On Critiques of ML

CS 229, Fall 2018

Chris Chute

November 30, 2018

Alternate Title

Alternate Title

So you studied ML, and you'll soon see your relatives.
Now what?

Holiday gatherings

Holiday gatherings

Holiday gatherings

Figure: "So... I hear they disproved AI." – My (adversarial) uncle

Overview

1. Brittleness
2. Interpretability
3. Interpretability
4. Expense: Data and compute
5. Expense: Data and compute
6. Community weaknesses

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].
- Adversarial examples: [1, 2, 3, 4].
- Theory: [2].
- How to construct: [2, 5].
- How to defend: [1, 6, 7, 8].
- Future: Still an open problem. How fundamental?

Adversarial examples

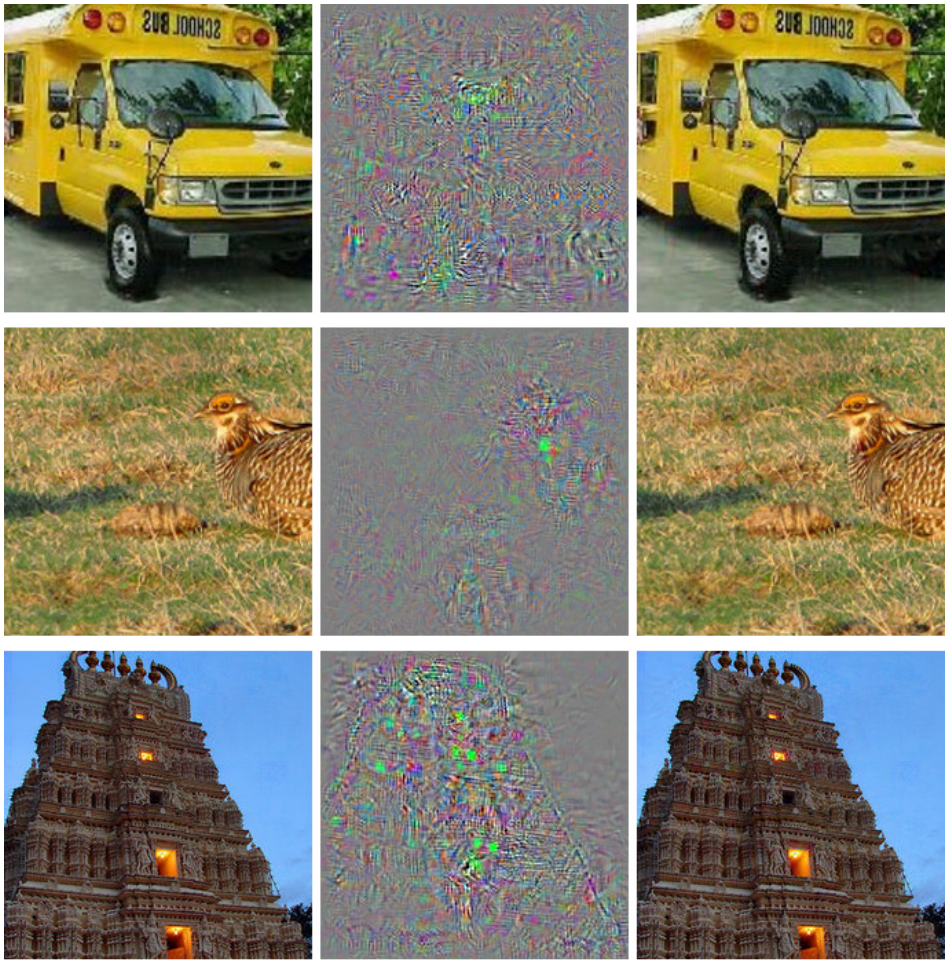

Figure: Left: Correctly classified image, center: perturbation, right: classified as *Ostrich.* Reproduced from [1].

Constructing adversarial examples

Fast gradient sign method [2]. Let θ be parameters, $\boldsymbol{x}$ input, y target, and $J(\theta, \boldsymbol{x}, y)$ cost.

