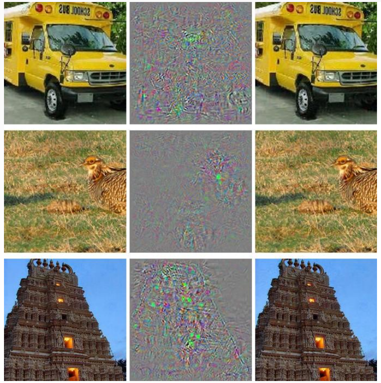
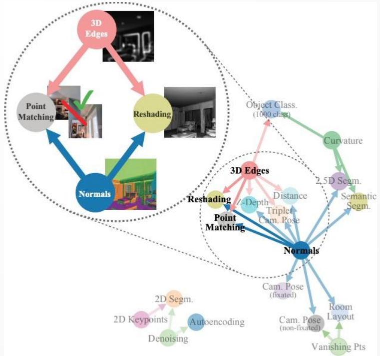
Technical and Societal Critiques of ML

CS 229, Spring 2020
by Angelica Sun

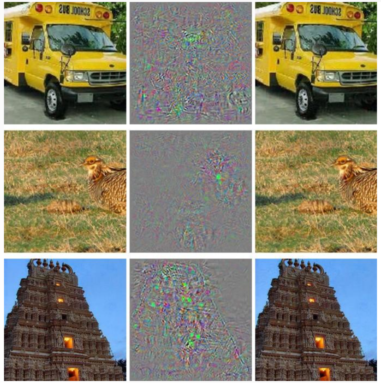
Slides adapted from
Chris Chute, Taide Ding, and Andrey Kurenkov

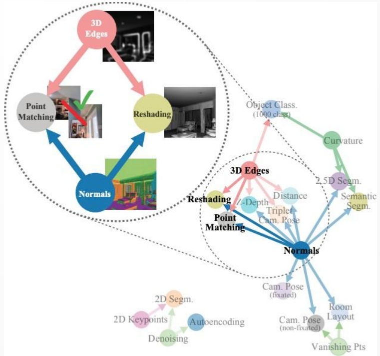
- Adversarial Examples
- Interpretability
- Expense: Data and Computation
- Community Weakness
- Ethical Concerns



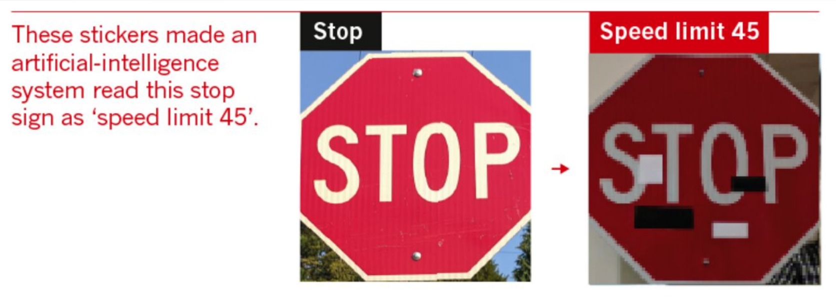


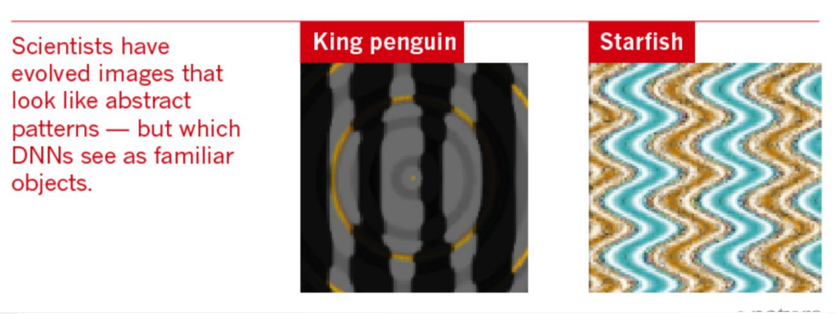
- **Adversarial Examples**
- Interpretability
- Expense: Data and Computation
- Community Weakness
- Ethical Concerns





