Technical and Societal Critiques of ML

CS 229, Fall 2019

slides by Chris Chute, Taide Ding, and Andrey Kurenkov

November 15, 2019
Overview

1. Adversarial examples
2. Interpretability
3. Expense: Data and compute
4. Community weaknesses
5. Ethical Concerns
Adversarial examples

Figure: Left: Correctly classified image. Right: classified as Ostrich. Reproduced from [1].
Adversarial examples

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

- How to construct: [2, 3].
- How to defend: [1, 4, 5, 6].

Still an open problem
Constructing adversarial examples

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

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

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

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

**Figure:** FGSM example, GoogLeNet trained on ImageNet, $\epsilon = .007$. Source: [2].
Constructing adversarial examples

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

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

**Figure:** FGSM example, GoogLeNet trained on ImageNet, \( \epsilon = .007 \). Source: [2].

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

- Change often indistinguishable to human eye.

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

- Change often indistinguishable to human eye.
- Adversarial examples **generalize** across architectures, training sets.
- Adversarial perturbations $\eta$ generalize across examples.
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

- Train on mixture of clean $x$, perturbed $\tilde{x}$ [1].
Defenses

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

$$w^T\tilde{x} = w^Tx + w^T\eta.$$
Defenses

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

$$w^T \tilde{x} = w^T x + w^T \eta.$$

- Arms race: generating adversarial examples with GANs (Ermon Lab: [3])
Switching gears: Interpretability.
Switching gears: Interpretability.

Considerations (from Lipton: "The Mythos of Model Interpretability" [8]):

1. Trust: Costs of relinquishing control - is the model right where humans are right?
2. Causality: Need to uncover causal relationships?
3. Transferability: generalizes to other distributions / novel environments?
4. Informativeness: not just answer, but context
5. 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.
Switching gears: Interpretability.

Considerations (from Lipton: “The Mythos of Model Interpretability” [8]):

1. **Trust**: Costs of relinquishing control - is the model right where humans are right?

2. **Causality**: Need to uncover causal relationships?

3. **Transferability**: Generalizes to other distributions / novel environments?

4. **Informativeness**: Not just answer, but context

5. **Fairness and ethics**: Will real-world effect be fair?
Interpretability

Switching gears: Interpretability.

Considerations (from Lipton: "The Mythos of Model Interpretability" [8]):

1. Trust: Costs of relinquishing control - is the model right where humans are right?
2. Causality: Need to uncover causal relationships?
Switching gears: Interpretability.

Considerations (from Lipton: “The Mythos of Model Interpretability” [8]):

1. Trust: Costs of relinquishing control - is the model right where humans are right?
2. Causality: Need to uncover causal relationships?
3. Transferability: generalizes to other distributions / novel environments?
Switching gears: Interpretability.

Considerations (from Lipton: "The Mythos of Model Interpretability" [8]):

1. Trust: Costs of relinquishing control - is the model right where humans are right?
2. Causality: Need to uncover causal relationships?
3. Transferability: generalizes to other distributions / novel environments?
4. Informativeness: not just answer, but context
Interpretability

Switching gears: Interpretability.

Considerations (from Lipton: “The Mythos of Model Interpretability” [8]):

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

Switching gears: Interpretability.

Considerations (from Lipton: “The Mythos of Model Interpretability” [8]):

1. Trust: Costs of relinquishing control - is the model right where humans are right?
2. Causality: Need to uncover causal relationships?
3. Transferability: generalizes to other distributions / novel environments?
4. Informativeness: not just answer, but context
5. 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
Fallacy 1. “Linear models are interpretable. Neural networks are black boxes.”
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: rich features to visualize, verbalize, cluster.
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: rich features to visualize, verbalize, cluster.
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.
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: rich features to visualize, verbalize, cluster.
Interpretability Definition 2: Post-hoc Explanation

- **Visualization.** e.g. render distributed representations in 2D with t-SNE [9].
Interpretability Definition 2: Post-hoc Explanation

- **Visualization.** e.g. render distributed representations in 2D with t-SNE [9].

- **Local explanation.** Popular: e.g., Saliency Maps [10], CAMs (class activation mapping) [11], Grad-CAMs [12], attention [13, 14].
Interpretability Definition 2: Post-hoc Explanation

- **Visualization.** e.g. render distributed representations in 2D with t-SNE [9].

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

Figure: Grad-CAMs.
Interpretability Definition 2: Post-hoc Explanation

- **Visualization.** e.g. render distributed representations in 2D with t-SNE [9].

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

![Grad-CAMs](image)

**Figure:** Grad-CAMs.

- **Explanation by example.** Run $k$-NN on representations.
Interpretability: Fallacies

- **Fallacy 2.** “All AI applications need to be transparent.”
Interpretability: Fallacies

- **Fallacy 2.** “All AI applications need to be transparent.”

*Figure: Is this a transparent algorithm? If not, why do you use it?*
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.
**Fallacy 3.** Always trust post-hoc explanation (e.g., CAMs).
Fallacy 3. Always trust post-hoc explanation (e.g., CAMs).
Post-hoc interpretations can be \textit{optimized to mislead}. 
**Interpretability: Fallacies**

- **Fallacy 3.** Always trust post-hoc explanation (e.g., CAMs).
- 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].
Never discuss “interpretability” without clarifying the definition.
Interpretability: Summary

- Never discuss “interpretability” without clarifying the definition.
- Beware of interpretability fallacies.
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.
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.
Switching gears: ML can be expensive.
Costly data collection and computation (in time and money).
Costly data collection and computation (in time and money).