Constructing adversarial examples

Fast gradient sign method [2]. Let $\boldsymbol{\theta}$ be parameters, $\boldsymbol{x}$ input, y target, and $J(\boldsymbol{\theta}, \boldsymbol{x}, y)$ cost. Then set $\tilde{x} := x + \boldsymbol{\eta}$ where

$$\boldsymbol{\eta} = \epsilon \cdot \text{sign}(\nabla_{\boldsymbol{x}} J(\boldsymbol{\theta}, \boldsymbol{x}, y)).$$

Constructing adversarial examples

Fast gradient sign method [2]. Let $\boldsymbol{\theta}$ be parameters, $\boldsymbol{x}$ input, y target, and $J(\boldsymbol{\theta}, \boldsymbol{x}, y)$ cost. Then set $\tilde{x} := x + \boldsymbol{\eta}$ where

$$\boldsymbol{\eta} = \epsilon \cdot \text{sign}(\nabla_{\boldsymbol{x}} J(\boldsymbol{\theta}, \boldsymbol{x}, y)).$$

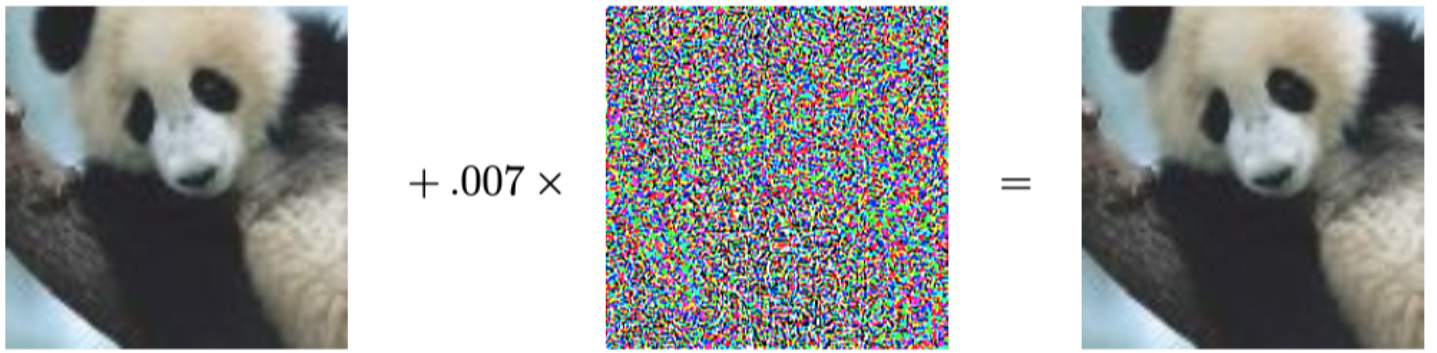

Figure: FGSM example, GoogLeNet trained on ImageNet, $\epsilon = .007$. Reproduced from [2].

Properties

- Change often indistinguishable to human eye.

Properties

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

Properties

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

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.

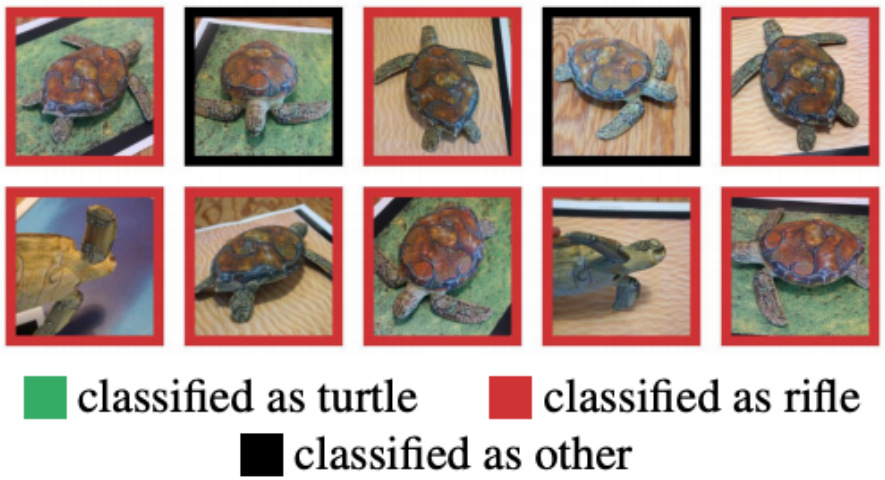

Figure: A turtle. Or is it a rifle? Reproduced from [4].

Defenses

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

Defenses

- Train on mixture of clean $\boldsymbol{x}$, perturbed $\tilde{\boldsymbol{x}}$ [1].
- Use **distillation** [6] as a defense [7]. *I.e.*, train second network to match high-temperature softmax activations of first one.

