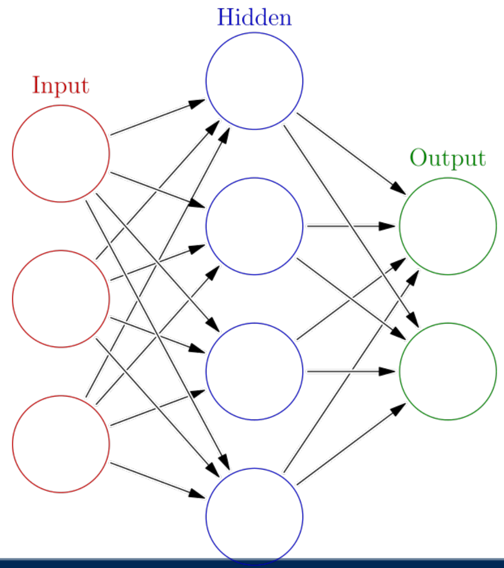
- The vector representations are well-defined up to rotation of the space, so the **individual units are unlikely to contain semantic information.**

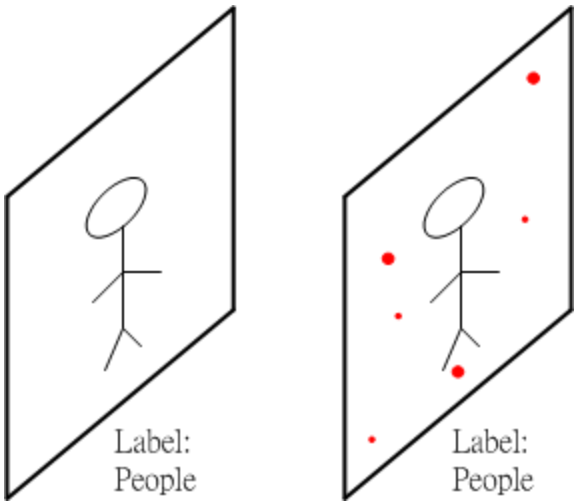


UCDAVIS

Blind Spots

- Networks level contains semantic info.
- Output is highly nonlinear function of its input.
- Encoded a non-local generalization prior over input to put non-significant probability to some region without misclassified.

