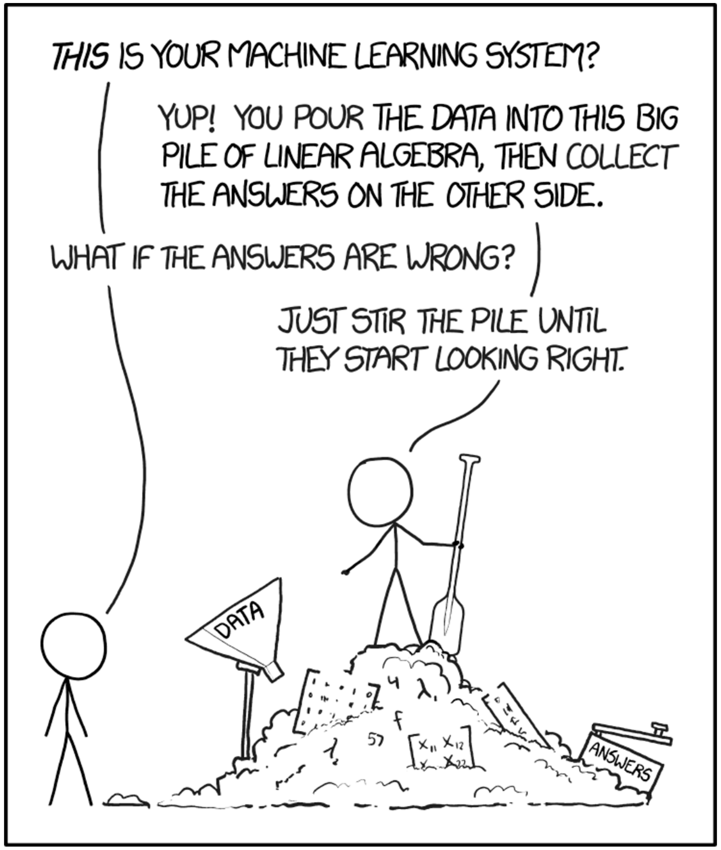
https://qdata.github.io/deep2Read/

Interpreting Deep Learning Models


THIS IS YOUR MACHINE LEARNING SYSTEM?
YUP! YOU POUR THE DATA INTO THIS BIG PILE OF LINEAR ALGEBRA, THEN COLLECT THE ANSWERS ON THE OTHER SIDE.
WHAT IF THE ANSWERS ARE WRONG?
JUST STIR THE PILE UNTIL THEY START LOOKING RIGHT.
DATA
ANSWERS

Presented by Eli Lifland, 2/28/2020

Overview

- What even is interpretability?
- Why do we need interpretations?
- Interpretation methods
- Is attention explanation?
- Problems with interpretations
- Where do we go from here?

What is interpretability in ML?

- Depends who you ask: interpretability is a “suitcase word,” meaning it’s used to represent many different meanings
- A general definition from Murdoch (2019):
    - Interpretable ML is the use of machine-learning models for the extraction of relevant knowledge about domain relationships contained in data

Why interpretations?

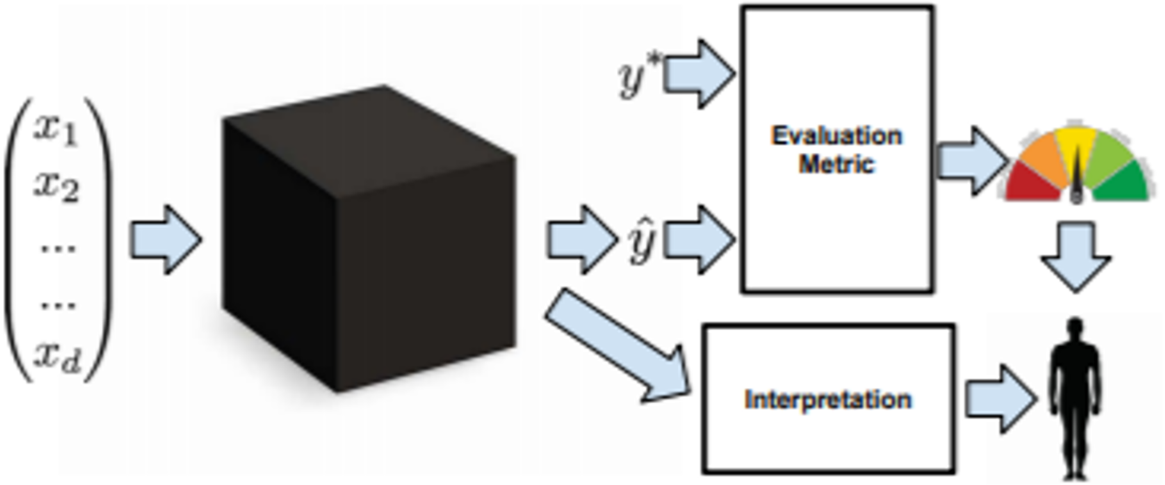


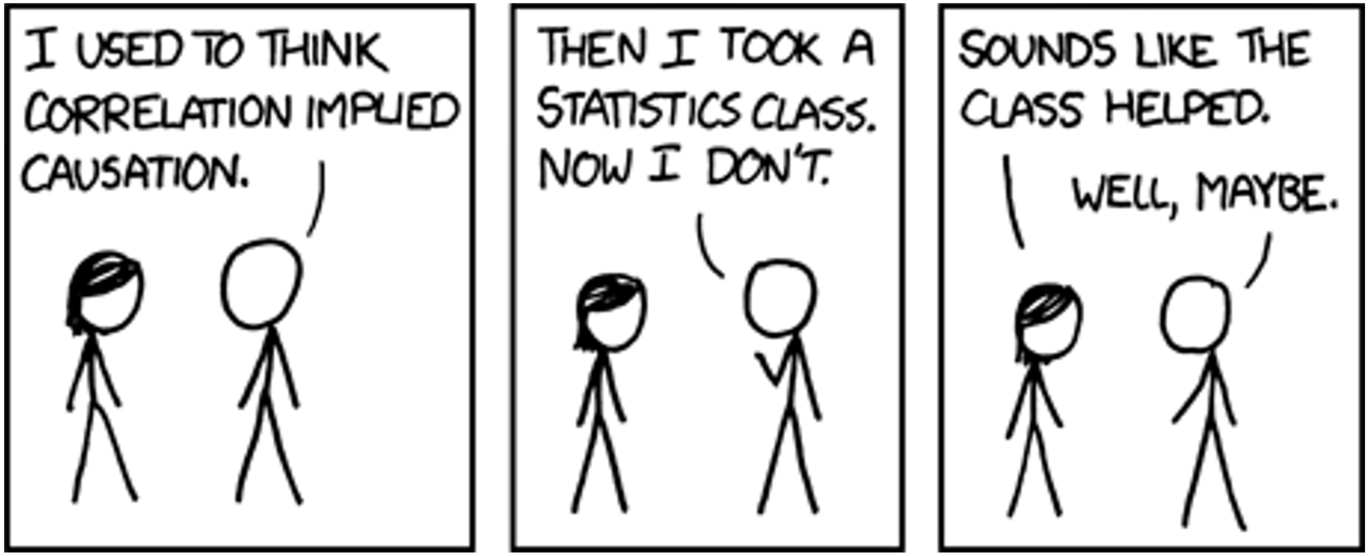
Figure 1. Typically, evaluation metrics require only predictions and *ground truth* labels. When stakeholders additionally demand *interpretability*, we might infer the existence of desiderata that cannot be captured in this fashion.

