Experimental Design for Machine Learning on Multimedia Data

See: http://www.icsi.berkeley.edu/~fractor/fall2019/

Prof. Gerald Friedland,
fractor@eecs.berkeley.edu
About Me….

▪ Adjunct Assistant Professor
▪ Data Scientist at National Lab
▪ Started work in Machine Learning in 2001
About Me…


http://www.siox.org
The Multimedia Commons (YFCC100M)

100M videos and images, and a growing pool of tools for research with easy access through Cloud Computing

Collaboration Between Academia and Industry:

Benchmarks & Grand Challenges:

Supported in part by NSF Grant 1251276
“BIGDATA: Small: DCM: DA: Collaborative Research: SMASH: Scalable Multimedia content Analysis in a High-level language”
Jupyter Integration
YFCC100M+MMCommons+Amazon’s MXNet

Now we can predict an image’s location. We use images from Placing Task 2016 dataset to evaluate the result.

```python
statinfo = os.stat(filepath)
print('Successfully downloaded', imname, statinfo.st_size, 'bytes.')
return filepath

In [77]: imname = 'd661b83d659af4d818ddd6edf54096.jpg'
   : result = predict_and_evaluate(imname)

('Ground truth: ', (36.016233, -114.737316))
rank=1, prob=0.462128, lat=36.0164119656, lng=-114.737289028, dist from groundtruth=0.020047 km
rank=2, prob=0.260691, lat=36.0141364684, lng=-114.733238705, dist from groundtruth=0.434543 km
rank=3, prob=0.236675, lat=36.0164436503, lng=-114.740746605, dist from groundtruth=0.309436 km
rank=4, prob=0.006943, lat=35.9723883996, lng=-113.791496647, dist from groundtruth=85.229922 km
rank=5, prob=0.006833, lat=36.8630814967, lng=-111.561140839, dist from groundtruth=299.302093 km
```
Why do we care?

• Consumer-Produced Multimedia allows empirical studies at never-before seen scale in various research disciplines such as sociology, medicine, economics, environmental sciences, computer science...

• Recent buzzword: BIGDATA
Problem

How can YOU effectively work on large scale multimedia data (without working at Google)?
Amazon EC2/HPC: Practical Question

▪ How much money (cpu time, memory, IO) do I need to budget for my deep learning experiment?

▪ State of the Art: No answer.
  For example, ImageNet models vary significantly:
  ▪ AlexNet: 238MB model, 2.27Bn Ops
  ▪ DarkNet: 28MB model, 0.96Bn Ops
  ▪ VGG-16: 528 MB, 30.94Bn Ops

Source: https://pjreddie.com/darknet/imagenet/
What is this class about?

- Introduction to systematic experimental design of Machine Learning Experiments
- Covers some theory but is also hands on.
- Covers different modalities: Visual, Audio, Tags, Sensor Data
- Covers different side topics, such as adversarial examples
Content of 2012 Class

- Visual methods for video analysis
- Acoustic methods for video analysis
- Meta-data and tag-based methods for video analysis
- Inferring from the social graph and collaborative filtering
- Information fusion and multimodal integration
- Coping with memory and computational issues
- Crowd sourcing for ground truth annotation
- Privacy issues and societal impact of video retrieval
Content of 2019 Class

• Less anecdotal
• More systematic
• Adds: Concepts for Machine Learning measurements
• See: Machine Learning Cheat Sheet and Design Process
Course Overview

The scientific process and how to think about it in the age of Machine Learning

• The machine learning scientific process

• Measurements beyond accuracy:
  – Capacity
  – Generalization

• Types of Training, Regularization, Occam’s Razor

• Reproducibility vs Repeatability

• Experimental Setup: Annotator Agreement, which machine learner to chose

• Adversarial Examples

• Evaluating success beyond accuracy

• Intrinsics of audio data

• Intrinsics of Image and Video data
Lecture Material


• More material as we go…
How do you receive Credit?

- Attend Lecture Regularly
- Measure out a project. More information: http://www.icsi.berkeley.edu/~fractor/fall2019/
- Weekly homework is optional but will improve your understanding, which will improve your grade!
- Final exam is optional unless you are MEng.
Typical Homework/Final Exam

▪ Calculate capacity for different networks
  ▪ How do you deal with convolutional layers?
  ▪ How does regularization count?

