Cheat sheet on Training and Evaluating Machine Learning models

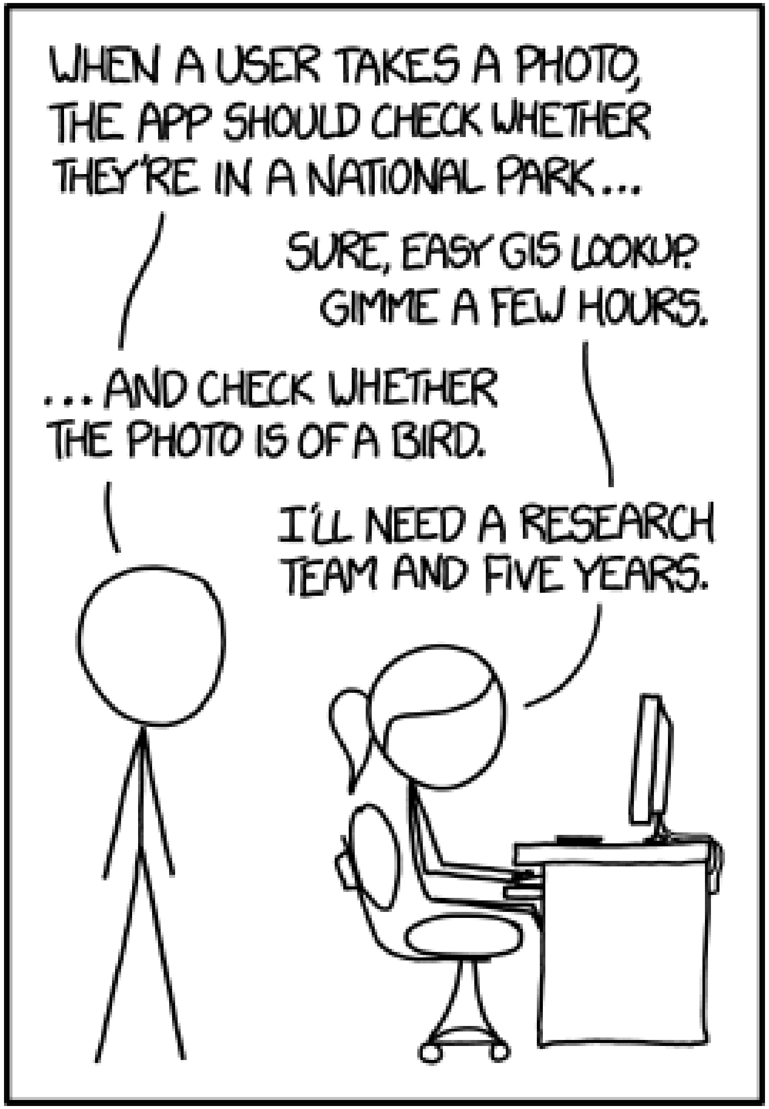
Including Deep Learning models

Outline

1. List possible approaches for improving model performance
2. In real application of Machine Learning for Data Science, which stage from the following list that often consume most of the time?
    a. Data preprocessing
    b. Model construction and training
    c. Model evaluation
3. What does “underfitting” mean? What are the main causes and how to solve this problem?
4. Likewise, What does “overfitting” mean? What are the main causes and how to solve this problem?

Improving model performance

1. Improve performance with data
2. Improve performance with algorithm
3. Improve performance with hyper parameter-tuning
4. Improve performance with ensembles


WHEN A USER TAKES A PHOTO, THE APP SHOULD CHECK WHETHER THEY'RE IN A NATIONAL PARK...
SURE, EASY GIS LOOKUP. GIMME A FEW HOURS.
...AND CHECK WHETHER THE PHOTO IS OF A BIRD.
I'LL NEED A RESEARCH TEAM AND FIVE YEARS.

IN CS, IT CAN BE HARD TO EXPLAIN THE DIFFERENCE BETWEEN THE EASY AND THE VIRTUALLY IMPOSSIBLE.