Why interpretations? Trust

- Why can’t we trust an accurate model?
- When training and deployment environments differ, we want to trust the model will still perform well
- Feel comfortable relinquishing control to model
  - Does it make mistakes in similar cases to human?
- Subjective notion of trust
  - Feel more at ease with well-understood model

Why interpretations? Causality

- As we know, correlation does not imply causation
- Help ensure model isn't learning spurious correlations
- Provides causal hypotheses which can be tested experimentally


I USED TO THINK CORRELATION IMPLIED CAUSATION.
THEN I TOOK A STATISTICS CLASS. NOW I DON'T.
SOUNDS LIKE THE CLASS HELPED.
WELL, MAYBE.

https://qdata.github.io/deep2Read/

Why interpretations? Transferability

- Generalization to distributional shift
- Scenarios where use of model alters environment
- Robustness to adversarial attacks
- Best to anticipate and understand failures to generalize

Why interpretations? Informativeness

- Provide more information than just the prediction
- Example: PhD student asks which venue best suits a paper, professor would not answer with the name of one conference
- Note information can be provided without revealing model's inner workings
  - Pointing out data points model saw as similar

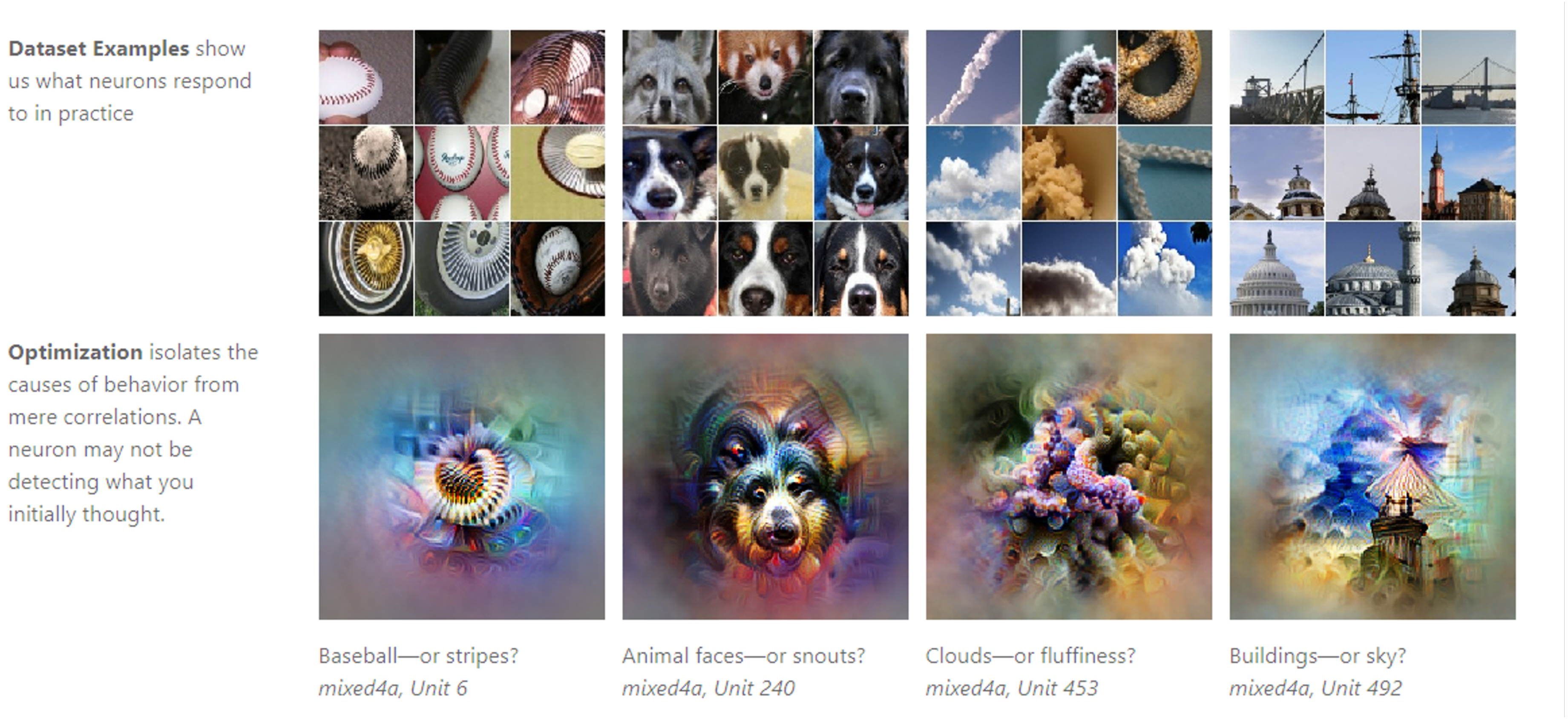
Why interpretations? Fairness

- Assessing whether decisions conform to ethical standards
- Are models using variables we don't want them to, reinforcing biases?
- European Union says:
    - People have right to explanation
    - Algorithmic decisions must be contestable

Interpretation Methods: Overview

- Post-hoc explanations
    - Feature visualization
    - Feature attribution
        - Instance-wise vs. model-wide
    - Feature visualization + attribution
    - Training example attribution
- Inherently interpretable models

Feature Visualization



Dataset Examples show us what neurons respond to in practice

Optimization isolates the causes of behavior from mere correlations. A neuron may not be detecting what you initially thought.

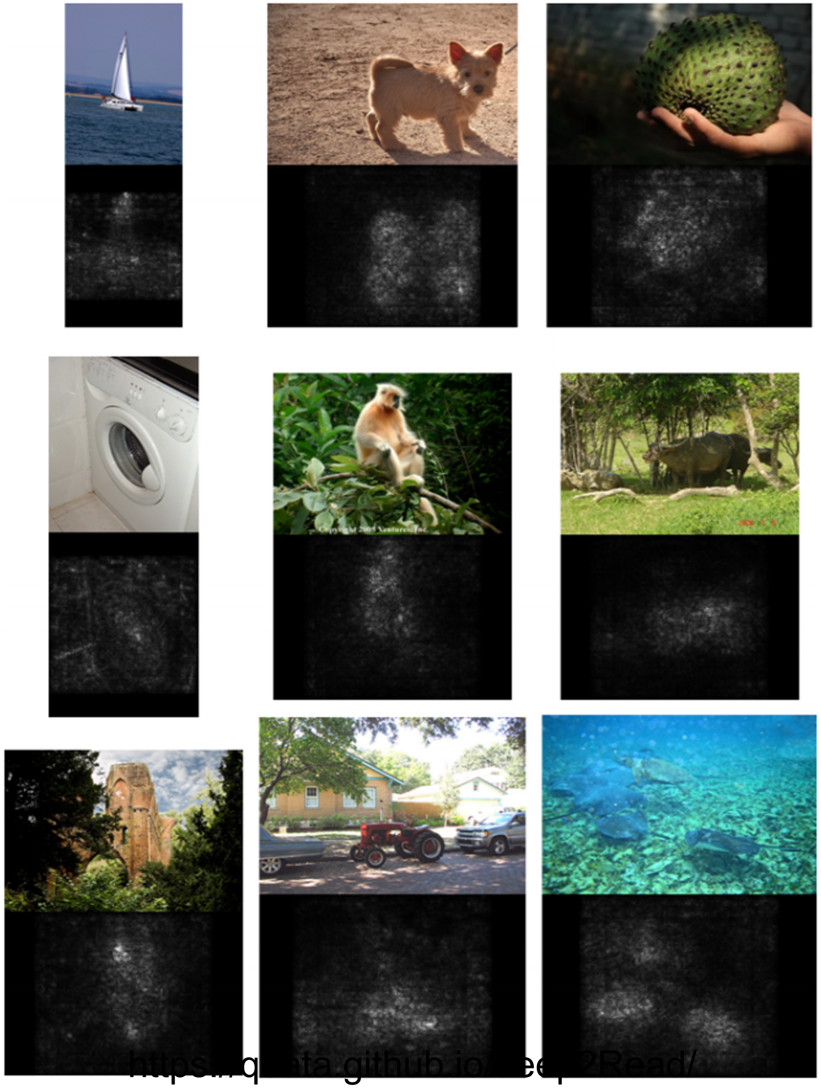
Baseball—or stripes?
mixed4a, Unit 6

Animal faces—or snouts?
mixed4a, Unit 240

Clouds—or fluffiness?
mixed4a, Unit 453

Buildings—or sky?
mixed4a, Unit 492

Instance-wise Feature Attribution: Saliency Maps

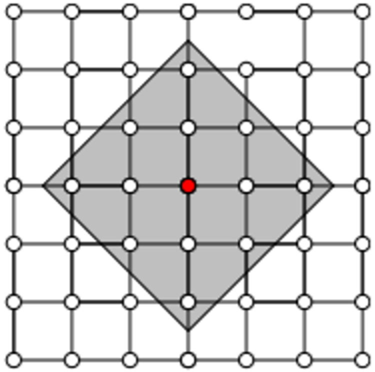


https://qdata.github.io/deep2Read/

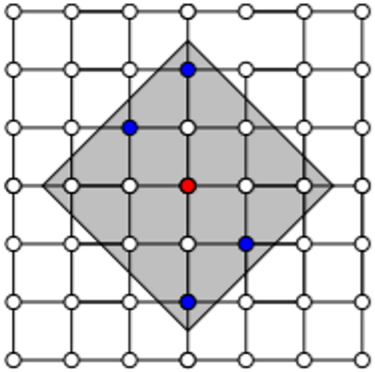
Feature Attribution: Shapley Approximation

