Machine Learning Systems Design

Lecture 12: Machine learning beyond accuracy

Stanford University

CS 329 | Chip Huyen

Sara Hooker
Research scientist

TESLA

Andrej Karpathy
Director of AI

Model Deployment Beyond Test Set Accuracy

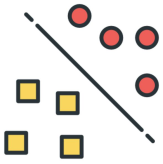
Stanford CS329

Sara Hooker
Google Brain

Google Research

My research agenda to-date has focused on:

- Going beyond test-set accuracy
- Training models that fulfill multiple desired criteria



Model Compression - compact machine learning models to work in resource constrained environments.

Fairness - imposes constraint on optimization that reflects societal norms of what is fair.

Model fragility and security - deploy secure models that protect user privacy.

Model Interpretability - reliable explanations for model behavior.

Model Deployment Beyond Test Set Accuracy

Accuracy without "true" learning.

The myth of the robust, interpretable, compact, fair, high test-set accuracy model.

Interpretability Tools

I'll mention research collaborations with my colleagues:
Nyalleng Moorosi, Gregory Clark, Samy Bengio, Emily Denton, Aaron Courville, Yann Dauphin, Andrea Frome, Chirag Agarwal, Daniel Souza, Dumitru Erhan.

Accuracy without “true” learning.

Google™

The Clever Hans Effect 1891 - 1907

Hans the horse:

- arithmetic functions
- identify colours
- Count the crowd

Myth of Clever Hans persisted 1891 - 1907

Experimental Design -
Can Hans answer a question if the human does not know the answer?

Hans answered correctly by picking up on microscopic clues.

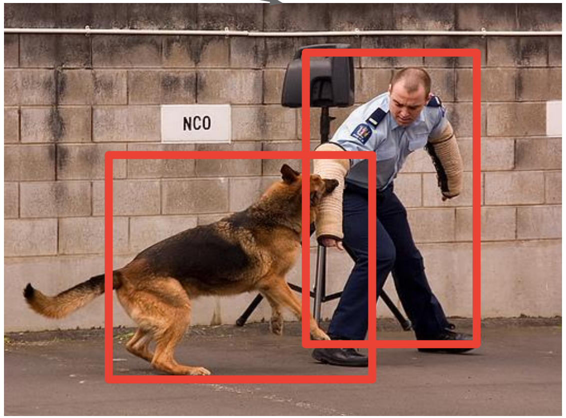
High accuracy without “true” learning.

Deep Neural Networks have resulted in a huge leap forward in top-line performance on image classification tasks.

Computer vision tasks

Image Classification

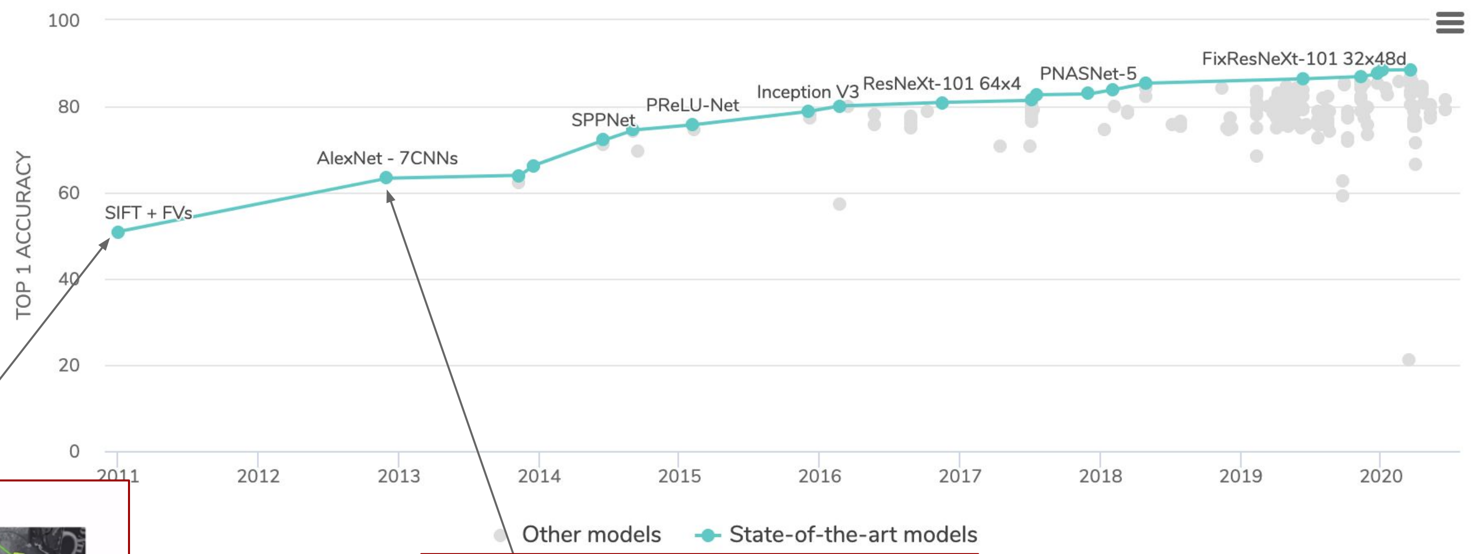
Object localization

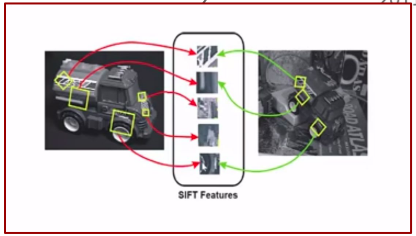


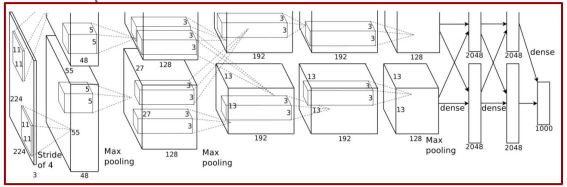
Object recognition

Performance on ImageNet

Image Classification on ImageNet



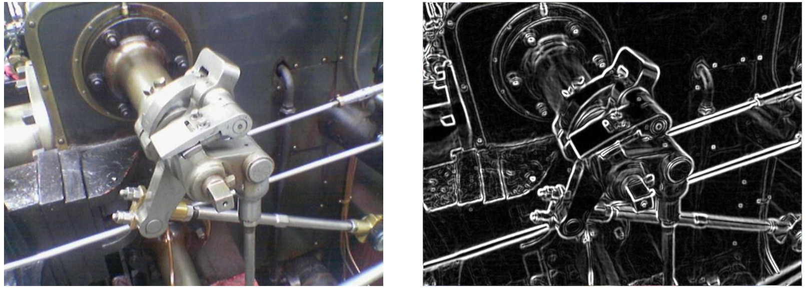




Google

Before 2012 - Hand engineered encoders were very interpretable but had difficulty generalizing well beyond a few narrow tasks.

Sobel Edge Filter



Hog Filter

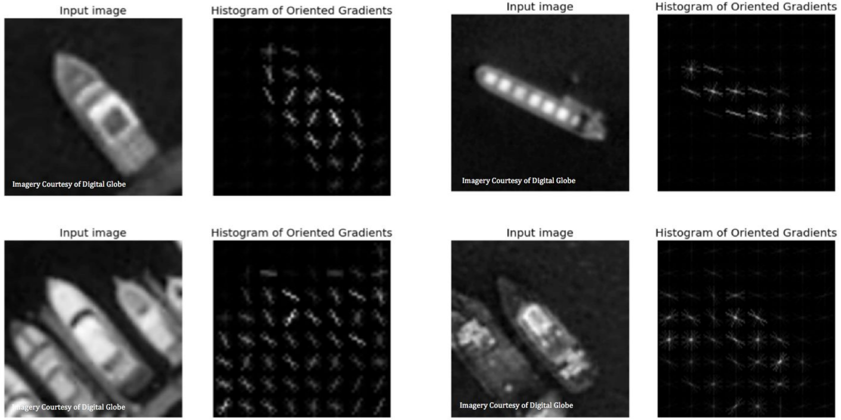
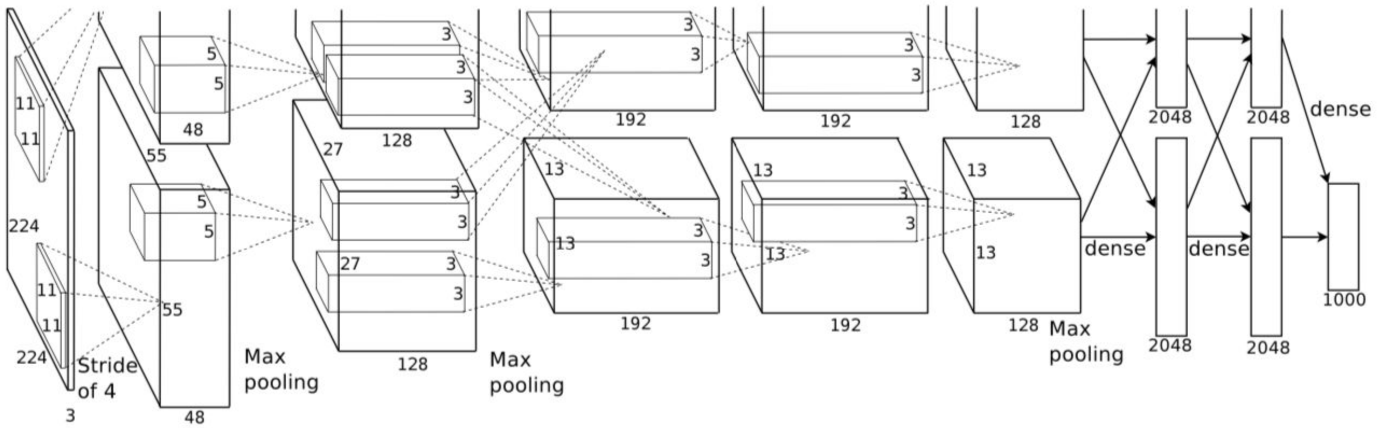


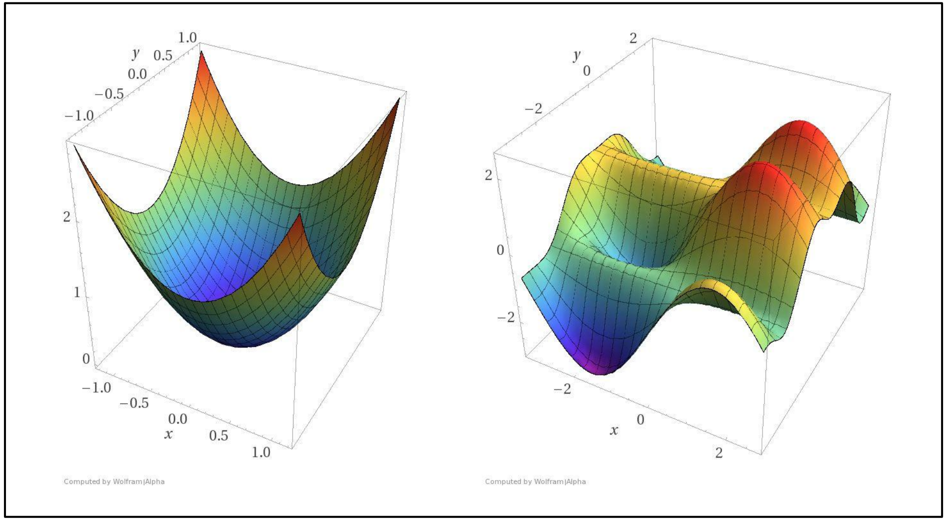
Image source: Digital Globe, Wikipedia

2012 - Enter convolutional neural networks:
- AlexNet swept the competition on ImageNet using CNNs. Error rate of 16.4% (runner up was at 26.2%).

Huge advantage = CNN's have dominated ever since.



Instead of telling the model what features to extract, the model learns what features are important for the task through feedback (minimizing the loss through backpropagation).



We give up full specification of the function - “black-box” because difficult to specify why model made a certain prediction.

Delegating learning of the function to the model can (and has) led to Clever Hans moments.

Cow

Limousine

Berry et al. ([paper link](#))
Hooker et al. 2019 ([paper link](#))

High accuracy without “true” learning.

Delegating learning of the function to the model can (and has) led to Clever Hans moments.

Sheep

A herd of sheep grazing on a lush green hillside
Tags: grazing, sheep, mountain, cattle, horse

Dog

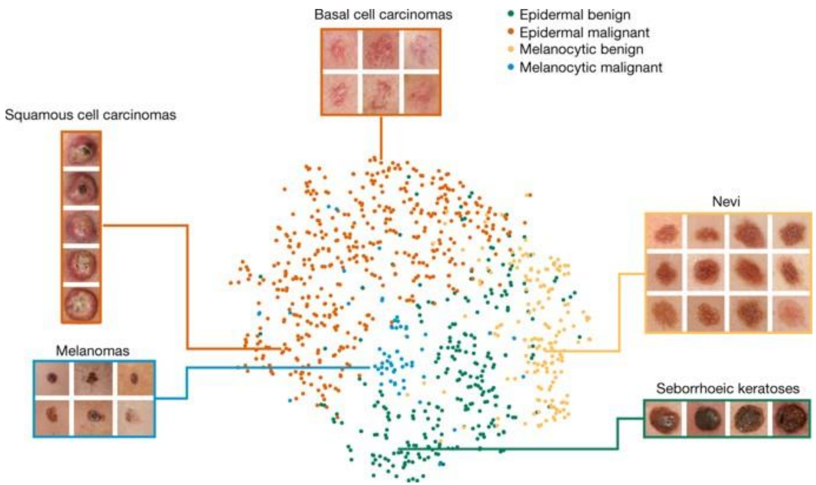
Left: A man is holding a dog in his hand
Right: A woman is holding a dog in her hand
Image: @SouperSarah

High accuracy without “true” learning.

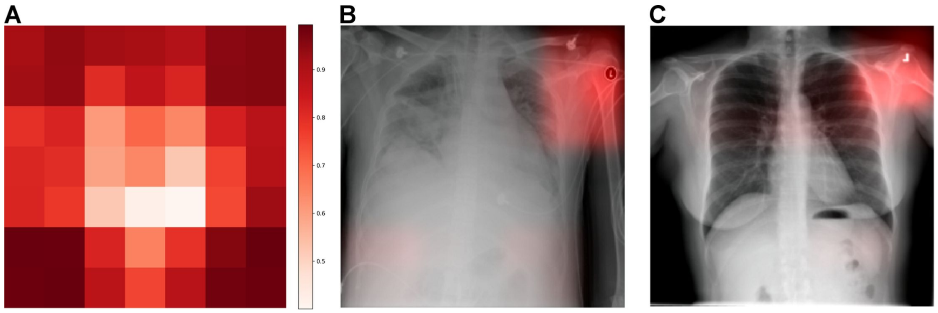
Blog [link]

When Cleverhans moments happen in sensitive domains, there can be a huge cost to human welfare.