Source: https://xkcd.com/1425/

Datasets Over Algorithms

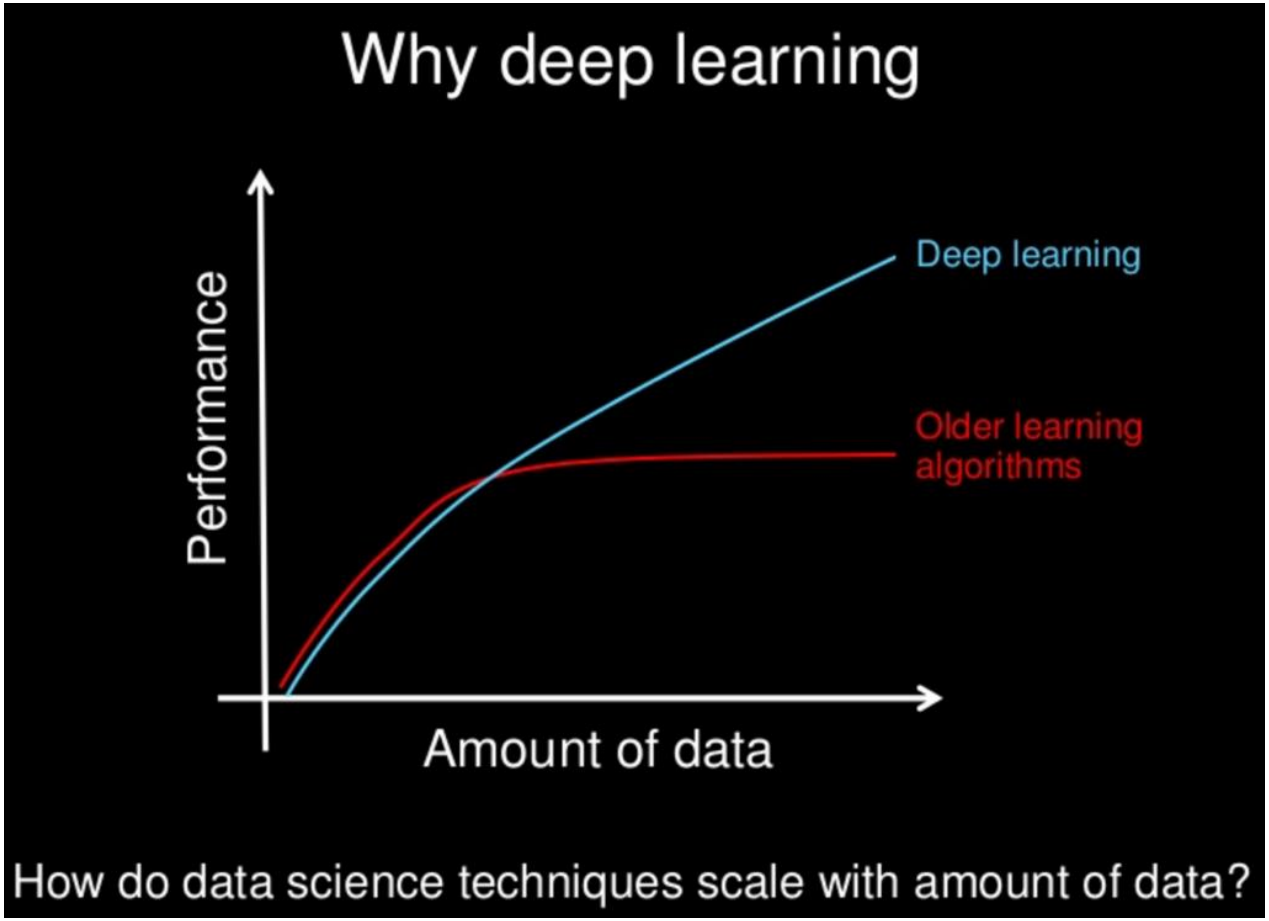
1967: At the dawn of AI, two of its founders anticipated that solving the problem of computer vision would take only a summer.

A half century later ….

- 1994: Speech recognition
- 1997: IBM’s Deep Blue
- 2005: Google translation
- 2011: IBM’s Watson
- 2014: GoogleNet
- 2015: Google’s Deep Mind