- Shapley value: method of determining the contributions of different players
- Treat the features as players and the game is prediction
- For set of features S, define marginal contribution of feature i to S as F(S) - F(S \ {i})
- The Shapley value is the average marginal contribution across all possible subsets S containing i
- But there are 2^n possible subsets
- Lots of approximation methods

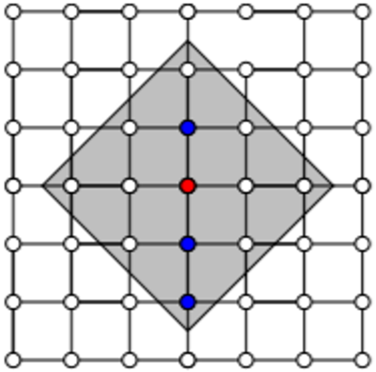
Feature Attribution: L-Shapley and C-Shapley



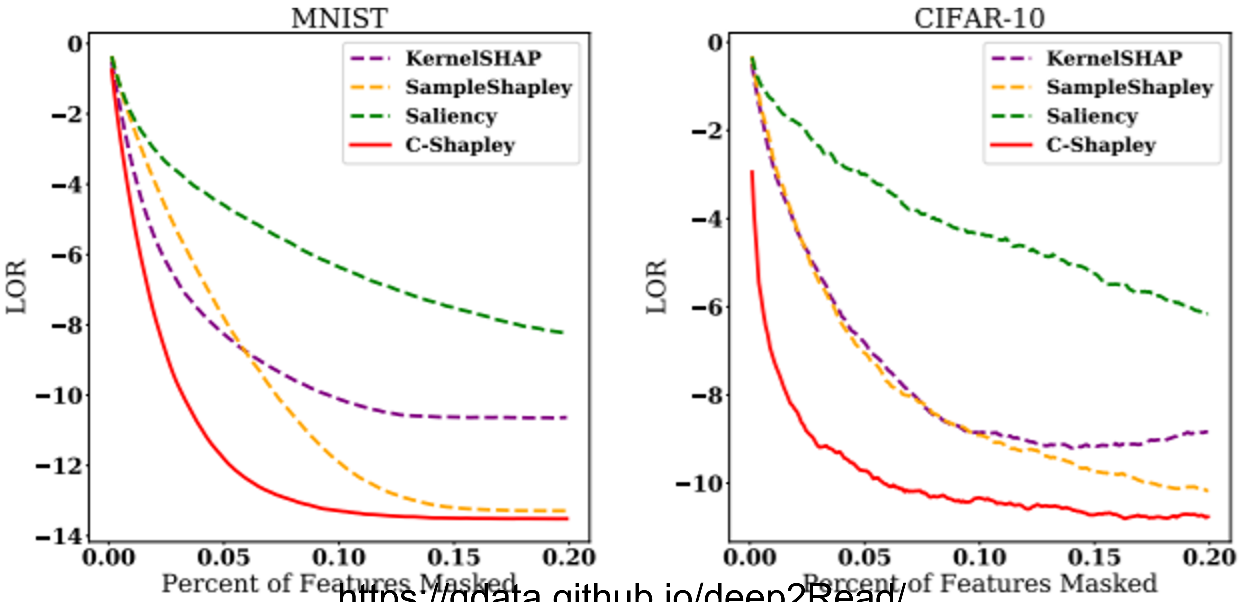
(a)



(b)

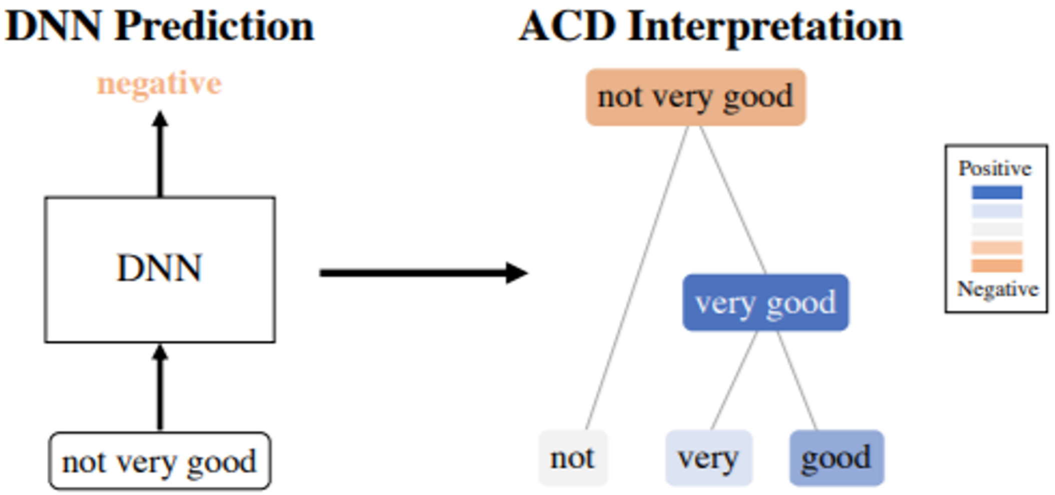


(c)



Feature Attribution: Contextual Decomposition (CD)

- CD decomposes logits into sum of importance measures of feature groups, other factors
- Captures both feature importance and interaction between features
- Can be used for hierarchical interpretations



https://qdata.github.io/deep2Read/

Feature Attribution: Contextual Decomposition (CD)

- CD also allows for easily penalizing certain features or groups of features

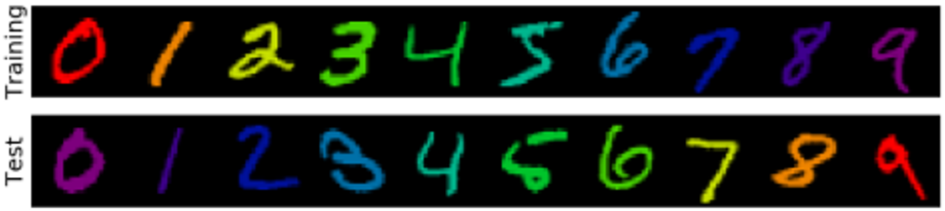


Figure 4: ColorMNIST: the test set shapes remain the same as the training set, but the colors are inverted. A vanilla network trained on this training set will get 0% accuracy on the test set.

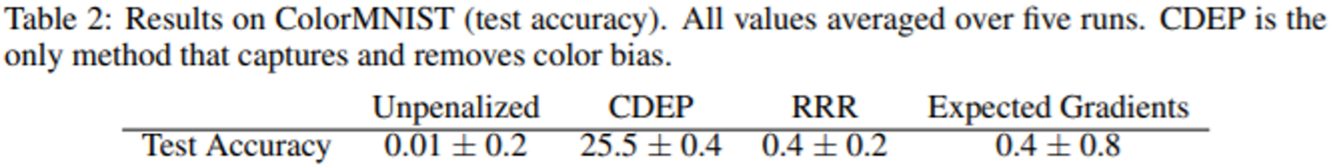
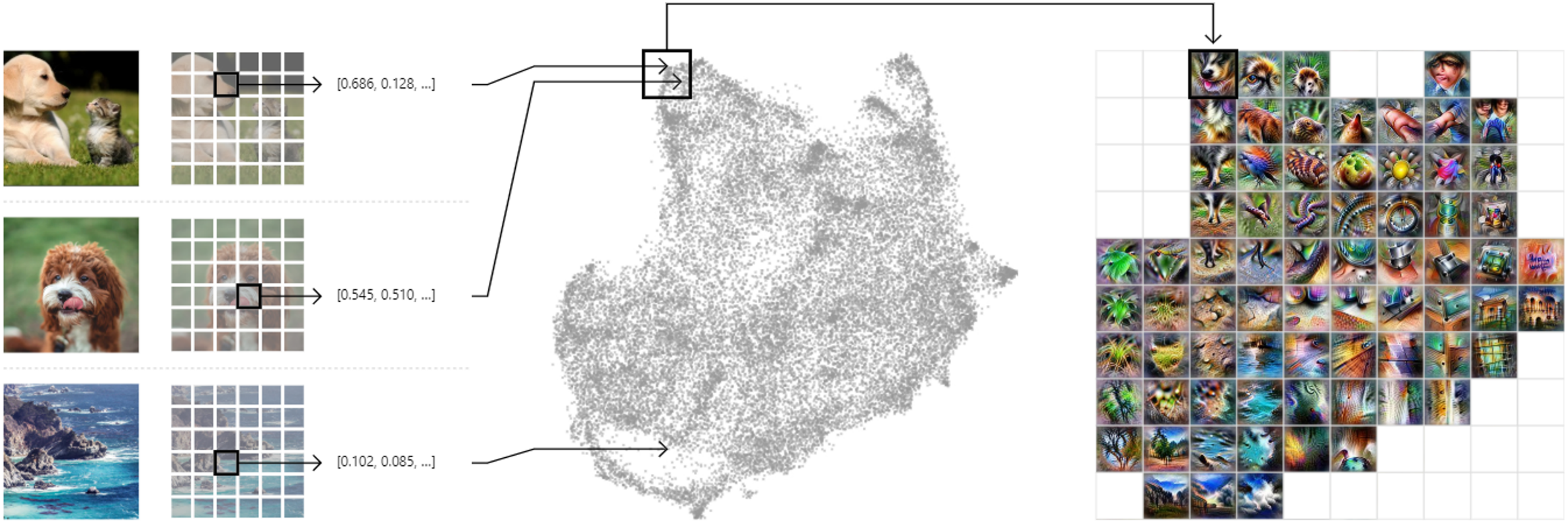


