Progress #3

김나현

COURSERA_machine_learning

the progress of the lecture : ~ Week 5

the progress of the assignment : ~ Week 4

Last week : the assignments weren't released,
so I only took lectures.

This week : I did the assignment that I couldn't do last week,
so I could find out exactly what I didn't know.

Reading Paper

Explaining and Harnessing Adversarial Examples

At ICLR 2015

Reading Paper – 1. abstract / Introduction

Early attempts : focused on nonlinearity and overfitting

This paper : focused on linear nature of ML models

Create an adversarial attack called FGSM

leveriging linear nature.

Reading Paper – 2. related work

L-BFGS(Limited-memory BFGS)

The same adversarial example is often misclassified by a variety of classifiers with different architectures.

3. The linear explanation OF A.E.

$$\boldsymbol{w}^\top \tilde{\boldsymbol{x}} = \boldsymbol{w}^\top \boldsymbol{x} + \boldsymbol{w}^\top \boldsymbol{\eta}.$$

The adversarial perturbation causes the activation to grow

By $\boldsymbol{w}^\top \boldsymbol{\eta}.$

4. Linear perturbation of non-linear models

neural networks are too linear to resist linear adversarial perturbation

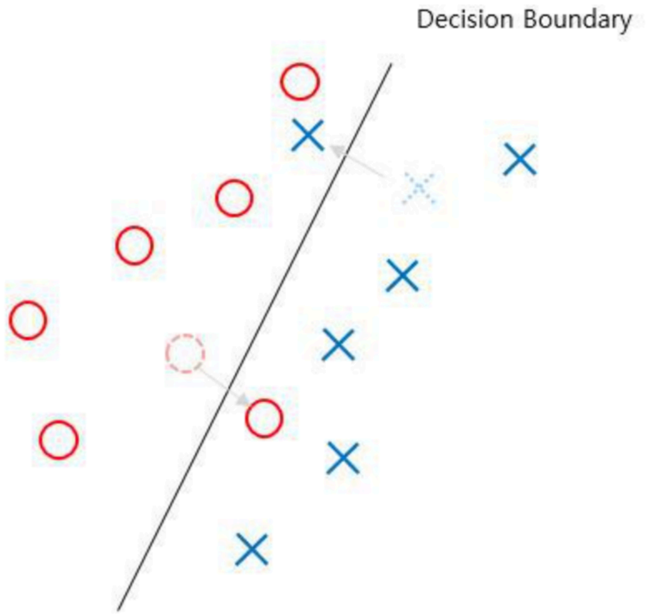
LSTMs, ReLUs, maxout networks

-> intentionally designed to behave in very linear ways

$$\boldsymbol{\eta} = \epsilon \operatorname{sign}\left(\nabla_{\boldsymbol{x}} J(\boldsymbol{\theta}, \boldsymbol{x}, y)\right).$$