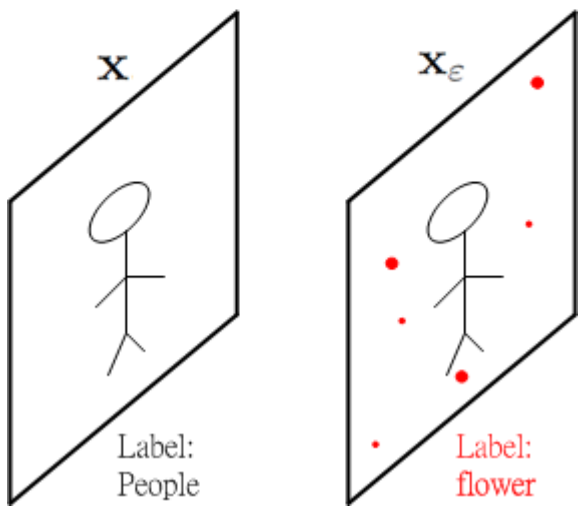




UCDAVIS

Local generalization

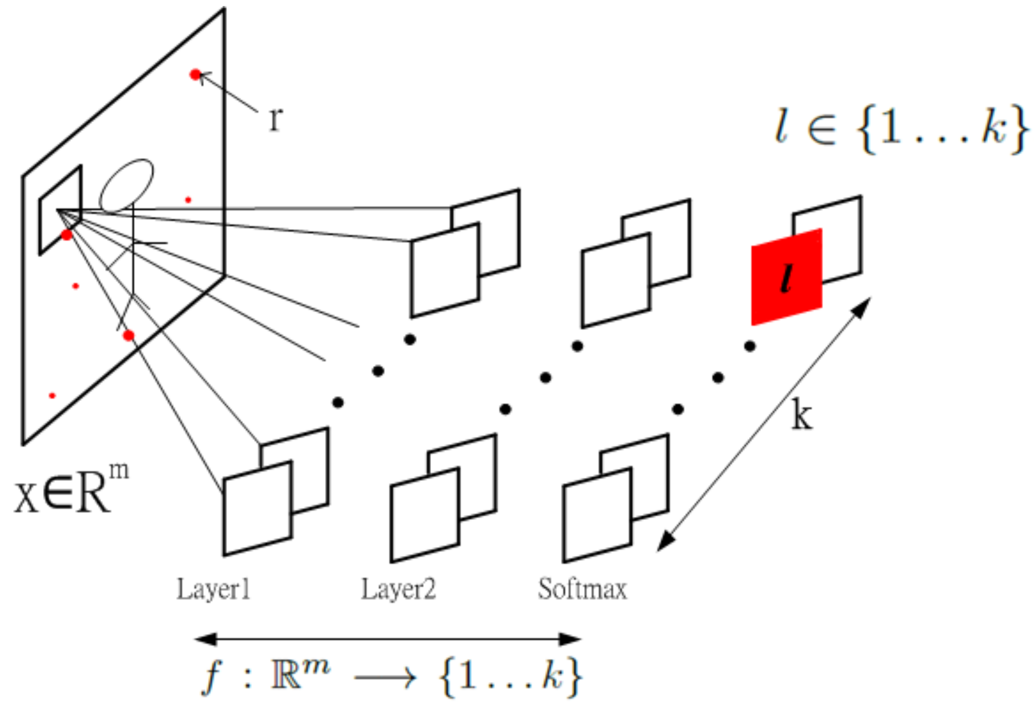
- Input $\mathbf{x}$, and generate $\mathbf{x}_\varepsilon$ which satisfies $||\mathbf{x} - \mathbf{x}_\varepsilon|| < \varepsilon$, ε is a small enough radius.
- $\mathbf{x}_\varepsilon$ should NOT change underlying class.
- This paper want find ***adversarial examples*** that cause **MISLABEL**.



(a) All misLabeled to ostrich (b)