Table 2: Results on ColorMNIST (test accuracy). All values averaged over five runs. CDEP is the only method that captures and removes color bias.

	Unpenalized	CDEP	RRR	Expected Gradients
Test Accuracy	0.01 ± 0.2	25.5 ± 0.4	0.4 ± 0.2	0.4 ± 0.8

Visualization and Attribution: Activation Atlas

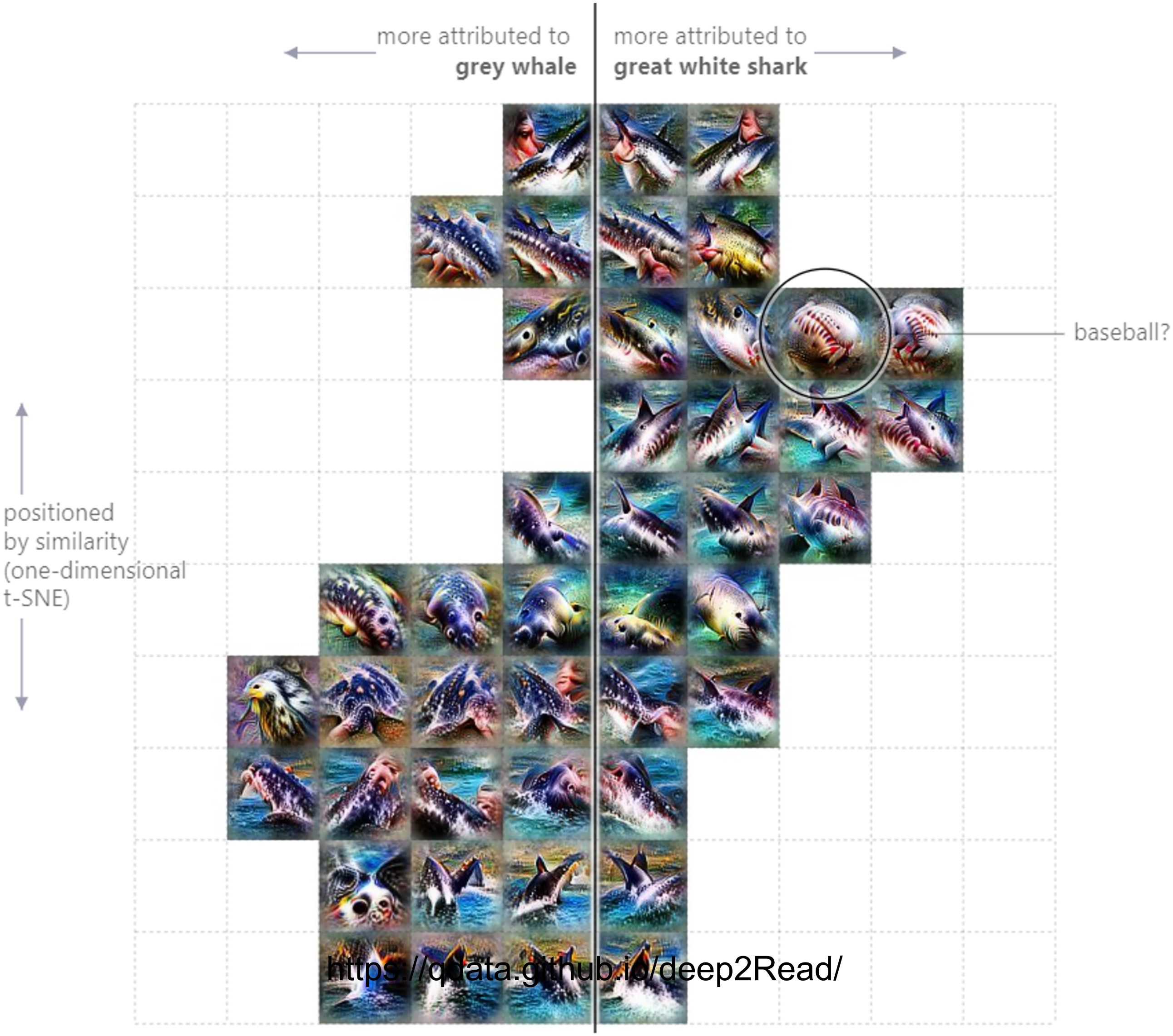


A randomized set of one million images is fed through the network, collecting one random spatial activation per image.

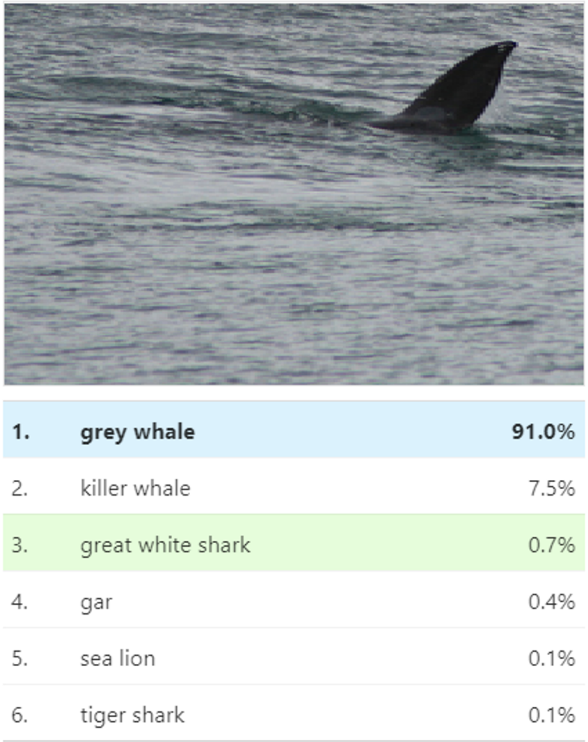
The activations are fed through UMAP to reduce them to two dimensions. They are then plotted, with similar activations placed near each other.

We then draw a grid and average the activations that fall within a cell and run feature inversion on the averaged activation. We also optionally size the grid cells according to the density of the number of activations that are averaged within.

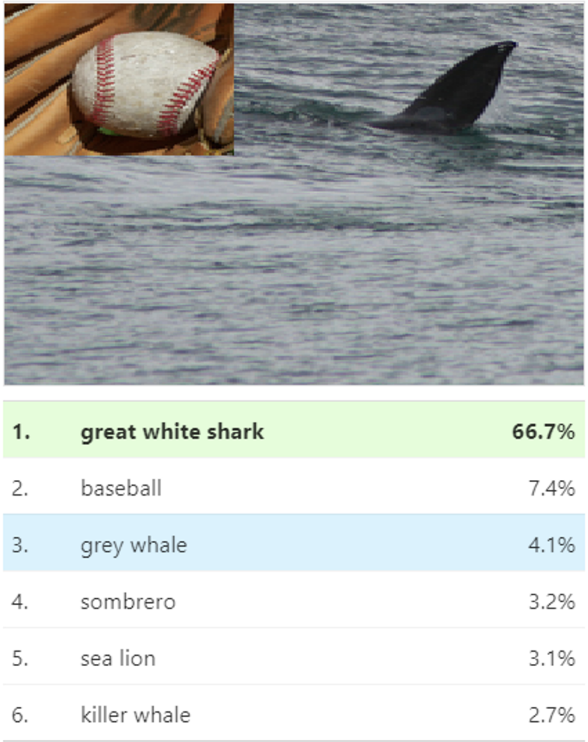
Visualization and Attribution: Activation Atlas


more attributed to grey whale
more attributed to great white shark
baseball?
positioned by similarity (one-dimensional t-SNE)

