Course Logistics

• Website: https://stellar.mit.edu/S/course/6/sp18/6.883/index.html
• Mailing list: 6883-all@lists.csail.mit.edu [Be sure to fill out the Google form]
• Prerequisites: algorithms (6.046); probability (6.042/6.041/6.008); ML (6.867)
• Format: Five modules (five lectures each)

1. Optimization and Generalization in Deep Learning
2. (Deep) Generative Models
3. Robust/Secure Machine Learning
4. Deep Reinforcement Learning
5. Societal Impact of Machine Learning

• Each module: Two intro lectures by the lectures & Three presentations by 2-3 student teams [45%]
• Crucial aspect: Class discussion [10%]
• Class projects: Explores questions raised in discussion (experiments and theory); done in 2-3 person student teams [45%]

[We will run a team matching process soon]
What will this class be about?

Goal: Build a principled and crisp overview of what deep learning can and cannot do, and what we do and do not know about it.

Science = theoretical models + empirical evaluation
What this class is **NOT**?

- Intro to machine learning/deep learning/Tensorflow/PyTorch/…
  → 6.867, 6.S198
  → http://www.coursera.org/learn/machine-learning
  → http://www.fast.ai/
  → http://neuralnetworksanddeeplearning.com/ (Book)
  → http://www.deeplearningbook.org/ (Book*) (* Parts I and II)
- A survey of state of the art deep learning techniques
  → Impossible (10s of papers uploaded every day)
- Tips on how to make your AI/deep learning startup cooler (Sorry!)

**Key skill we want you to develop:**
“Critical thinking” about deep learning (and ML/AI, in general)
History of Deep Learning (abridged)

- Four pioneers:
  - Hinton
  - LeCun
  - Schmidhuber
  - Bengio
Humble beginnings

- Perceptron [Rosenblatt ‘58]

- Criticism of Perceptrons (XOR affair) [Minsky Papert ‘69]
  \[\rightarrow\] Effectively causes a “deep learning winter”
(Early) Spring

- Back-propagation  [Rumelhart et al. ’86, LeCun ‘85, Parker ‘85]

- Convolutional layers  [LeCun et al. ‘90]

- Recurrent Neural Networks/Long Short-Term Memory (LSTM)  [Hochreiter Schmidhuber ‘97]
Summer

- **2006**: First big success: speech recognition
- **2012**: Breakthrough in computer vision: AlexNet [Krizhevsky et al. ‘12]
- **2015**: Deep learning-based vision models outperform humans
What enabled this success?

• Better architectures (e.g., ReLUs) and regularization techniques (e.g. Dropout)

• Sufficiently large datasets

• Enough computational power
Geist of deep learning

[2018] War erupts for tickets
[2019] AI researchers discover time travel

I was winning ImageNet

Until a deeper model came along
Module I: Optimization and Generalization in Deep Learning
Supervised Machine Learning

\[ f^* = \text{concept to learn} \]
Supervised Machine Learning

\[ f^* = \text{concept to learn} \]
Supervised Machine Learning

\[ f(\theta) = \text{classifier (parametrized by } \theta) \]

Choice of (the family) \( f(\cdot) \) is crucial

Too simple \( \rightarrow \) underfitting

Training: Recover (approx. of) \( f^* \)
by finding parameters \( \theta^* \)
s.t. \( f(\theta^*) \) fits the training data

\( f^* = \text{concept to learn} \)
Supervised Machine Learning

\[ f^* = \text{concept to learn} \]

**Training:** Recover (approx. of) \( f^* \) by finding parameters \( \theta^* \) s.t. \( f(\theta^*) \) fits the training data

\[ f(\theta) = \text{classifier (parametrized by } \theta) \]

Choice of (the family) \( f(\cdot) \) is crucial

- Too simple \( \rightarrow \) underfitting
- Too flexible \( \rightarrow \) overfitting
Supervised Machine Learning

Training: Recover (approx. of) $f^*$ by finding parameters $\theta^*$ s.t. $f(\theta^*)$ fits the training data

