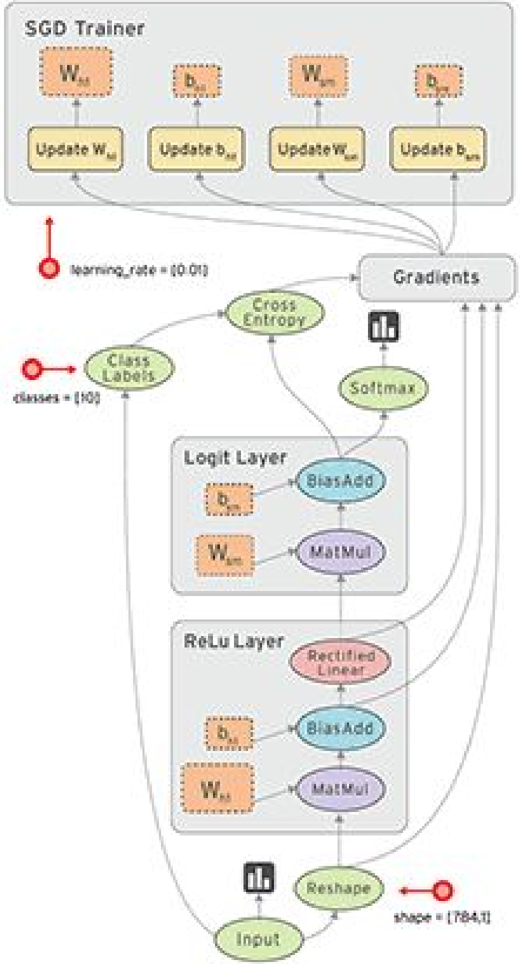
TensorFlow Workshop on Adversarial Learning

Illia Polosukhin, @ilblackdragon

Goal

- Intuition how to use various tool to make models more resistant to various problems.
- Learn how to use basics of TF.Learn
- Give some real life examples