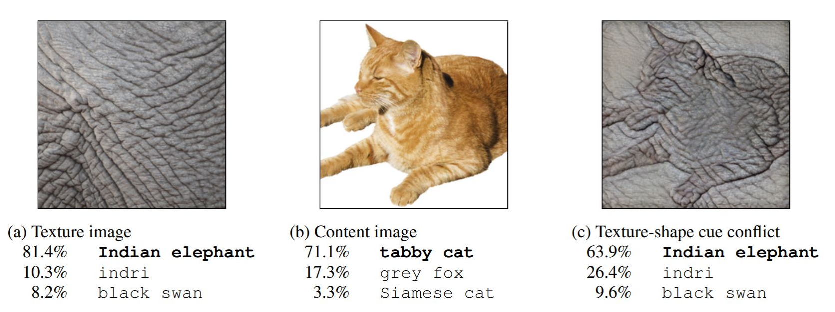
Eykholt, Kevin, et al. "Robust physical-world attacks on deep learning visual classification." *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2018.





Nguyen, Anh, Jason Yosinski, and Jeff Clune. "Deep neural networks are easily fooled: High confidence predictions for unrecognizable images." *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2015.

Geirhos, Robert, et al. "ImageNet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness." *arXiv preprint arXiv:1811.12231* (2018).



Goh, Gabriel, et al. "Multimodal neurons in artificial neural networks." *Distill* 6.3 (2021): e30.

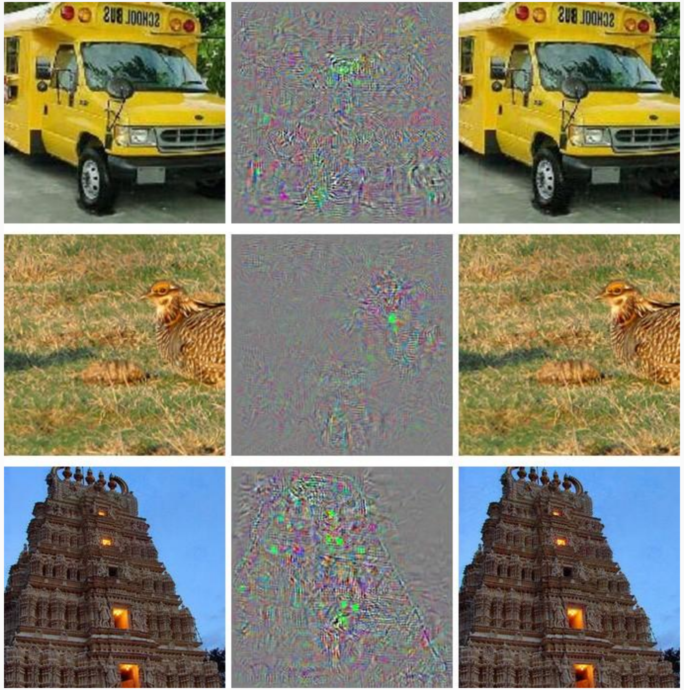


Figure: Left: Correctly classified image. Right: classified as **Ostrich.** Reproduced from [1].

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 $\lvert r \rvert < E$ 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 ***η*** in the direction of the gradient

$$\mathbf{x}_{Adv} = \mathbf{x} + \boldsymbol{\eta} = \mathbf{x} + E \cdot \text{sign}(\nabla_{\mathbf{x}} J(\boldsymbol{\theta}, \mathbf{x}, y)).$$

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

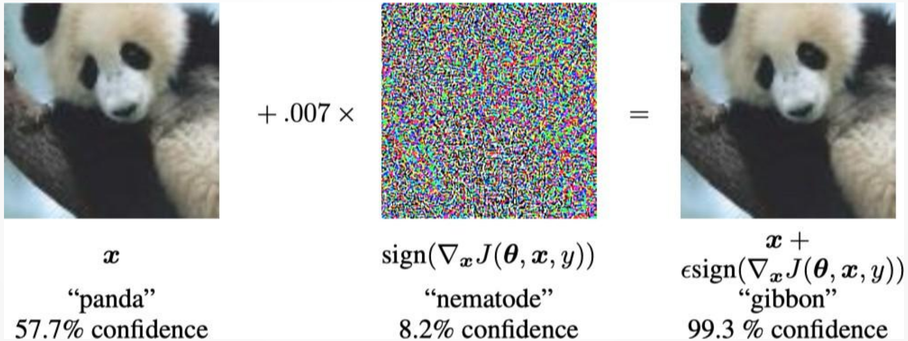


Figure: FGSM example, GoogLeNet trained on ImageNet, E = .007. Source: [2].