Skin lesions



Pneumonia



High accuracy without "true" learning.

Esteva et al. (link)
Zech et al. 2018 (link)
AlBadaway et al. 2018 (link)

Top line metrics often hide critical model behavior.

In deployment settings, necessary to go beyond top-1, top-5 to ensure desirable model behavior.

How **does** my model perform...

Classification accuracy / precision-recall curve / logarithmic loss / area under the curve / mean squared error / mean absolute error / F1 score / standard deviation / variance / confidence intervals / KL divergence / false positive rate / false negative rate / <insert **metric** here>

How **might** my model perform...

on a sample of test data / on cross-slices of test data / on an individual data point / if a datapoint is perturbed / if model thresholds were different / if optimized differently / across all values of a feature / when compared to a different model / on different data points within a neighborhood of data points / <insert **question** here>

Google

Test-set accuracy does not guarantee that the trained function fulfills other properties we may care about.

$$loss = \sum_{i=1}^{B} \mathcal{L}(y_i, \hat{y}_i)$$

test-set accuracy - extract a representation for the task that is generalizable to unseen data.

Compactness

Interpretability

Fairness

Robustness

Typical loss functions in machine learning (MSE, Hinge-Loss and CE) impose no preference for functions that are compact, interpretability, fairness and robust.

$$loss = \sum_{i=1}^{B} \mathcal{L}(y_i, \hat{y}_i)$$

Compactness

Interpretability

Fairness

Adversarial Robustness

Deployment models to fulfill multiple desiderata.

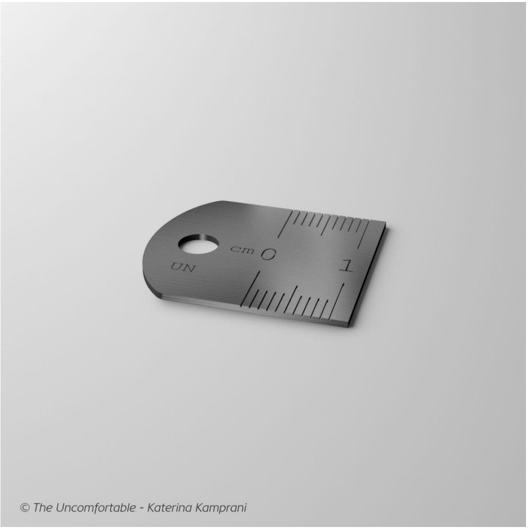
test-set accuracy - extract a representation for the task that is generalizable to unseen data.

Model Compression	Cheap - fast to evaluate Compact - minimal memory
Interpretability	Understandable - Model function performance meaningful to humans.
Adversarial Robustness	Not vulnerable to non-meaningful changes in data distribution.
Fairness	Reflect preferences about how model should behave on subsets of protected features.

Training Models to Fulfill Multiple Desiderata

Chapter 1: Fairness

Google



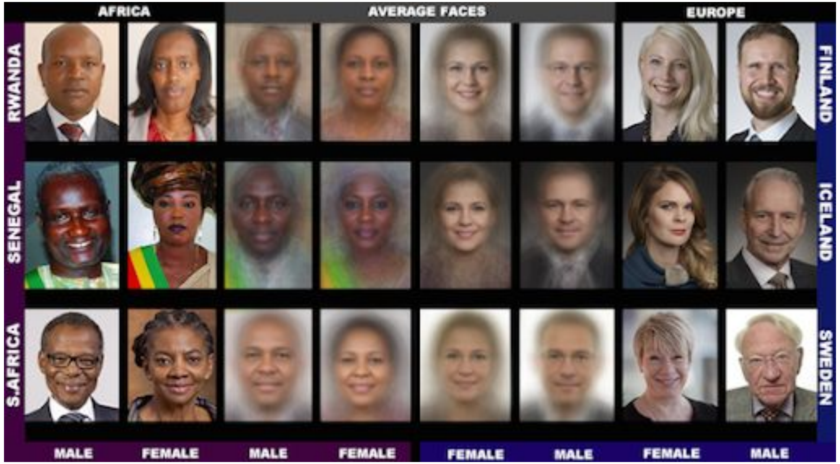
[Katerina Kamprani]

What if discomfort is not uniform, but targeted?

Google

[Further reading]

Algorithmic bias - errors that create unfair outcomes.



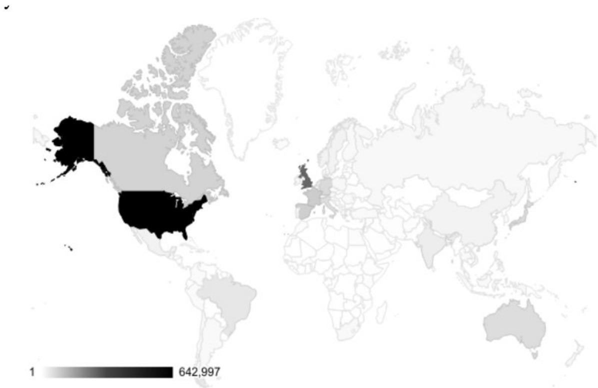


Figure 2: Distribution of the geographically identifiable images in the Open Images data set, by country. Almost a third of the data in our sample was US-based, and 60% of the data was from the six most represented countries across North America and Europe.

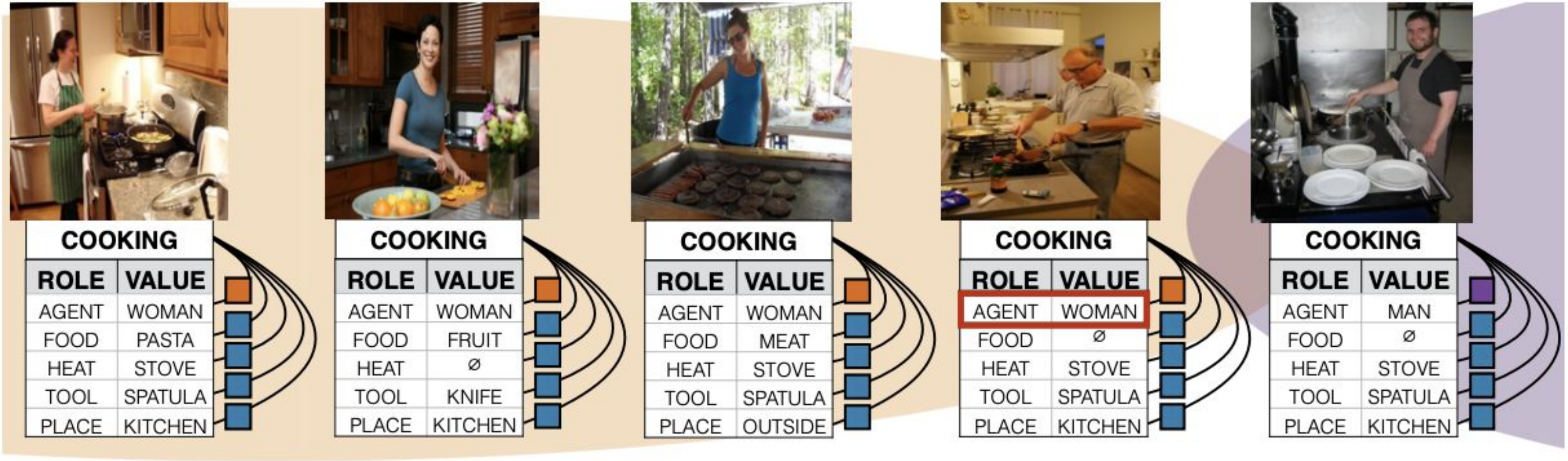
Gender shades (link)
Shankar et al. (link)

How a model treats underrepresented features often coincide with notions of fairness.

Geographic bias in how we collect our datasets. Shankar et al. (2017) show models perform far worse on locales undersampled in the training set.

No Classification without Representation: Assessing Geodiversity Issues in Open Data Sets for the Developing World (Shankar et al. (link))

Undersampling/oversampling leads to undesirable spurious correlations. Zhao, Jieyu et al. (2017) show Activity recognition datasets exhibit stereotype-aligned gender biases.



Men also like shopping (and cooking too).

Zhao, Jieyu et al. (2017).

Fairness

```
Preferences about how our trained model should
behave on subset of sensitive or protected
features.
```

Legally protected features:
Certain attributes are protected by law. For example, in the US it is illegal to discriminate based upon race, color, religion, sex, national origin, disability.

Legal framework will differ by country.

Sensitive features:
Income, eye color, hair, skin color, accent, locale.

These features may not be protected by law, but are often correlated with protected attributes .

Your choice of tool to audit and mitigate algorithmic bias will depend upon whether you know:

- the sensitive features which are adversely impacted
- have comprehensive labels for these features

- Unknown bias
- Incomplete or no labels for sensitive features

- Known concern
- Comprehensive labels

Your choice of tool to audit and mitigate algorithmic bias will depend upon whether you know:

- the sensitive features which are adversely impacted
- have comprehensive labels for these features

1.

- Unknown bias
- Incomplete or no labels for sensitive features

- Known concern
- Comprehensive labels

1. With known and comprehensive labels - track impact using intersectional metrics

What is it?
Statistically evaluate model performance (e.g. accuracy, error rates) by "subgroup"
e.g. skin tone, gender, age

Requires
Good, "balanced" test sets that are representative of the actual use-case(s) for the model in production

Fitzpatrick Skin Type	Male	Female	Non-binary
Type I			
Type II			
Type III			
Type IV			
Type V			
Type VI			

Acc/FNP/FPR/other

Example of intersectional audit

Gender Shades - Evaluated classifiers' performance across genders, skin types, and intersection of gender and skin type

Classifier	Metric	All	F	M	Darker	Lighter	DF	DM	LF	LM
MSFT	PPV(%)	93.7	89.3	97.4	87.1	99.3	79.2	94.0	98.3	**100**
	Error Rate(%)	6.3	10.7	2.6	12.9	0.7	**20.8**	6.0	1.7	0.0
	TPR (%)	93.7	96.5	91.7	87.1	99.3	92.1	83.7	**100**	98.7
	FPR (%)	6.3	8.3	3.5	12.9	0.7	**16.3**	7.9	1.3	0.0
Face++	PPV(%)	90.0	78.7	99.3	83.5	95.3	65.5	**99.3**	94.0	99.2
	Error Rate(%)	10.0	21.3	0.7	16.5	4.7	**34.5**	0.7	6.0	0.8
	TPR (%)	90.0	98.9	85.1	83.5	95.3	98.8	76.6	**98.9**	92.9
	FPR (%)	10.0	14.9	1.1	16.5	4.7	**23.4**	1.2	7.1	1.1
IBM	PPV(%)	87.9	79.7	94.4	77.6	96.8	65.3	88.0	92.9	**99.7**
	Error Rate(%)	12.1	20.3	5.6	22.4	3.2	**34.7**	12.0	7.1	0.3
	TPR (%)	87.9	92.1	85.2	77.6	96.8	82.3	74.8	**99.6**	94.8
	FPR (%)	12.1	14.8	7.9	22.4	3.2	**25.2**	17.7	5.20	0.4

Table 4: Gender classification performance as measured by the positive predictive value (PPV), error rate (1-PPV), true positive rate (TPR), and false positive rate (FPR) of the 3 evaluated commercial classifiers on the PPB dataset. All classifiers have the highest error rates for darker-skinned females (ranging from 20.8% for Microsoft to 34.7% for IBM).

Paper

When labels are known and complete - opens up range of remedies to mitigate impact

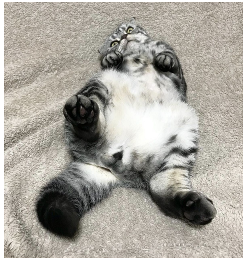
Data-Based

1. Re-balance or re-weight sensitive features to balance training set.

2. Remove problematic feature from training set (**not always feasible**)

Even with comprehensive labels removing or modifying problematic feature from training set is **not always feasible**





Toy Task: Sleeping or awake?

If *species* is a protected attribute, how do modify the dataset to remove it.

There may also be cases where removing a protected or sensitive feature degrades model performance on that subset.

Toy Task: Sleeping or awake?

If *species* is a protected attribute, how do modify the dataset to remove it.

However, complete labels give us much more freedom and control in modifying the training set by re-balancing/re-weighting.

When labels are known and complete - range of remedies to mitigate impact.

Data-Based

1. Re-balance or re-weight sensitive features to balance training set.

2. Remove problematic feature from training set (**not always feasible due to proxy variables**)

Model-Based

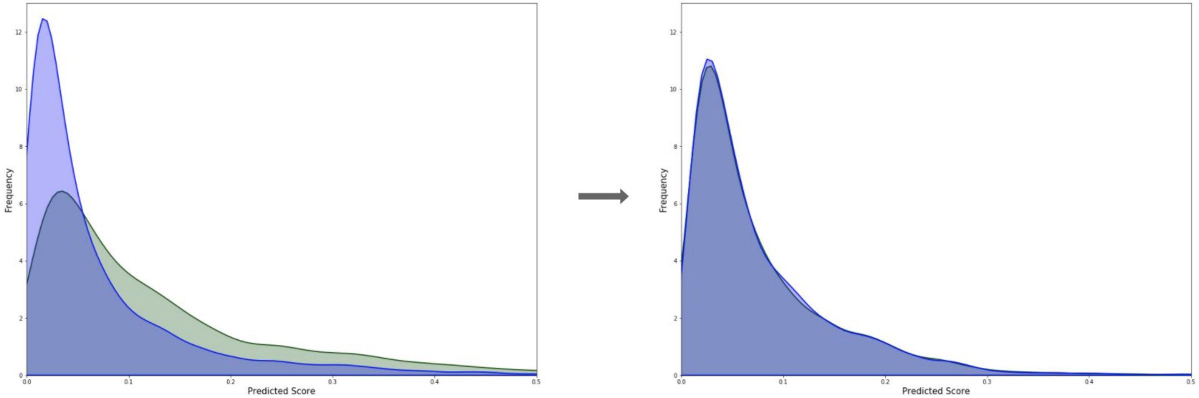
1. Min diff - penalizes model for differences in treatment of distributions

2. Rate constraint - guaranteeing recall or another rate metric is at least [x%] on a subset.

Growing software support for training with constraints.

How does MinDiff work?

Given two sets of examples from our dataset, MinDiff penalizes the model during training for differences in the distribution of scores between the two sets. The less distinguishable the two sets are based on prediction scores, the smaller the penalty that will be applied.



What about where we don’t have complete labels for the sensitive attribute we care about?

