Towards falsifiable interpretability research

Ari Morcos

CVPR 2021 Tutorial on Interpretable ML for CV

Matthew Leavitt
Former Facebook AI Resident

FACEBOOK AI

Interpretability can be viewed as building human intuition and understanding for complex "black box" models

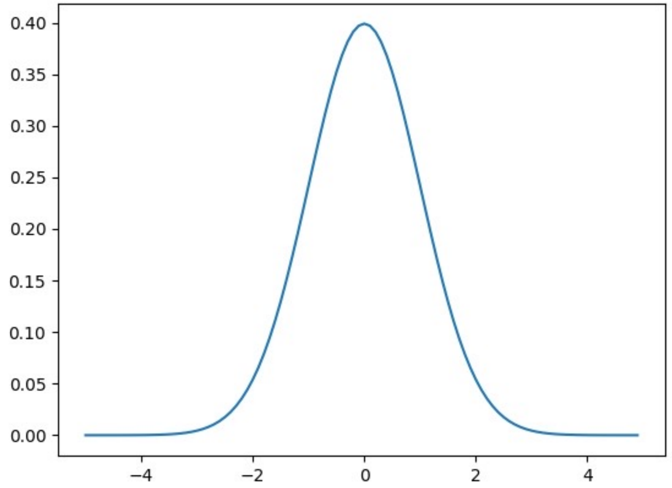
Intuition is critical for understanding, but unverified intuition can be misleading and lead us astray

In so far as a scientific statement speaks about reality, it must be falsifiable; and in so far as it is not falsifiable, it does not speak about reality.

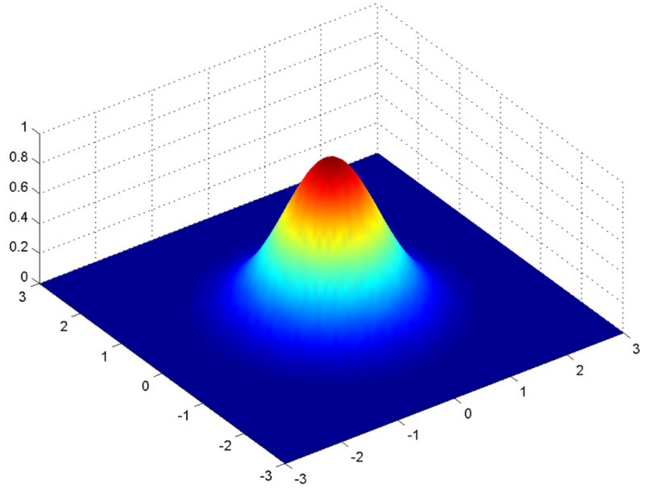
- Karl Popper
  The Logic of Scientific Discovery, 1959

On the perils of unverified intuition – the Gaussian distribution in high-dimensional spaces