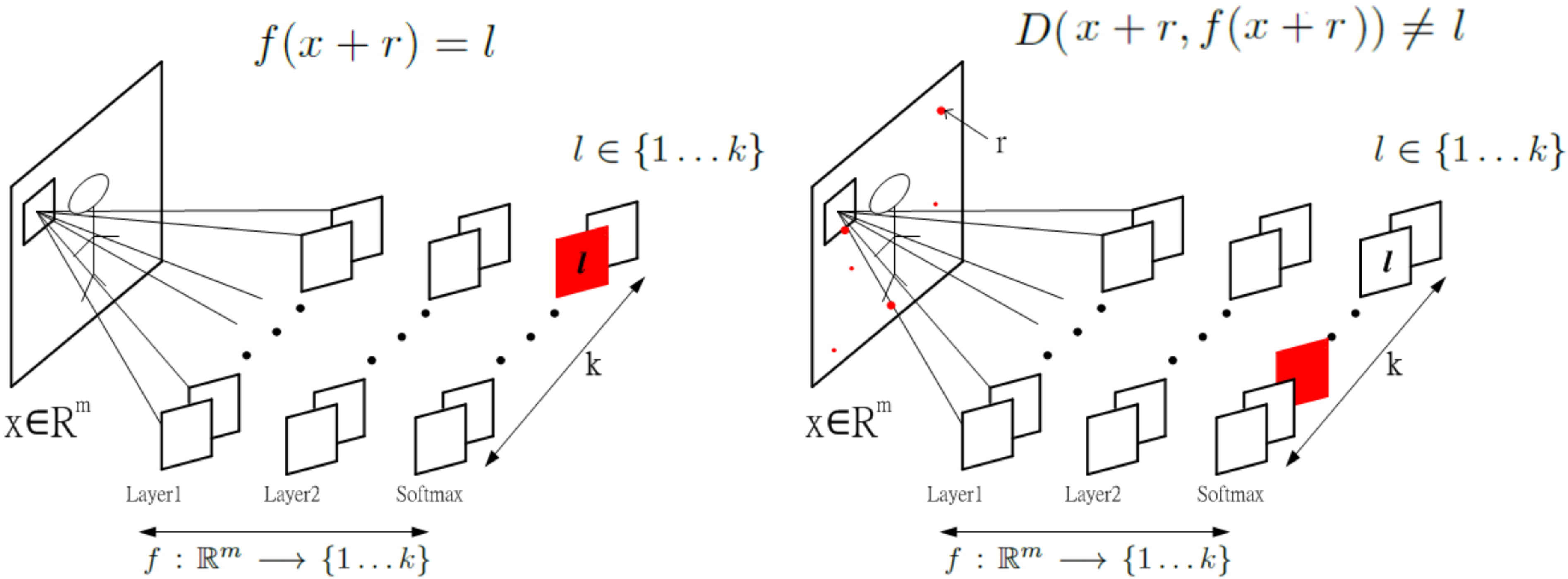
Function define

- Minimize $\|r\|_2$ subject to:
  1. $f(x+r) = l$
  2. $x + r \in [0,1]^m$



- A minimizer by D(x,l), if x+r is close to x, then x is classified as l by f. Like, D(x,f(x))=f(x)
- Using a box-constrained L-BFGS for optimization
- Our Goal: minimized r to get $D(x+r, f(x+r)) \neq l$