Fast gradient sign method properties

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

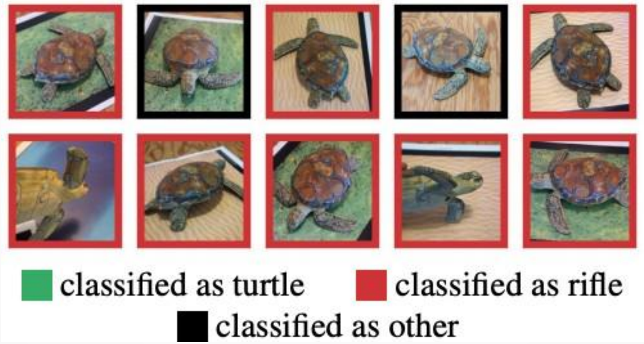


Figure: A turtle. Or is it a rifle? Reproduced from [7].

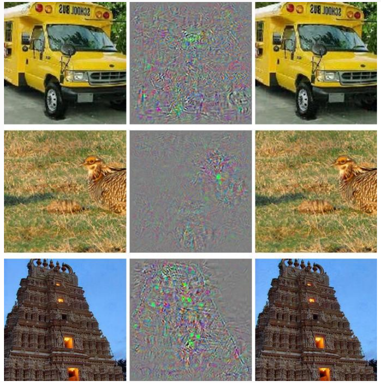
Defenses Techniques

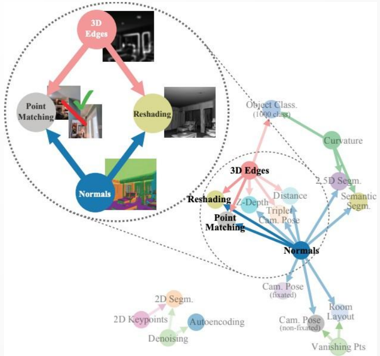
- Train on mixture of clean $\boldsymbol{x}$ and perturbed $\tilde{\boldsymbol{x}}$ [1].
- Use **distillation** [4] as a defense [5]. Instead of training with hard (one-hot) labels, train with a high-temperature softmax output of another NN trained with hard labels
- Many other defenses: [6]. But... Goodfellow et al. [2] claims fundamental problem with linear models (and high-dimensional input):

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

- 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.”

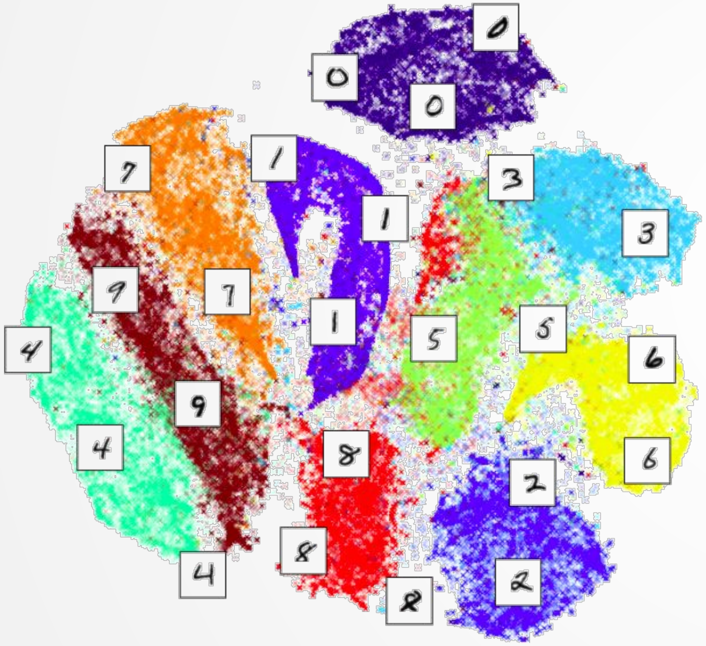
What is “interpretable”? Two possible perspectives:

- **Algorithmic transparency:** decomposable, understandable, can easily assign interpretations to parameters
- **Post-hoc interpretation:** Text, visualization, local explanation, explanation by example.

