

TensorFlow Workshop on Adversarial Learning

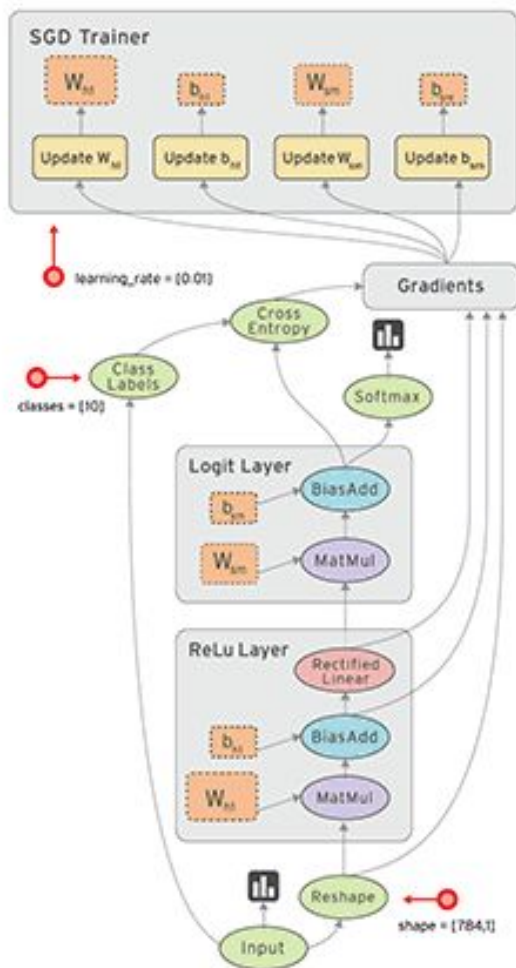
Illia Polosukhin, @ilblackdragon

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



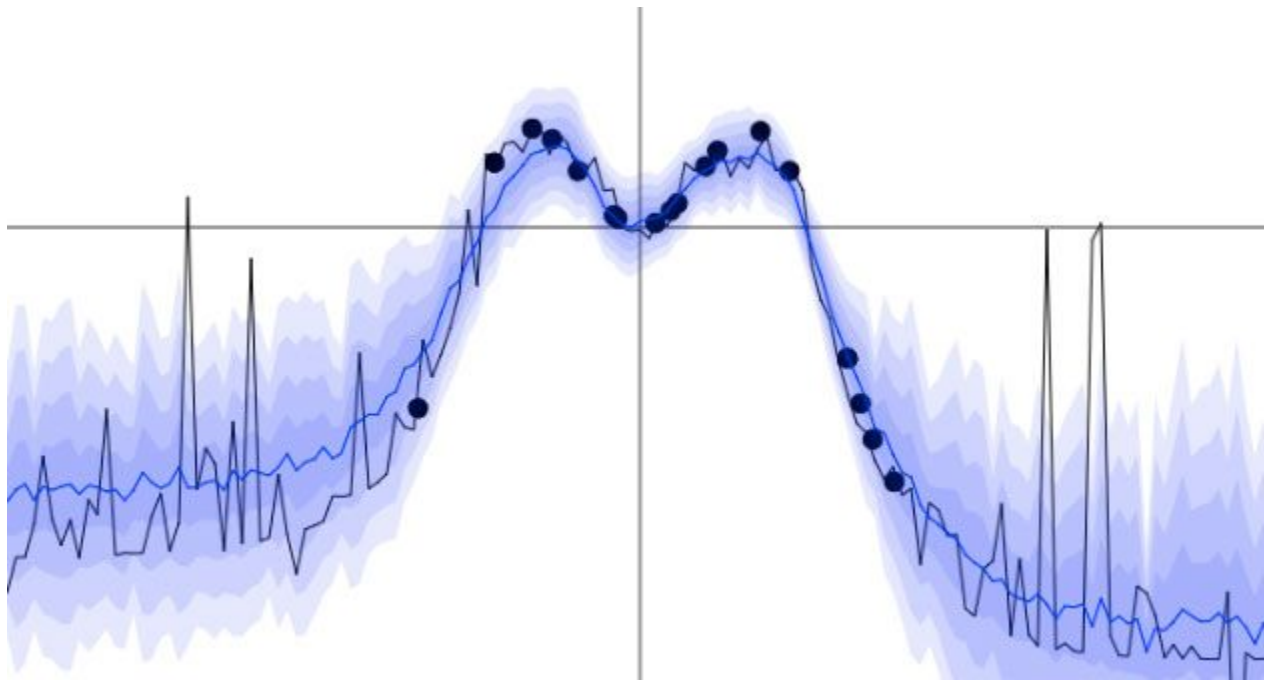
Follow the
examples

[https://github.com/il
blackdragon/adversa
rial_workshop](https://github.com/ilblackdragon/adversarial_workshop)

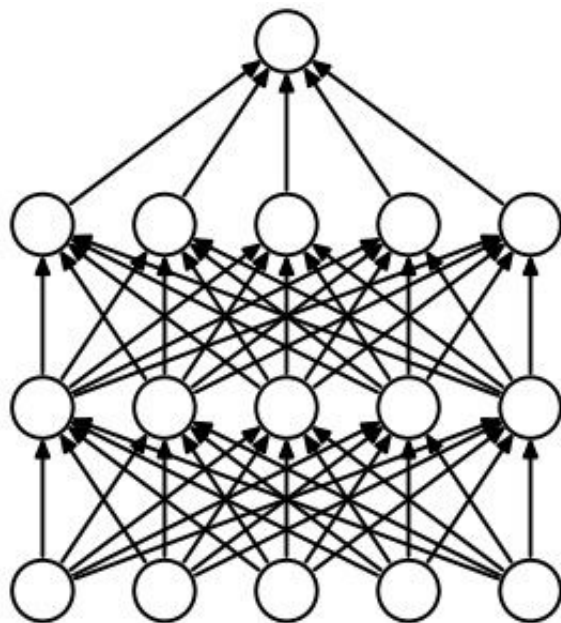


Uncertainty

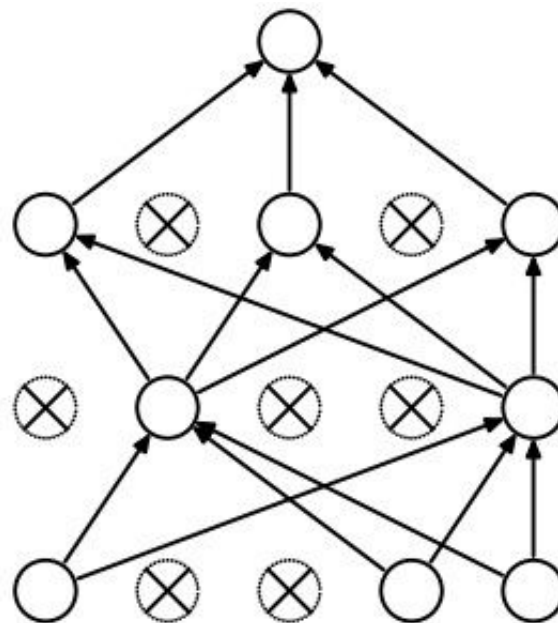
Uncertainty



Dropout



(a) Standard Neural Net



(b) After applying dropout.

Adversarial Examples

Adversarial Examples



x

“panda”

57.7% confidence

$+ .007 \times$



$\text{sign}(\nabla_x J(\theta, x, y))$

“nematode”

8.2% confidence

$=$



$x + \epsilon \text{sign}(\nabla_x J(\theta, x, y))$

“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.

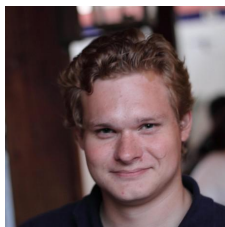


Triggering model

- 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