Goal

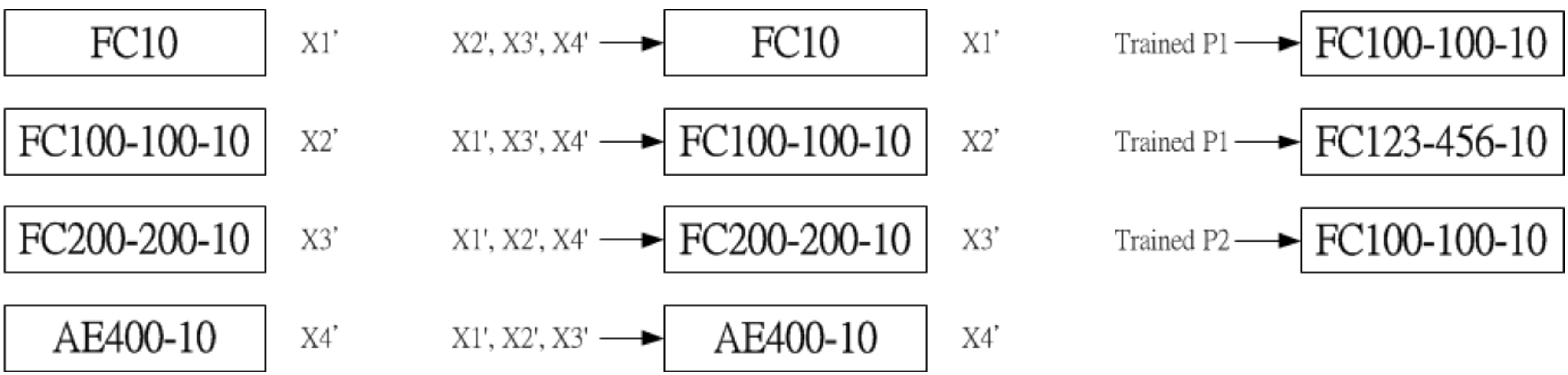


- Minimize $c|r| + \text{loss}_f(x+r, l)$ subject to $x + r \in [0, 1]^m$

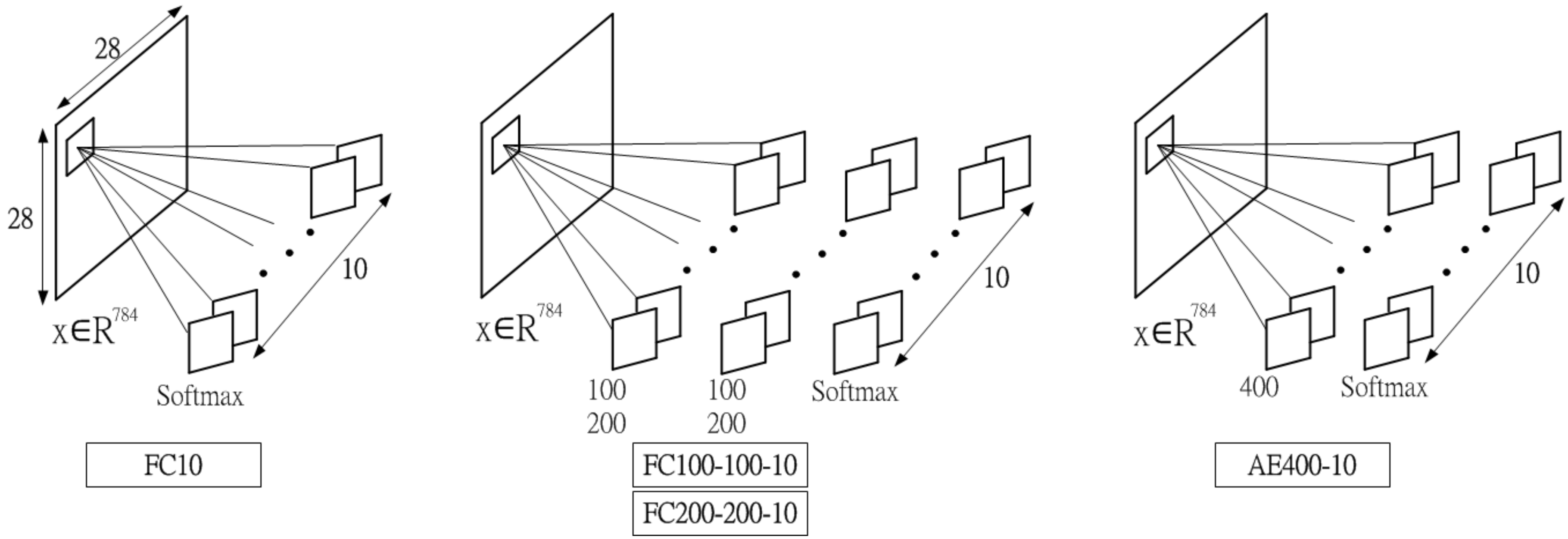
UCDAVIS

Properties for distortion function D

- All networks(MNIST, AlexNet) can generate visually indistinguishable adversarial example.
- Cross model generalization: misclassified by other networks.
- Cross training-set generalization: misclassified by other disjoint training set.



Tests to generate adversarial instance on MNIST(1/2)