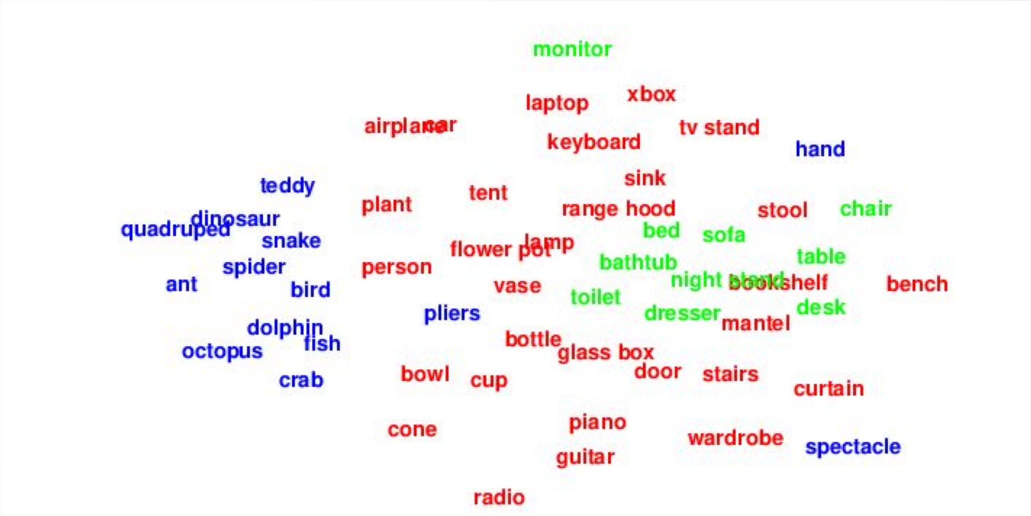
Linear models win on algorithmic transparency.

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].



MNIST classification:
https://nlml.github.io/in-raw-numpy/in-raw-numpy-t-sne/



Word embeddings
https://www.researchgate.net/publication/331397000_Zero-shot_Learning_of_3D_Point_Cloud_Objects

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

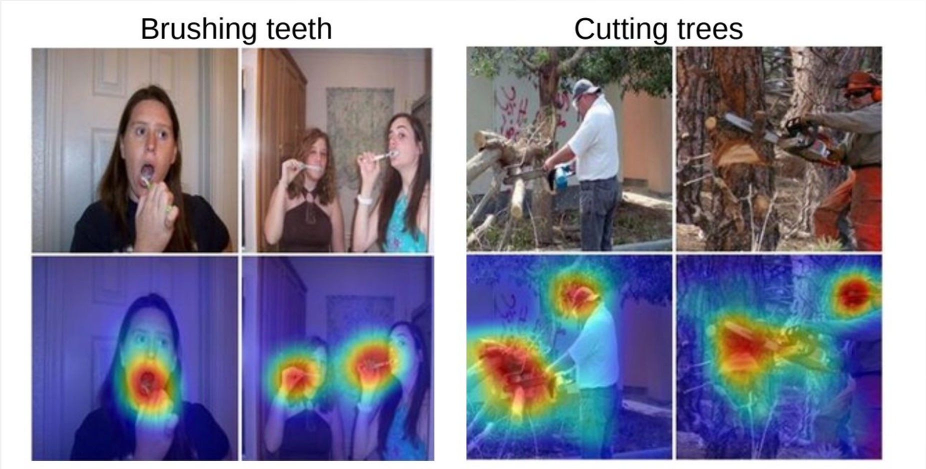


Figure: CAMs.

