UNINTUITIVE PROPERTIES OF DEEP NEURAL NETWORKS

Joan Serrà
@serrjoa

Telefónica Research, Barcelona
January 2018

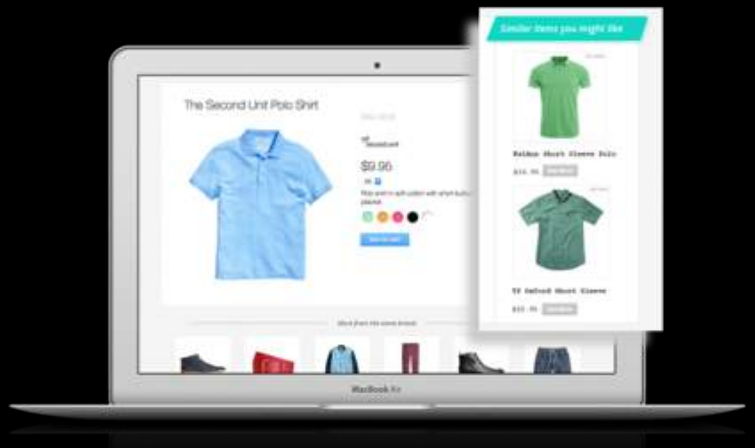
Telefónica
Investigación y Desarrollo

UNIVERSITAT POLITÈCNICA
DE CATALUNYA
BARCELONATECH

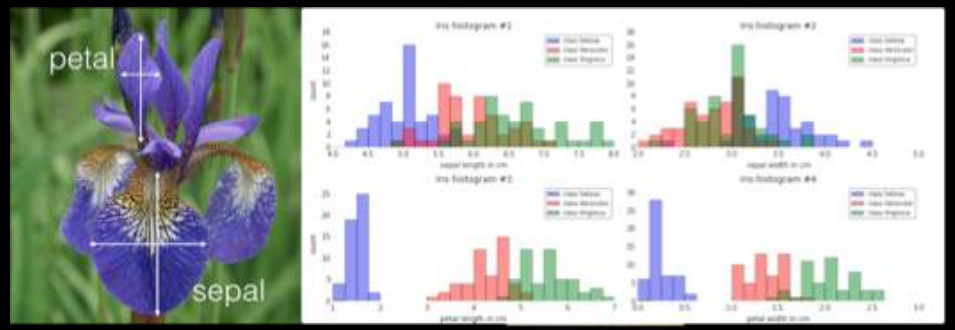
telecom
BCN

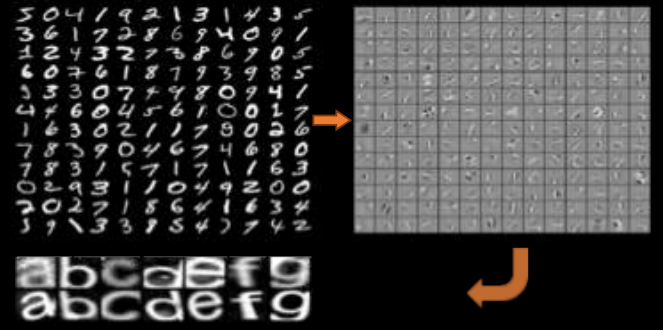
INTRODUCTION

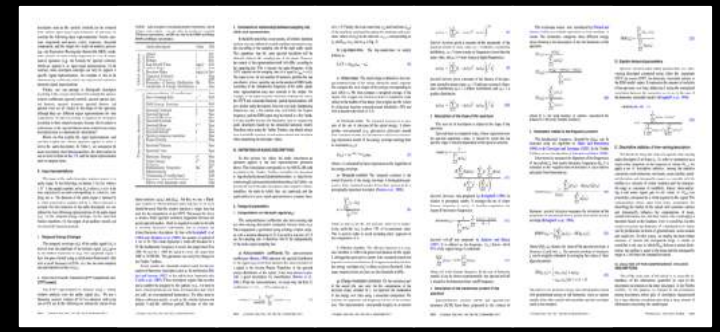
INTRODUCTION:
THE DEEP LEARNING REVOLUTION

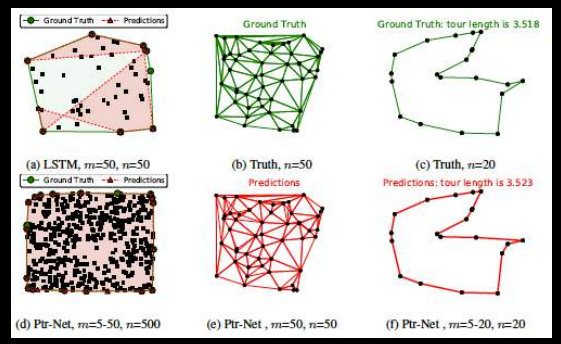


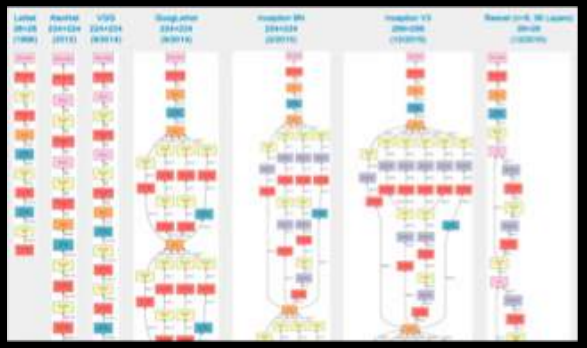
INTRODUCTION:
IN A SINGLE FRAMEWORK!