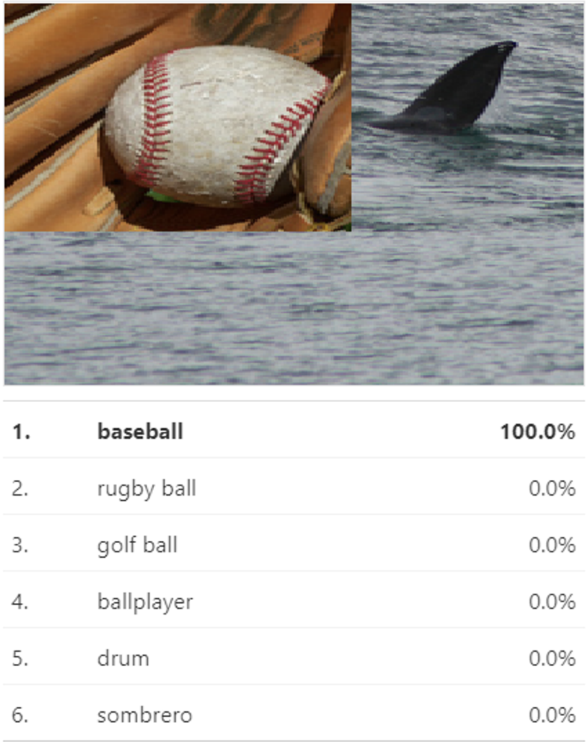
Visualization and Attribution: Activation Atlas



1.	**grey whale**	**91.0%**
2.	killer whale	7.5%
3.	great white shark	0.7%
4.	gar	0.4%
5.	sea lion	0.1%
6.	tiger shark	0.1%



1.	**great white shark**	**66.7%**
2.	baseball	7.4%
3.	grey whale	4.1%
4.	sombrero	3.2%
5.	sea lion	3.1%
6.	killer whale	2.7%

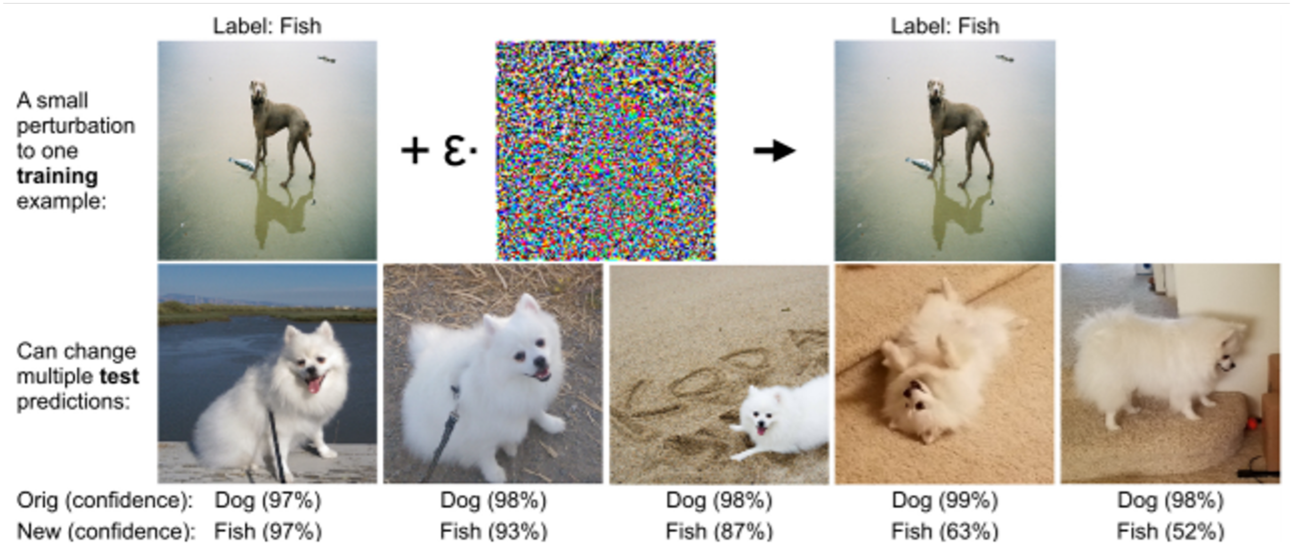


1.	**baseball**	**100.0%**
2.	rugby ball	0.0%
3.	golf ball	0.0%
4.	ballplayer	0.0%
5.	drum	0.0%
6.	sombrero	0.0%

Training Example Attribution: Influence Function

- Influence of upweighting training example z on loss at test point z_{test}:

$$-\nabla_\theta L(z_{\text{test}}, \hat{\theta})^\top H_{\hat{\theta}}^{-1} \nabla_\theta L(z, \hat{\theta}).$$

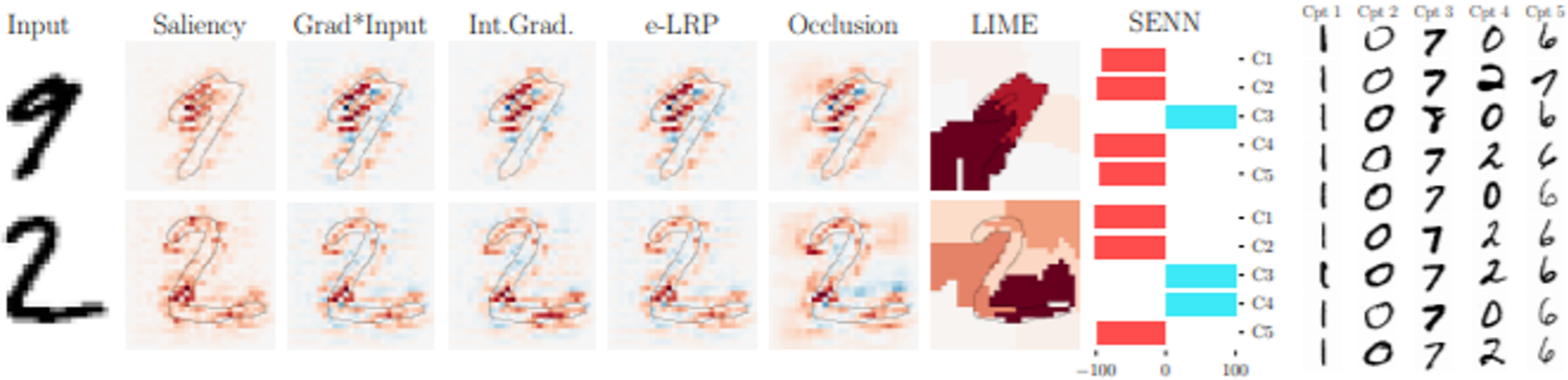

Label: Fish
A small perturbation to one **training** example:
+ ε·
Label: Fish
Can change multiple **test** predictions:
Orig (confidence): Dog (97%) Dog (98%) Dog (98%) Dog (99%) Dog (98%)
New (confidence): Fish (97%) Fish (93%) Fish (87%) Fish (63%) Fish (52%)

https://qdata.github.io/deep2Read/

Inherently Interpretable Models: Desiderata for Transparency

- Simulatability: Human able to internally simulate and reason about entire decision-making process
  - Examples: linear regression, decision trees
- Decomposability: Each part of model has an intuitive explanation
- Algorithmic Transparency: Do we understand the shape of the error surface?