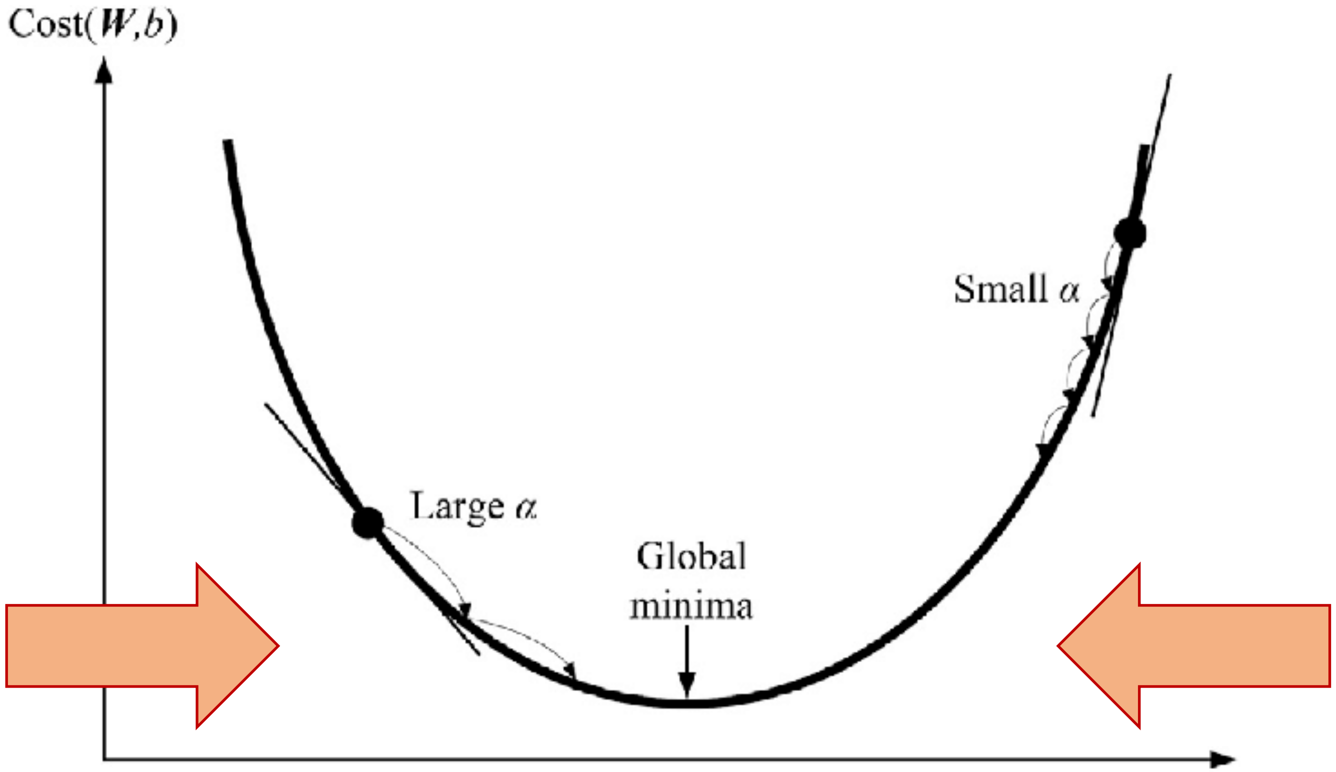
“fast gradient sign method”

of generating adversarial examples



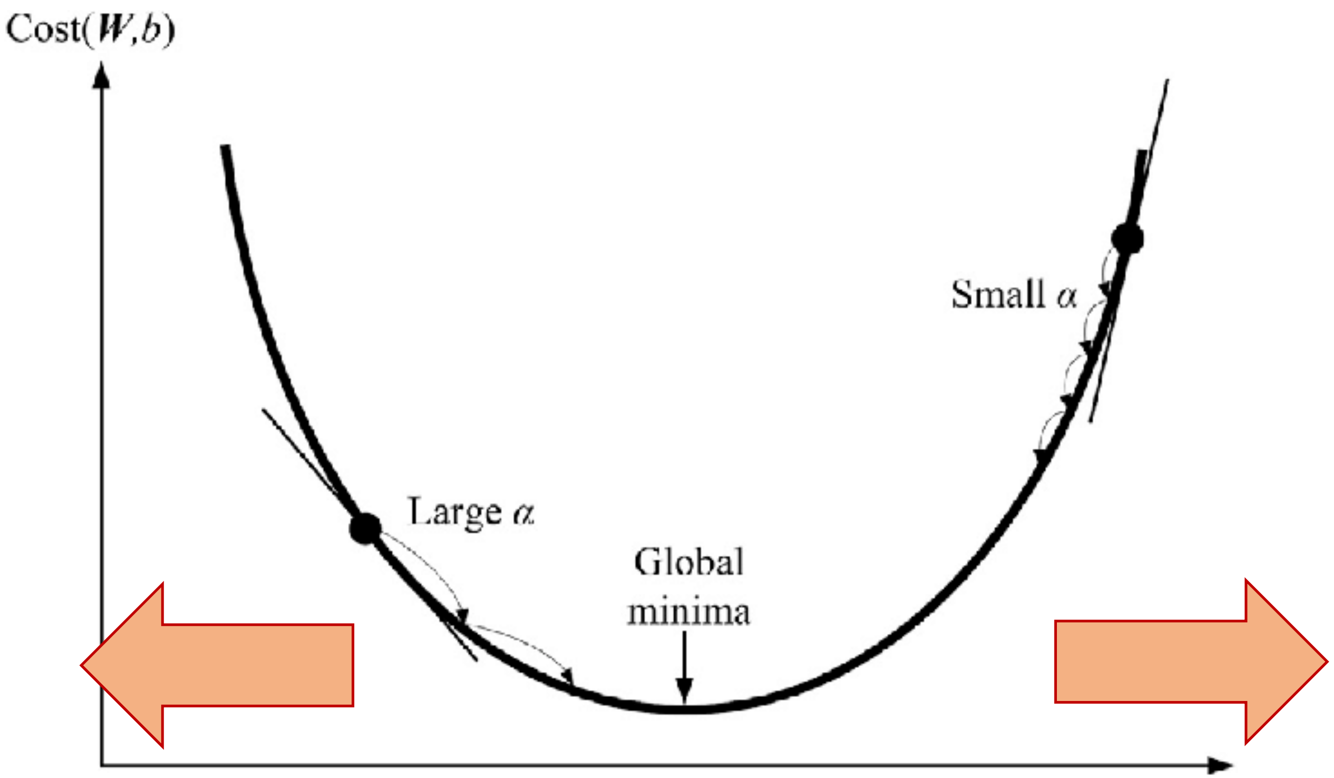
4. Linear perturbation of non-linear models

