Technical and Societal Critiques of ML

CS 229, Spring 2020
by Angelica Sun

Slides adapted from
Chris Chute, Taide Ding, and Andrey Kurenkov
- Adversarial Examples
- Interpretability
- Expense: Data and Computation
- Community Weakness
- Ethical Concerns
- Adversarial Examples
- Interpretability
- Expense: Data and Computation
- Community Weakness
- Ethical Concerns


Figure: Left: Correctly classified image. Right: classified as Ostrich. Reproduced from [1].
The **Smoothness Assumption** in conventional kernel methods:

“For a small enough radius $E > 0$ in the vicinity of a given training input $x$, an $x + r$ satisfying $|r| < E$ will get assigned a high probability of the correct class by the model” [1].

Does NOT hold true in deep neural networks

For more references:

Constructing adversarial examples: [2, 3].
Defending against them: [1, 4, 5, 6].
Constructing adversarial examples

Fast gradient sign method [2]. Given input $x$, add noise $\eta$ in the direction of the gradient

$$x_{\text{Adv}} = x + \eta = x + E \cdot \text{sign}(\nabla_x J(\theta, x, y)).$$

**Intuition:** By perturbing the example in the direction of the gradient, you increase the cost function w.r.t. the correct label most efficiently.

*Figure:* FGSM example, GoogLeNet trained on ImageNet, $E = .007$. Source: [2].
Fast gradient sign method properties

- Change often indistinguishable to human eye.
- Adversarial examples generalize across architectures, training sets.
  Adversarial perturbations $\eta$ generalize across examples.
- Can construct in the physical world (e.g. stop signs)

Figure: A turtle. Or is it a rifle? Reproduced from [7].
Defenses Techniques

- Train on mixture of clean $x$ and perturbed $x^\sim$ [1].
- Many other defenses: [6]. But... Goodfellow et al. [2] claims fundamental problem with linear models (and high-dimensional input):

$$w^T x^\sim = w^T x + w^T \eta.$$ 

- Arms race: generating adversarial examples with GANs (Ermon Lab: [3]).
- Adversarial Examples
- Interpretability
- Expense: Data and Computation
- Community Weakness
- Ethical Concerns
Interpretability

Considerations from Lipton, “The Mythos of Model Interpretability” [8]

- Trust: Costs of relinquishing control - is the model right where humans are right?
- Causality: Need to uncover causal relationships?
- Transferability: generalizes to other distributions / novel environments?
  Informativeness: not just answer, but context
- Fairness and ethics: Will real-world effect be fair?

Main problem: Evaluation metrics that only look at predictions and ground truth labels don’t always capture the above considerations.
Interpretability: Fallacies

**Fallacy 1**

“Linear models are interpretable. Neural networks are black boxes.”

**What is “interpretable”?** Two possible perspectives:

**Algorithmic transparency:** decomposable, understandable, can easily assign interpretations to parameters

**Post-hoc interpretation:** Text, visualization, local explanation, explanation by example.

Linear models win on algorithmic transparency.

Neural networks win on post-hoc interpretation, with rich features to visualize, verbalize, and cluster.
Visualization. e.g. render distributed representations in 2D with t-SNE [9].


Word embeddings
Local explanation. Popular: e.g., Saliency Maps [10], CAMs (class activation mapping) [11], Grad-CAMs [12], attention [13, 14].

Figure: CAMs.

Figure: Grad-CAMs.
Explanation by example. Run $k$-NN on representations.

Interpretability: Fallacies

**Fallacy 2**

“All AI applications need to be transparent.”

**Figure:** Is this a transparent algorithm? If not, why do you use it?

Transparency as a hard rule can exclude useful models that do complex tasks better than us.
Interpretability: Fallacies

Fallacy 3

“Always trust post-hoc explanation.”

- Post-hoc interpretations can be optimized to mislead.
- E.g., in college admissions, post-hoc explanations of leadership and originality disguise racial, gender discrimination [15].
Interpretability Summary

- Never discuss “interpretability” without clarifying the definition.
- Beware of interpretability fallacies.
- Find your domain-specific definition of interpretability, then use the tools available.
- Align evaluation metrics with what is qualitatively important.
● Adversarial Examples
● Interpretability
● Expense: Data and Computation
● Community Weakness
● Ethical Concerns
Costly data collection and computation (in time and money).

Since Deep Learning, compute use has been increasing faster than Moore’s Law!

- Humans for data annotation
- Cloud computation
- Long hours of training
- Electricity cost
- …

Popular media: *Training a single AI model can emit as much carbon as five cars in their lifetimes*
Solution to data expense: Unsupervised [16, 17] and semi-supervised approaches [18].