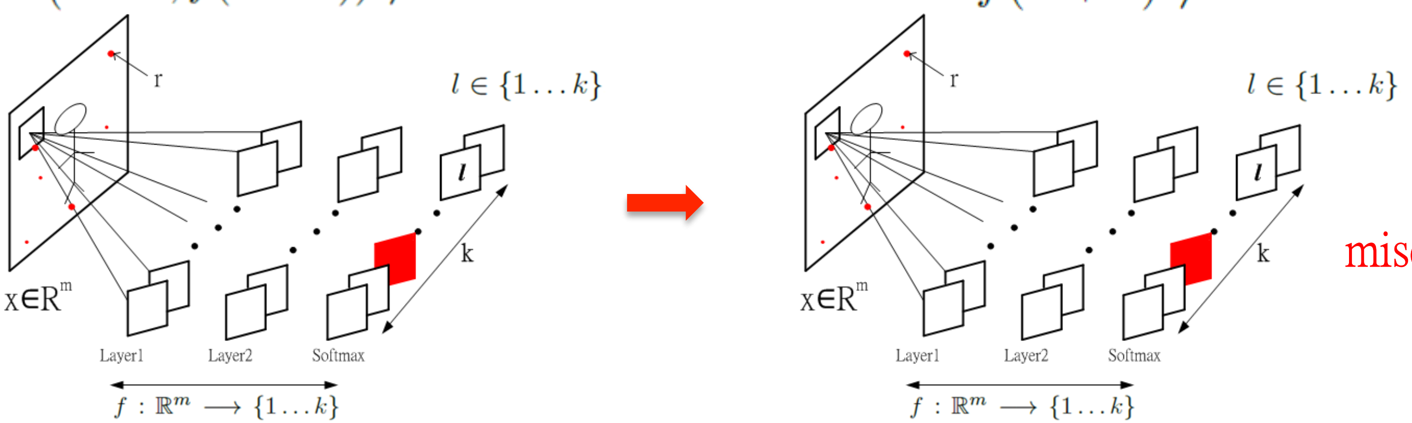
UCDAVIS

Tests to generate adversarial instance on MNIST(2/2)

Model Name	Description	Training error	Test error	Av. min. distortion
FC10(10^{-4})	Softmax with $\lambda = 10^{-4}$	6.7%	7.4%	0.062
FC10(10^{-2})	Softmax with $\lambda = 10^{-2}$	10%	9.4%	0.1
FC10(1)	Softmax with $\lambda = 1$	21.2%	20%	0.14
FC100-100-10	Sigmoid network $\lambda = 10^{-5}, 10^{-5}, 10^{-6}$	0%	1.64%	0.058
FC200-200-10	Sigmoid network $\lambda = 10^{-5}, 10^{-5}, 10^{-6}$	0%	1.54%	0.065
AE400-10	Autoencoder with Softmax $\lambda = 10^{-6}$	0.57%	1.9%	0.086

$D(x+r, f(x+r)) \neq l$ $\qquad\qquad$ $f(x+r) \neq l$

Train 60K images
Test 10K images

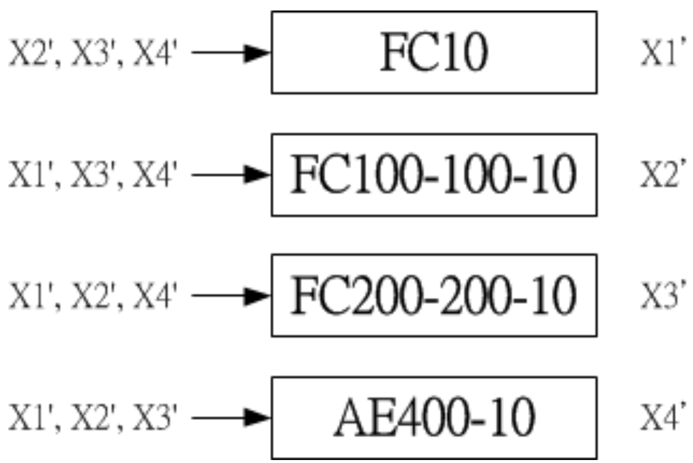


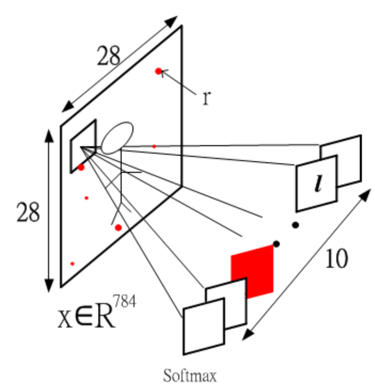
100%
misclassified

Produce r cause image 100% misclassified

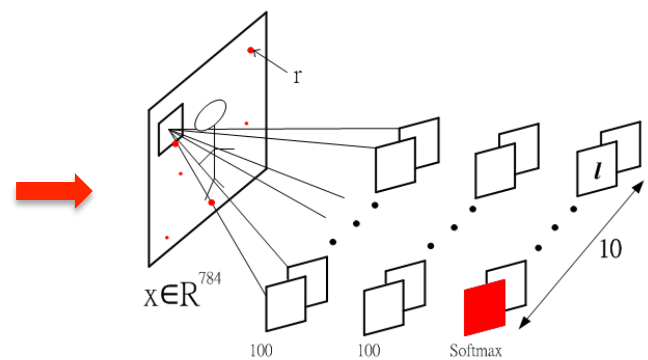
Cross model generalization on MNIST(1/2)