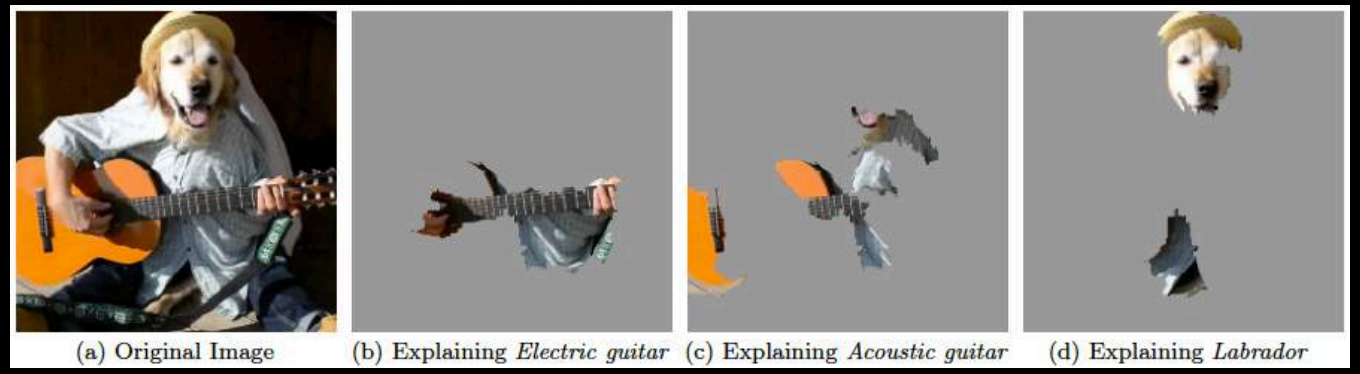


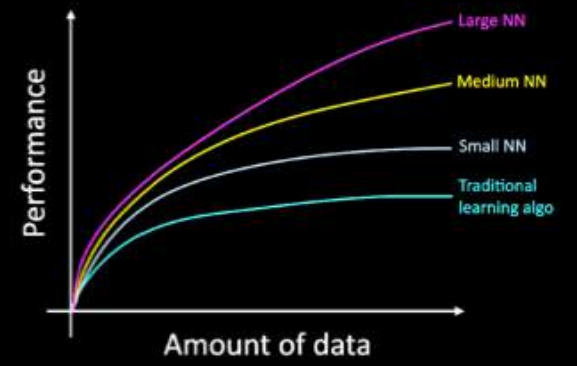


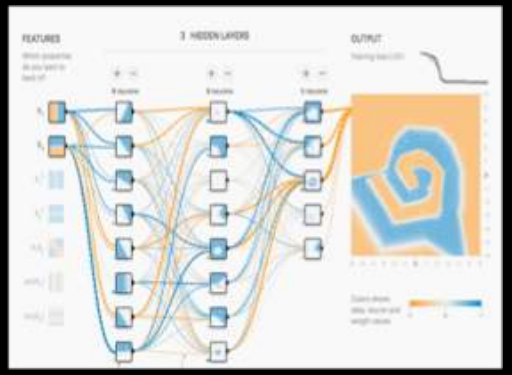










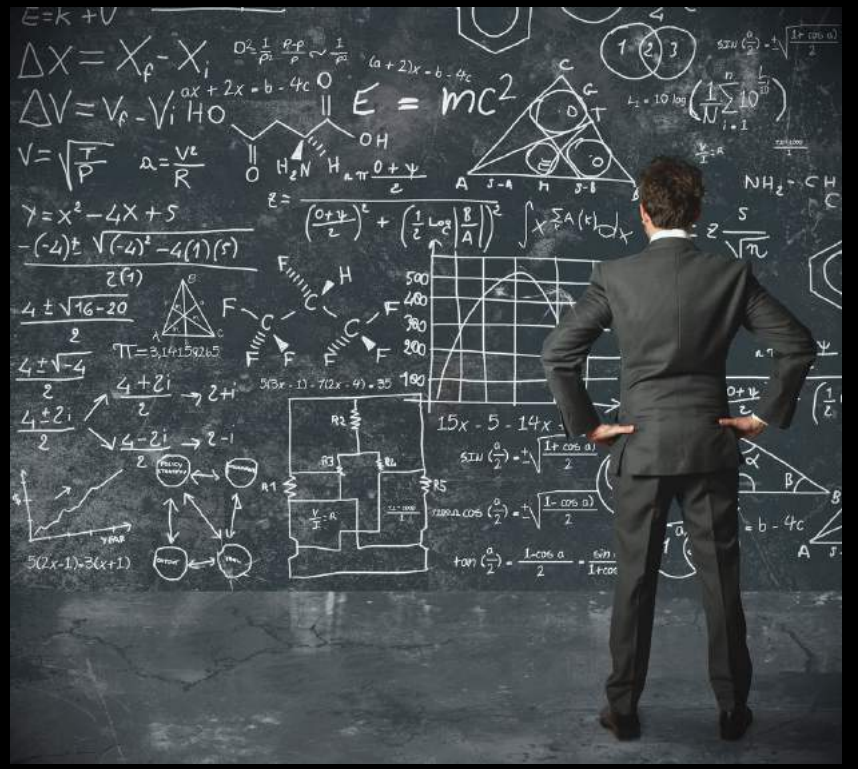


INTRODUCTION:
WHY NOW?