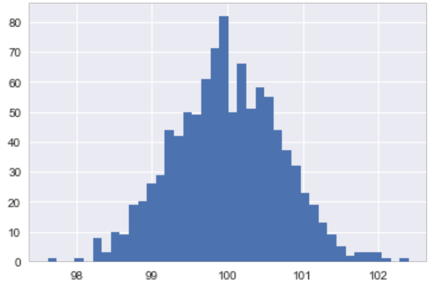
1-dimensional Gaussian



2-dimensional Gaussian



10k-dimensional Gaussian



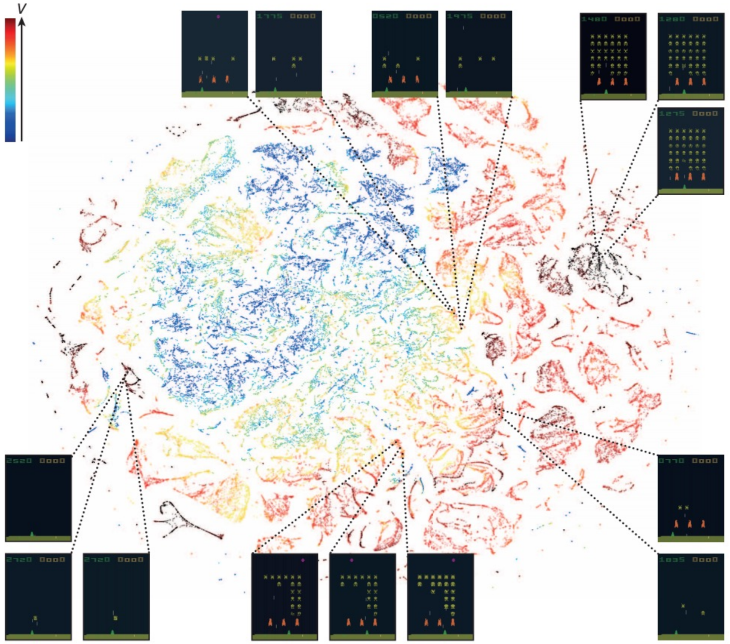
Sample norm

For an accessible description, see Ferenc Huszár's blog on the topic:
https://www.inference.vc/high-dimensional-gaussian-distributions-are-soap-bubble/

Building intuition with ML visualizations

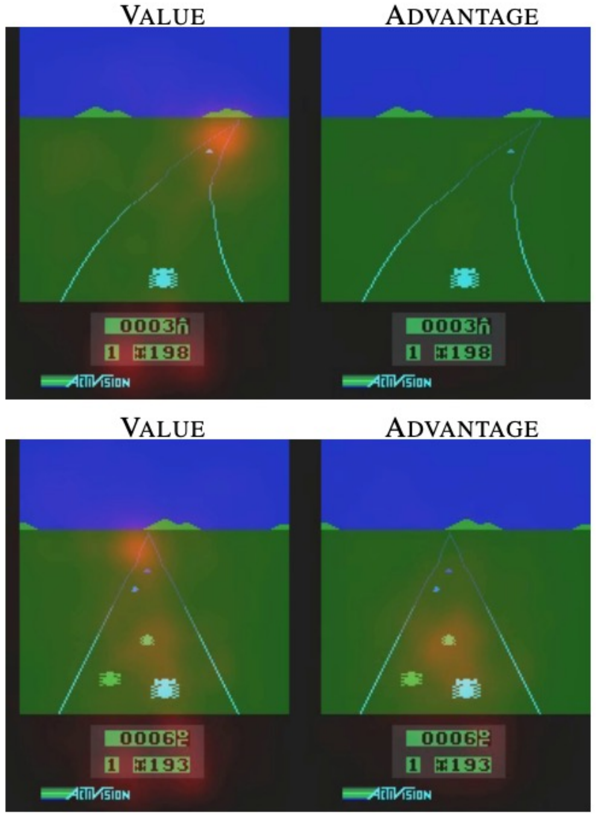
t-SNE

(van der Maaten and Hinton, *JMLR* 2008)



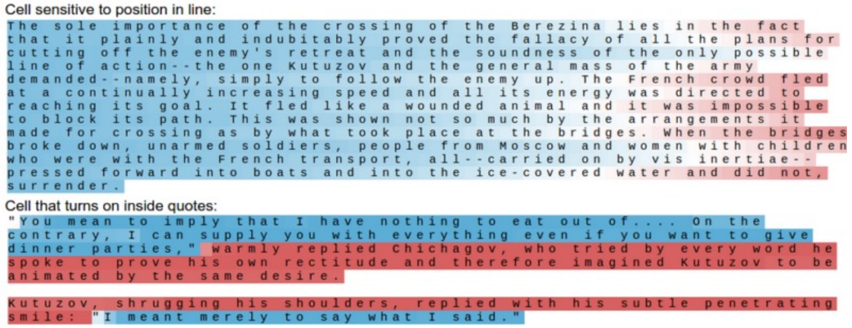
Mnih et al., *Nature* 2015

Saliency

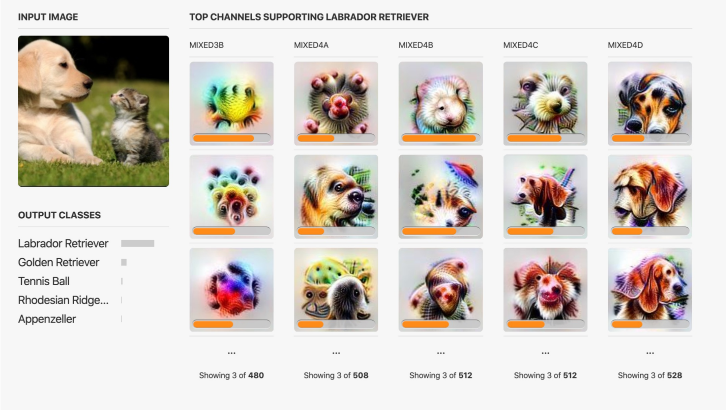


Wang et al., *ICML* 2016

Single neuron visualization



Karpathy et al., 2015



Olah et al., *Distill* 2017

Outline

- Impediments to falsifiable interpretability research
- Case study 1: The misdirection of saliency
- Case study 2: Understanding networks through easy-to-interpret neurons
- Building better hypotheses