▪ Given a task:
  ▪ Describe what happens if you have too many neurons
  ▪ Describe what happens if you layer too deep
  ▪ Describe what happens if you use features
Project

▪ Chose a project, either yours or somebody else’s or some project of the past.
▪ Write a report to answers 10 questions that
  ▪ require you to measure best-case accuracy, capacity, generalization and other quantities
  ▪ and then ask you to judge the success of the project
  ▪ and comment on its reproducibility.

Due 1 week after the end of the semester.
Help

- Jaeyoung Choi (EC2)
- Rishi Puri (TA)
Questions?
A Warning!

What we think we know:

• Neural Networks can be trained to be more intelligent than humans e.g., beat Go masters

• Deep Learning is better than „shallow“ Learning

• There is no data like more data

• AI is going to take over the world soon

• Let’s pray to AI!

It is what we think we know already that often prevents us from learning.

Claude Bernard
A game...

- Continue the sequence:
  - 2, 4, 6, 8, ....
  - 6, 5, 1, 3, .....  
- What is the next number?
  - 100000 (sequence 1)
  - 100000 (sequence 2)
- Why?
The Scientific Method

Data Science: The Science of Automating the Scientific Method

The Scientific Method: Practical (traditional)

\[ E = mc^2 \]
The Scientific Method: Practical (new)

\[ E = mc^2 \]
Thought Framework: Machine Learning

- Intelligence: *The ability to adapt* (Binet and Simon, 1904)

- Machine learning *adapts a finite state machine* $M$ *to an unknown function based on observations.*

- **Input:** $n$ rows of observations (instances) in a table with header:
  
  $$(x_1, x_2, \ldots, x_m, f(\vec{x}))$$

  where $f(\vec{x})$ is a column with labels we call target function.

- **Output:** State machine $M$ that maps a point
  
  $$(x_1, x_2, \ldots, x_m) \implies f(\vec{x})$$
Thought Framework: Machine Learning

Assume

\[ x_i \in \mathbb{R}, f(\overrightarrow{x}) \in \{0,1\} \]

(binary classifier)

Question:

How many state transitions does \( M \) need to model the training data?
Refresh: Memory Arithmetic

• **Information is reduction of uncertainty:**
  \[ H = -\log_2 P = -\log_2 \frac{1}{\#states} = \log_2 \#states \]
  measured in bits.

• Information: \( \log_2 \#states \) (positive bits)
  Uncertainty: \( \log_2 P = \log_2 \frac{1}{\#states} \) (negative bits)

• If states are not equiprobable, *Shannon Entropy*
  provides tighter bound.
  Math: Assumptions needed! (infinity, distribution)
  Engineering: Estimate using binning
Thought Framework: Machine Learning

Assume

\[ x_i \in \mathbb{R}, f(\vec{x}) \in \{0,1\} \]

(binary classifier)

Question:

How many state transitions does \( M \) need to model the training data?

Maximally: \#rows (lookup table)
Minimally: ?
**Thought Framework: Machine Learning**

- **Intellectual Capacity**: The number of unique target functions a machine learner is able to represent (as a function of the number of model parameters).

- **Memory Equivalent Capacity (MEC)**: A machine learner’s intellectual capacity is memory-equivalent to $N$ bits when the machine learner is able to represent all $2^N$ binary labeling functions of $N$ uniformly random inputs.

- At MEC or higher, $M$ is able to **memorize** all possible state transitions from the input to the output.
Important trick

Memorization is worst-case generalization

• If we deduce nothing from data, the only thing we can do is memorize the observations verbatim.
• Using as many parameters as needed for memorization is therefore an indicator that the machine learner did not deduce anything (overfitting).
• Reducing parameters below memorization capacity will, in the best case, make the machine learner forget what’s not relevant with regards to the target function: generalization.
Generalization in Machine Learning

Memorization is worst-case generalization.

For binary classifiers:

\[
G = \frac{\# \text{correctly classified instances}}{\text{Memory Equivalent Capacity}} \text{ [bits/bit]}
\]

- \(G < 1\) \(\Rightarrow\) \(M\) needs more training/data (not even memorizing)
- \(G = 1\) \(\Rightarrow\) \(M\) is memorizing = overfitting
- \(1 < G < G_{\text{MEM}}\) \(\Rightarrow\) \(M\) could be implementing a lossless compression (and still overfit)
- \(G > G_{\text{MEM}}\) \(\Rightarrow\) \(M\) is generalizing (no chance for overfitting)
Hands-On Intuition: Experimental Design for TensorFlow

http://tfmeter.icsi.berkeley.edu

No Homework this week!