Data availability

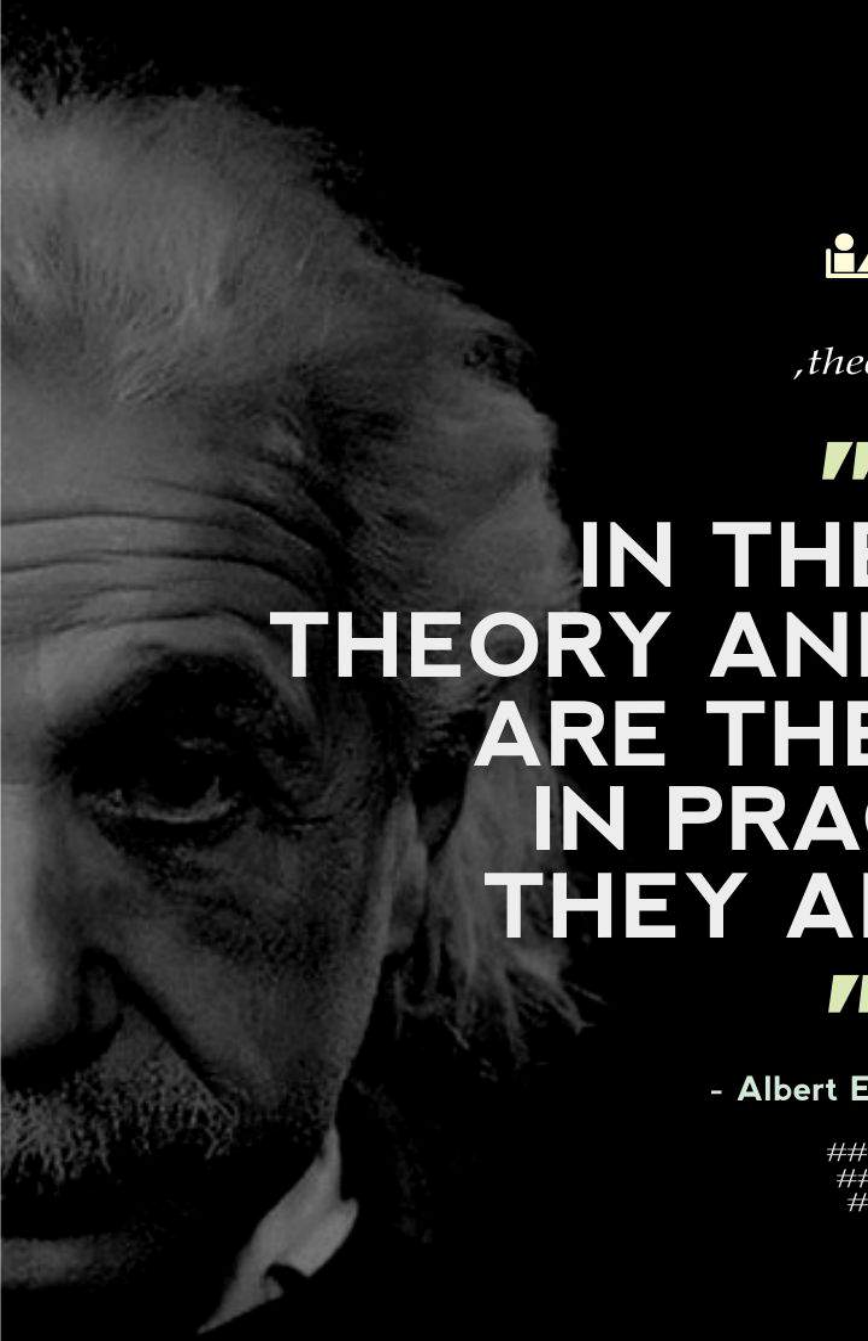
Computing power

New techniques



INTRODUCTION:
METHODS OR TRICKS?

- Unsupervised pre-training
- Convolutions
- ReLUs instead of sigmoids
- Batch normalization
- Skip connections
- Attention
- Dropout
- Plain SGD or Adam?
- Small or large batches?
- Gradient clipping or annealing?
- Weight initialization?
- Deep or wide?
- Erase previous knowledge?
- …



,theory

"IN THEORY, THEORY AND PRACTICE ARE THE SAME. IN PRACTICE, THEY ARE NOT"

- Albert Einstein -

####
##
#

UNINTUITIVE PROPERTIES

NEURAL NETWORKS:

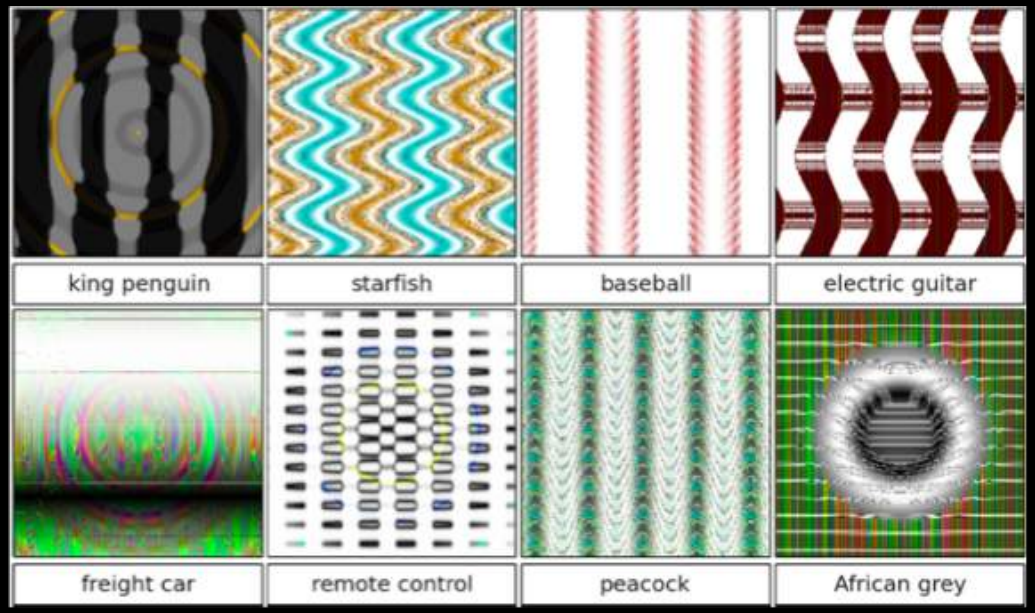
THEY CAN MAKE DUMB ERRORS

- A network can misclassify an image by just applying “a certain hardly perceptible perturbation” (Szegedy et al., 2014)



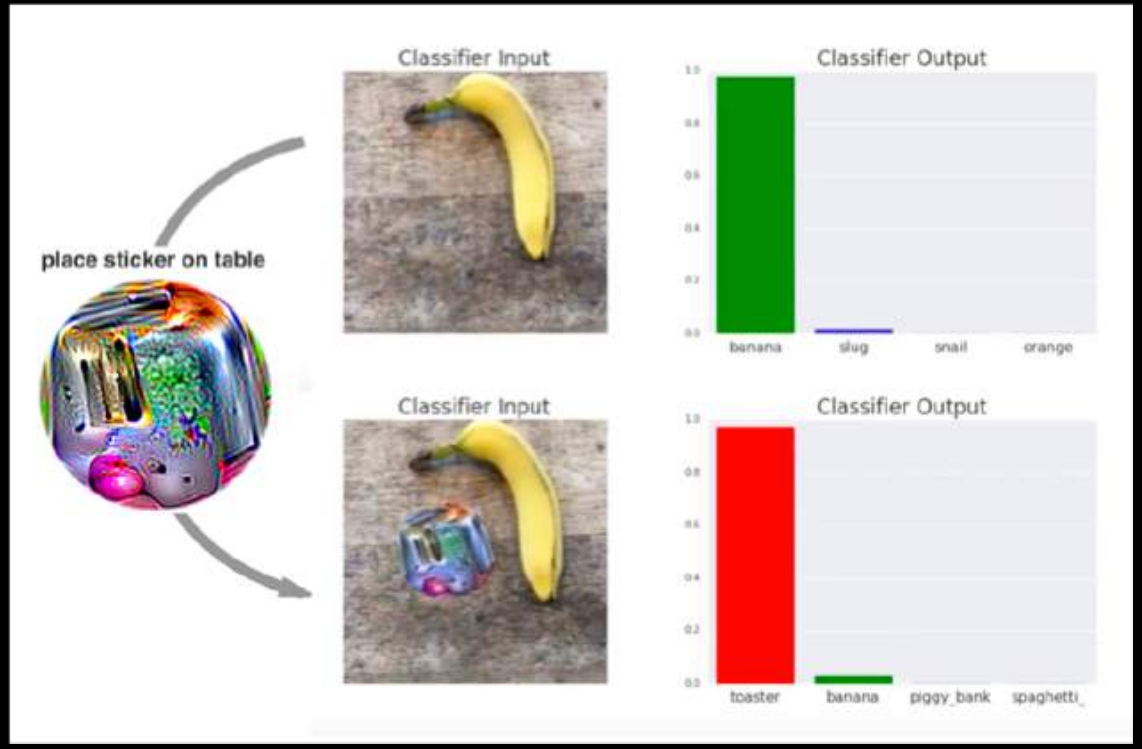
NEURAL NETWORKS:
THEY CAN MAKE DUMB ERRORS

- A network can misclassify an image by just applying “a certain hardly perceptible perturbation” (Szegedy et al., 2014)
- A network can misclassify a totally unrecognizable, artificial image with 99.99% confidence (Nguyen et al., 2015)