Goal

Build intuition so that we can better understand DNNs, but ensure that this intuition is grounded in rigorous, falsifiable experiments

Impediments to falsifiable interpretability research

1. Underemphasis on clear, specific, testable, and most critically, falsifiable hypotheses

2. Interpretability methods are often not verified as being important for DNN function

3. The double-edged sword of visualization
   - Visualization can be extremely helpful for building understanding, especially during exploration
   - However, it can also lead to strong feelings of comprehension regardless of how accurate the visualization may be, especially since visualizations often feature individual examples

Impediments to falsifiable interpretability research - continued

4. Lack of quantification
    - Relates directly to lack of verification and the double-edged sword of visualization
    - Especially important in deep learning where similar models can have dramatically different properties across different hyperparameters
    - Without quantification, visualization can serve as a Rorschach test for a researcher

5. For interpretability methods designed with the intent to provide actionable explanations to humans, it is critical to conduct controlled experiments to evaluate the utility of these methods

Case study 1:
The misdirection of saliency

Introduction to saliency/feature attribution visualizations

Which portions of the input are most responsible for a given classification decision?

Typically based on some form of the gradient of a given output wrt the input

Simonyan et al., *ICLR Workshop* 2013

Improved saliency methods

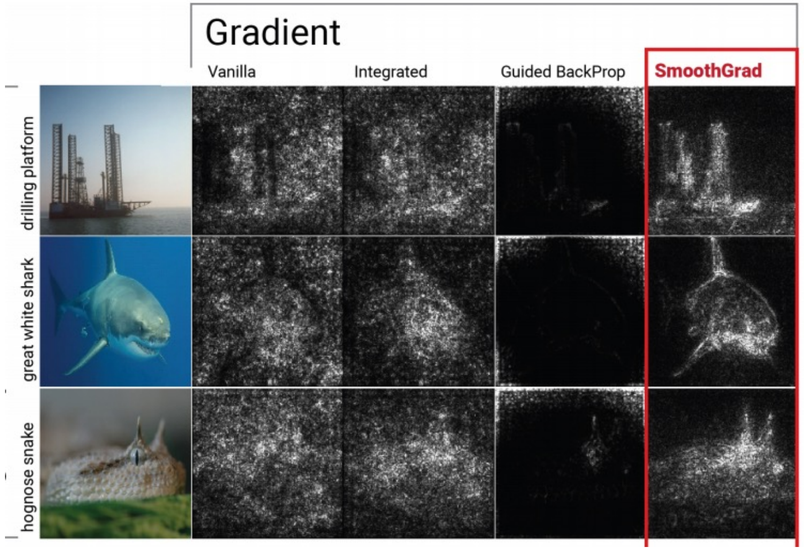
Input image

Guided Grad-CAM
“Cat”