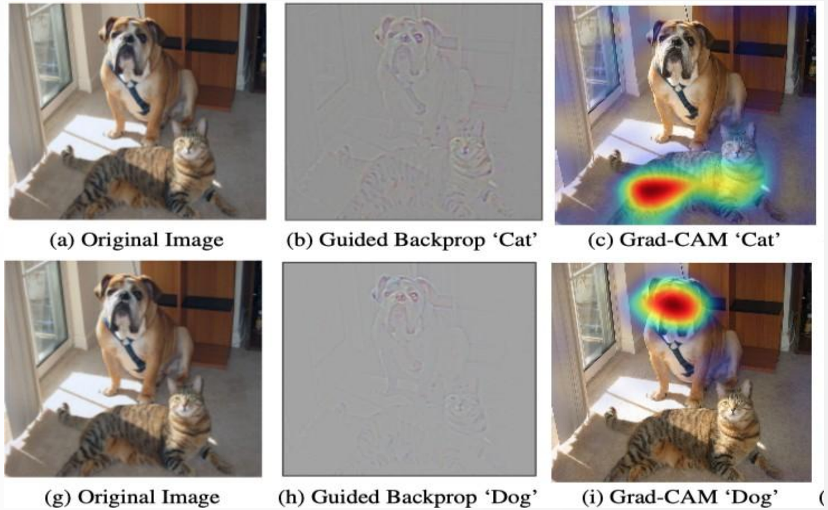
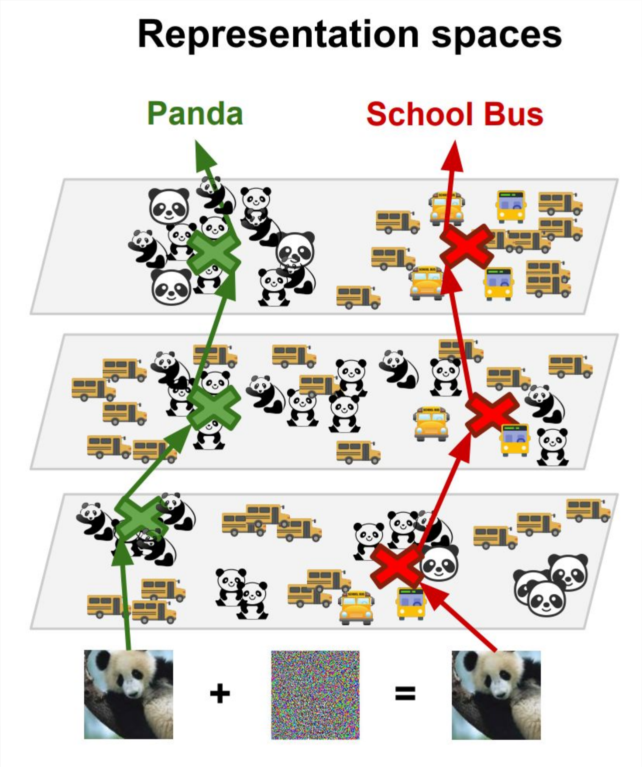


Figure: Grad-CAMs.

Explanation by example. Run *k*-NN on representations.



Papernot, Nicolas, and Patrick McDaniel. "Deep k-nearest neighbors: Towards confident, interpretable and robust deep learning." *arXiv preprint arXiv:1803.04765* (2018).

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

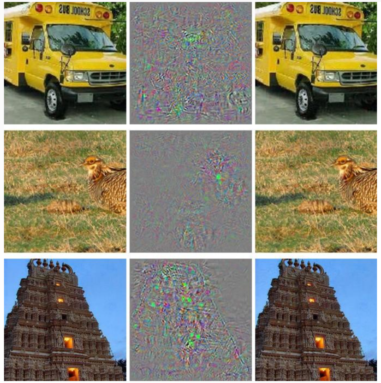
"Always trust post-hoc explanation."

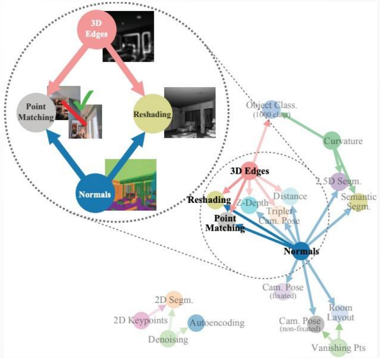
- Post-hoc interpretations can be **optimized to mislead.**
- *E.g.*, in college admissions, post-hoc explanations of *leadership* and *originality* disguise racial, gender discrimination [15].