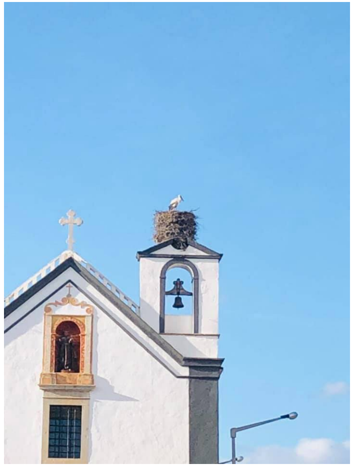
2.

- Unknown bias
  Incomplete or no labels for sensitive features

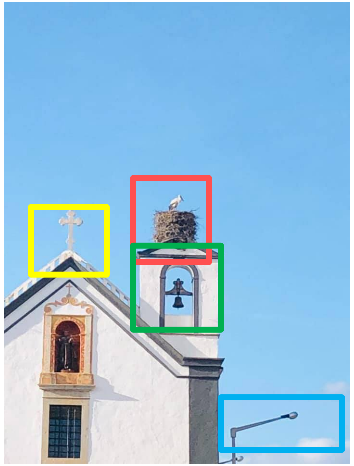
- Known concern
- Comprehensive labels

For high dimensional datasets:

- Labelling becomes expensive at scale, very difficult to do comprehensive labelling.



church



Bird, nest, street lamp, cross, statue, window, window grid.

Google

For high dimensional datasets:

- Hard to label all proxy variables that correspond with sensitive feature

Task: Sleeping or awake?

While *species* is the protected attribute, many other variables may be proxy variables (indoor/outdoor background).

For high dimensional datasets:

- Labelling becomes expensive at scale, very difficult to do comprehensive labelling.
- Hard to label all proxy variables that correspond with sensitive feature.

Additional difficulties in data collection:

- There may be legal obstacles/additional sensitivity around collecting labels on protected identities like race or gender.

In the absence of labelled data, auditing tools play an important role in surfacing what most needs human auditing.

Surfaces a tractable subset of the most challenging/least challenging examples for human inspection. Avoids time consuming need to inspect every example.

Data Cleaning

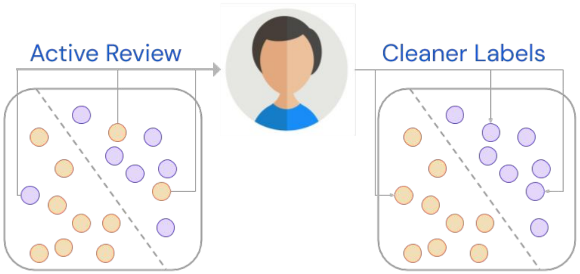
Isolating subset for relabelling

Identify issues with fairness

Global feature importance - Ranks dataset examples by which are most challenging.

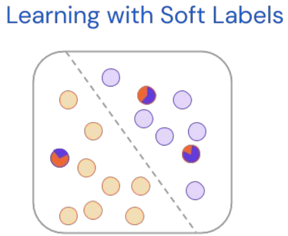
1

Use it to clean/audit the dataset



2

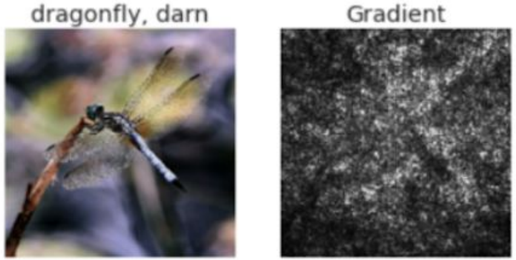
Use it to improve training.



Variance of Gradients (VoG) is an example of a global ranking tool.

$$VOG_i = \frac{1}{N}\sqrt{\left(\frac{1}{N_t}\sum_{t=1}^{N_t}(S_{ti} - \mu_i)^2\right)}$$

gradient



Compute average variance in gradients (VOG) for an image over training.

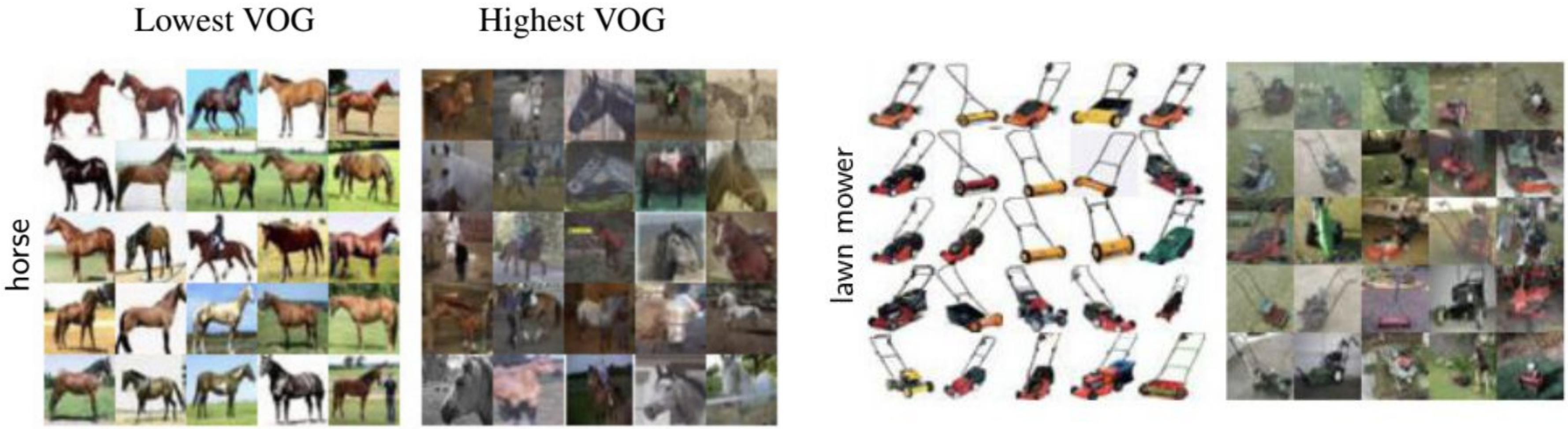
0 epochs

90 epochs

Estimating Example Difficulty using Variance of Gradients, Agarwal, Souza and Hooker, 2020

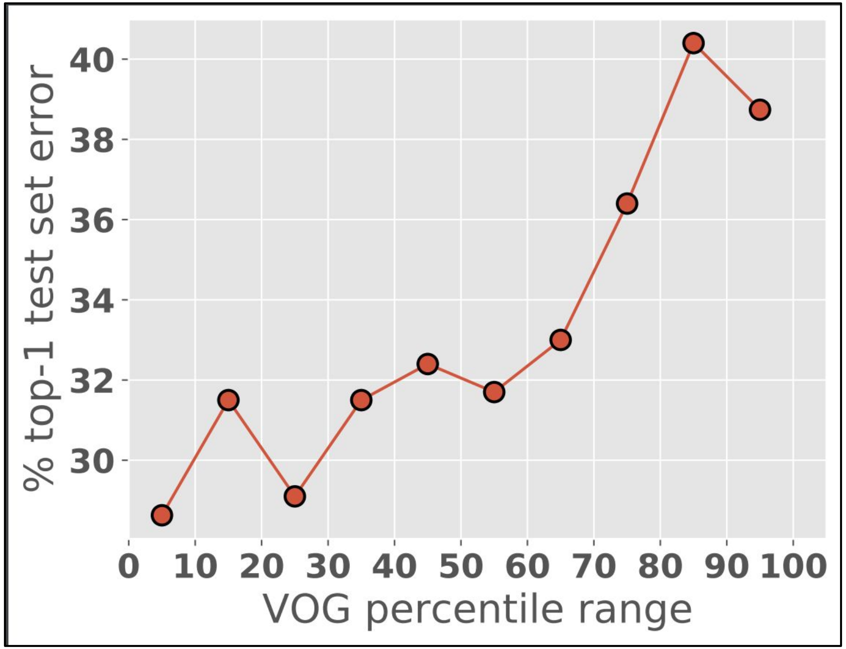
VoG computes a relative ranking of each class.

What examples does the model find challenging or easy to learn?

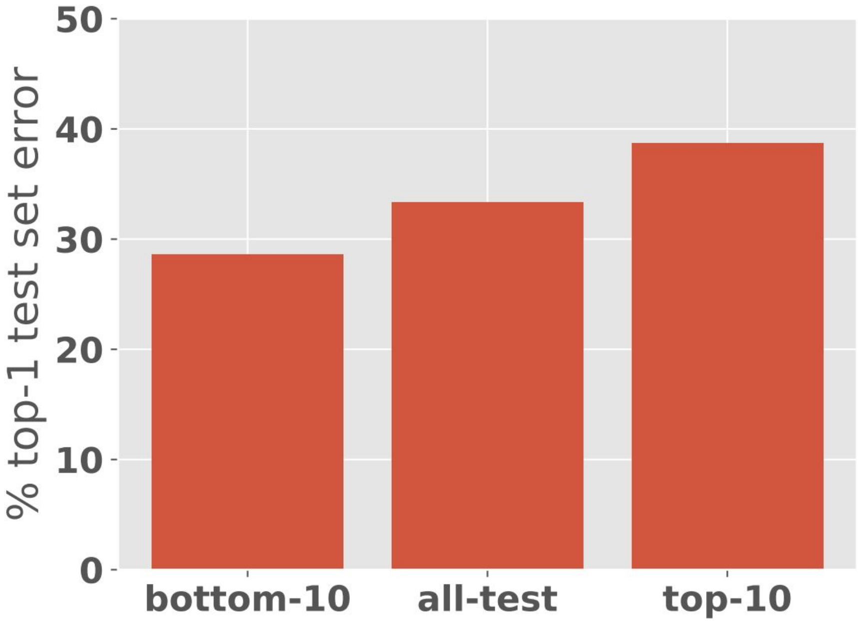


Estimating Example Difficulty using Variance of Gradients, Agarwal, Souza and Hooker, 2020

VOG effectively discriminates between easy & challenging examples.



CIFAR-100
(Across all percentiles)



CIFAR-100
<10th, all, >90th percentile

Understand how feature importance forms over the course of training.

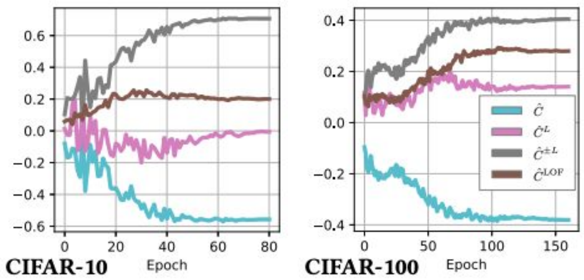
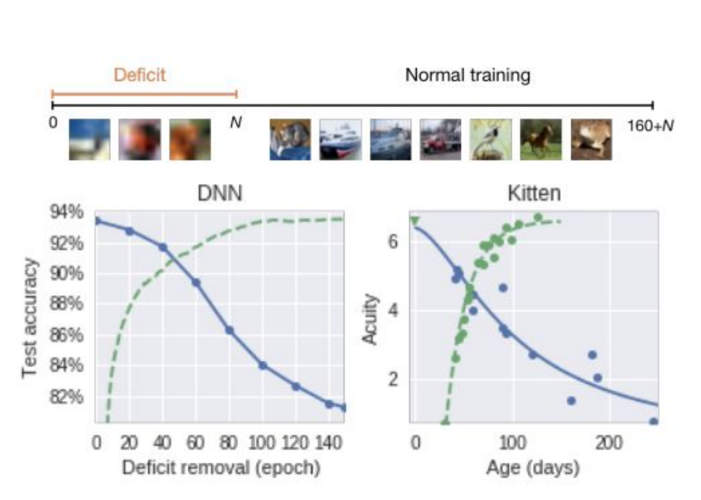


Figure 6: Spearman rank correlation between C-score and distance-based score on hidden representations.

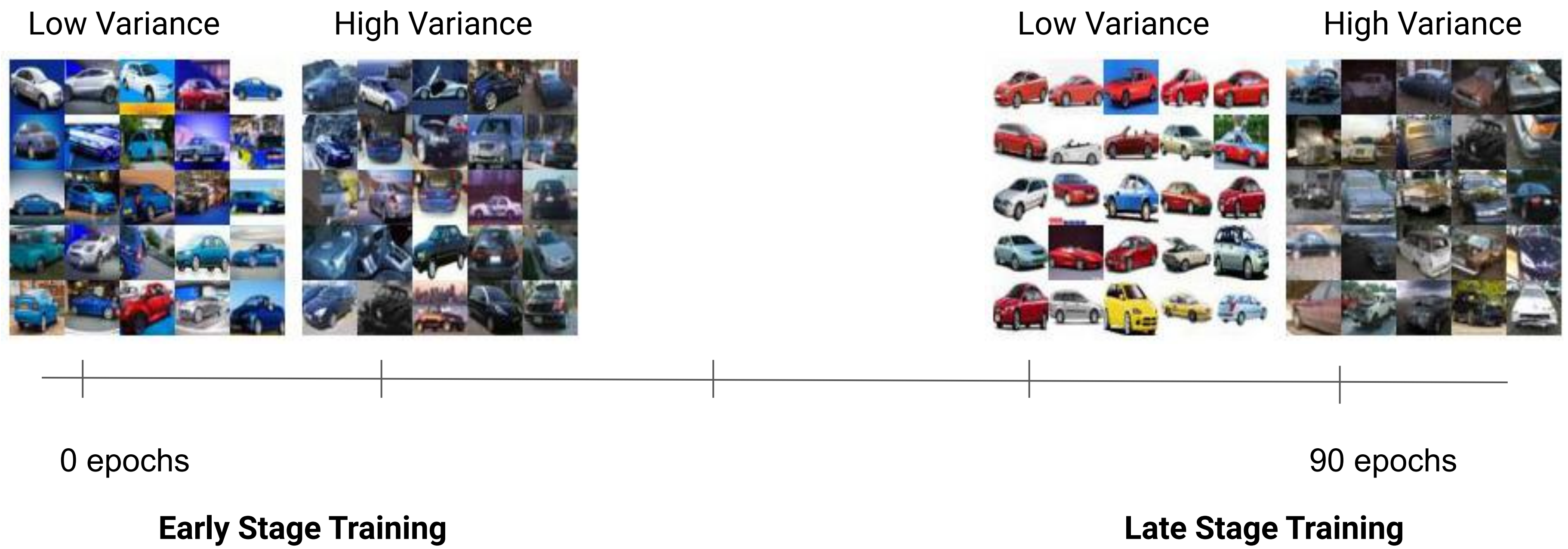


Recent research suggests there are distinct stages to training. Valuable opportunity to understand what features emerge when.