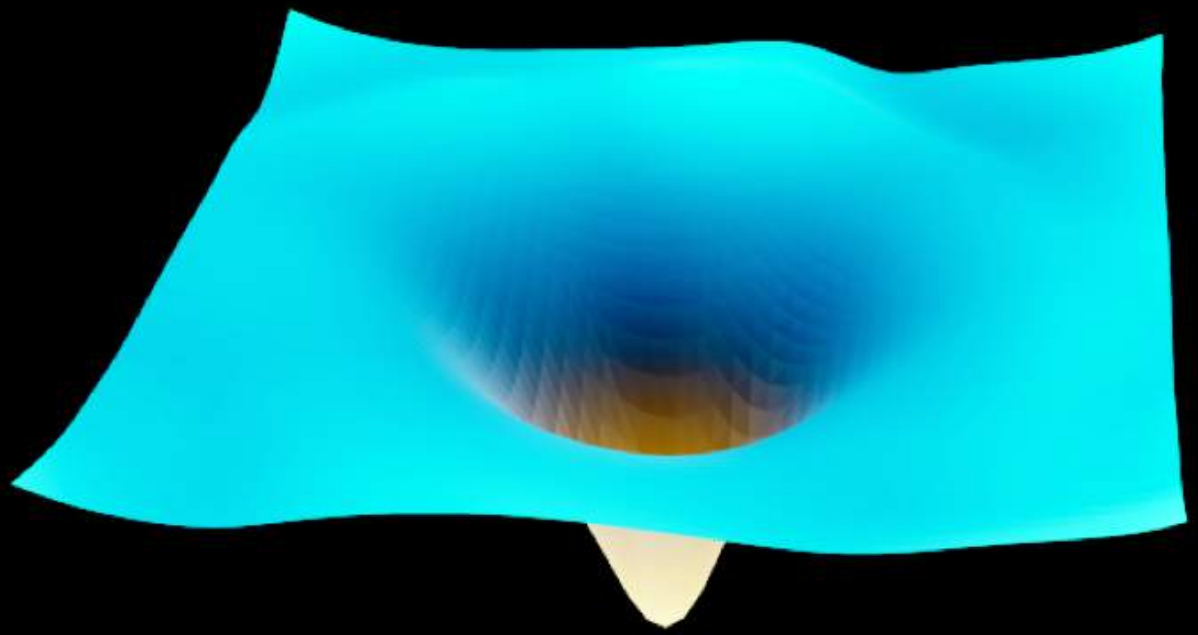
NEURAL NETWORKS:
THEY CAN MAKE DUMB ERRORS

- A network can misclassify an image by just applying "a certain hardly perceptible perturbation" (Szegedy et al., 2014)
- A network can misclassify a totally unrecognizable, artificial image with 99.99% confidence (Nguyen et al., 2015)
- There exist "universal, robust, [and] targeted adversarial image patches" (Brown et al., 2017)



NEURAL NETWORKS:
THEY HAVE A WEIRD LOSS SPACE

$$L = g(\mathbf{y}, f(\mathbf{x}, \mathbf{W}))$$



(Li et al., 2017)

NEURAL NETWORKS:

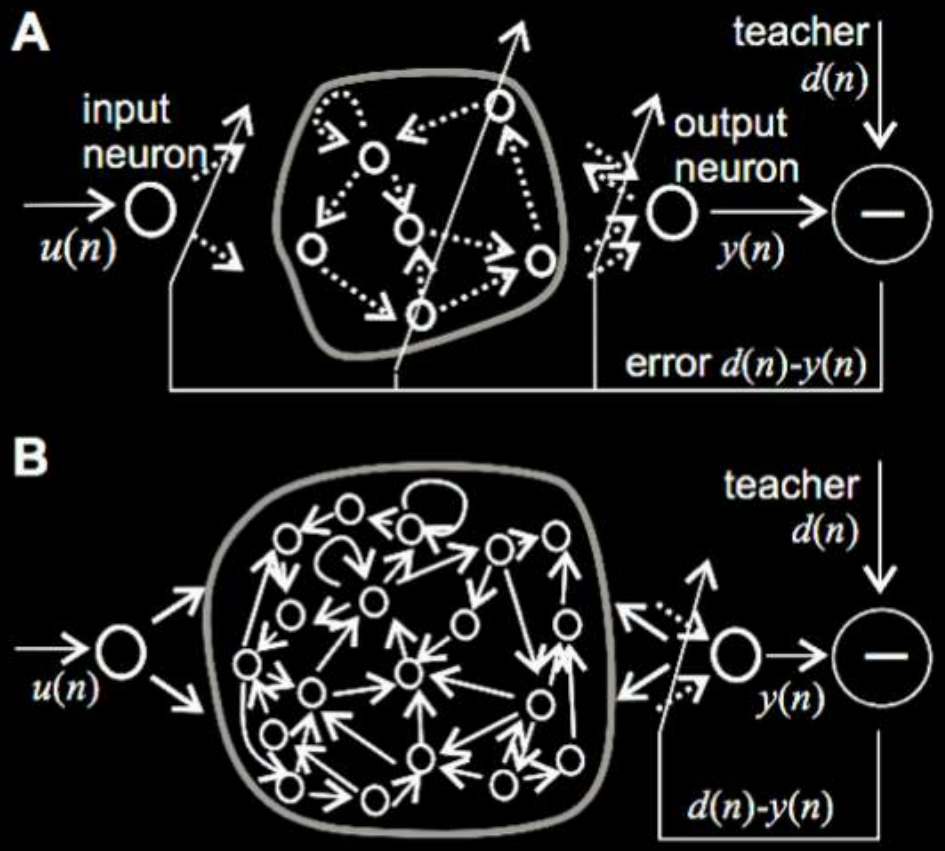
YET THEY WORK BEST WHEN 'BADLY' TRAINED

Why random initializations & the simplest search algorithms still work?