Interpretability Summary

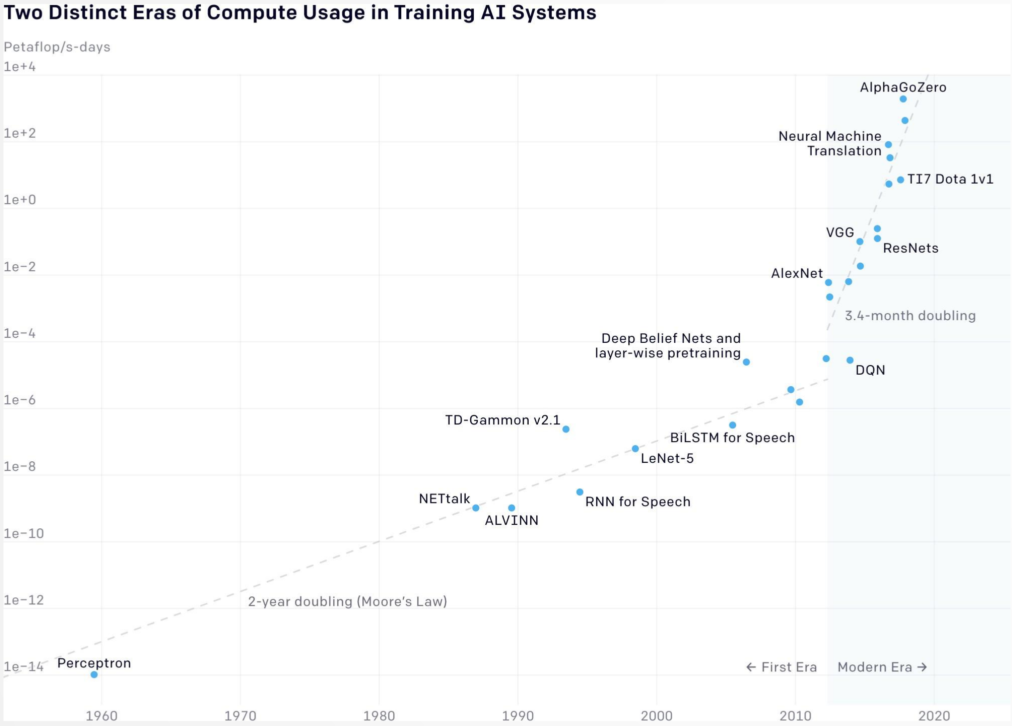
- Never discuss “interpretability” without clarifying the definition.
- Beware of interpretability fallacies.
- Find your domain-specific definition of interpretability, then use the tools available.
- Align evaluation metrics with what is qualitatively important

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





Costly data collection and computation (in time and money).



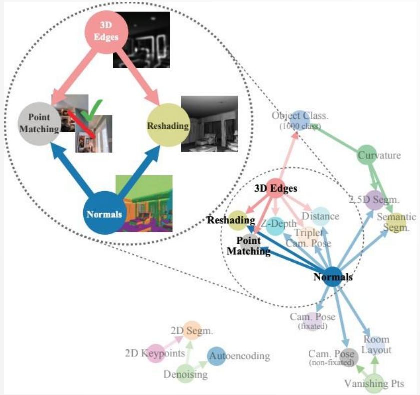
Since Deep Learning, compute use has been increasing faster than Moore's Law!

- 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].

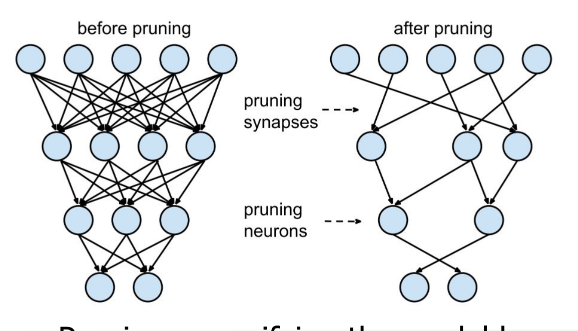
- Transfer learning [17, 19]. Pretrain on related tasks.
- Use public datasets,
- Download model parameters from internet.
- E.g. Many computer vision models use a backbone pretrained on ImageNet.



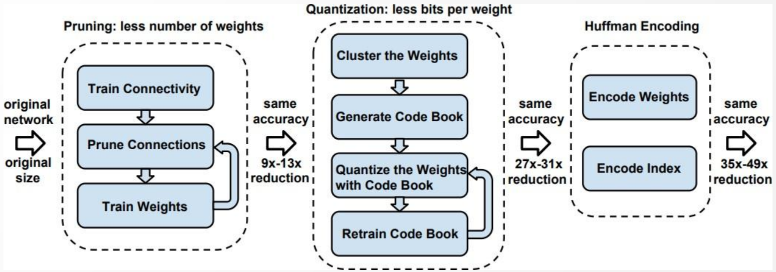
Recent work from Stanford researchers, Taskonomy [20], models the structure of space of visual tasks to guide transfer learning.

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.

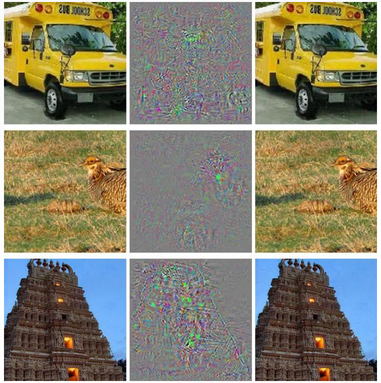


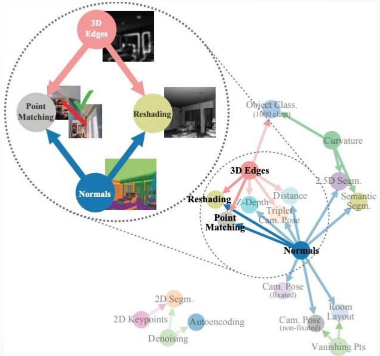
Pruning: sparsifying the model by removing unimportant weights