+ rotating x by a small angle in the direction of the gradient reliably produces adversarial examples.



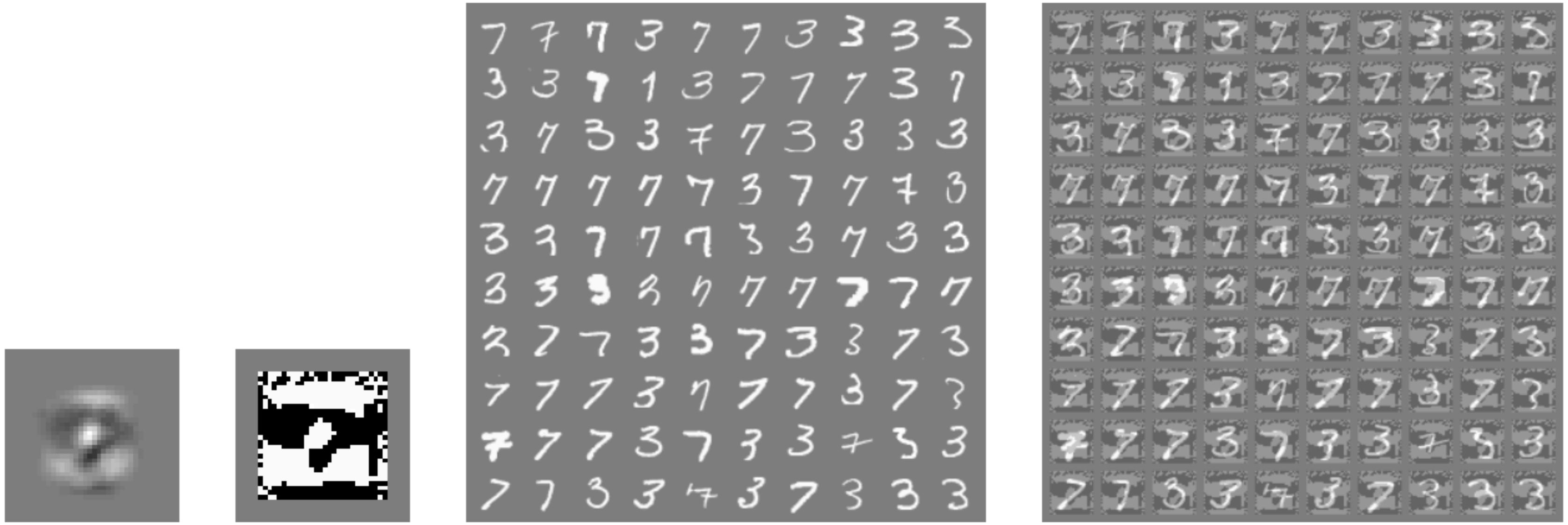
4. Linear perturbation of non-linear models

+ rotating x by a small angle in the direction of the gradient reliably produces adversarial examples.