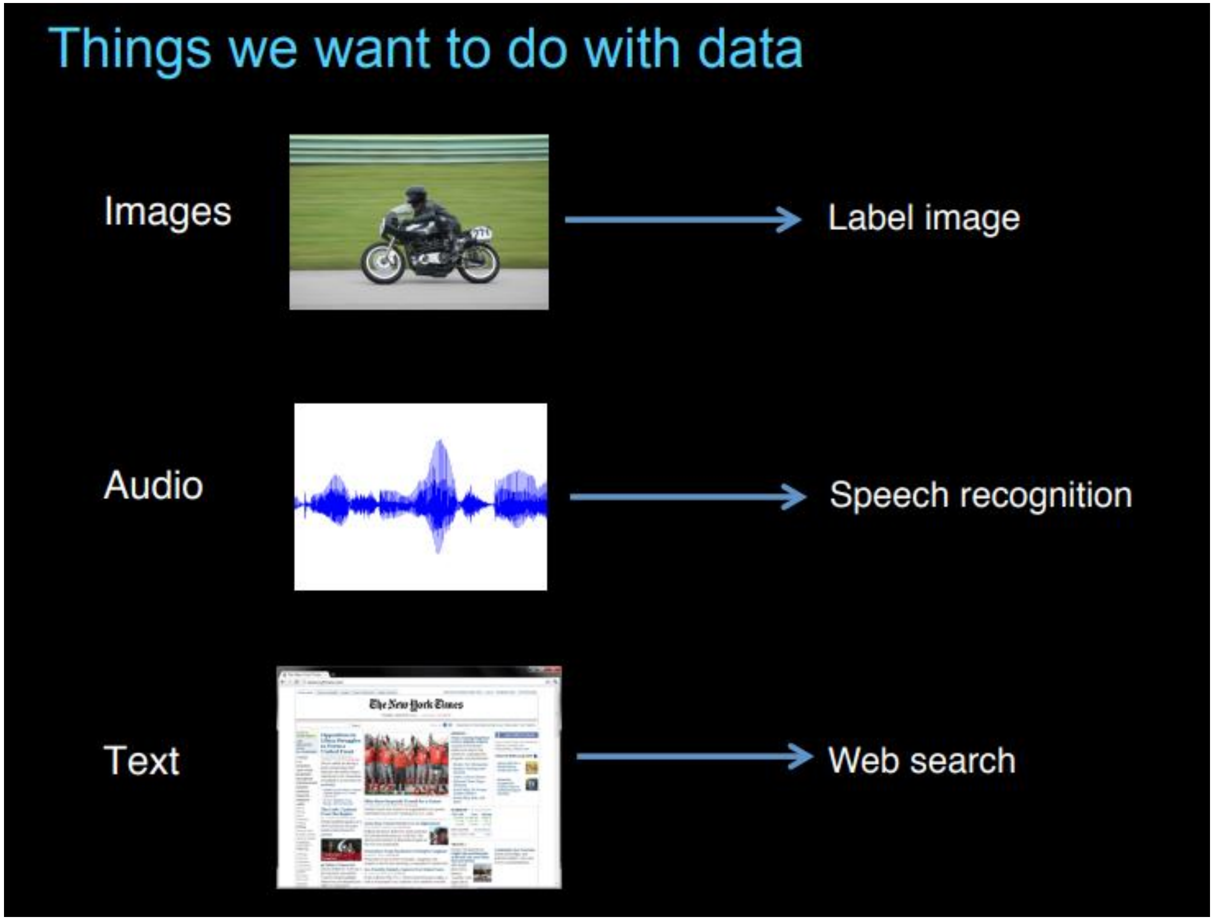
Datasets over algorithms

1. Improve performance with data



Source: CS229-Deep Learning by Andrew Ng.

1. Improve performance with data



Things we want to do with data

Images → Label image

Audio → Speech recognition

Text → Web search

Source: CS229-Deep Learning by Andrew Ng.

1. Improve performance with data

- Get More Data.
- Data augmentation and generation
- Data normalization and representation

  Sub problems: feature importance, feature extraction, feature selection, feature construction, feature learning, feature transformation
- Reframe problem

How to define your Machine Learning Problem

“A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P, if its performance at tasks in T, as measured by P, improves with experience E.” (Tom Mitchell)

Problem: *I need a program that will tell me which tweets will get retweets*

- **Task** (*T*): Classify a tweet that has not been published as going to get retweets or not.
- **Experience** (*E*): A corpus of tweets for an account where some have retweets and some do not.
- **Performance** (*P*): Classification accuracy, the number of tweets predicted correctly out of all tweets considered as a percentage.

Reframe the problem

2. Improve performance with algorithm

- Resampling methods (on training and evaluating stage)
- Evaluation metric
- Baseline performance
- Spot check: No Free Lunch Theory
- Literature review

3. Improve performance with hyper parameter tuning

- Diagnostics.
- Weight Initialization.
- Learning Rate.
- Activation Functions.
- Network Topology.
- Batches and Epochs.
- Regularization.
- Optimization and Loss.
- Early Stopping.

4. Improve performance with ensembles

- Combine Models.
- Combine Views.
- Stacking

References

A Few Useful Things to Know about Machine Learning

https://homes.cs.washington.edu/~pedrod/papers/cacm12.pdf

Why Deep Learning?

http://cs229.stanford.edu/materials/CS229-DeepLearning.pdf