- Transfer learning [17, 19]. Pretrain on related tasks.
- Use public datasets,
- Download model parameters from internet.
- E.g. Many computer vision models use a backbone pretrained on ImageNet.

Recent work from Stanford researchers, Taskonomy [20], models the structure of space of visual tasks to guide transfer learning.
Solution to compute expense: Obtaining smaller models

- Compression [21].
  - Codebook
  - Pruning
  - Distillation
- Low-bit Quantization [22].
- Specialized hardware [23, 24]. GPUs are inefficient. More efficiency with FPGA, TPU.

Pruning: sparsifying the model by removing unimportant weights

Deep compression: Pruning connections, quantizing weights, and Huffman coding (shorter codes for higher frequencies of occurrence)
● Adversarial Examples
● Interpretability
● Expense: Data and Computation
● Community Weakness
● Ethical Concerns
Community Weakness

- Cycle of hype and winter [25].
- Lack of rigor and worries of troubling scholarship trends [26, 27].
  - Many incorrect theories invented to explain observations, rather than derived from theoretical foundations [28, 29].
  - Suggestion of [28]: Spend more time doing experiments to find root cause for unexpected results, rather than chasing performance.
- Barriers to entry (funding and data) and tech monopoly?
- Side-effects of industry-driven research?
- Adversarial Examples
- Interpretability
- Expense: Data and Computation
- Community Weakness
- Ethical Concerns
Ethical Concerns

- ML captures social biases in datasets
- **NYT (11/11/19):** "We Teach A.I. Systems *Everything,* Including Our Biases"
  - "BERT is more likely to associate the word “programmer” with men than with women.”
  - "If a tweet or headline contained the word “Trump,” the tool almost always judged it to be negative, no matter how positive the sentiment.”

- **NYT (06/17/19) "Exposing the Bias Embedded in Tech”**
  - Imbalance in training data leads to negative societal consequences: Xbox Kinect (2010) worked less well for women and children (trained on 18-35 year old men)
  - Facial recognition more accurate with lighter-skinned men
  - A.I. resume readers penalized occurrences of “women” and “women’s colleges”

- **Pro Publica (2016) ”Machine Bias” - race and AI risk assessments / bail calculations**
  - The COMPAS controversy
Ethical Concerns

- ML captures social biases in datasets
- Like any technology, ML can be used in ways whose legality / ethics are questionable
Questionable Use of AI

CNN (01/2019): "When seeing is no longer believing - Inside the Pentagon’s race against deepfake videos"

DeepFake Eroding trustworthiness of video evidence

VICE (06/27/19): "Creator of DeepNude, App That Undresses Photos of Women, Takes It Offline"

Legality and legal rights over deepfakes

Which one is real?

Need legal frameworks for holding AI users accountable
Ethical Concerns

- ML captures social biases in datasets
- Like any technology, ML can be used in ways whose legality / ethics are questionable
- More discussions on privacy and security, e.g.
  - Large-scale web scraping for dataset collection
  - Ethical decisions: what should a autopilot car do in a trolley dilemma?
  - Privacy issues e.g. face detection
  - ...

All for today:

- Adversarial Examples
- Interpretability
- Expense: Data and Computation
- Community Weakness
- Ethical Concerns

ML is a dynamic field with wide-reaching societal impact. Take your critics and stakeholders seriously!
References

2 Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. 

3 Yang Song, Rui Shu, Nate Kushman, and Stefano Ermon. Generative adversarial examples. 

4 Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. 


8 Zachary C Lipton. The mythos of model interpretability. 


19  Kaiming He, Ross Girshick, and Piotr Dollár.
Rethinking imagenet pretraining.

20  Amir R Zamir, Alexander Sax, William Shen, Leonidas Guibas,
Jitendra Malik, and Silvio Savarese.
Taskonomy: Disentangling task transfer learning.
*In Proceedings of the IEEE Conference on Computer Vision and

21  Song Han, Huizi Mao, and William J Dally.
Deep compression: Compressing deep neural networks with pruning,
trained quantization and huffman coding.
Itay Hubara, Matthieu Courbariaux, Daniel Soudry, Ran El-Yaniv, and Yoshua Bengio.

*Field-programmable gate arrays*, volume 180.

Norm Jouppi.
Google supercharges machine learning tasks with tpu custom chip.
*Google Blog, May, 18, 2016.*


[29] Shibani Santurkar, Dimitris Tsipras, Andrew Ilyas, and Aleksander Madry.
How does batch normalization help optimization?(no, it is not about internal covariate shift).