5. adversarial traning of linear models VS weight decay

somewhat similar



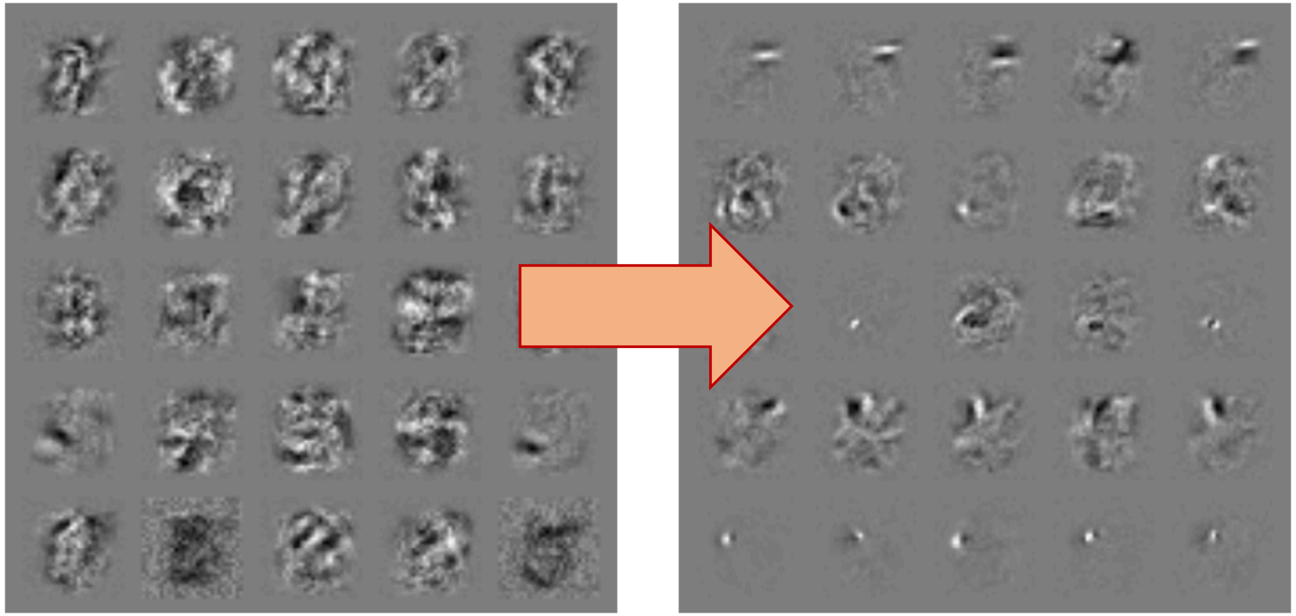
(a) (b) (c) (d)

The fast gradient sign method applied to logistic regression

Weight decay overestimates the damage achievable with perturbation even more in the case of a deep network with multiple hidden units

6. Adversarial traning of deep network

without adversarial training, this same kind of model had an error rate of 89.4% on adversarial examples based on the fast gradient sign method. With adversarial training, the error rate fell to 17.9%.