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

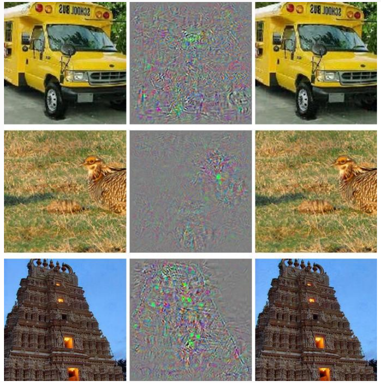


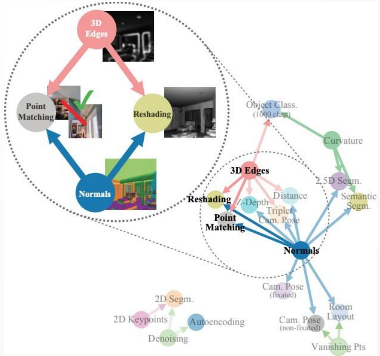


Community Weakness

- 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**

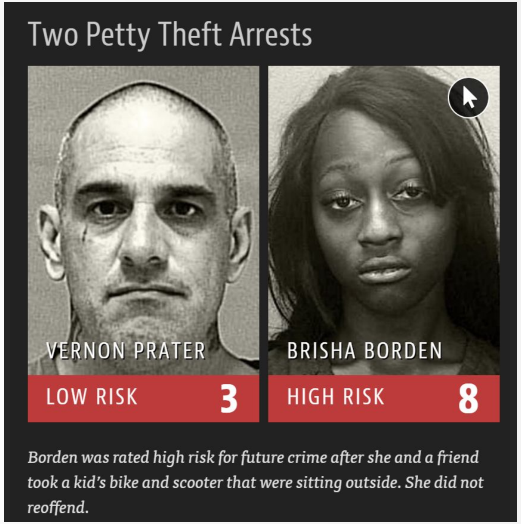


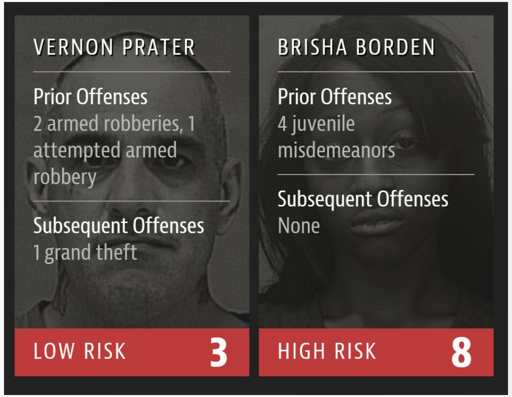


Ethical Concerns

- 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
    - AI 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?

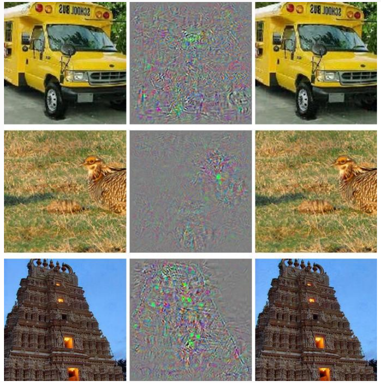
Need legal frameworks for holding AI users accountable

Ethical Concerns

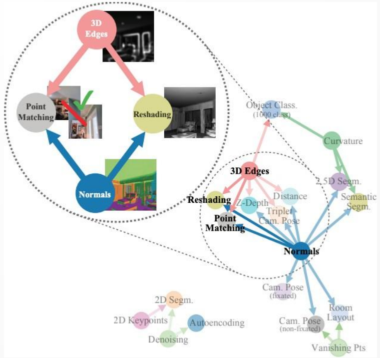
- ML captures social biases in datasets
- Like any technology, ML can be used in ways whose legality / ethics are questionable
- More discussions on privacy and security, e.g.
  - Large-scale web scraping for dataset collection
  - Ethical decisions: what should a autopilot car do in a trolley dilemma?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

1 Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus.
Intriguing properties of neural networks.
arXiv preprint arXiv:1312.6199, 2013.