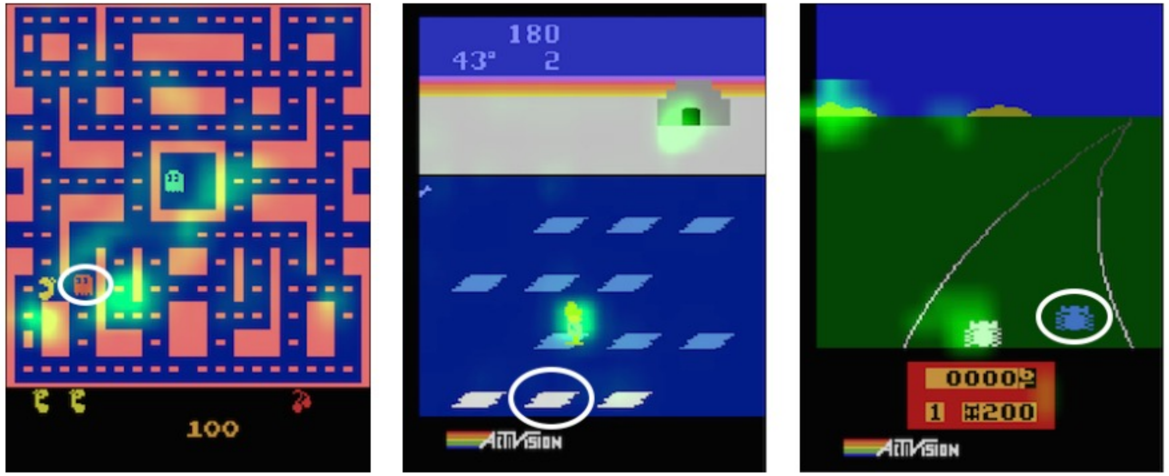
Guided Grad-CAM
“Dog”

Grad-CAM
(Selvaraju et al., ICCV 2017)

SmoothGrad
(Smilkov et al., *ICML Workshop on Visualization for DL* 2017)