TensorFlow Basics

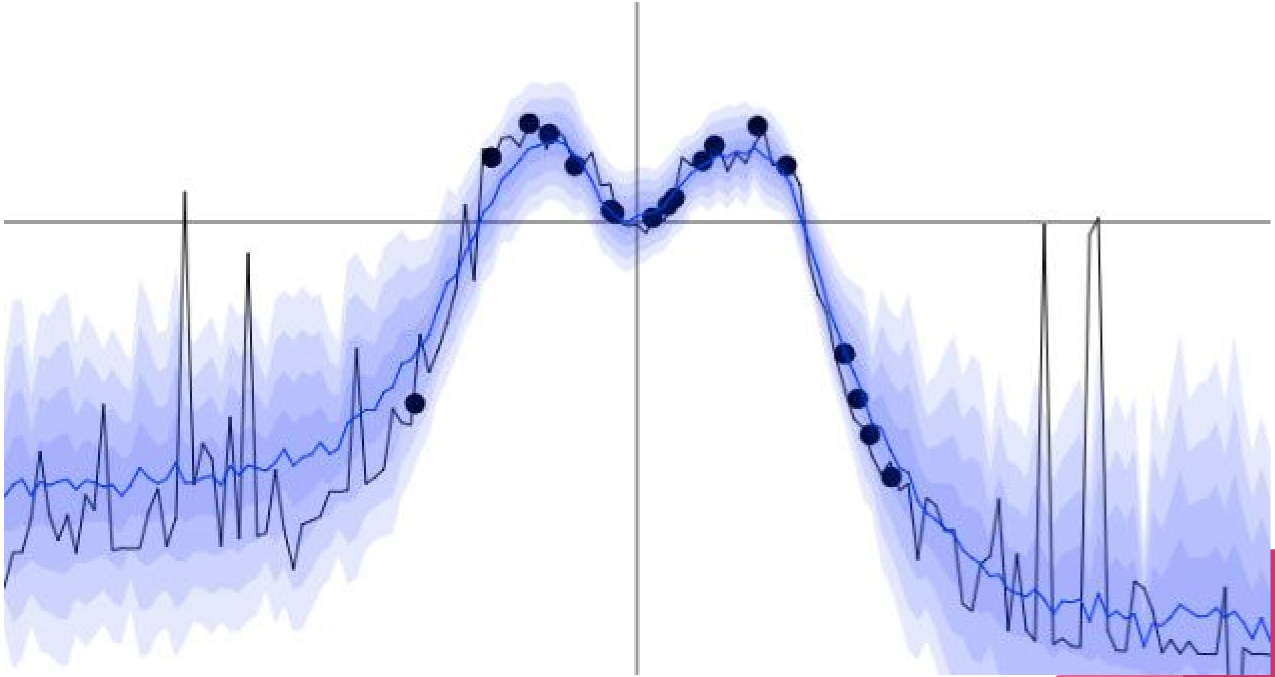

SGD Trainer
W_hl
b_hl
W_sm
b_sm
Update W_hl
Update b_hl
Update W_sm
Update b_sm
learning_rate = [0.01]
Gradients
Cross Entropy
Class Labels
classes = [10]
Softmax
Logit Layer
BiasAdd
b_sm
W_sm
MatMul
ReLu Layer
Rectified Linear
BiasAdd
b_hl
W_hl
MatMul
Reshape
shape = [784,1]
Input

Follow the examples

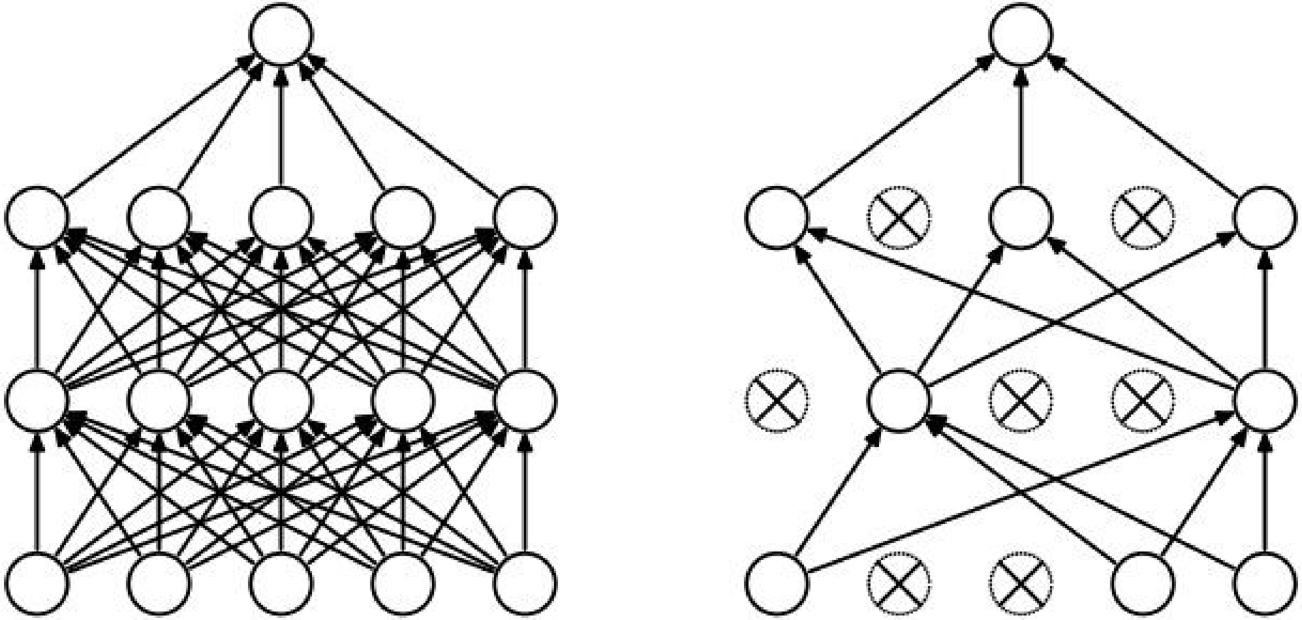
https://github.com/ilblackdragon/adversarial_workshop