	FC10(10^{-4})	FC10(10^{-2})	FC10(1)	FC100-100-10	FC200-200-10	AE400-10	Av. distortion
FC10(10^{-4})	100%	11.7%	22.7%	2%	3.9%	2.7%	0.062
FC10(10^{-2})	87.1%	100%	35.2%	35.9%	27.3%	9.8%	0.1
FC10(1)	71.9%	76.2%	100%	48.1%	47%	34.4%	0.14
FC100-100-10	28.9%	13.7%	21.1%	100%	6.6%	2%	0.058
FC200-200-10	38.2%	14%	23.8%	20.3%	100%	2.7%	0.065
AE400-10	23.4%	16%	24.8%	9.4%	6.6%	100%	0.086
Gaussian noise, stddev=0.1	5.0%	10.1%	18.3%	0%	0%	0.8%	0.1
Gaussian noise, stddev=0.3	15.6%	11.3%	22.7%	5%	4.3%	3.1%	0.3





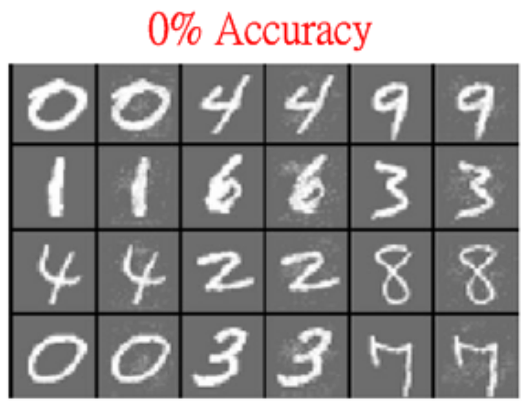
FC10



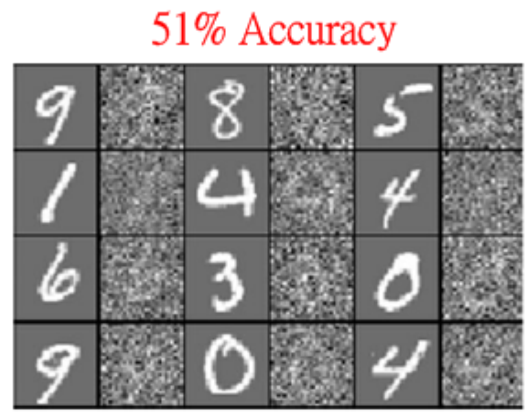
FC100-100-10

Cross model generalization on MNIST(2/2)

(a) Even columns: adversarial examples for a linear (FC) classifier (stddev=0.06)



(b) Even columns: adversarial examples for a 200-200-10 sigmoid network (stddev=0.063)



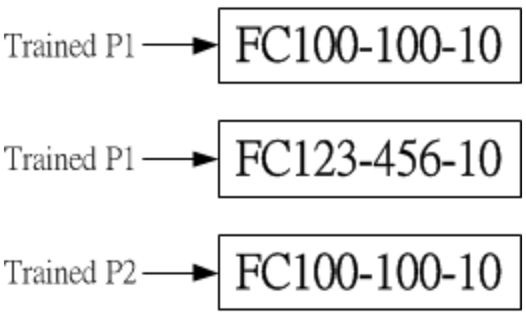
(c) Randomly distorted samples by Gaussian noise with stddev=1. Accuracy: 51%.