2 Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy.
Explaining and harnessing adversarial examples.
arXiv preprint arXiv:1412.6572v3, 2015.

3 Yang Song, Rui Shu, Nate Kushman, and Stefano Ermon.
Generative adversarial examples.
arXiv preprint arXiv:1805.07894, 2018.

4 Geoffrey Hinton, Oriol Vinyals, and Jeff Dean.
Distilling the knowledge in a neural network.
arXiv preprint arXiv:1503.02531, 2015.

5 Nicolas Papernot, Patrick McDaniel, Xi Wu, Somesh Jha, and Ananthram Swami.
Distillation as a defense to adversarial perturbations against deep neural networks.
In *2016 IEEE Symposium on Security and Privacy (SP)*, pages 582–597. IEEE, 2016.

6 Yang Song, Taesup Kim, Sebastian Nowozin, Stefano Ermon, and Nate Kushman.
Pixeldefend: Leveraging generative models to understand and defend against adversarial examples.
arXiv preprint arXiv:1710.10766, 2017.

7 Anish Athalye and Ilya Sutskever.
Synthesizing robust adversarial examples.
arXiv preprint arXiv:1707.07397, 2017.

8 Zachary C Lipton.
The mythos of model interpretability.
arXiv preprint arXiv:1606.03490, 2016.

9 Laurens van der Maaten and Geoffrey Hinton.
Visualizing data using t-sne.
Journal of machine learning research, 9(Nov):2579–2605, 2008.

10 Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman.
Deep inside convolutional networks: Visualising image classification models and saliency maps.
arXiv preprint arXiv:1312.6034, 2013.

11 Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba.
Learning deep features for discriminative localization.
In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2921–2929, 2016.

12 Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, Dhruv Batra, et al.
Grad-cam: Visual explanations from deep networks via gradient-based localization.
In *ICCV*, pages 618–626, 2017.

13 Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. Show, attend and tell: Neural image caption generation with visual attention. In *International conference on machine learning*, pages 2048–2057, 2015.

14 Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, L- ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, pages 5998–6008, 2017.

15 Opinion — is harvard unfair to asian-americans? - the new york times. https://www.nytimes.com/2014/11/25/opinion/is-harvard-unfair-to-asian-americans.html?_r=0, 2014.

16 Honglak Lee, Peter Pham, Yan Largman, and Andrew Y Ng.
Unsupervised feature learning for audio classification using convolutional deep belief networks.
In *Advances in neural information processing systems*, pages 1096–1104, 2009.

17 Dumitru Erhan, Yoshua Bengio, Aaron Courville, Pierre-Antoine Manzagol, Pascal Vincent, and Samy Bengio.
Why does unsupervised pre-training help deep learning?
Journal of Machine Learning Research, 11(Feb):625–660, 2010.

18 Diederik P Kingma, Shakir Mohamed, Danilo Jimenez Rezende, and Max Welling.
Semi-supervised learning with deep generative models.
In *Advances in Neural Information Processing Systems*, pages 3581–3589, 2014.

19 Kaiming He, Ross Girshick, and Piotr Doll´ar.
Rethinking imagenet pretraining.
arXiv preprint arXiv:1811.08883, 2018.

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 Pattern Recognition*, pages 3712–3722, 2018.

21 Song Han, Huizi Mao, and William J Dally.
Deep compression: Compressing deep neural networks with pruning, trained quantization and huffman coding.
arXiv preprint arXiv:1510.00149, 2015.

22 Itay Hubara, Matthieu Courbariaux, Daniel Soudry, Ran El-Yaniv, and Yoshua Bengio.
Quantized neural networks: Training neural networks with low precision weights and activations.
The Journal of Machine Learning Research, 18(1):6869–6898, 2017.

23 Stephen D Brown, Robert J Francis, Jonathan Rose, and Zvonko G Vranesic.
Field-programmable gate arrays, volume 180.
Springer Science & Business Media, 2012.

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

25 Ian Goodfellow, Yoshua Bengio, Aaron Courville, and Yoshua Bengio. *Deep learning*, volume 1. MIT press Cambridge, 2016.

26 Zachary C Lipton and Jacob Steinhardt. Troubling trends in machine learning scholarship. *arXiv preprint arXiv:1807.03341*, 2018.

27 Theories of deep learning (stats 385). https://stats385.github.io/readings, 2017.

28 Ali Rahimi. 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).
arXiv preprint arXiv:1805.11604, 2018.