7 DIFFERENT KINDS OF MODEL CAPACITY

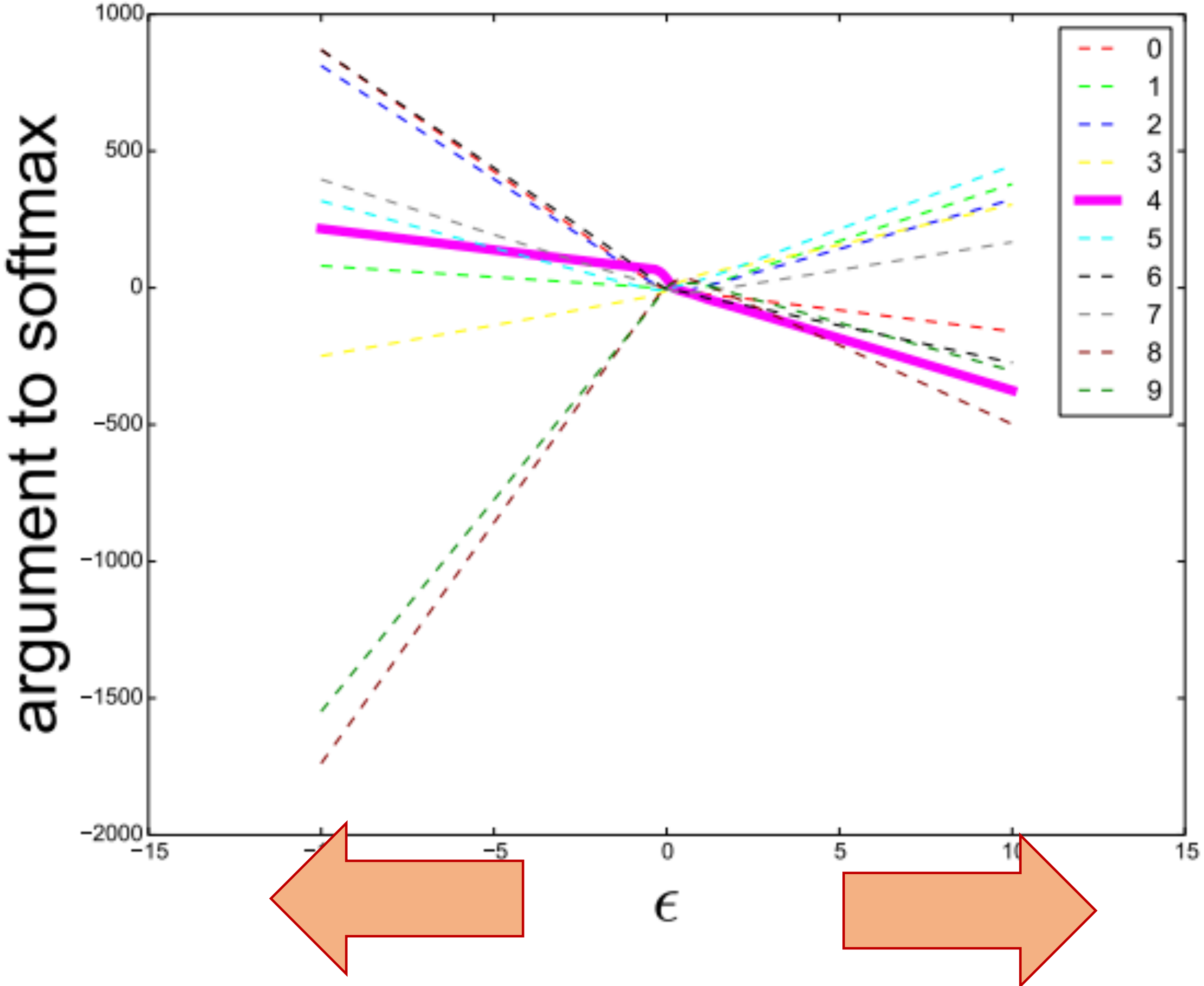
RBF networks are naturally immune to adversarial examples Explanations based on extreme non-linearity

8 WHY DO ADVERSARIAL EXAMPLES GENERALIZE

adversarial examples occur in contiguous regions, not in fine pockets.

the unnormalized log probabilities for each class are conspicuously piecewise linear with ϵ

and the wrong classifications are stable across a wide region of ϵ values.



8 WHY DO ADVERSARIAL EXAMPLES GENERALIZE?

neural networks trained with current methodologies
all resemble the linear classifier learned on the same training set

-> learn approximately the same classification weights

The stability of the underlying classification weights in turn results in the stability of adversarial examples.

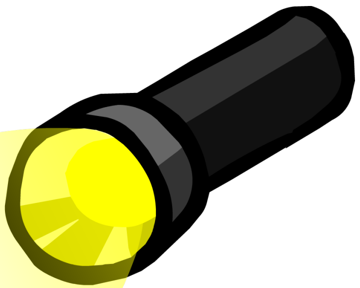
Comparing two papers

Explaining and Harnessing Adversarial Examples

VS

Adversarial Examples Are Not Bugs,

They Are Features



Different perspectives