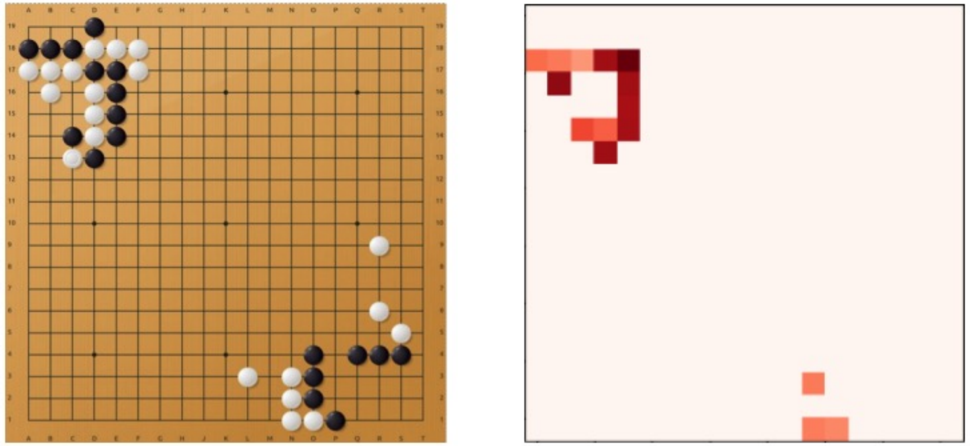


Saliency for model explainability in RL



(a) MsPacman (b) Frostbite (c) Enduro

Greydanus et al., *ICML* 2018

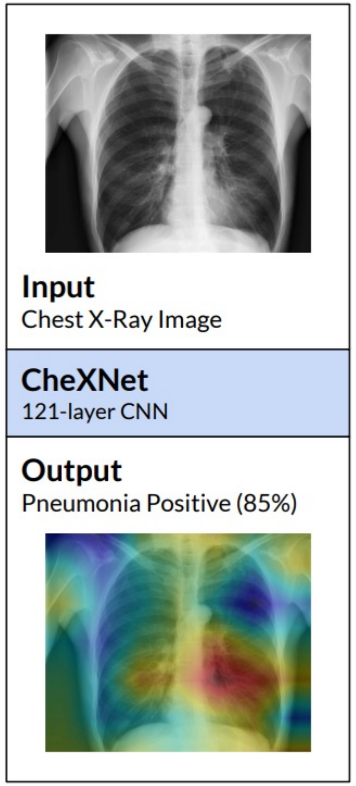


(a) Original Position (b) SARFA

Puri et al., *ICLR* 2020

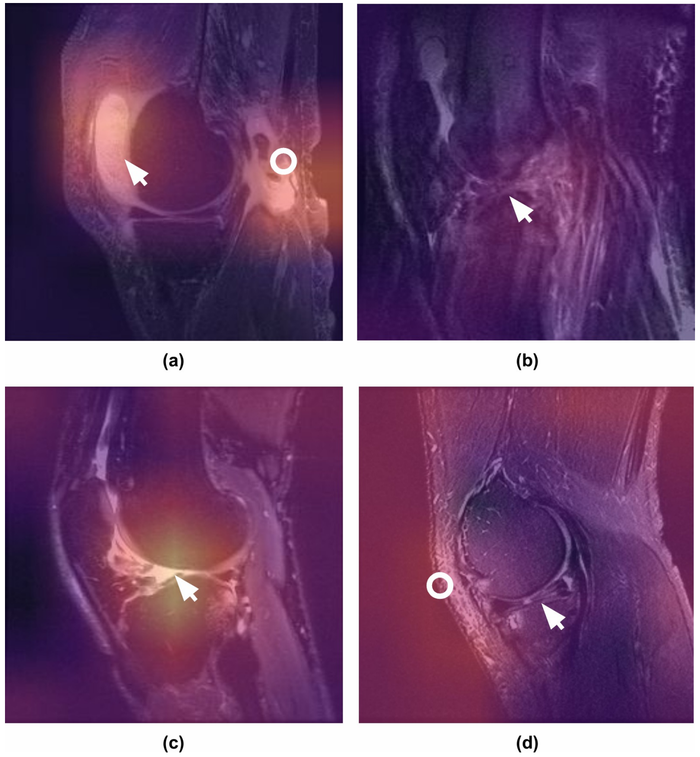
Saliency in medical imaging

Pneumonia detection from chest X-Rays



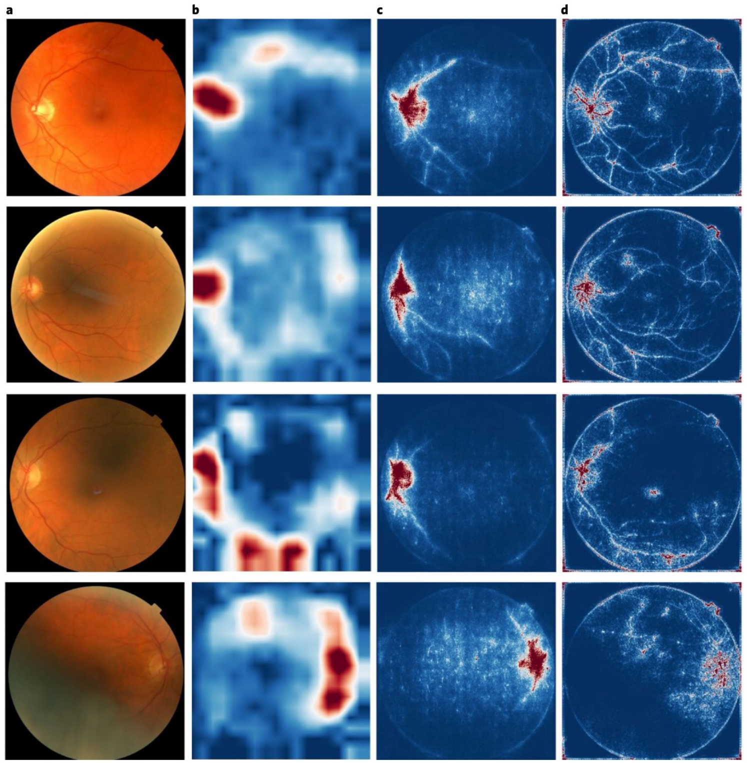
Rajpurkar et al., 2017

Interpretation of knee MRIs



Bien et al., *PLOS Medicine* 2018

Detecting anaemia from retinal fundus images



Mitani et al., *Nat Biomed Eng* 2020

Insights from integrated gradients: what axioms should saliency methods satisfy?