Characterizing Structural Regularities of Labeled Data in Overparameterized Models, 2020 (link)

Critical Learning Periods in Deep Neural Networks, 2017 (link)

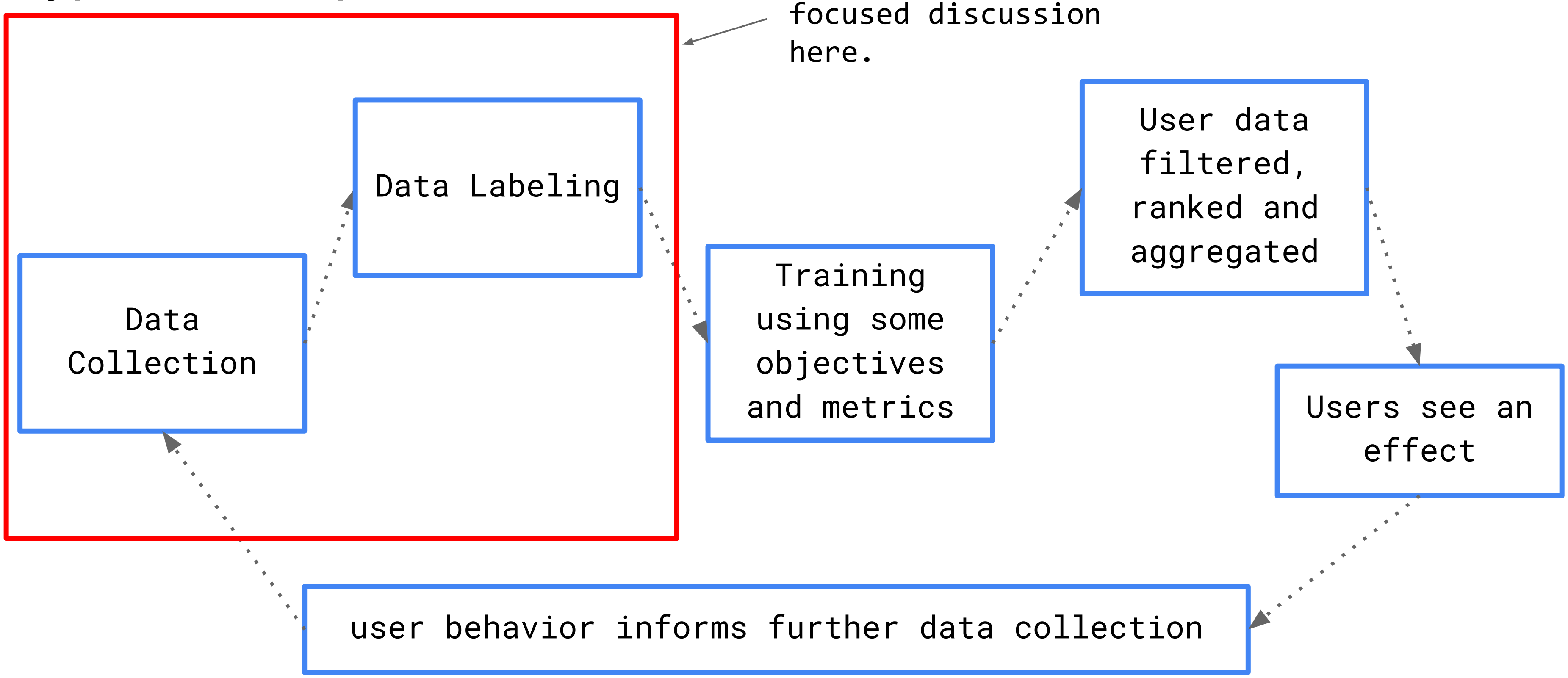
Easy examples are learnt early in training, hard examples require memorization later in training.



Estimating Example Difficulty using Variance of Gradients, Agarwal, Souza and Hooker, 2020

Typical ML Pipeline

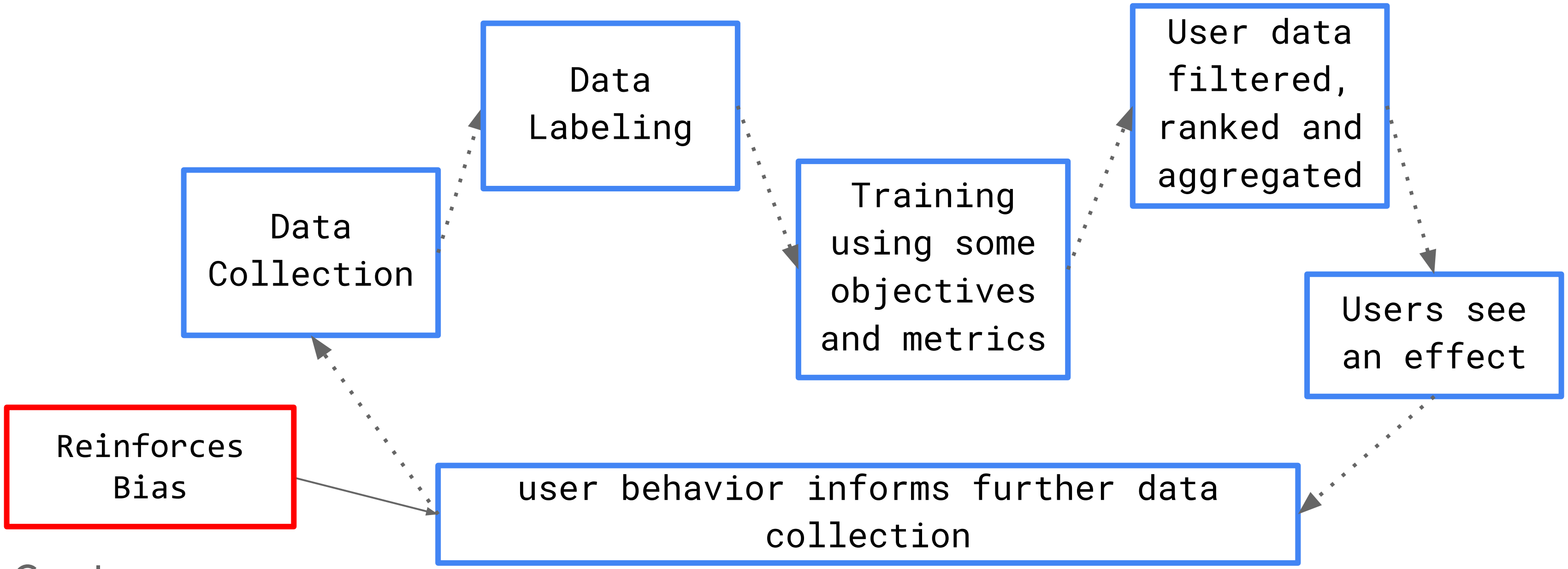
So far, we have focused discussion here.



In deployment settings fairness is rarely static.

Problems often have:

- Feedback loops that amplify disparate harm



Google

In deployment settings fairness is rarely static.

Problems often have:

- Feedback loops that amplify disparate harm
    - Intervention impacts future distribution of data.

Google

In deployment settings fairness is rarely static.

Problems often have:

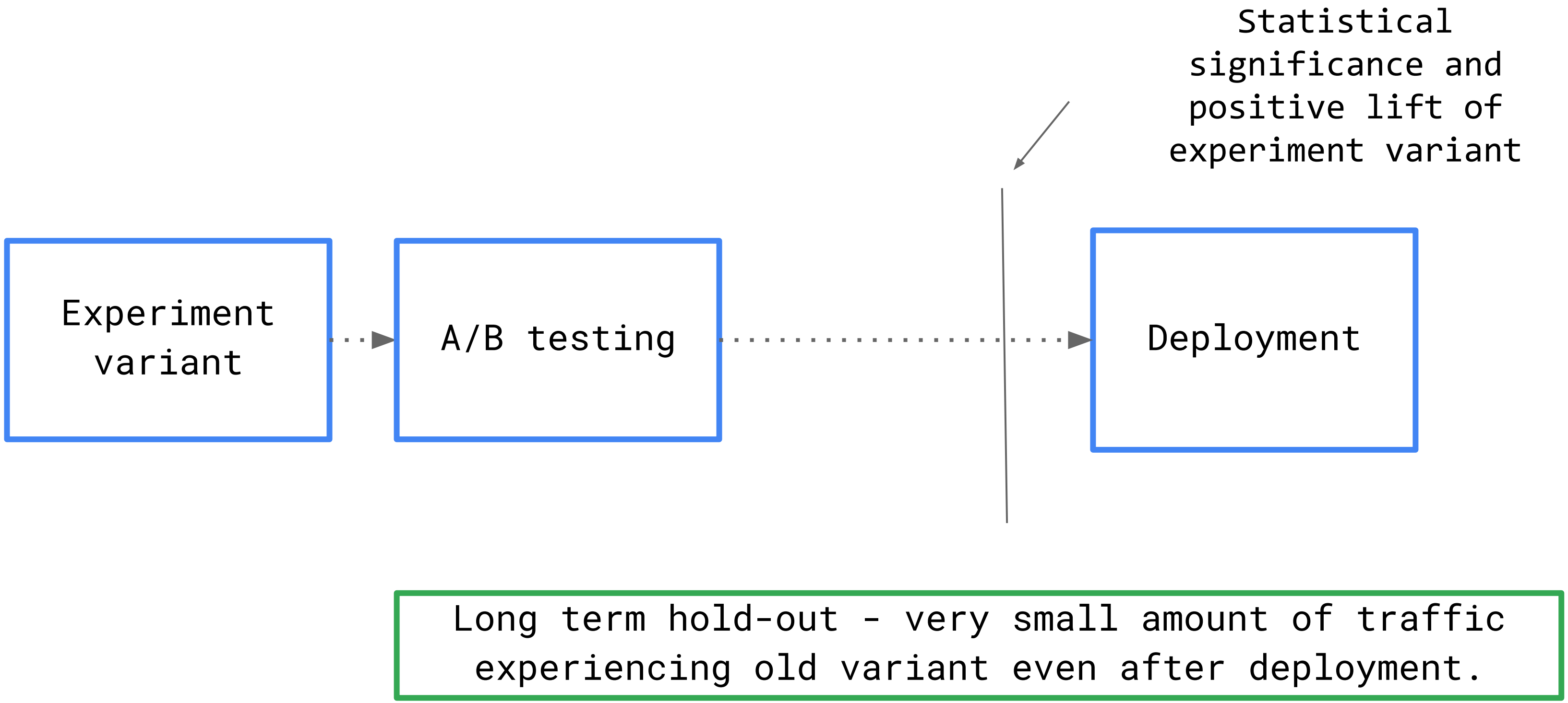
- Feedback loops that amplify disparate harm
- Involve long term outcomes
    - i.e long term user retention

In deployment settings fairness is rarely static.

Problems often have:

- Feedback loops that amplify disparate harm
- Involve long term outcomes
- Have complex dynamics that are hard to fully codify
    - i.e. recommendation box interactions

The importance of long-term holdouts in A/B testing frameworks



Training Models to Fulfill Multiple Desiderata

Chapter 2: Robustness

Robustness - Sensitivity of model behavior to deviations from the training set.

Google

Robustness testing in deployment settings

Is...

- A non-statistical test to gain a relative understanding of how model performance changes under certain distribution shifts or on certain subsets of the distribution
- Should involve a clear understanding of the distribution shift that is being modelled.

Is not ...

- Meant to capture all possible failure modes
- Meant to be a precise measure of model performance once deployed

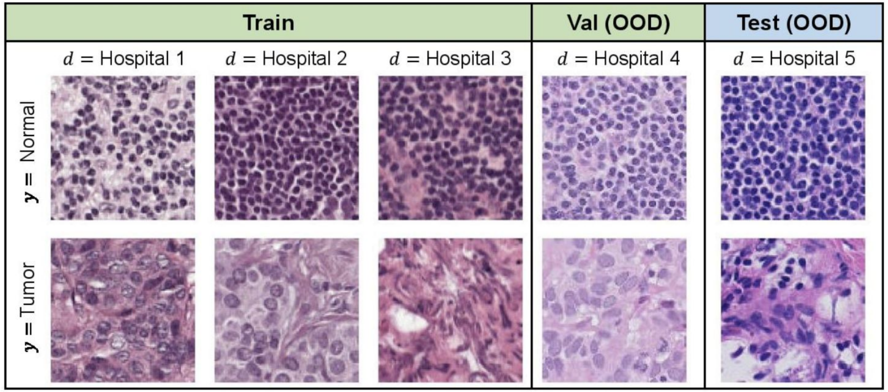
1. Academic benchmarks for robustness testing - ImageNet-A and ImageNet-C

[ImageNet-A](): Natural adversarial examples 7,500 examples from iNaturalist, Flickr, DuckDuckGo

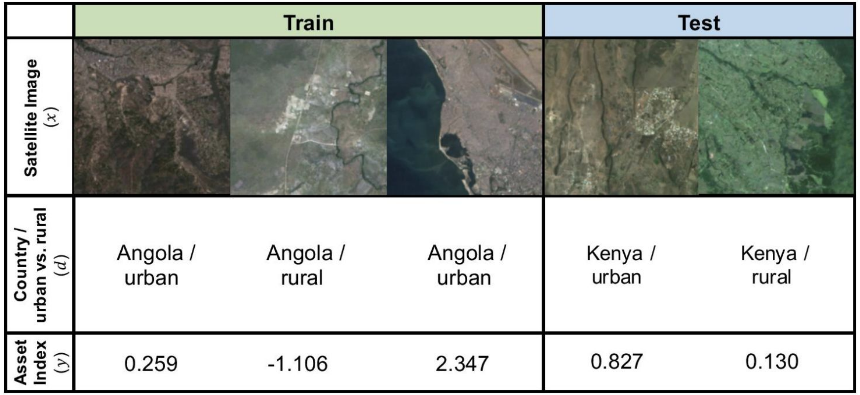
[ImageNet-C](): Set of corruptions applied to ImageNet test image.

2. Academic benchmarks for robustness testing - WILDS benchmark

Camelyon17



PovertyMap



[WILDS benchmark]

Google

3. Craft a robustness benchmark specific to your deployment task.

Set aside subsets of data (not to be including to training) that differ in known ways from the training set distribution.

- From a time range that differs from the training dataset range.
- From a different geography than the training dataset locale.
- From users who use a different language or device.

3 Craft a robustness benchmark specific to your task.

Valuable way to audit for algorithmic bias when you only have labels for a limited subset of the dataset with the sensitive feature you want to track.

From a time range that differs from the training dataset range.

From a different geography than the training dataset locale.

From users who use a different language or device.

The myth of the fair, robust, compact, private, high performance model.

Chapter 3: Trade-offs

Flawed assumption -- when we optimize for a desirable property, all other properties are held static.