Inherently Interpretable Models: SENN



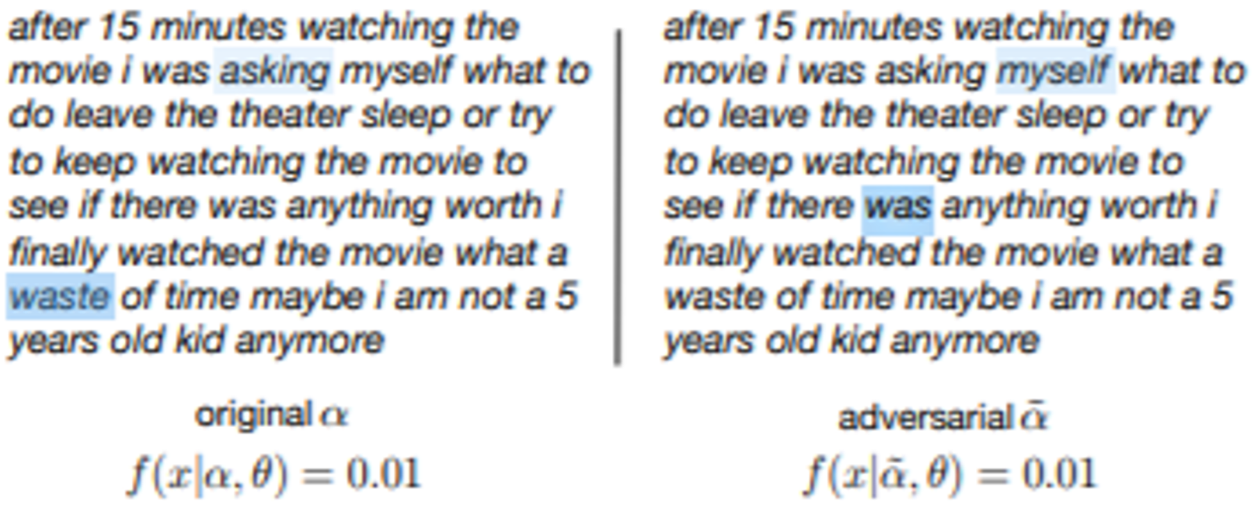
https://qdata.github.io/deep2Read/

Are attention weights explanation?

- Many in NLP have claimed attention weights provide insight into what model is looking at
- They imply the weights provide an explanation for why the model makes its decision
- But can attention weights really be viewed as an explanation?

"Attention is Not Explanation"

- Attention does not provide meaning explanation since:
  - ○ 1. Attention weights are frequently uncorrelated with feature importance scores
  - ○ 2. Can identify adversarial attention distributions which yield same prediction, very different explanation



after 15 minutes watching the movie i was asking myself what to do leave the theater sleep or try to keep watching the movie to see if there was anything worth i finally watched the movie what a waste of time maybe i am not a 5 years old kid anymore

original α

$f(x|\alpha, \theta) = 0.01$

after 15 minutes watching the movie i was asking myself what to do leave the theater sleep or try to keep watching the movie to see if there was anything worth i finally watched the movie what a waste of time maybe i am not a 5 years old kid anymore

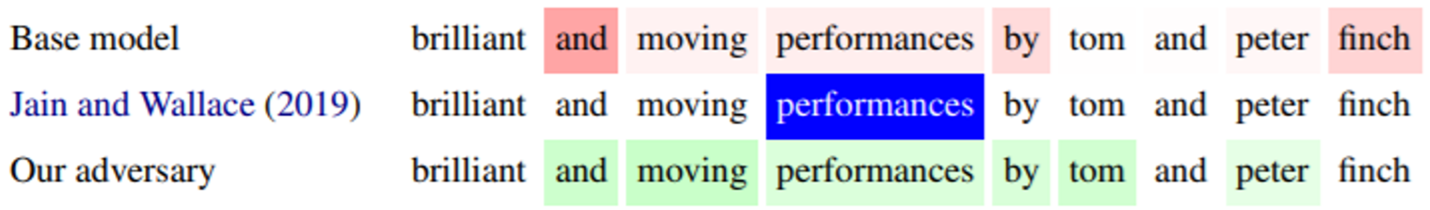
adversarial $\tilde{\alpha}$

$f(x|\tilde{\alpha}, \theta) = 0.01$

https://qdata.github.io/deep2Read/

“Attention is Not Not Explanation”

- Argues Claim 2 does not advance thesis, since:
    - Attention distribution is not a primitive
        - Must train model with adversarial objective
    - Existence doesn’t entail exclusivity
        - Provides *an* explanation, not *the* explanation



Base model	brilliant	and	moving	performances	by	tom	and	peter	finch
Jain and Wallace (2019)	brilliant	and	moving	performances	by	tom	and	peter	finch
Our adversary	brilliant	and	moving	performances	by	tom	and	peter	finch

Figure 2: Attention maps for an IMDb instance (all predicted as positive with score $>$ 0.998), showing that in practice it is difficult to learn a distant adversary which is consistent on all instances in the training set.

“Attention is Not Not Explanation”

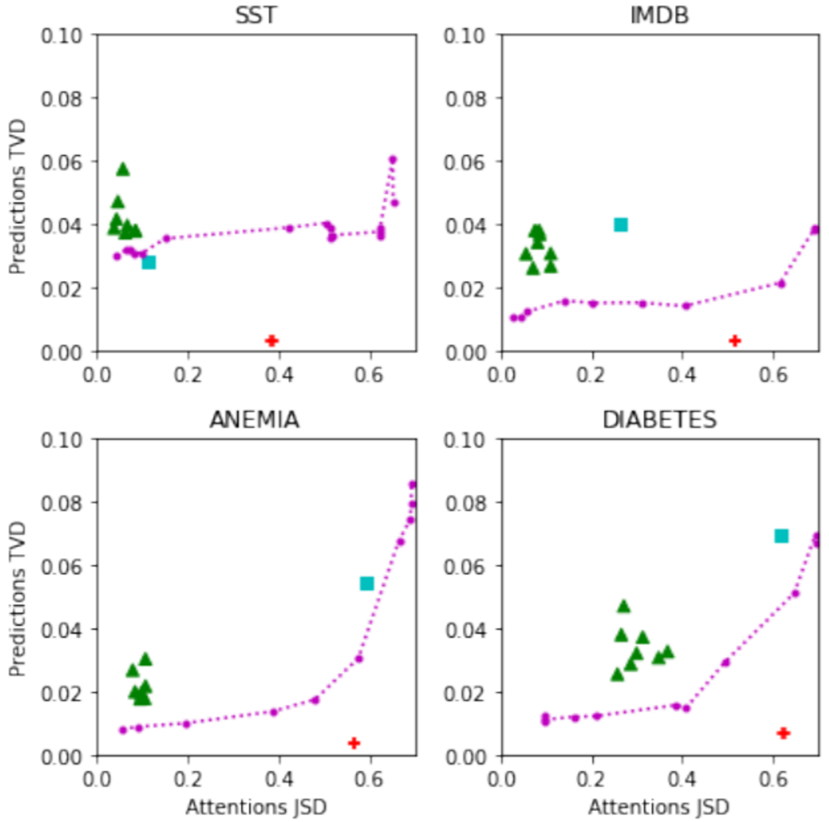
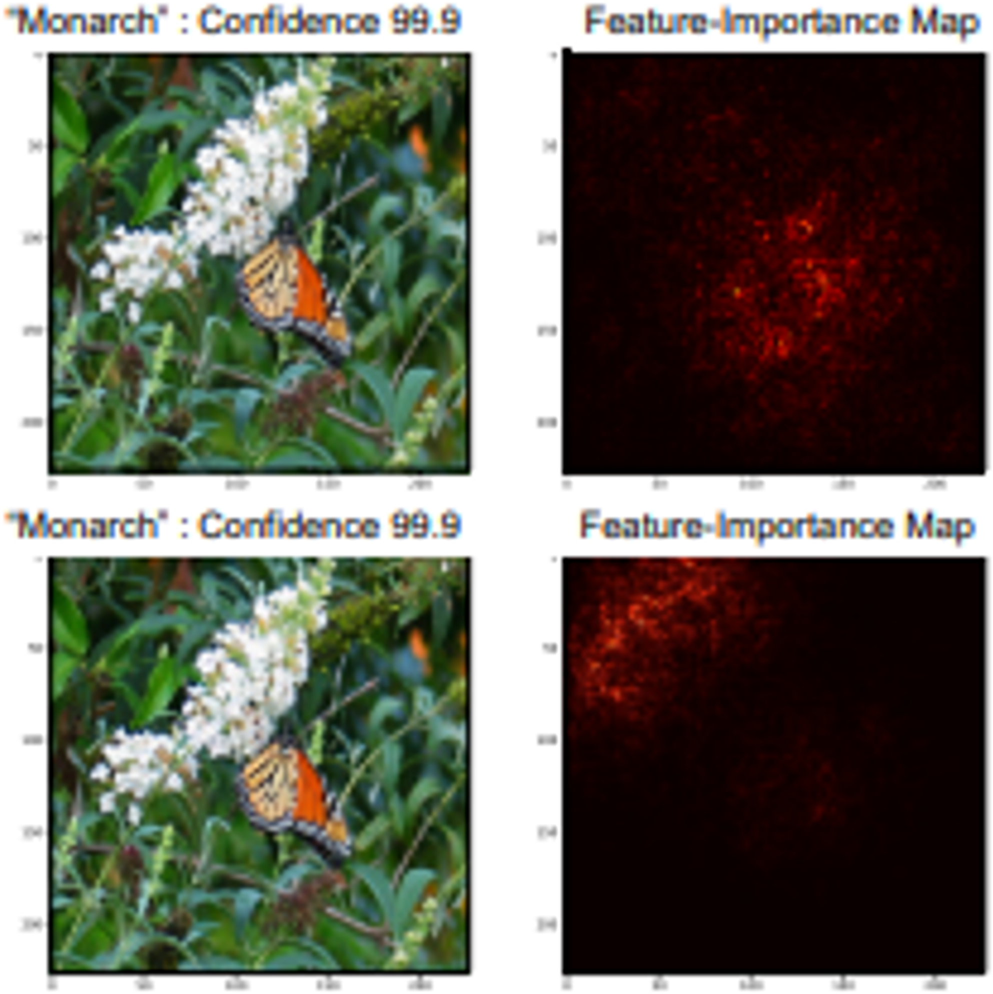