Axiom 1: Sensitivity

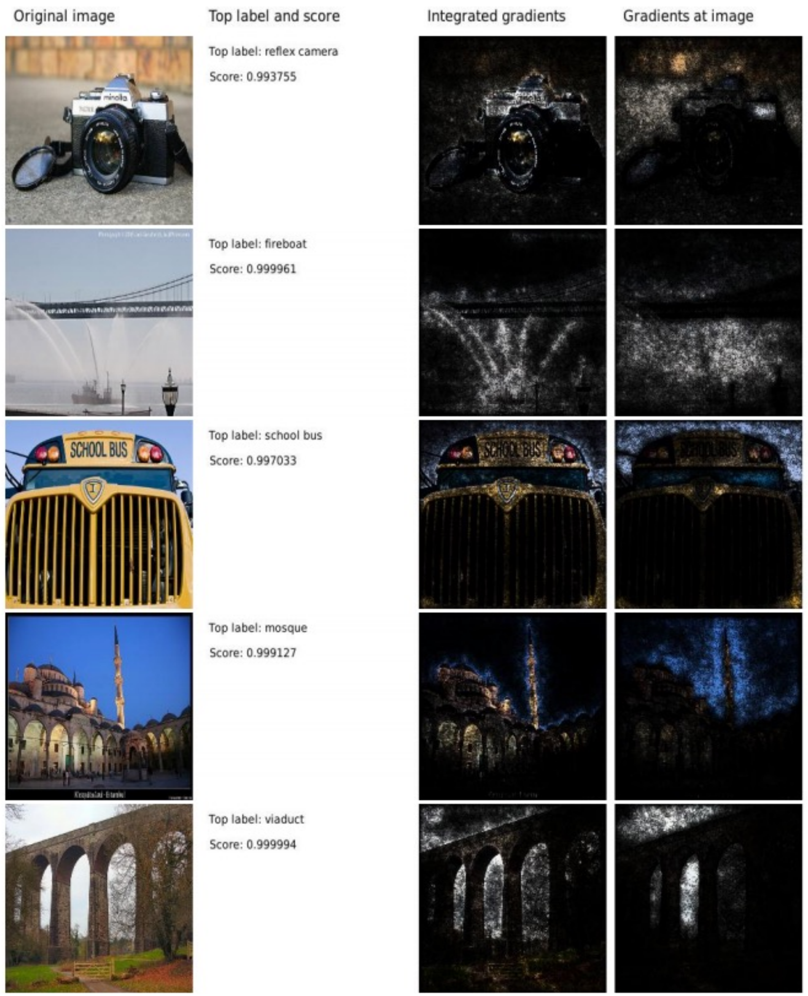
- If two inputs differ in only one feature but result in different predictions, the differing feature should have non-zero saliency
- Standard gradient-based saliency fails this axiom because ReLUs can result in zero gradient despite different inputs if, for example, the pre-ReLU activation is less than 0

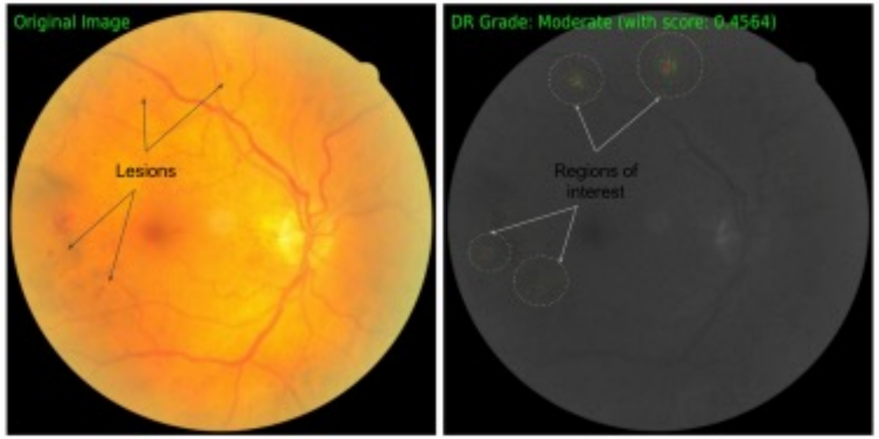
Axiom 2: Implementation invariance

- If two networks are *functionally equivalent* (produce identical output in response to all inputs), their saliency maps should also be identical
  - In other words, the saliency map should be invariant to the implementation of the function
- Standard gradient-based saliency satisfies this condition, but many more complicated saliency methods do not

Sundararajan et al., *ICML* 2017

Insights from integrated gradients: what axioms should saliency methods satisfy?