Defenses

- Train on mixture of clean $\boldsymbol{x}$, perturbed $\tilde{\boldsymbol{x}}$ [1].
- Use **distillation** [6] as a defense [7]. *I.e.*, train second network to match high-temperature softmax activations of first one.
- Many others [8]. But... [2] claims fundamental problem with linear models (and high-dimensional input):

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

- Continue to find new attacks that defeat previous defenses (*e.g.*, [5]).

Interpretability

- Switching gears: Interpretability.

Interpretability

- Switching gears: Interpretability.
- Desiderata for interpretability:

Interpretability

- Switching gears: Interpretability.
- Desiderata for interpretability:
  1. Trust: OK relinquishing control?

Interpretability

- Switching gears: Interpretability.
- Desiderata for interpretability:
  1. Trust: OK relinquishing control?
  2. Causality: Uncover causal relationships?

Interpretability

- Switching gears: Interpretability.
- Desiderata for interpretability:
  1. Trust: OK relinquishing control?
  2. Causality: Uncover causal relationships?
  3. Transferability: Works on other distributions?

Interpretability

- Switching gears: Interpretability.
- Desiderata for interpretability:
  1. Trust: OK relinquishing control?
  2. Causality: Uncover causal relationships?
  3. Transferability: Works on other distributions?
  4. Informativeness: How much info. do we get?

Interpretability

- Switching gears: Interpretability.
- Desiderata for interpretability:
  1. Trust: OK relinquishing control?
  2. Causality: Uncover causal relationships?
  3. Transferability: Works on other distributions?
  4. Informativeness: How much info. do we get?
  5. Fairness and ethics: Will real-world effect be fair?
- Many ideas from [9].

Interpretability

- Switching gears: Interpretability.
- Desiderata for interpretability:
  1. Trust: OK relinquishing control?
  2. Causality: Uncover causal relationships?
  3. Transferability: Works on other distributions?
  4. Informativeness: How much info. do we get?
  5. Fairness and ethics: Will real-world effect be fair?
- Many ideas from [9].

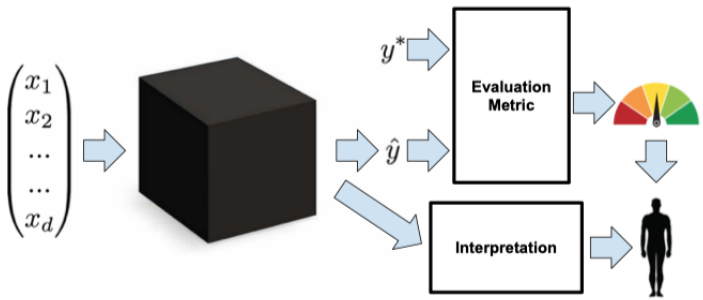

Figure: Reproduced from [9]. Main problem: Evaluation only requires y^*, $\hat{y}$. Often difficult to capture real-world costs (*e.g.*, ethics, legality).

Interpretability: Fallacies

- **Fallacy 1.** "Linear models are interpretable. Neural networks are black boxes."

Interpretability: Fallacies

- **Fallacy 1.** "Linear models are interpretable. Neural networks are black boxes."
- Any discussion of what is "interpretable" must fix a definition:

Interpretability: Fallacies

- **Fallacy 1.** "Linear models are interpretable. Neural networks are black boxes."
- Any discussion of what is "interpretable" must fix a definition:
  - Transparent: Simulatable, decomposable, understandable algorithm.

Interpretability: Fallacies

- **Fallacy 1.** "Linear models are interpretable. Neural networks are black boxes."
- Any discussion of what is "interpretable" must fix a definition:
  - Transparent: Simulatable, decomposable, understandable algorithm.
  - Post-hoc interpretation: Text, visualization, local explanation, explanation by example.

Interpretability: Fallacies

- **Fallacy 1.** "Linear models are interpretable. Neural networks are black boxes."
- Any discussion of what is "interpretable" must fix a definition:
  - Transparent: Simulatable, decomposable, understandable algorithm.
  - 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 1: Transparency

- **Simulatable.**

Interpretability Definition 1: Transparency

- **Simulatable.**
- **Decomposable.**

Interpretability Definition 1: Transparency

- **Simulatable.**
- **Decomposable.**
- **Understandable algorithm.**

Interpretability Definition 2: Post-hoc Explanation

- **Text.** *E.g.*, Auxiliary RNN to produce sentence.