In complicated systems, it is hard to vary one variable in isolation or foresee all implications.

From iron curtain to green belt

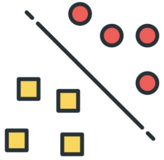
[European green belt]

Google

It is unrealistic to assume optimizing for one property holds all others static.

How we often talk about different properties in the literature.

Fairness - imposes constraint on optimization that reflects societal norms of what is fair.



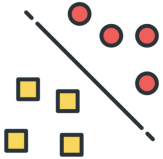
Model Compression - compact machine learning models to work in resource constrained environments.

Model fragility and security - deploy secure models that protect user privacy.

Optimizing for one objective will entail trade-offs with others.

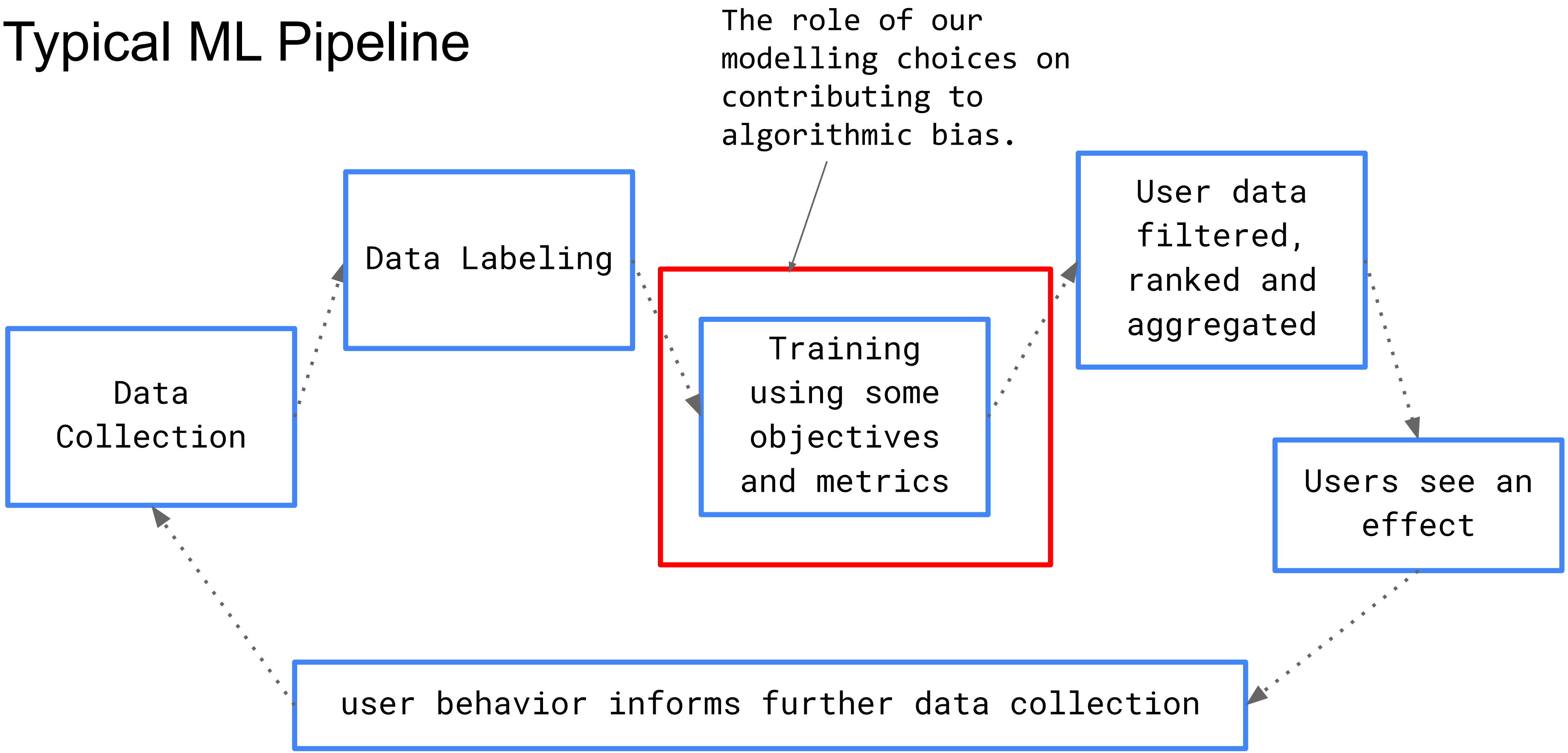
Fairness - imposes constraint on optimization that reflects societal norms of what is fair.

Model fragility and security - deploy secure models that protect user privacy.



Model Compression - compact machine learning models to work in resource constrained environments.

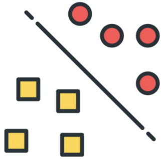
Typical ML Pipeline



Case Study: How does model compression trade-off against other properties we care about such as robustness and fairness?

Model Interpretability - reliable explanations for model behavior.

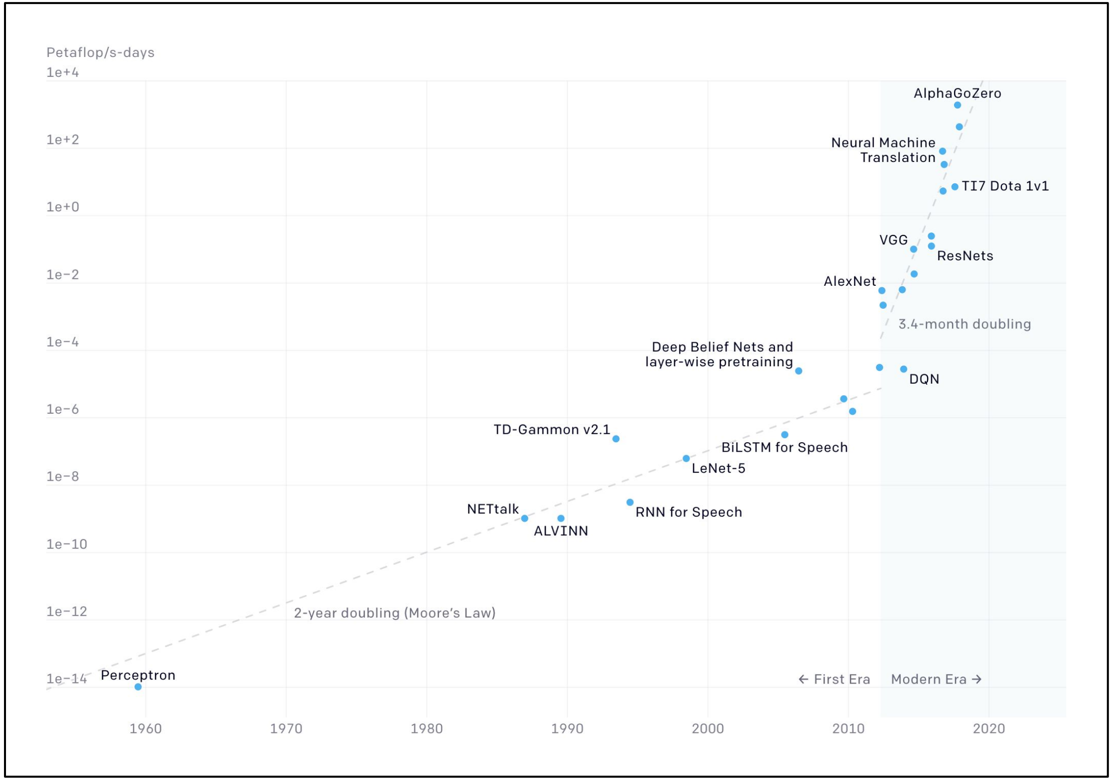
Fairness - imposes constraint on optimization that reflects societal norms of what is fair.



Model Compression - compact machine learning models to work in resource constrained environments.

Model fragility and security - deploy secure models that protect user privacy.

A “bigger is better” race in the number of model parameters has gripped the field of machine learning.



Canziani et al., 2016, Open AI 2019

Bigger models complicates democratization of AI models to resource constrained environments.

As you increase size of networks:

- More memory to store
- Higher latency for each forward pass in training + inference time

ML at the edge:

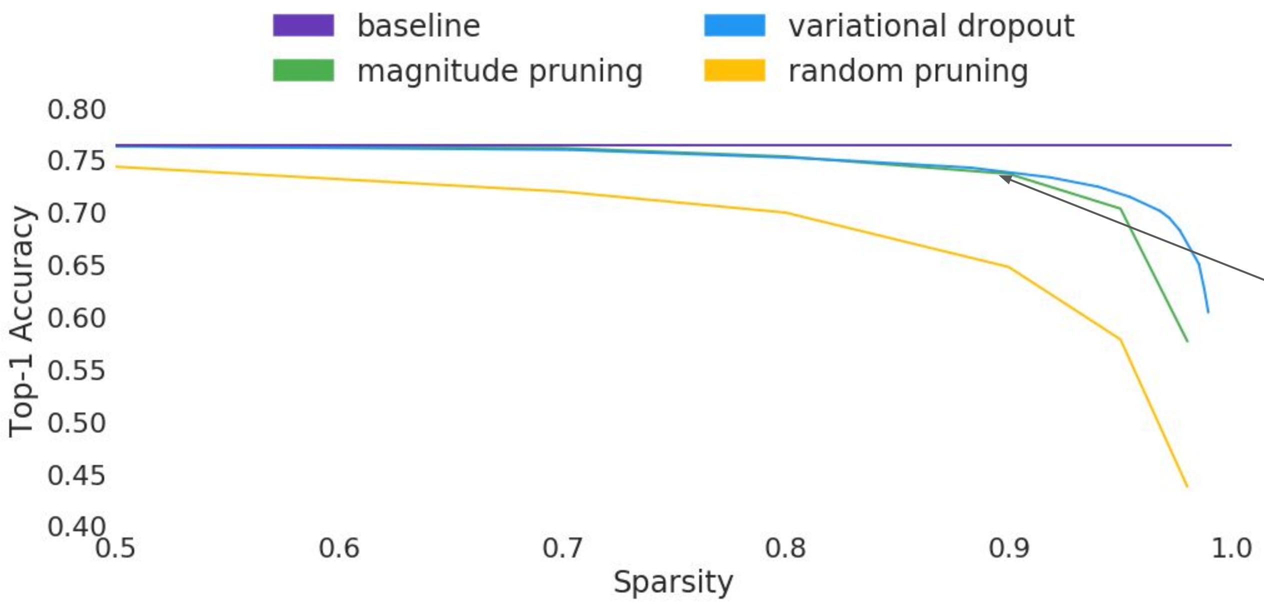
- Many different devices, hardware constraints
- Many different resource constraints - memory, compute
- Power, connectivity varies

Benefits of Compressed Models

- High Preservation of Top-1 Accuracy
- Low Latency
- Low Power Usage
- Portability etc...

Celeb-A			
Fraction Pruned	**Top 1**	**Quantization**	**Top 1**
No Compression	94.73	-	-
0.3	94.75	hybrid int8	94.65
0.5	94.81	fixed-point int8	94.65
0.7	94.44	-	-
0.9	94.07	-	-
0.95	93.39	-	-
0.99	90.98	-	-

Compression techniques like pruning and quantization remove weights from a network with remarkably little impact to top-line metrics.



With 90% of the weights removed, a ResNet-50 only loses ~3% of performance

[[The State of Sparsity in Deep Neural Networks, 2019, Gale, Elsen, Hooker]]

How can networks with radically different structures and number of parameters have comparable performance?

0% pruning
76.70%

50% pruning
76.20%

One possibility is that test-set accuracy is not a precise enough measure to capture how pruning impacts the generalization properties of the model.

In this work, we go beyond test-set accuracy.

Here, we ask - How does model behavior diverge as we vary level of compression?

1. Measure sensitivity to certain types of distributional shifts. (natural adversarial examples and corruptions)

2. Measure divergence in class level and exemplar classification performance.

Experimental Framework

Train populations of models with minimal differences in test-set accuracy to different end sparsities [0%, 30%, 50%, 70%, 90%, **95%, 99%**].

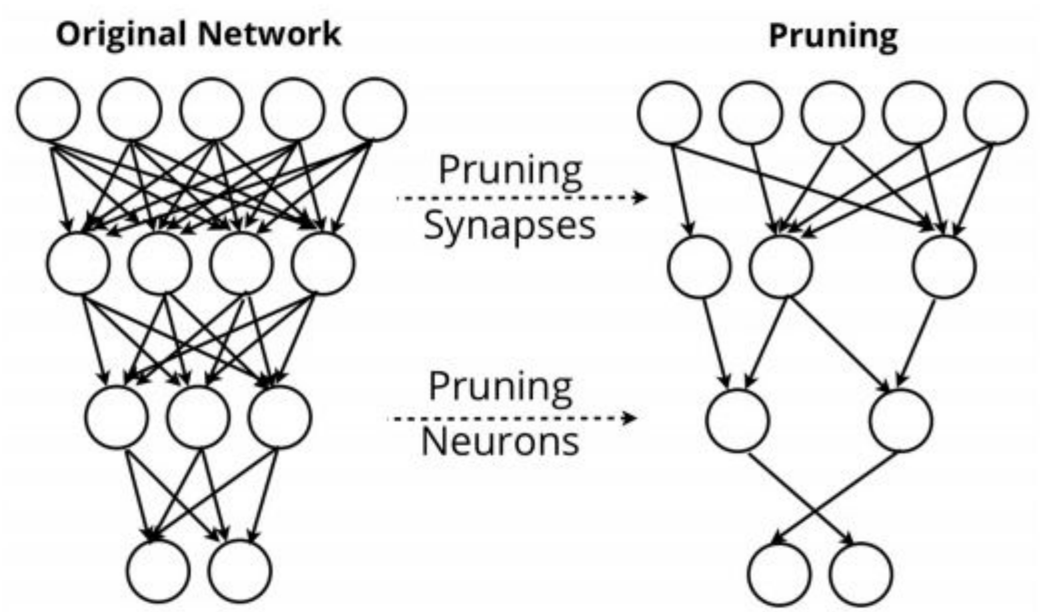
0 %

Overparameterized
Dense Model

90 %

Model with 90%
weights removed

Sparsity of 90% means that by the end of training the model only has 10% of all weights remaining. Apply mask of 0 to remaining weights.



Initial weight matrix

After activations have been removed.

Image source

Some nice properties of this empirical set-up:

Models all achieve similar regime of top-line performance.

We can precisely vary how radically the weight representation differs - by controlling end sparsity.

0 %

Overparameterized
Dense Model

90 %

Model with 90%
weights removed

Key results upfront: top level metrics hide critical differences in generalization between compressed and compressed populations of models.