Solution 1: Unsupervised [16, 17] and semi-supervised approaches [18].
Transfer learning [17, 19]. Pretrain on related tasks.
Expense: Data

- Transfer learning [17, 19]. Pretrain on related tasks.
- Use public datasets, e.g., ImageNet.
Expense: Data

- Transfer learning [17, 19]. Pretrain on related tasks.
- Use public datasets, e.g., ImageNet.
- Download model parameters from internet.
Expense: Data

- Transfer learning [17, 19]. Pretrain on related tasks.
- Use public datasets, e.g., ImageNet.
- Download model parameters from internet.
- Recent work from Stanford researchers: Taskonomy [20].

**Figure:** Taskonomy: “taxonomy of tasks” to guide transfer learning - modeling the structure of space of visual tasks
Expense: Compute

Since Deep Learning, compute use has been increasing faster than Moore’s Law! Popular media: Training a single AI model can emit as much carbon as five cars in their lifetimes.

Figure: From OpenAI
Expense: Compute

- Compression [21].
Expense: Compute

- Compression [21].
- Quantization [22].
Expense: Compute

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

Figure: Deep compression: Pruning connections, quantizing weights, and Huffman coding (shorter codes for higher frequencies of occurrence) (50× gains).
Data: Transfer learning, public datasets, unsupervised pretraining. Newer techniques coming out frequently.

Compute: Compression, quantization, specialized hardware.
Community weaknesses

- Cycle of hype and winter [25].
Community weaknesses

- Cycle of hype and winter [25].
- Lack of rigor and worries of troubling scholarship trends [26, 27].
Community weaknesses

- 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].

Barriers to entry (funding and data).

Side-effects of industry-driven research?
Community weaknesses

- 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.
Community weaknesses

- 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).
Community weaknesses

- 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).

Side-effects of industry-driven research?
Last gear switch: some things to be mindful of wrt ML.
Ethical Concerns

- ML captures social biases in dataset
- Like any technology, ML can be used in ways whose legality / ethics are questionable
ML captures social biases in dataset

NYT (11/11/19): ”We Teach A.I. Systems Everything, Including Our Biases”
ML captures social biases in dataset

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.”
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.”
ML captures social biases in dataset

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”
ML captures social biases in dataset

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
ML captures social biases in dataset

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)
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
ML captures social biases in dataset

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
  - AI resume readers penalized occurrences of "women" and "women’s colleges"
ML captures social biases in dataset

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
  - AI resume readers penalized occurrences of ”women” and ”women’s colleges”

Pro Publica (2016) ”Machine Bias” - race and AI risk assessments / bail calculations
CNN (01/2019): “When seeing is no longer believing - Inside the Pentagon’s race against deepfake videos”

- Eroding trustworthiness of video evidence
CNN (01/2019): ”When seeing is no longer believing - Inside the Pentagon’s race against deepfake videos”
- 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
Questionable Use of AI

CNN (01/2019): ”When seeing is no longer believing - Inside the Pentagon’s race against deepfake videos”
  - 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
NYT (04/14/19): ”One Month, 500,000 Face Scans: How China Is Using A.I. to Profile a Minority”
  - Billion-dollar companies with government contracts for mass surveillance
CNN (01/2019): ”When seeing is no longer believing - Inside the Pentagon’s race against deepfake videos”

- 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

NYT (04/14/19): ”One Month, 500,000 Face Scans: How China Is Using A.I. to Profile a Minority”

- Billion-dollar companies with government contracts for mass surveillance

**Legal frameworks for holding AI users accountable are needed**
Conclusion

Technical and societal critiques of AI: we’ve only scratched the surface.

1. Adversarial examples
2. Interpretability
3. Expense: Data and compute
4. Community weaknesses
5. Ethical Concerns

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


Distillation as a defense to adversarial perturbations against deep neural networks.

Pixeldefend: Leveraging generative models to understand and defend against adversarial examples.

Synthesizing robust adversarial examples.
[8] Zachary C Lipton. 
The mythos of model interpretability. 

Visualizing data using t-sne. 

Deep inside convolutional networks: Visualising image classification models and saliency maps. 
Learning deep features for discriminative localization.


Rethinking imagenet pretraining.  

Taskonomy: Disentangling task transfer learning.  

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

*Field-programmable gate arrays*, volume 180.

Google supercharges machine learning tasks with tpu custom chip.
*Google Blog, May, 18, 2016.*
*Deep learning*, volume 1.  

[26] Zachary C Lipton and Jacob Steinhardt.  
Troubling trends in machine learning scholarship.  

[27] Theories of deep learning (stats 385).  

Ai is the new alchemy (nips 2017 talk).  
https://www.youtube.com/watch?v=Qi1Yry33TQE, December 2017.
[29] Shibani Santurkar, Dimitris Tsipras, Andrew Ilyas, and Aleksander Madry. 
How does batch normalization help optimization? (no, it is not about internal covariate shift).