Figure 5: Averaged per-instance test set JSD and TVD from base model for each model variant. JSD is bounded at ~ 0.693. ▲: random seed; ■: uniform weights; dotted line: our adversarial setup as λ is varied; +: adversarial setup from Jain and Wallace (2019).

https://qdata.github.io/deep2Read/

Interpretation Robustness: Interpretation is Fragile

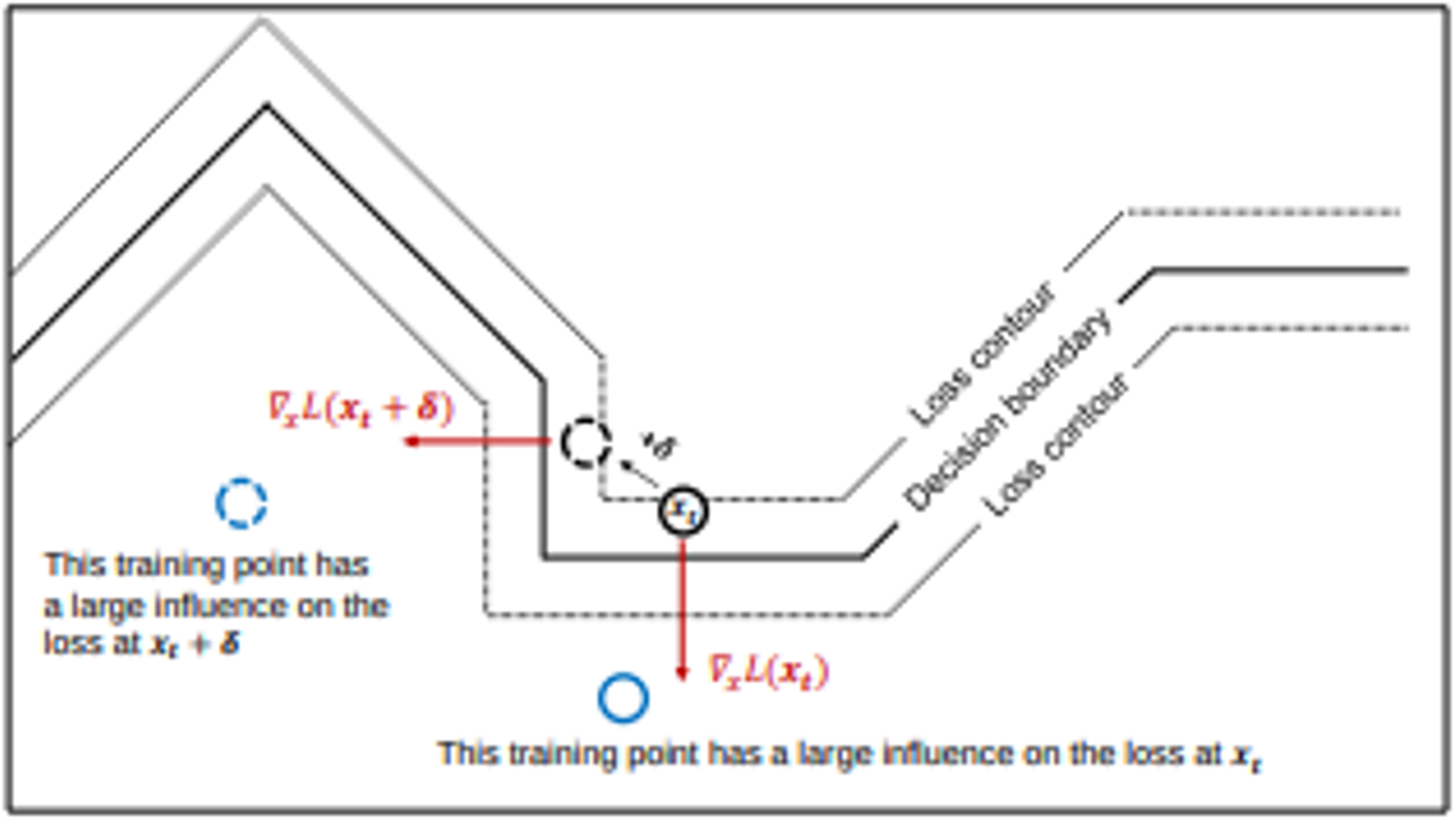


https://qdata.github.io/deep2Read/

Interpretation Robustness: Interpretation is Fragile

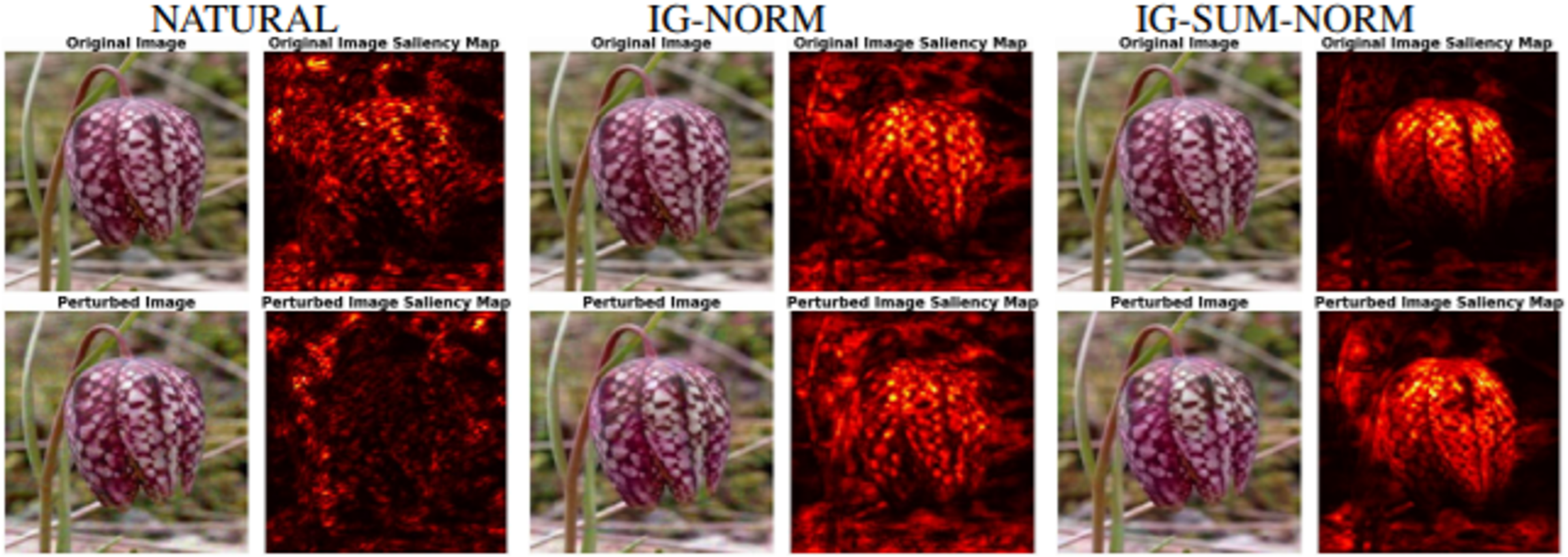
- Series of steps in direction which maximizes differentiable dissimilarity function between original, perturbed interpretation
  - Top-k attack: Decreases relative importance of k most important features
  - Mass-center attack for image data: maximizes spatial displacement of center of mass of feature importance map
  - Targeted attack for image data: Increases concentration of feature importance scores in pre-defined region of image