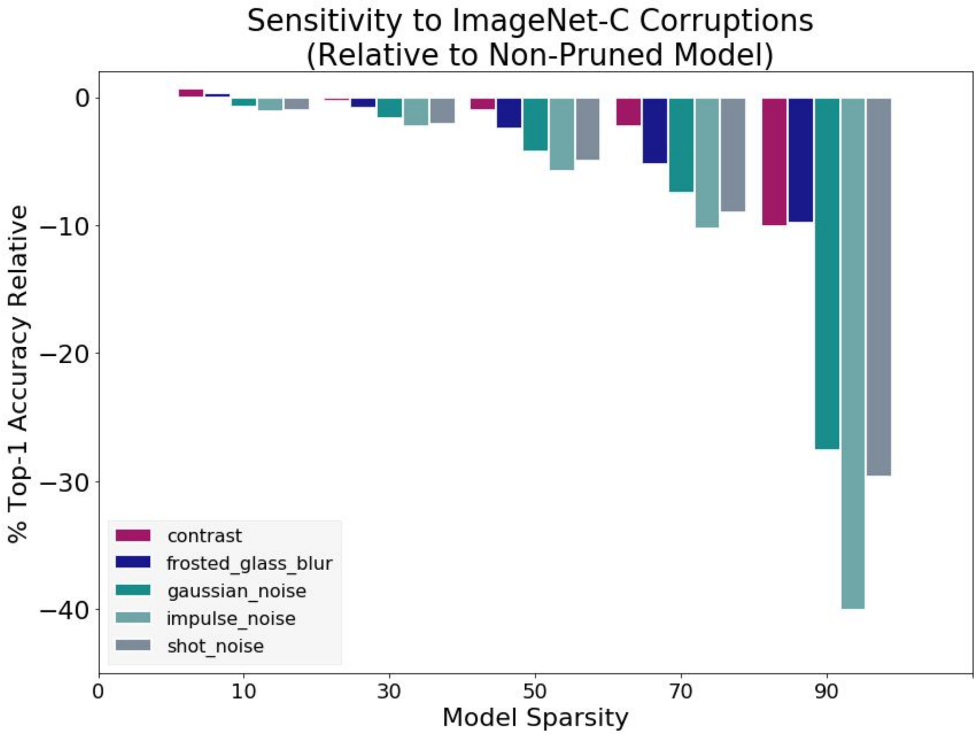
1. Compressed models have amplified sensitivity to adversarial examples and common corruptions.

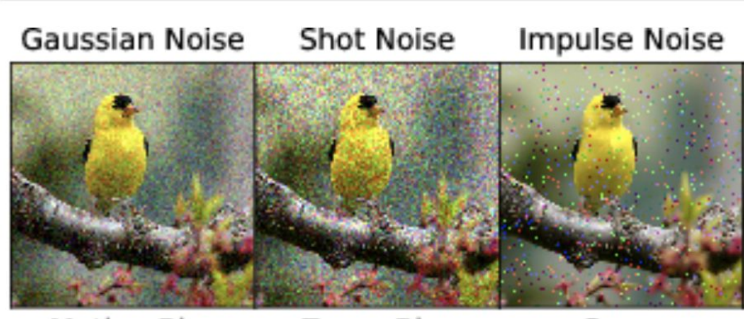
2. Varying capacity disproportionately and systematically impact a small subset of classes and exemplars.

→ Why is a narrow part of the data distribution far more sensitive to varying capacity?

Compression trade-off with robustness

A. Sensitivity to natural adversarial images ImageNet-C



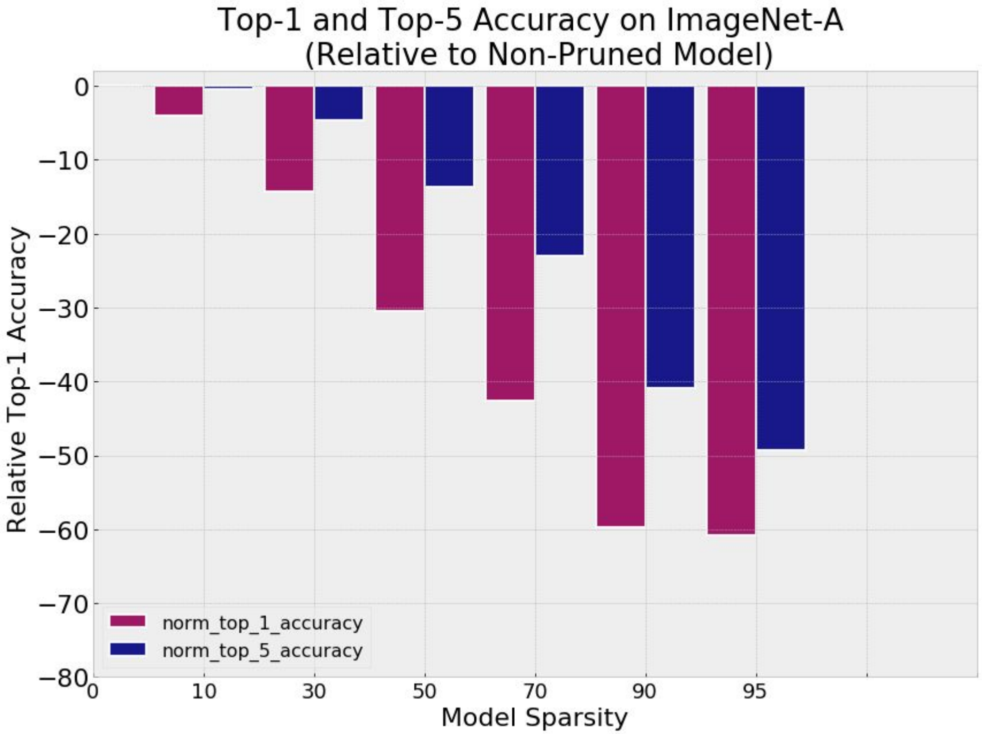


Amplification of sensitivity to some perturbations are far more pronounced than others.

Sparse models are particularly sensitive to noise.

[Hooker et al.](#)

A. Sensitivity to natural adversarial images ImageNet-A



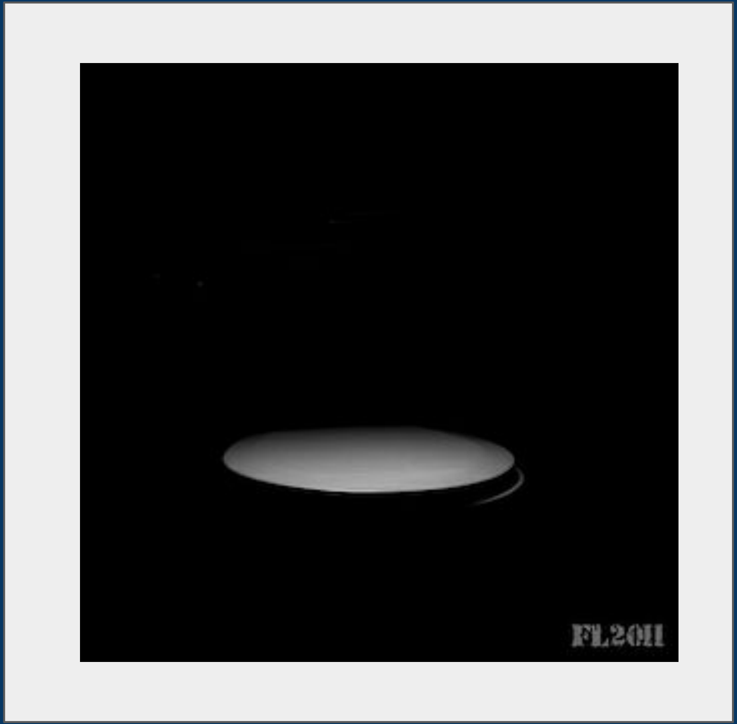
ImageNet-A: Natural adversarial examples
7,500 examples from iNaturalist, Flickr, DuckDuckGo

Hooker et al.

Compression trade-off with algorithmic bias

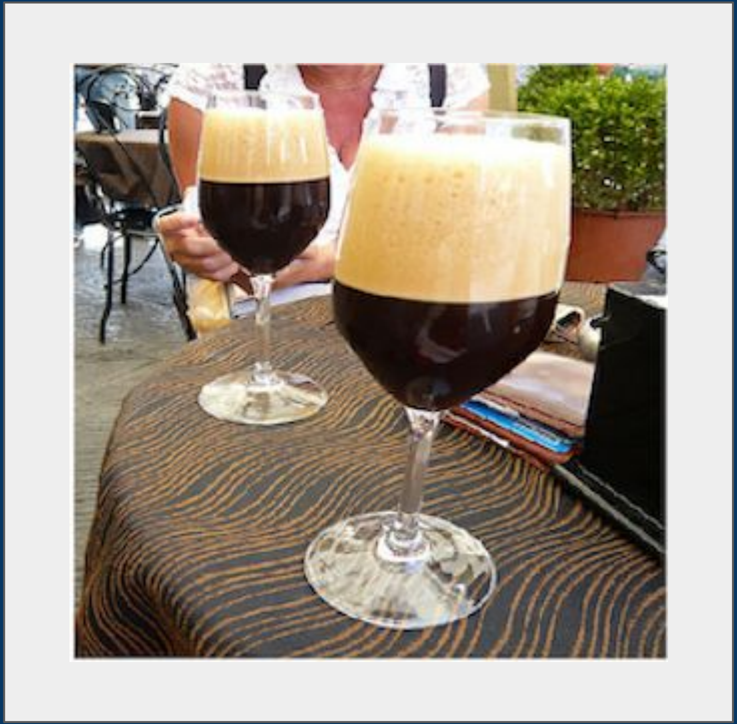
Pruning Identified Exemplars (PIEs)

are images where predictive behavior diverges between a population of independently trained compressed and non-compressed models.



ImageNet test-set.
True label?


toilet seat
Non-PIE
PIE



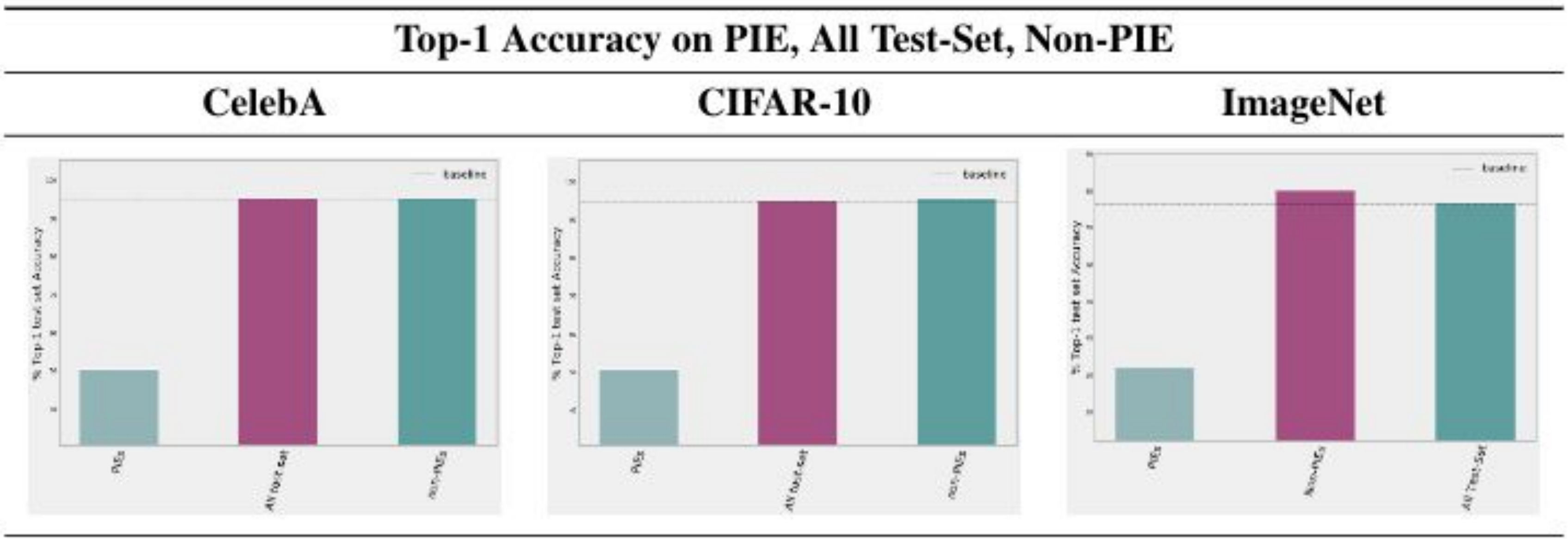
ImageNet test-set.
True label?

espresso

Non-PIE

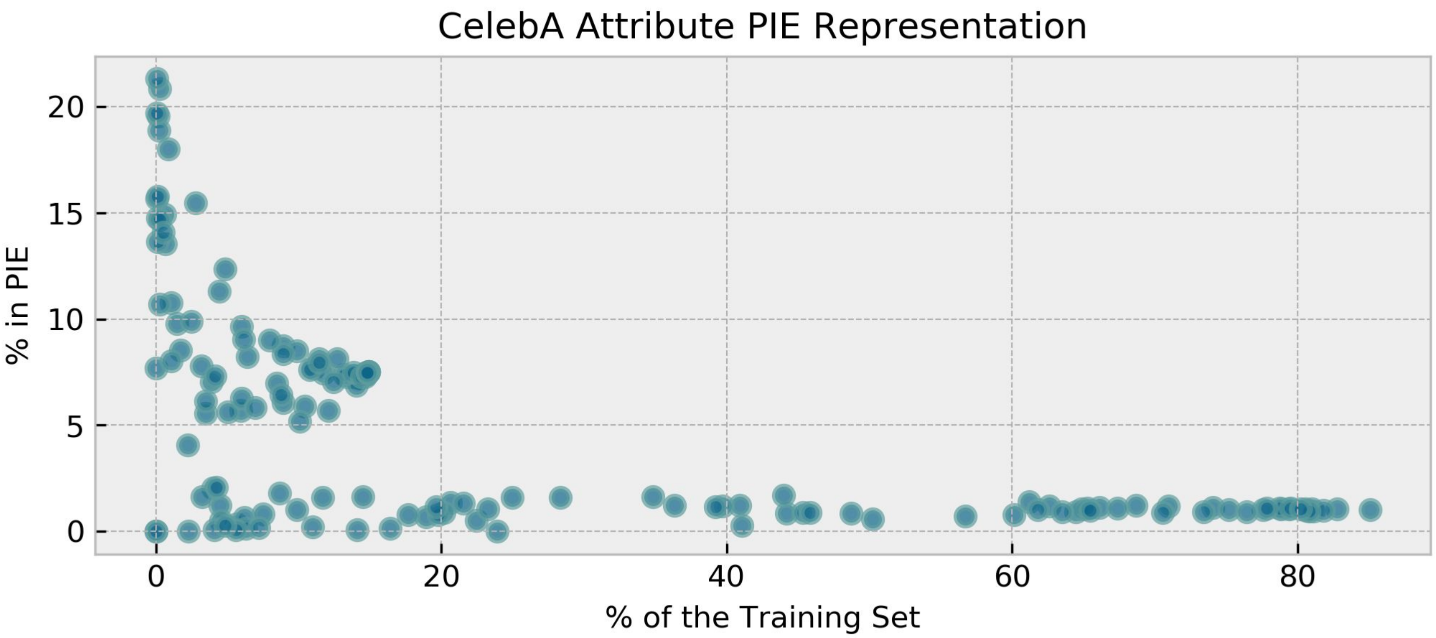
PIE

PIEs are also more challenging for algorithms to classify.