- Flat minima generalize better (Schmidhuber, 1993; Keskar et al., 2017)
- The blessing of dimensionality for saddle points (Dauphin et al., 2014)
- Obstacles not encountered during minimum search (Goodfellow et al., 2015)
- All minima are "global" minima (Kawaguchi, 2016)
- Architectural decisions "convexify" the space (Li et al., 2017)

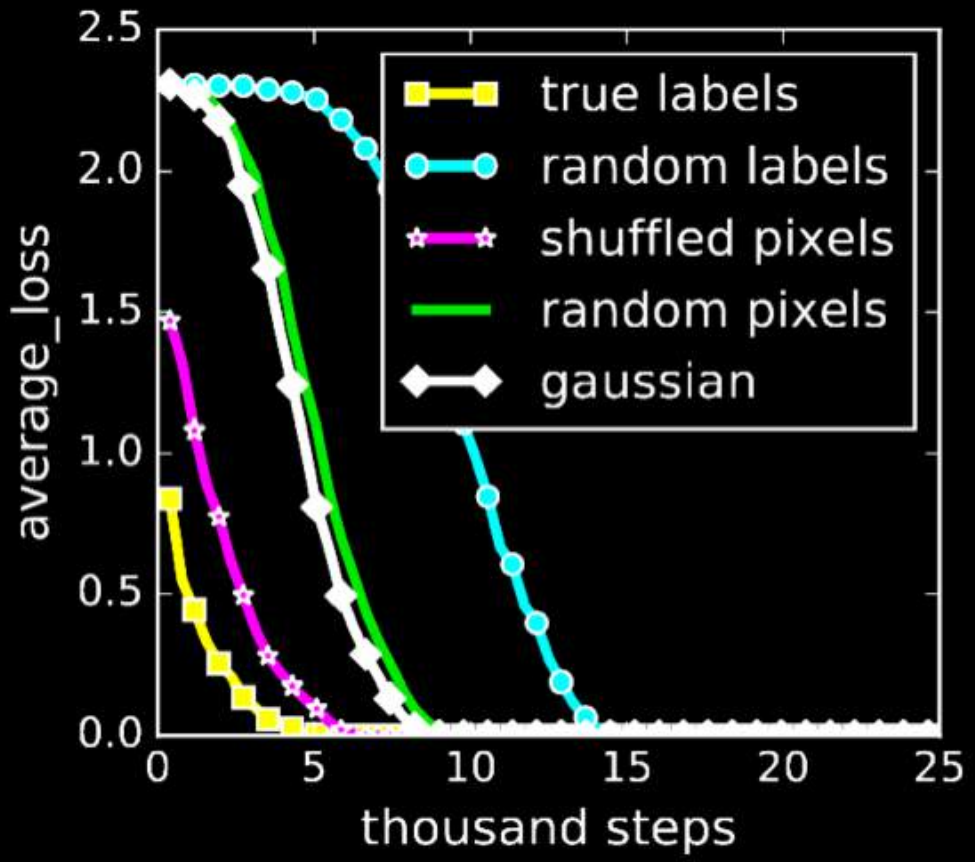
NEURAL NETWORKS:
OR EVEN WITH NO TRAINING

- Only training the last layer:
  - Liquid state machines (Maas et al., 2002)
  - Echo state networks (Jaeger & Haas, 2001)
- Random first convolutional layer (Saxe et al., 2011)
- Only training a tiny fraction of weights (Rosenfeld & Tsotsos, 2018)



NEURAL NETWORKS:
THEY CAN EASILY MEMORIZE

- Universal function approximators
  (Cybenko, 1989)
- Can model any finite-sized data set
  (Zhang et al., 2017), even if it is composed of
  - Random labels
  - Shuffled pixels
  - Random pixels
  - Noise



NEURAL NETWORKS:
YET THEY GENERALIZE TO UNSEEN DATA

Why do they generalize so well?

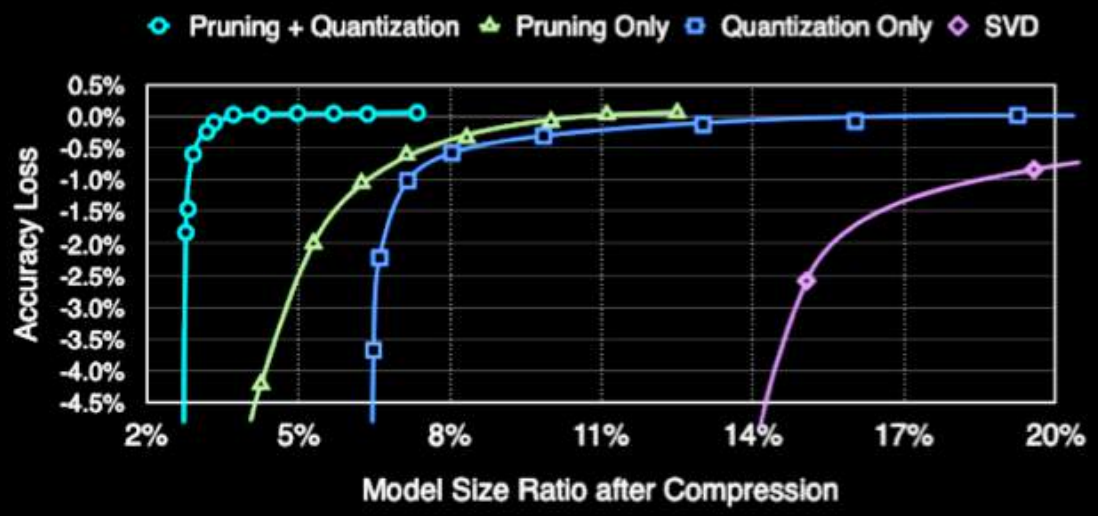
Millions of parameters (orders of magnitude higher than the number of training samples!)

They challenge conventional wisdom:

- ✗ Prefer simple models to achieve good generalization
- ✗ A more-or-less explicit form of regularization needs to take place

NEURAL NETWORKS:
THEY CAN BE COMPRESSED

- To ridiculous sizes (Han et al., 2015)
- Many architectures and data sets
- Using pruning, quantization, etc.