Interpretability Definition 2: Post-hoc Explanation

- **Text.** *E.g.,* Auxiliary RNN to produce sentence.
- **Visualization.** E.g., render distributed representations in 2D with t-SNE [10].
- **Local explanation.** Popular: *e.g.,* Saliency Maps [11], CAMs [12], Grad-CAMs [13], attention [14, 15].

Interpretability Definition 2: Post-hoc Explanation

- **Text.** *E.g.,* Auxiliary RNN to produce sentence.
- **Visualization.** E.g., render distributed representations in 2D with t-SNE [10].
- **Local explanation.** Popular: *e.g.,* Saliency Maps [11], CAMs [12], Grad-CAMs [13], attention [14, 15].

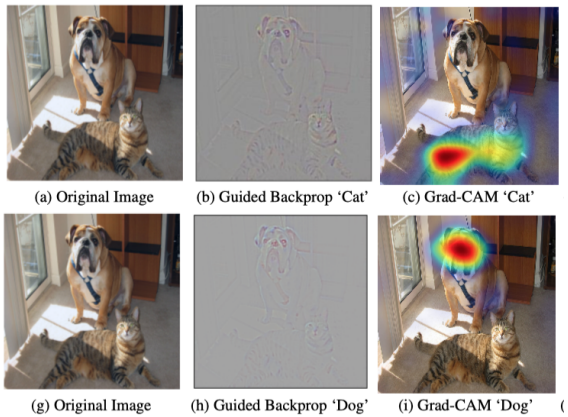

Figure: Grad-CAMs.

Interpretability Definition 2: Post-hoc Explanation

- **Text.** *E.g.,* Auxiliary RNN to produce sentence.
- **Visualization.** E.g., render distributed representations in 2D with t-SNE [10].
- **Local explanation.** Popular: *e.g.,* Saliency Maps [11], CAMs [12], Grad-CAMs [13], attention [14, 15].

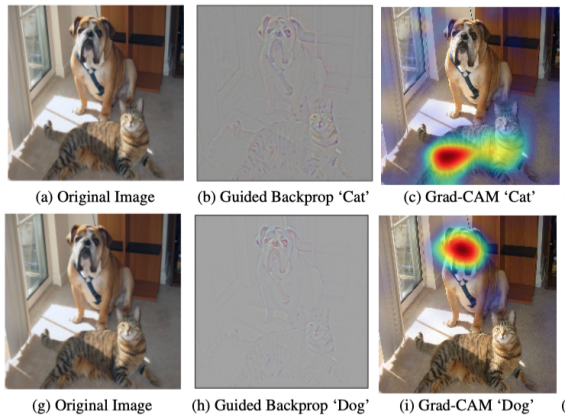

Figure: Grad-CAMs.

- **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?

Interpretability: Fallacies

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

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

- Full transparency can preclude models that surpass our ability on complex tasks.

Interpretability: Fallacies

- **Fallacy 3.** Always trust post-hoc explanation (*e.g.*, CAMs).

Interpretability: Fallacies

- **Fallacy 3.** Always trust post-hoc explanation (*e.g.*, CAMs).
- Post-hoc interpretations can be 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 [16].

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.
- Try to solve the core problem: Align loss with downstream task. *E.g.*, segmentation over classification.

Expense: Data and compute

Switching gears: ML can be expensive.

Expense: Data

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

Expense: Data

- Costly data collection and computation (in time and money).
- Solution 1: Unsupervised [17, 18] and semi-supervised approaches [19].

Expense: Data

- Case study: Unsupervised pre-training [18].