Uncertainty

Uncertainty



Dropout

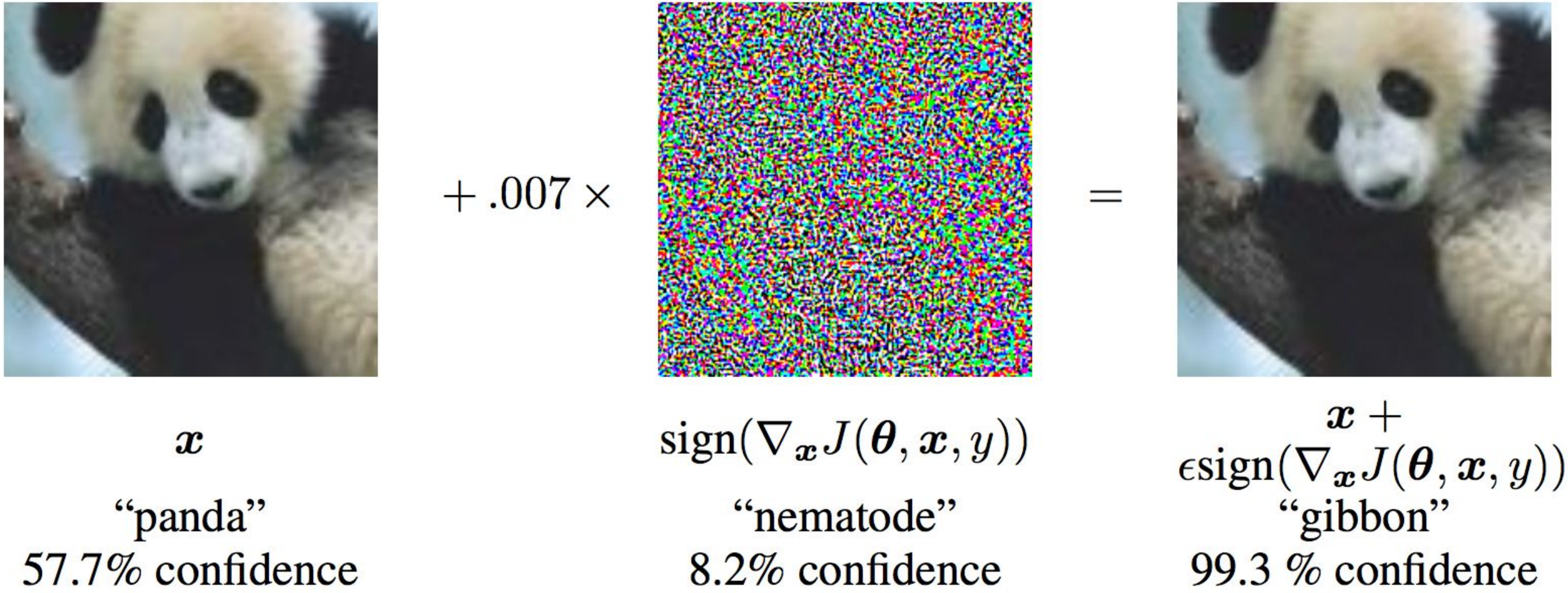


(a) Standard Neural Net

(b) After applying dropout.

Adversarial Examples

Adversarial Examples


$+ .007 \times$
$=$
$\boldsymbol{x}$
“panda”
57.7% confidence
$\text{sign}(\nabla_{\boldsymbol{x}} J(\boldsymbol{\theta}, \boldsymbol{x}, y))$
“nematode”
8.2% confidence
$\boldsymbol{x} +$
$\epsilon\text{sign}(\nabla_{\boldsymbol{x}} J(\boldsymbol{\theta}, \boldsymbol{x}, y))$
“gibbon”
99.3 % confidence

Adversarial Examples for Text

Nearest neighbours for ‘**good**’

Baseline	**Random**	**Adversarial**
great	great	decent
decent	decent	great
bad	excellent	nice
Good	Good	entertaining
fine	bad	interesting

Adversarial Examples for Text

Nearest neighbours for ‘**bad**’

Baseline	**Random**	**Adversarial**
terrible	terrible	terrible
awful	awful	awful
horrible	horrible	horrible
good	good	poor
Bad	poor	BAD

Defensive modelling

Blacklist & Whitelist

Whitelist:

- Pro: Provides an easy way to control what user will see
- Con: Doesn’t allow to easily explore

Blacklist:

- Pro: Fast way to turn of terrible things
- Con: Always retroactive and not generalizable

Uncertainty

- Use models that have uncertainty (like dropout-based)

- Estimate on a test set dependency between uncertainty and accuracy.

Triggering model

- Collect examples of things you want your model to trigger and not to trigger (e.g. whitelist and blacklist)
- Train a simple model that decides if your main model should be executed.

Questions?

Illia Polosukhin
@ilblackdragon

http://github.com/tensorflow/tensorflow

http://tensorflow.org

http://medium.com/@ilblackdragon

https://github.com/ilblackdragon/adversarial_workshop