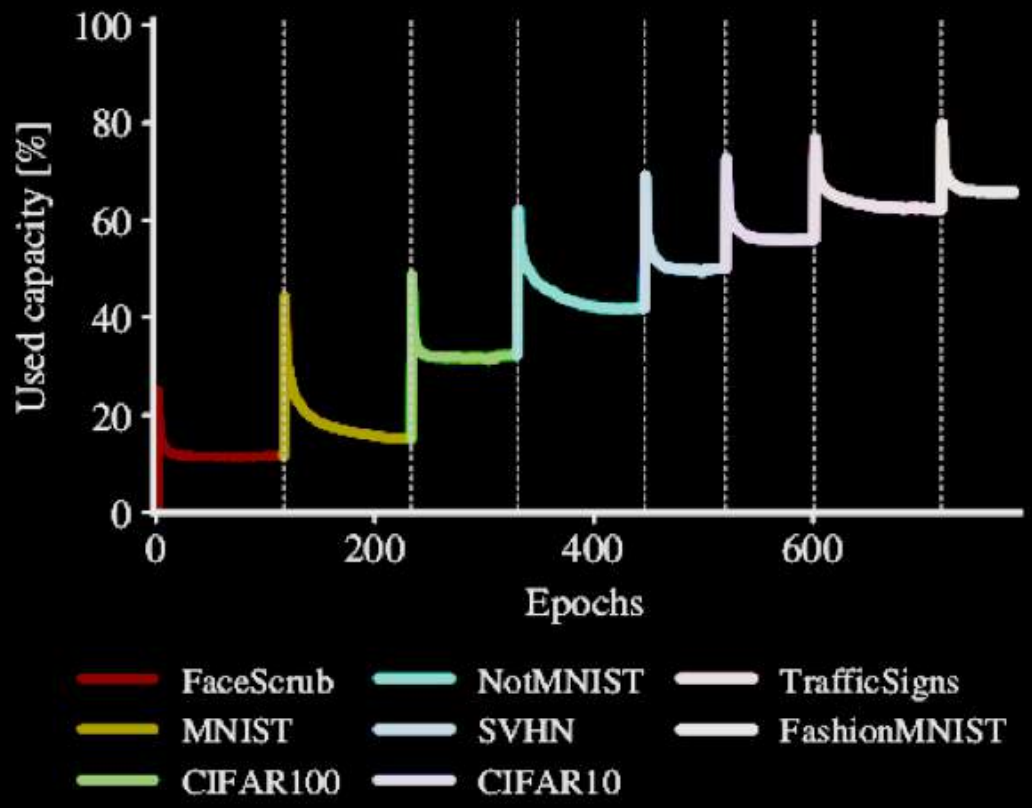
NEURAL NETWORKS:
THEY CAN BE COMPRESSED

- To ridiculous sizes (Han et al., 2015)
- Many architectures and data sets
- Using pruning, quantization, etc.
- Not only a single network but multiple ones: distillation (Hinton et al., 2014)

System	Test Frame Accuracy	WER
Baseline	58.9%	10.9%
10xEnsemble	61.1%	10.7%
Distilled Single model	60.8%	10.7%

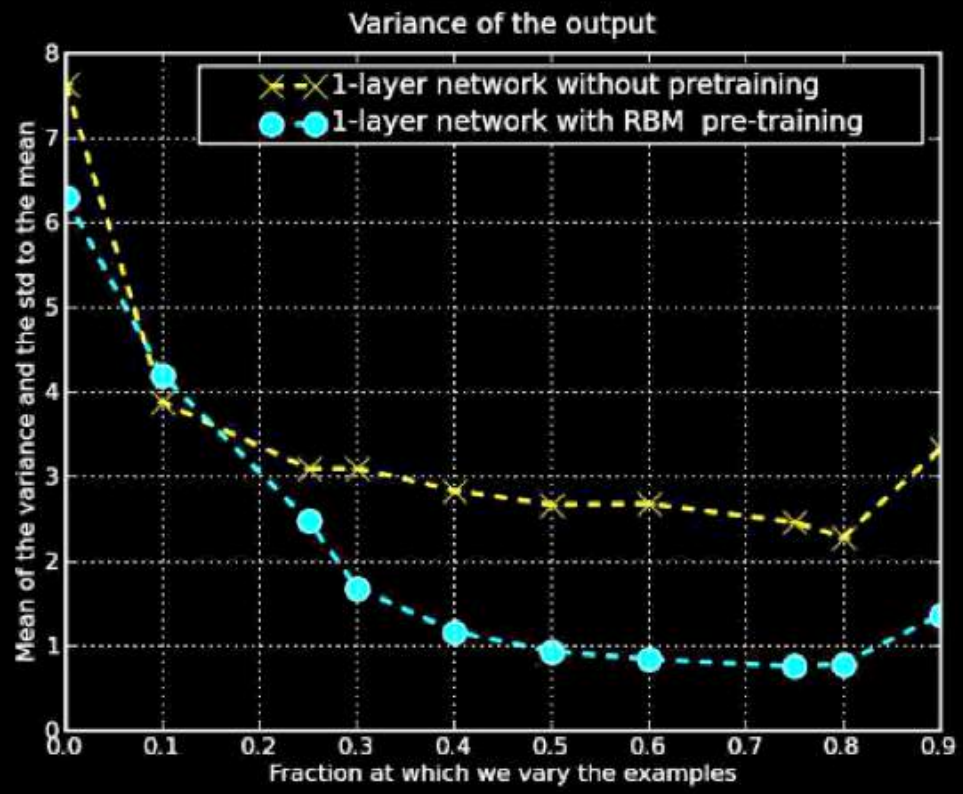
NEURAL NETWORKS:
THEY CAN BE COMPRESSED

- To ridiculous sizes (Han et al., 2015)
- Many architectures and data sets
- Using pruning, quantization, etc.
- Not only a single network but multiple ones: distillation (Hinton et al., 2014)
- Even at learning time! (Serrà et al., 2018)