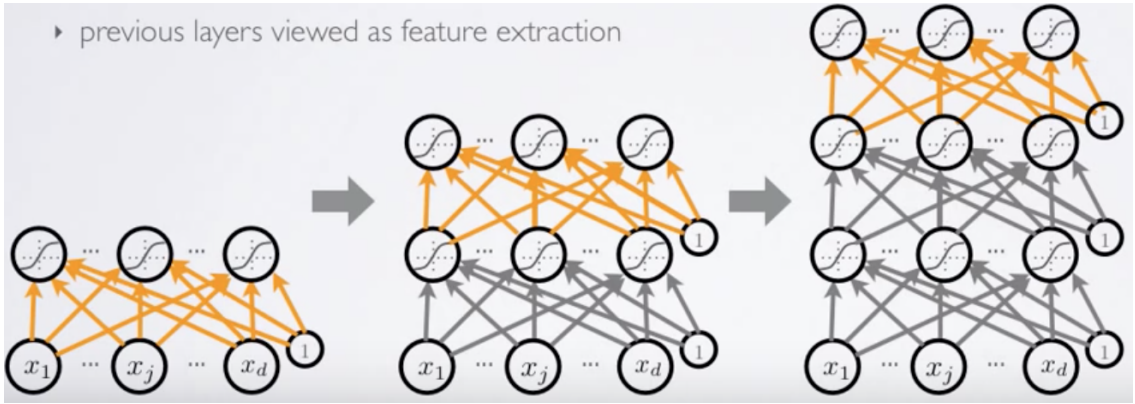

Figure: Layer-wise unsupervised pre-training. Author: Hugo Larochelle.

Expense: Data

- Case study: Data distillation [20].

Expense: Data

- Case study: Data distillation [20].

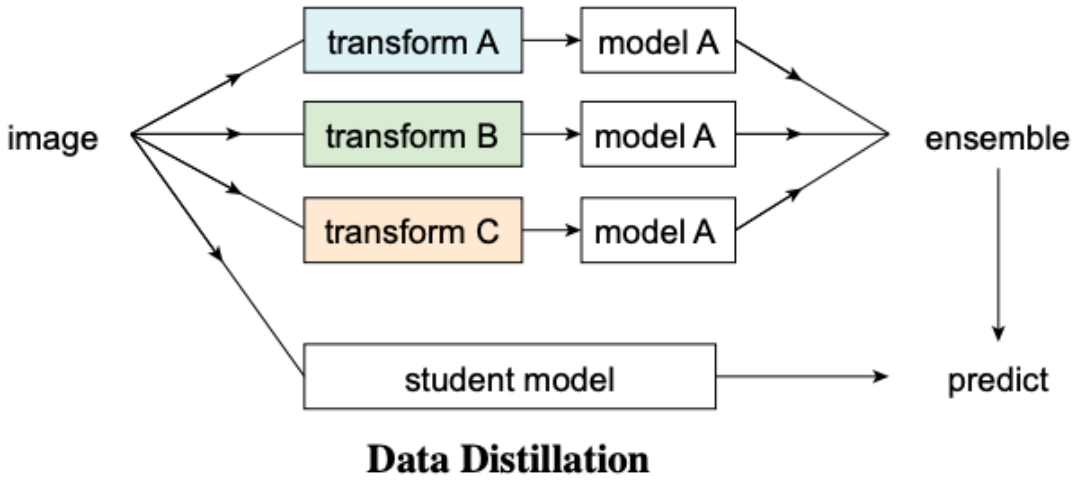

Figure: Expanding your training set with data distillation.

Expense: Data

- Transfer learning [18, 21]. Pretrain on related tasks.

Expense: Data

- Transfer learning [18, 21]. Pretrain on related tasks.
- Use public datasets, e.g., ImageNet.

Expense: Data

- Transfer learning [18, 21]. Pretrain on related tasks.
- Use public datasets, e.g., ImageNet.
- Download model parameters from internet.

Expense: Data

- Transfer learning [18, 21]. Pretrain on related tasks.
- Use public datasets, e.g., ImageNet.
- Download model parameters from internet.
- Recent work from Stanford researchers: Taskonomy [22].