Interpretation Robustness: Interpretation is Fragile


$\nabla_x L(x_t + \delta)$
$+\delta$
x_t
Loss contour
Decision boundary
Loss contour
This training point has a large influence on the loss at $x_t + \delta$
$\nabla_x L(x_t)$
This training point has a large influence on the loss at x_t

https://qdata.github.io/deep2Read/

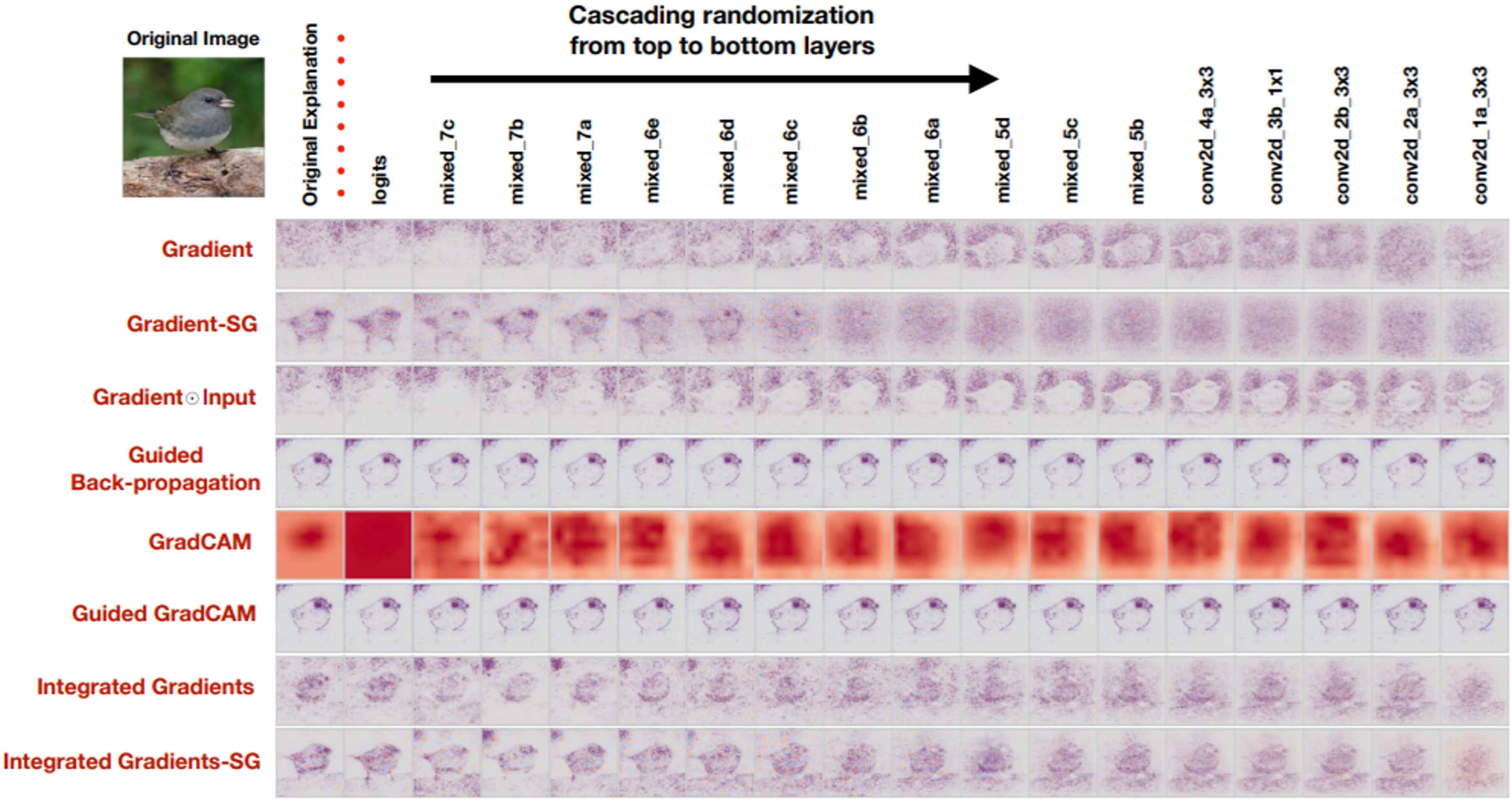
Interpretation Robustness: Robust Attribution Regularization


NATURAL
Original Image
Original Image Saliency Map
Perturbed Image
Perturbed Image Saliency Map
IG-NORM
Original Image
Original Image Saliency Map
Perturbed Image
Perturbed Image Saliency Map
IG-SUM-NORM
Original Image
Original Image Saliency Map
Perturbed Image
Perturbed Image Saliency Map

NATURAL	IG-NORM	IG-SUM-NORM
Top-1000 Intersection: 0.1% Kendall's Correlation: 0.2607	Top-1000 Intersection: 58.8% Kendall's Correlation: 0.6736	Top-1000 Intersection: 60.1% Kendall's Correlation: 0.6951

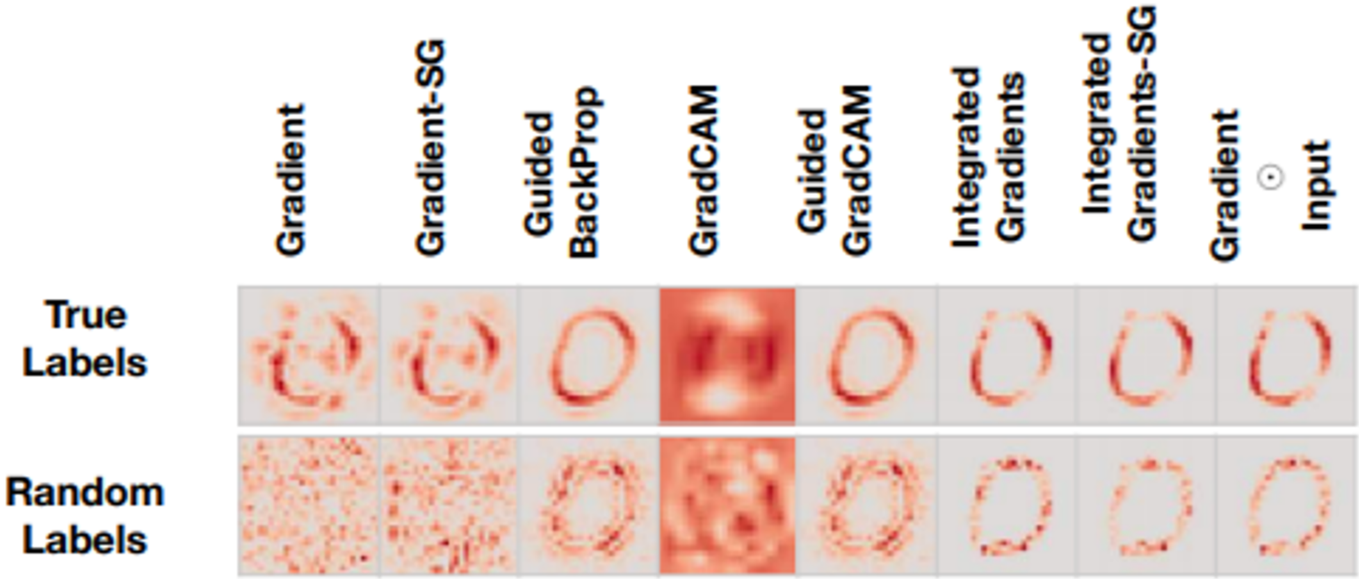
$$\underset{\theta}{\text{minimize}} \quad \underset{(\boldsymbol{x},y)\sim P}{\mathbb{E}}\left[\max_{\boldsymbol{x}'\in N(\boldsymbol{x},\varepsilon)}\left\{\ell(\boldsymbol{x}',y;\theta)+\beta\|\,\mathrm{IG}^{\ell_y}(\boldsymbol{x},\boldsymbol{x}')\|_1\right\}\right]$$

Sanity Checks For Saliency Maps



https://qdata.github.io/deep2Read/

Sanity Checks For Saliency Maps



https://qdata.github.io/deep2Read/

Where do we go from here?

- Consensus, clarity on definition of suitcase words such as “interpretation” and “explanation”
- Better and more standardized evaluation methods
- Applying techniques similar to the Activation Atlas to text, audio
- Acknowledgment and discussion of tradeoffs
  - Inherently interpretable models vs. post-hoc explanations
  - PDR
- Continued exploration of better methods of both feature visualization and attribution, particularly understanding feature interaction