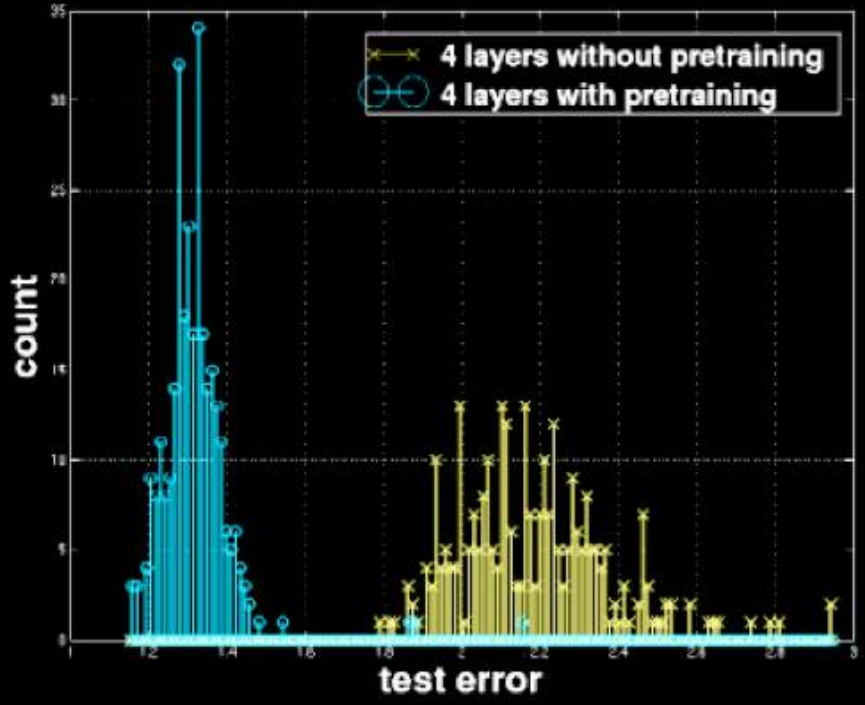
NEURAL NETWORKS:
THEY ARE INFLUENCED BY INIT & ORDER

- Weight initializations are tricky (LeCun et al., 2002)
- Different data orderings yield different results
- Early examples have more influence (Erhan et al., 2010)



NEURAL NETWORKS:
THEY ARE INFLUENCED BY INIT & ORDER

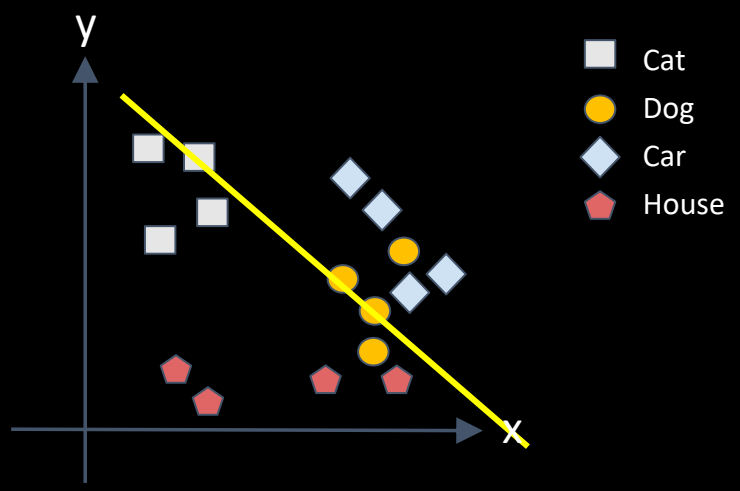
- Weight initializations are tricky (LeCun et al., 2002)
- Different data orderings yield different results
- Early examples have more influence (Erhan et al., 2010)
- Network pre-training (Hinton et al., 2006) or transfer (Yosinski et al., 2014) can be a killer



NEURAL NETWORKS:
THEY FORGET WHAT THEY LEARN

Simple case:

- 2 features (x,y)
- 2 degrees of freedom (line in 2D plane)
- Weights (parameters) forget previous classification



NEURAL NETWORKS:
THEY FORGET WHAT THEY LEARN

- Use of memories (Lopez-Paz & Ranzato, 2017)
- Rehearsal or ‘dreaming’ parallels (Shin et al., 2017)
- Constraining the plasticity of neurons (Kirkpatrick et al., 2017)
- Attention strategies (Serrà et al., 2018)

But backprop may be the hurdle… (McCloskey & Cohen, 1989)

CONCLUSION

CONCLUSION:
MIND THE GAP

PRACTICE-THEORY GAP:
A BAD THING?

Healthy

- Vibrant, dynamic, rapidly evolving field!
- Fosters critical discussion
- Democratizes the field

We have many things to understand!

- Theory should catch up at some point
- However, also much evidence will be empirically driven