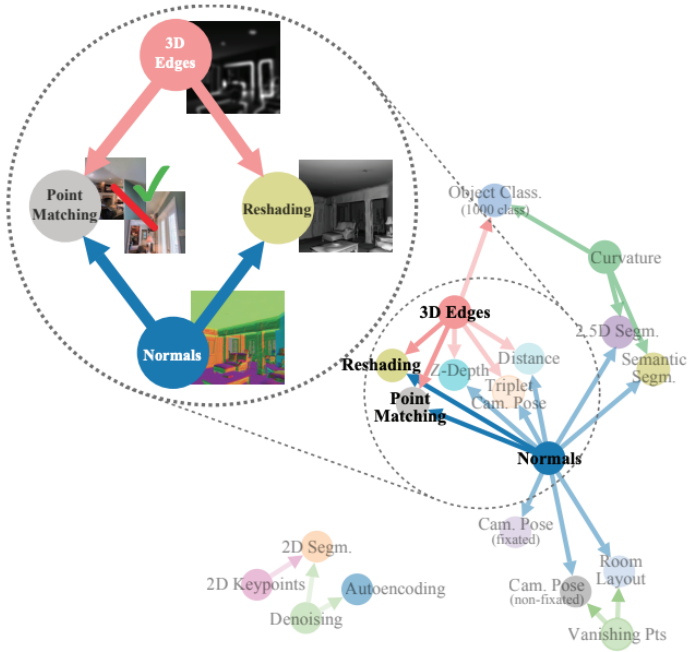

Figure: Taskonomy: "taxonomy of tasks" to guide transfer learning.

Expense: Compute

- Compression [23].

Expense: Compute

- Compression [23].
- Quantization [24]. Why use `float32` for all your weights?

Expense: Compute

- Compression [23].
- Quantization [24]. Why use `float32` for all your weights?
- Specialized hardware [25, 26]. GPUs are inefficient. More efficiency with FPGA, TPU.

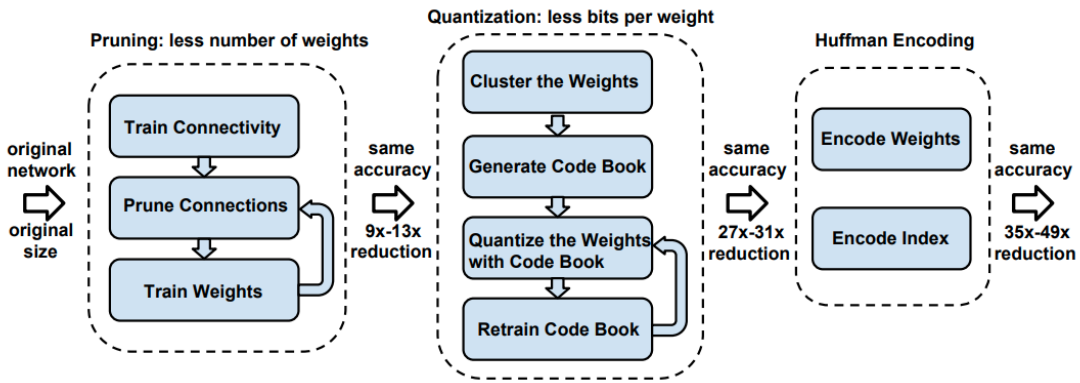

Figure: Deep compression: Pruning, quantization, and Huffman coding. 50× gains.

Expense: Compute

- Efficient models [27, 28].
- Knowledge distillation [6, 29].

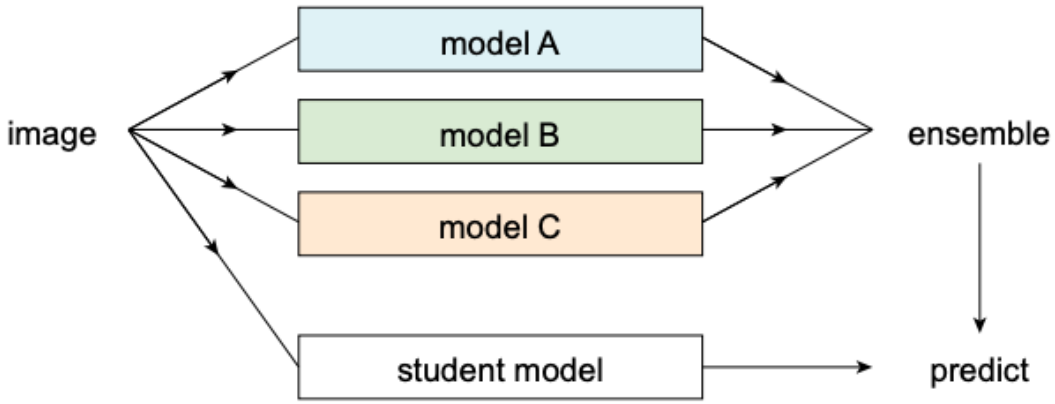

Figure: Knowledge distillation.

Summary: Expense

- Data: Transfer learning, public datasets, unsupervised pretraining. Newer technizues coming out frequently.
- Compute: Compression, quantization, specialized hardware.