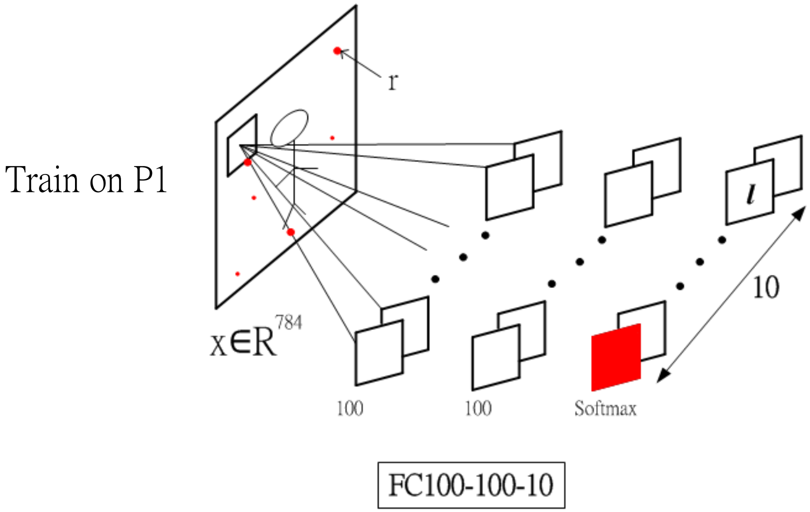
Cross training-set generalization on MNIST

- Cross training-set generalization – baseline

MNIST 60K images
Half P1 and half P2

Model	Error on P_1	Error on P_2	Error on Test	Min Av. Distortion
FC100-100-10: 100-100-10 trained on P_1	0%	2.4%	2%	0.062
FC123-456-10: 123-456-10 trained on P_1	0%	2.5%	2.1%	0.059
FC100-100-10' trained on P_2	2.3%	0%	2.1%	0.058



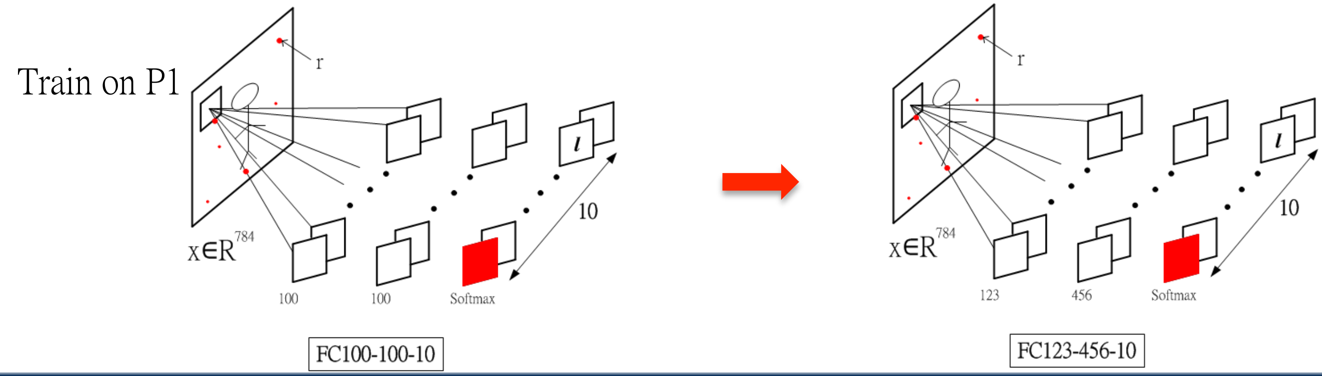


Cross training-set generalization on MNIST

- Cross training-set generalization error rate(magnify distortion)

	FC100-100-10	FC123-456-10	FC100-100-10'
Distorted for FC100-100-10 (av. stddev=0.062)	100%	26.2%	5.9%
Distorted for FC123-456-10 (av. stddev=0.059)	6.25%	100%	5.1%
Distorted for FC100-100-10' (av. stddev=0.058)	8.2%	8.2%	100%
Gaussian noise with stddev=0.06	2.2%	2.6%	2.4%
Distorted for FC100-100-10 amplified to stddev=0.1	100%	98%	43%
Distorted for FC123-456-10 amplified to stddev=0.1	96%	100%	22%
Distorted for FC100-100-10' amplified to stddev=0.1	27%	50%	100%
Gaussian noise with stddev=0.1	2.6%	2.8%	2.7%

$$x + 0.1\frac{x'-x}{\|x'-x\|_2}$$



UCDAVIS

Conclusion

- Space rather than the individual units contain of the semantic information.
- Adversarial examples(imperceptible perturbation) misclassify by the network no matter type of network, cross model and cross training-set