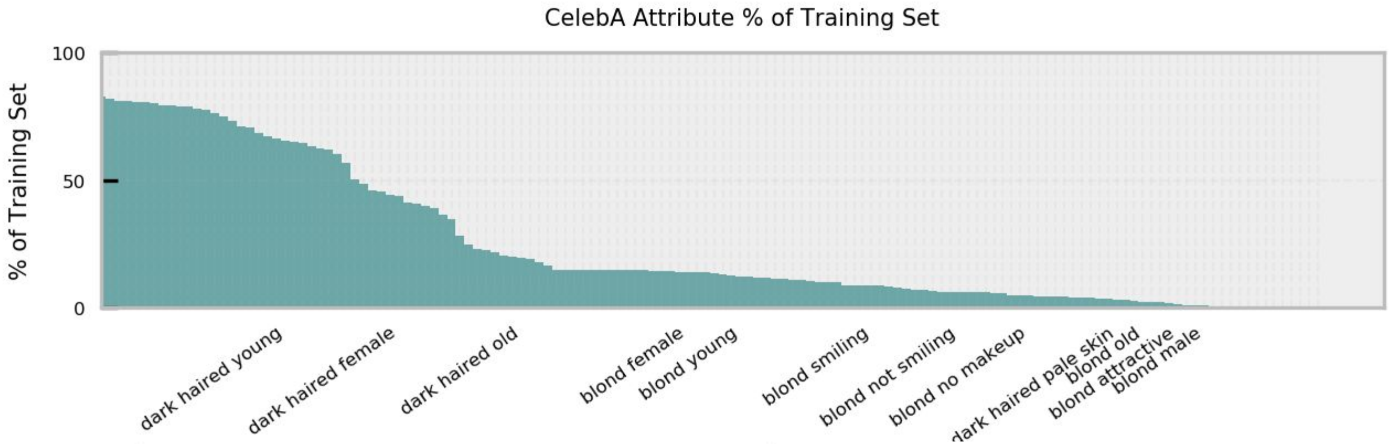
- Restricting inference to PIEs drastically degrades model performance.
- For ImageNet, removing PIEs from test-set improves top-1 accuracy beyond baseline.

PIEs over-index on the long-tail of underrepresented attributes.



Attribute Proportion of CelebA Training Data vs. relative representation in PIE

Compression disproportionately impacts underrepresented features.



Celeb-A

Y = {Blonde, Dark Haired}

Training set: 162,770

High Frequency Sub-Groups

Low Frequency Sub-Groups

Dark Haired Males
66,874
44%

Dark Haired Females
71,628
41%

Blonde Females
22,880
14%

Blond Male
1,387
0.85%

Blonde Old
4,037
2.48%

Pruning amplifies algorithmic bias when the underrepresented feature is protected (age/gender)

Model	Metric	Aggregate	Unitary M	Unitary F	Unitary Y	Unitary O	Intersectional MY	Intersectional MO	Intersectional FY	Intersectional FO
Baseline (0% pruning)	Error	5.30%	2.37%	7.15%	5.17%	5.73%	2.28%	2.50%	5.17%	5.73%
	FPR	2.73%	0.93%	4.12%	2.59%	3.18%	0.81%	1.12%	2.59%	3.18%
	FNR	22.03%	62.65%	19.09%	21.35%	24.47%	60.45%	66.87%	21.35%	24.47%
Normalized Difference Between **1)** Compressed and **2)** Non-Compressed Baseline										
Compressed (95% pruning)	Error	24.63%	24.49%	24.67%	20.64%	35.84%	7.96%	49.12%	20.64%	35.84%
	FPR	12.72%	49.54%	6.32%	3.35%	36.02%	5.37%	101.88%	3.35%	36.02%
	FNR	34.22%	8.41%	40.30%	33.83%	35.39%	9.21%	6.98%	33.83%	35.39%

Table 3: Performance metrics disaggregated across `Male` (M), `not Male` (F), `Young` (Y), and `not Young` (O) sub-groups. For all error rates reported, we average performance over 10 models. **Top Row**: Baseline error rates, **Bottom Row:** Relative change in error rate between baseline models and models pruned to 95% sparsity,

Google

[[Hooker et al. 2019, Hooker, Moorosi et al, 2020]]

Case study 2: Privacy trade-off with fairness.

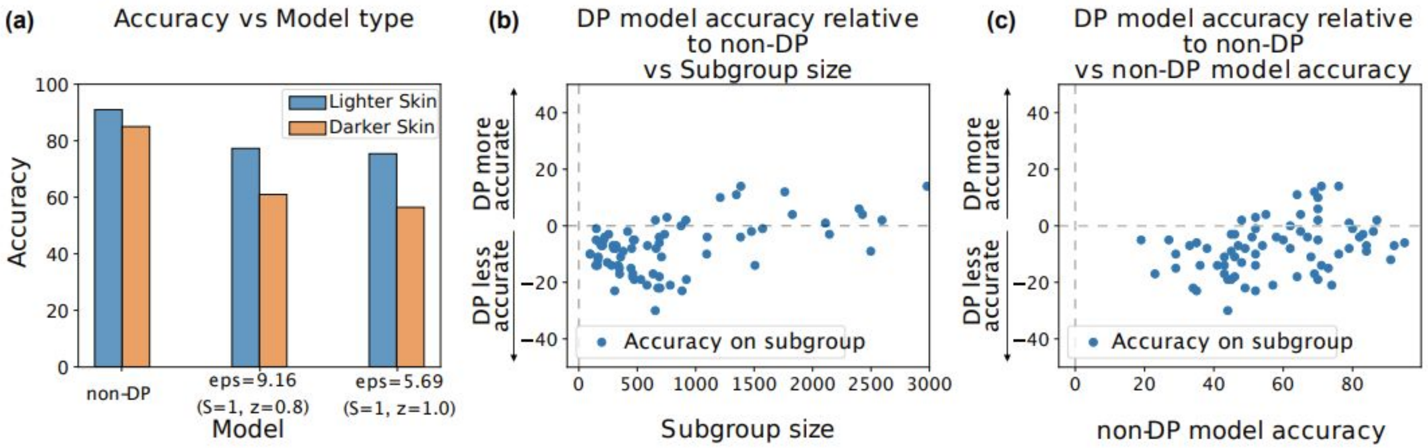


Figure 1: Gender and age classification on facial images.

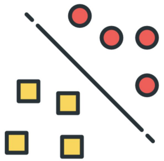
[Bagdasaryan et al.](#)

Beyond "Algorithmic bias is a data problem."

Algorithms do not simply impartially reflect biases. Choices we make when we model can amplify or minimize harm.

This is because disparate harm is not held static while other properties are optimized.

Fairness - imposes constraint on optimization that reflects societal norms of what is fair.



Model Compression - compact machine learning models to work in resource constrained environments.

Model fragility and security - deploy secure models that protect user privacy.

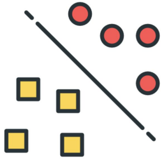
The known unknowns

Chapter 4: Interpretability

Interpretability tools aim to provide insight into model behavior. Enable auditing of other desirable properties such as fairness and robustness.

Model Interpretability - reliable explanations for model behavior.

Fairness - imposes constraint on optimization that reflects societal norms of what is fair.



Model Compression - compact machine learning models to work in resource constrained environments.

Model fragility and security - deploy secure models that protect user privacy.

Emphasis we place on interpretability will depend on multiple factors

Sensitive domain

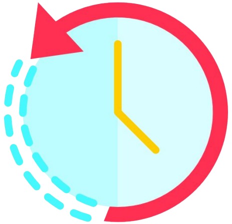
Can the model adversely impact human welfare?

Trade-off with other model desiderata

Does improving interpretability jeopardize other desirable properties i.e. model security or privacy?

Historical performance

Is historical data of model behavior in different test conditions limited?



Criteria for what is meaningful as an interpretable tool will deep upon our vantage point and downstream tasks

Domain Expert

End Consumer

Vantage point also impacts the type of interpretability tooling that is most useful.

End user: Will always want to know the explanation for their data point.

Specialist: Will want to place the model explanation within relative context. Both an individual explanation and global ranking desirable.

Deployment engineer: Will want to gain insight into domain shift, surface examples which are most challenging. Automatically surface candidates for additional annotation. Audit any model errors.

A local explanation often fails to provide enough context for actionable downstream decision making.

Google

Understanding how model behavior aligns/diverges from human knowledge has become even **more** paramount.

1) We have chosen functional forms that delegates feature representation to the model - harder to extract feature importance estimates.

2) Models are widely deployed in settings where human welfare can be impacted adversely.

3) The size of modern day datasets mean it is critical we provide tools which surface what is most critical for human inspection.

Interpretability does not require explaining everything about a model.

- Goal is to gain intuition into model behavior
- We are unlikely to ever sign off an a model as interpretable.

Google

Building Interpretable Models

Most research has focused here.

1. Model Distillation
2. Regularization of weights during training to condition heat map properties.

Post-hoc Interpretability Methods

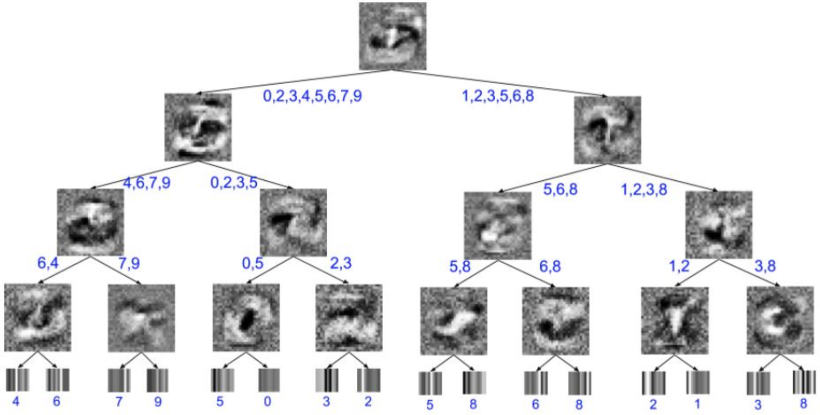
1. Neuron/weight importance
2. Input feature importance
3. Outlier detection

Visualization/Human guided investigations

1: Model Distillation

Distill the knowledge of a large neural network into a functional form considered more interpretable.

(note: hard to compete in accuracy)

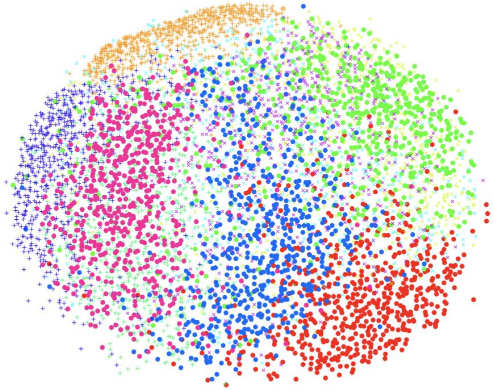


Distilling a Neural Network Into a Soft Decision Tree [[Frosst and Hinton , 2017]].

[[Ba et al. 2014, Hinton et al., 2015, Frosst and Hinton , 2017, Wang and Rudin, 2015, Tan et al. 2018]]

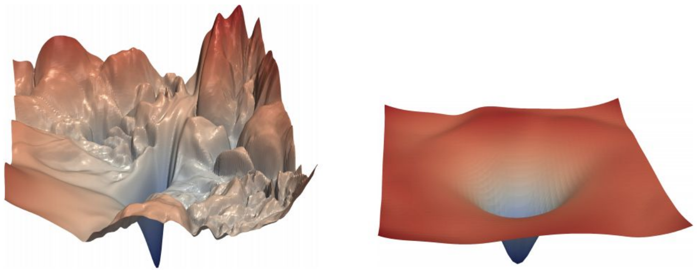
Google

2: Visualization tools reduce high dimensionality of deep neural networks



(b) Visualization by Sammon mapping.

Figure 2: Visualizations of 6,000 handwritten digits from the MNIST data set.



(a) without skip connections

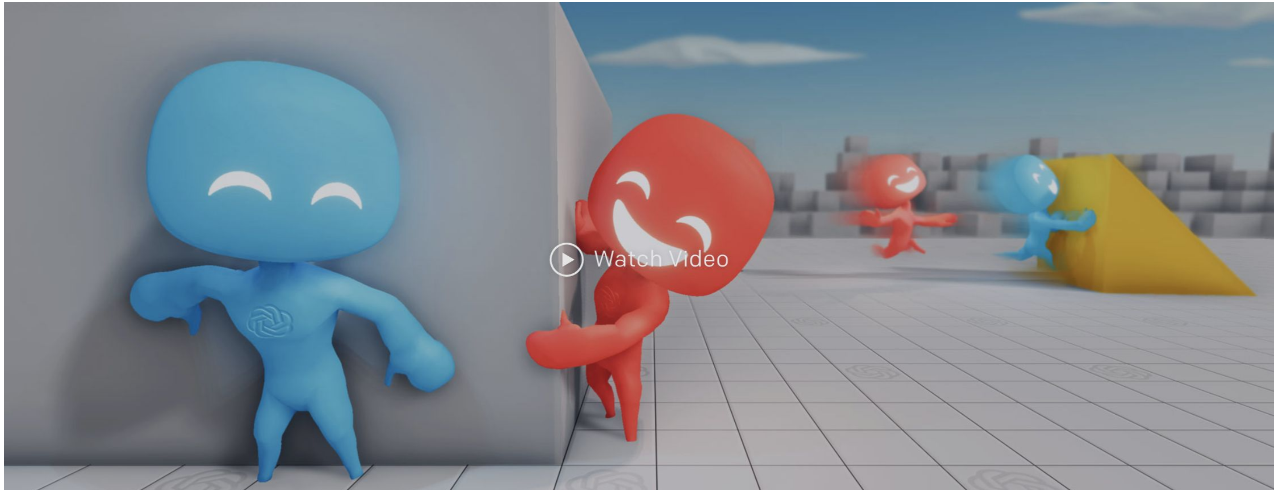
(b) with skip connections

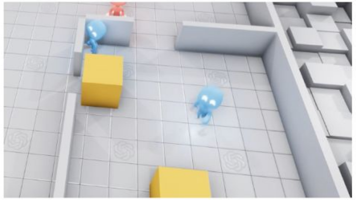
Figure 1: The loss surfaces of ResNet-56 with/without skip connections. The proposed filter normalization scheme is used to enable comparisons of sharpness/flatness between the two figures.

t-Distributed Stochastic Neighbor Embedding (t-SNE)
[[van der Maaten and Hinton, 2008]

Visualizing the loss landscape of deep neural networks
[[paper]]

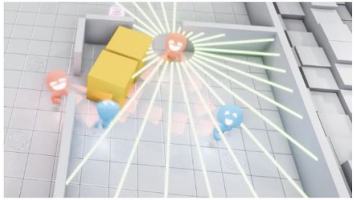
3: Agent Based Exploration



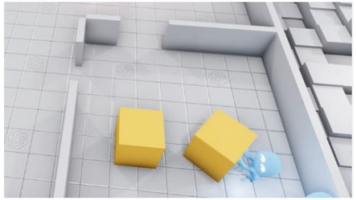


The agents can **move** by setting a force on themselves in the x and y directions as well as rotate along the z-axis.