Community weaknesses

- Cycle of hype and winter [30].

Community weaknesses

- Cycle of hype and winter [30].
- Lack of rigor and worries of troubling scholarship trends [31, 32].
  - Many incorrect theories invented to explain observations, rather than derived from theoretical foundations [33, 34].
  - Suggestion of [33]: Spend more time doing experiments to find root cause for unexpected results, rather than chasing performance.

Community weaknesses

- Cycle of hype and winter [30].
- Lack of rigor and worries of troubling scholarship trends [31, 32].
  - Many incorrect theories invented to explain observations, rather than derived from theoretical foundations [33, 34].
  - Suggestion of [33]: Spend more time doing experiments to find root cause for unexpected results, rather than chasing performance.
- Lack of equal representation. Example efforts to counteract: [35, 36].

Community weaknesses

- Cycle of hype and winter [30].
- Lack of rigor and worries of troubling scholarship trends [31, 32].
  - Many incorrect theories invented to explain observations, rather than derived from theoretical foundations [33, 34].
  - Suggestion of [33]: Spend more time doing experiments to find root cause for unexpected results, rather than chasing performance.
- Lack of equal representation. Example efforts to counteract: [35, 36].
- Barriers to entry (funding and data).

Conclusion

1. Brittleness
2. Interpretability
3. Interpretability
4. Expense: Data and compute
5. Expense: Data and compute
6. Community weaknesses

"Max Planck said, 'Science progresses one funeral at a time.' The future depends on some graduate student who is deeply suspicious of everything I have said." —Geoff Hinton [37]

References I

[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] Robin Jia and Percy Liang.
Adversarial examples for evaluating reading comprehension systems.
arXiv preprint arXiv:1707.07328, 2017.

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

References II

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

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

[7] 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.

References III

[8] 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.

[9] Zachary C Lipton.
The mythos of model interpretability.
arXiv preprint arXiv:1606.03490, 2016.

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

References IV

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

[12] 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.

[13] 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.

References V

[14] 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.

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

[16] 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.

References VI

[17] 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.

[18] 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.

[19] 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.

References VII

[20] Ilija Radosavovic, Piotr Dollár, Ross Girshick, Georgia Gkioxari, and Kaiming He.
Data distillation: Towards omni-supervised learning.
arXiv preprint arXiv:1712.04440, 2017.

[21] Kaiming He, Ross Girshick, and Piotr Dollr.
Rethinking imagenet pretraining.
arXiv preprint arXiv:1811.08883, 2018.

[22] 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.

References VIII

[23] 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.

[24] 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.

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

References IX

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

[27] Forrest N Iandola, Song Han, Matthew W Moskewicz, Khalid Ashraf, William J Dally, and Kurt Keutzer.
Squeezenet: Alexnet-level accuracy with 50x fewer parameters and¡ 0.5 mb model size.
arXiv preprint arXiv:1602.07360, 2016.

[28] Ningning Ma, Xiangyu Zhang, Hai-Tao Zheng, and Jian Sun.
Shufflenet v2: Practical guidelines for efficient cnn architecture design.
arXiv preprint arXiv:1807.11164, 2018.

References X

[29] Aaron van den Oord, Yazhe Li, Igor Babuschkin, Karen Simonyan, Oriol Vinyals, Koray Kavukcuoglu, George van den Driessche, Edward Lockhart, Luis C Cobo, Florian Stimberg, et al.
Parallel wavenet: Fast high-fidelity speech synthesis.
arXiv preprint arXiv:1711.10433, 2017.

[30] Ian Goodfellow, Yoshua Bengio, Aaron Courville, and Yoshua Bengio.
Deep learning, volume 1.
MIT press Cambridge, 2016.

[31] Zachary C Lipton and Jacob Steinhardt.
Troubling trends in machine learning scholarship.
arXiv preprint arXiv:1807.03341, 2018.

[32] Theories of deep learning (stats 385).
`https://stats385.github.io/readings`, 2017.

References XI

[33] Ali Rahimi.
Ai is the new alchemy (nips 2017 talk).
https://www.youtube.com/watch?v=Qi1Yry33TQE, December 2017.

[34] 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.

[35] Black in ai.
https://blackinai.github.io/.

[36] Home - wimlds.
http://wimlds.org/.

References XII

[37] Steve LeVine.
Artificial intelligence pioneer says we need to start over, Sep 2017.