$f(\theta) = \text{classifier (parametrized by } \theta)$

Choice of (the family) $f(\cdot)$ is crucial

Too simple $\rightarrow$ underfitting
Too flexible $\rightarrow$ overfitting

"Classic" ML developed a rich and successful theory to understand this phenomenon
Generalization in Deep Learning

Gentlemen, our learner overgeneralizes because the VC-Dimension of our Kernel is too high. Get some experts and minimize the structural risk in a new one. Rework our loss function, make the next kernel stable, unbiased and consider using a soft margin.
Deep neural networks are very expressive, why don’t they overfit?
Optimization in Deep Learning

**Our true goal:** To minimize (wrt $\theta$) the **population risk**

$$E_{(x,y) \sim D} \left[ \text{loss}(f(\theta, x), y) \right]$$

**What we actually do:** Minimize (wrt $\theta$) the **empirical risk**

$$\sum_i \text{loss}(f(\theta, x_i), y_i)$$

where $\{(x_i, y_i)\}_i$ are the training data points

→ In case of neural networks, empirical risk is a continuous and (mostly) differentiable function

→ Can use gradient descent method (back-propagation) to solve it!
Optimization in Deep Learning

\[ \min_{\theta} \sum_i \text{loss}(f(\theta, x_i), y_i) \]

→ **Issue 1:** There is a lot of terms in this sum
→ Use **stochastic** gradient descent (SGD) instead of grad. descent (SGD = the workhorse of deep learning)

→ **Issue 2:** This problem is very non-convex
→ Still, we seem to reliably* converge to good solutions. Why?

*In fact: Stochasticity of SGD seems to be a “feature”, not a deficiency. (Hypothesis: “Implicit regularization”.)
Module II: Deep Generative Models
Unsupervised Machine Learning

• **Goal:** Learn from **unlabeled** data by understanding its structure

  **Popular approach:** Try to fit the data to some **generative model**

• **Example:** Fit the distribution to a mixture of Gaussians
Deep Generative Models

• Neural networks constitute (parametric) models too!

• Variational Autoencoders (VAEs) [Kingma Welling ’13, Rezende et al. ’14]

• Generative Adversarial Networks (GANs) [Goodfellow et al. ’14]

Questions:
• What are/should be the guarantees these models aim to satisfy?
• Do existing constructions work? Can they ever?
• How would we measure their success?
Module III: Robust/Secure ML
Recent Progress in ML

Have we *really* achieved human-level performance?
Adversarial Examples

Too fragile?

Too contrived?

Translations + rotations (shifts by <10% pixels, <30° rotations)

CIFAR10: 93% → 8% accuracy
ImageNet: 76% → 31% accuracy


[Engstrom, Tsipras, Schmidt, M., 2017]
Why Does It Matter?

- **Security** (currently, everything is “broken”)
  [Sharif, Bhagavatula, Bauer, Reiter, 2016]

- **Safety** (“benign” noise can be a problem too)

- **Understanding “failure modes” of current vision models**
  (they are not as “human-like” as we might have expected)

**Crucial question:** Can you really rely on your (deep) ML model?
What Do We Do Now?

- **Problem:** Adversarial examples are not at odds with our current notion of generalization

- Time to re-think what we mean by generalization?

- There is a number of other problems/questions, such as data poisoning, model theft,…

- **Again:** This is not only about security/safety but also about understanding how ML/deep learning works (and fails!)
Module IV: (Deep) Reinforcement Learning
Reinforcement Learning (RL)

What if the Agent was a (deep) neural network?

Questions:
- How to train such agent (exploration vs. exploitation)?
- What are the fundamental limits on efficiency of this approach?
- How to ensure that the agent does what we really intend it to do?
Module V: Societal Impacts of ML
Machine learning is entering (and taking control of) every aspects of our life

• Should we be worried?

• Potential concerns:
  → Interpretability (Can we understand ML models “reasoning”?)
  → Reliability (Can I trust the prediction of an ML model?)
  → Fairness (Is the ML model behaving in a “fair” way?)
  → Privacy (Is the ML model protecting our privacy?)
  → AI Alignment (Can we teach AI to share our values?)
  → AI Safety (If we build a super-human AI, will it destroy us?)
  → (Your suggestion here)