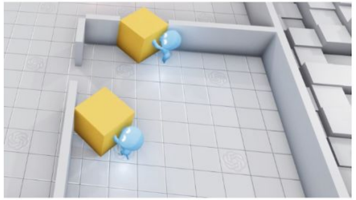
The agents can **see** objects in their line of sight and within a frontal cone.



The agents can **sense** distance to objects, walls, and other agents around them using a lidar-like sensor.



The agents can **grab and move** objects in front of them.



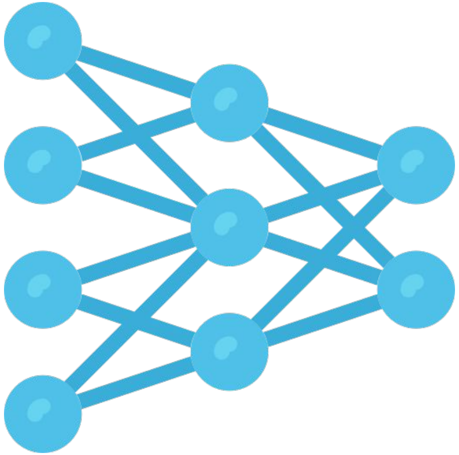
The agents can **lock** objects in place. Only the team that locked an object can unlock it.

Baker et al, 2019 (link)

4: Estimates of feature importance

Local Feature Importance

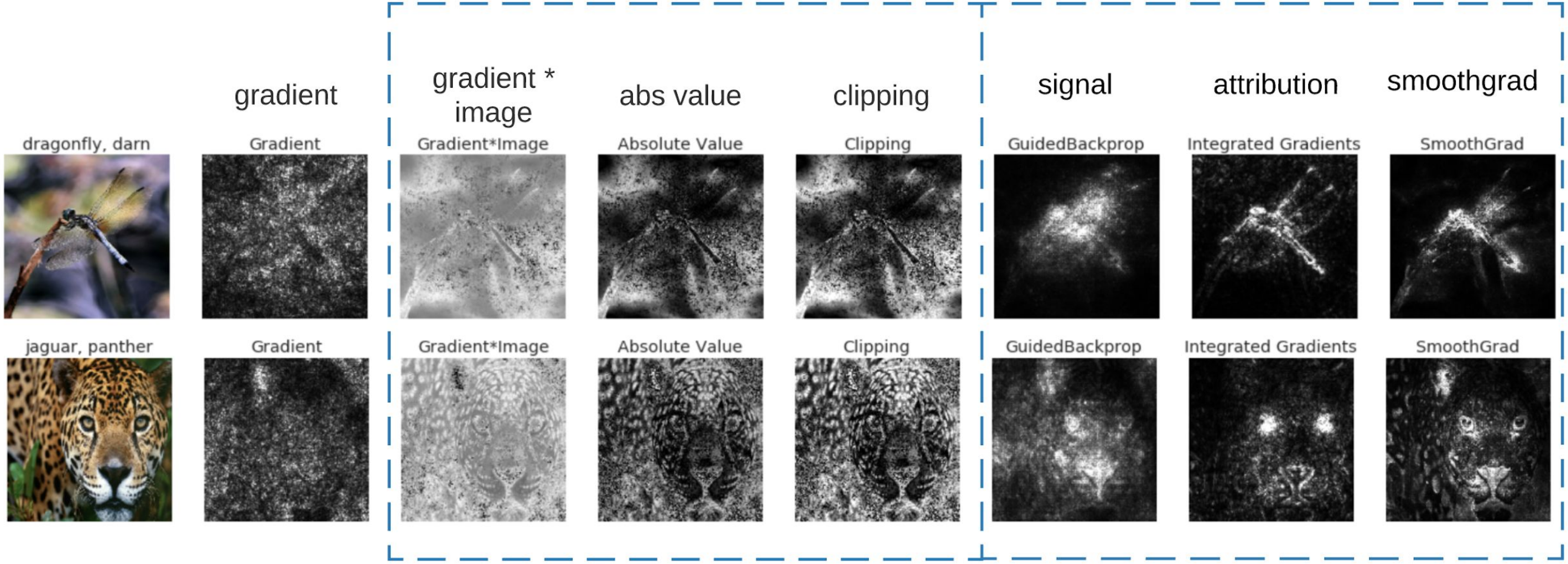
Global Feature Importance



Weights and Activations

4.1: Local Feature Importance

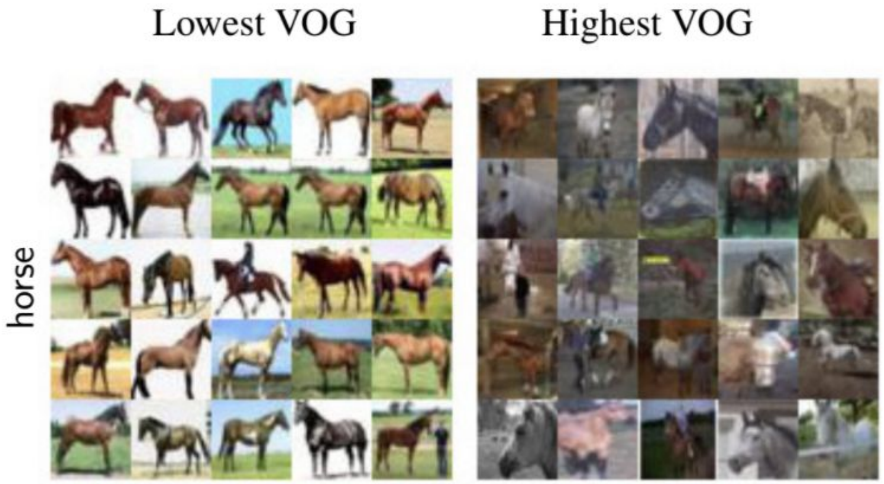
Estimates the feature importance of the attributes in a data example to a **single** model prediction.



[Erhan et al., 2009, Simonyen et al.,2013, Springenberg et al.,2015, Fong and Vedaldi 2017, Sundararajan et al. 2017, Smilkov et al., 2017., many more...]

4.2: Global Feature Importance

Estimates the feature importance of the attributes to the overall decision boundary.
What examples does the model find challenging or easy to learn?



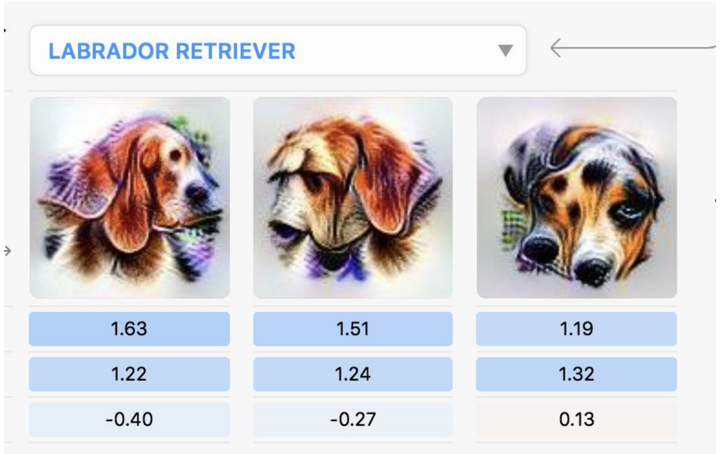
Estimating Example Difficulty using Variance of Gradients, Agarwal* and Hooker*, 2020

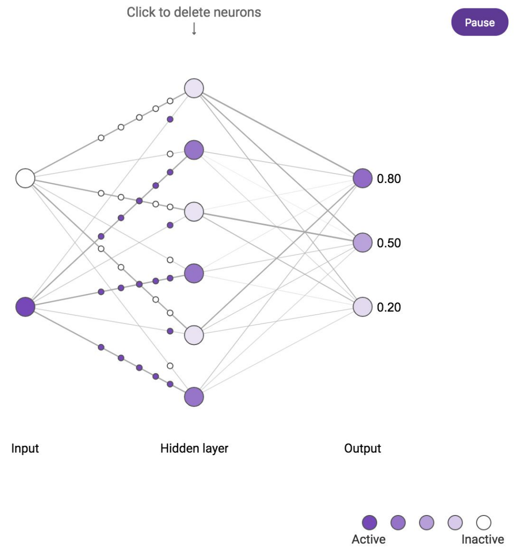
What does a compressed deep neural network forget? Hooker et al. 2020

Google

4.3: Weight and Activations

Estimates the role or importance of individual neurons or weights.





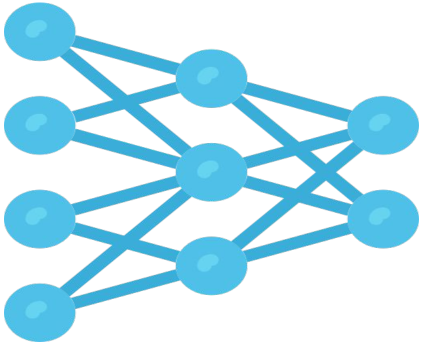
Neuron interpretation [[Olah,C et al, 2017]

Weight/layer ablation studies [[Morcos A. et al., 2018]]

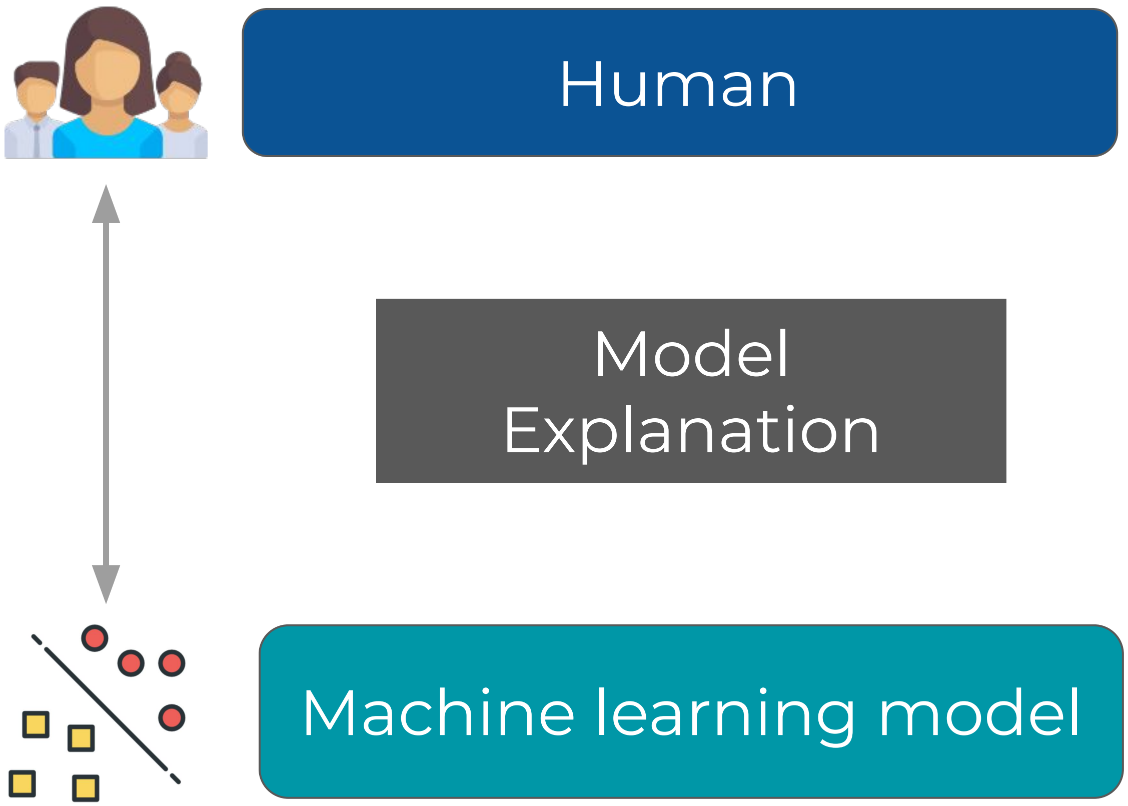
A large amount of interpretability research for deep neural networks has focused on local feature importance.

Local Feature Importance

Global Feature Importance



Weights and Activations



An interpretable explanation of a model prediction must be both: **meaningful** to a human + an **accurate** reflection of the model.

Key open challenges in interpretability:

1) Meaningful does not equate with reliable - identifying failure points in explanations.

2) Disproportionate emphasis on feature importance at the end/after training.

3) Providing both global and local explanations of model behavior that are scalable to deployment settings.

Closing Thoughts (and Q&A)

Thanks for the invite Chip!

Questions?

Estimating Example Difficulty using Variance of Gradients Chirag Agarwal*, Sara Hooker* [[link]]

What do compressed deep neural networks forget?, Sara Hooker, Aaron Courville, Gregory Clark, Yann Dauphin, Andrea Frome [[link]]

Characterizing Bias in Compressed Models Sara Hooker*, Nyalleng Moorosi*, Gregory Clark, Samy Bengio, Emily Denton [[link]]

More work -- links in the slides. Feel free to email me for a copy.

Final takeaways:

Beyond test-set accuracy - It is not always possible to measure the trade-offs between criteria using test-set accuracy alone.

The myth of the compact, private, interpretable, fair model - Desiderata are not independent of each other. Training beyond test set accuracy requires trade-offs in our model preferences.

Relative vs local feature importance - human understanding is relative, promising work to surface subset of data points that are more/less challenging to aid understanding.

.

Email: shooker@google.com