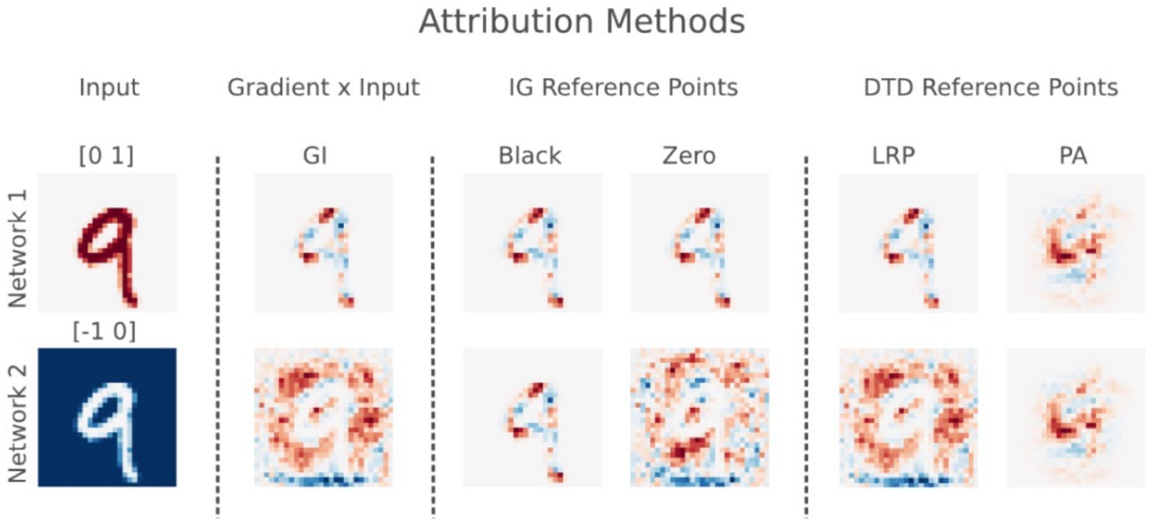


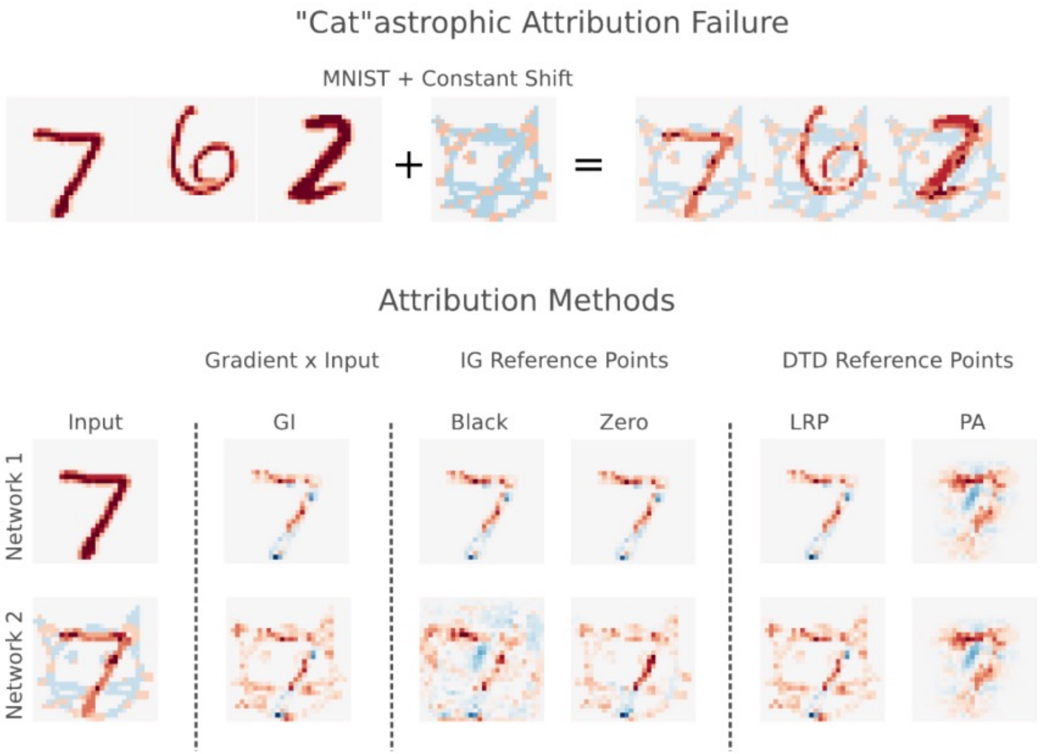
Sundararajan et al., *ICML* 2017

Many saliency methods are not invariant to input transformations

Axiom: input invariance

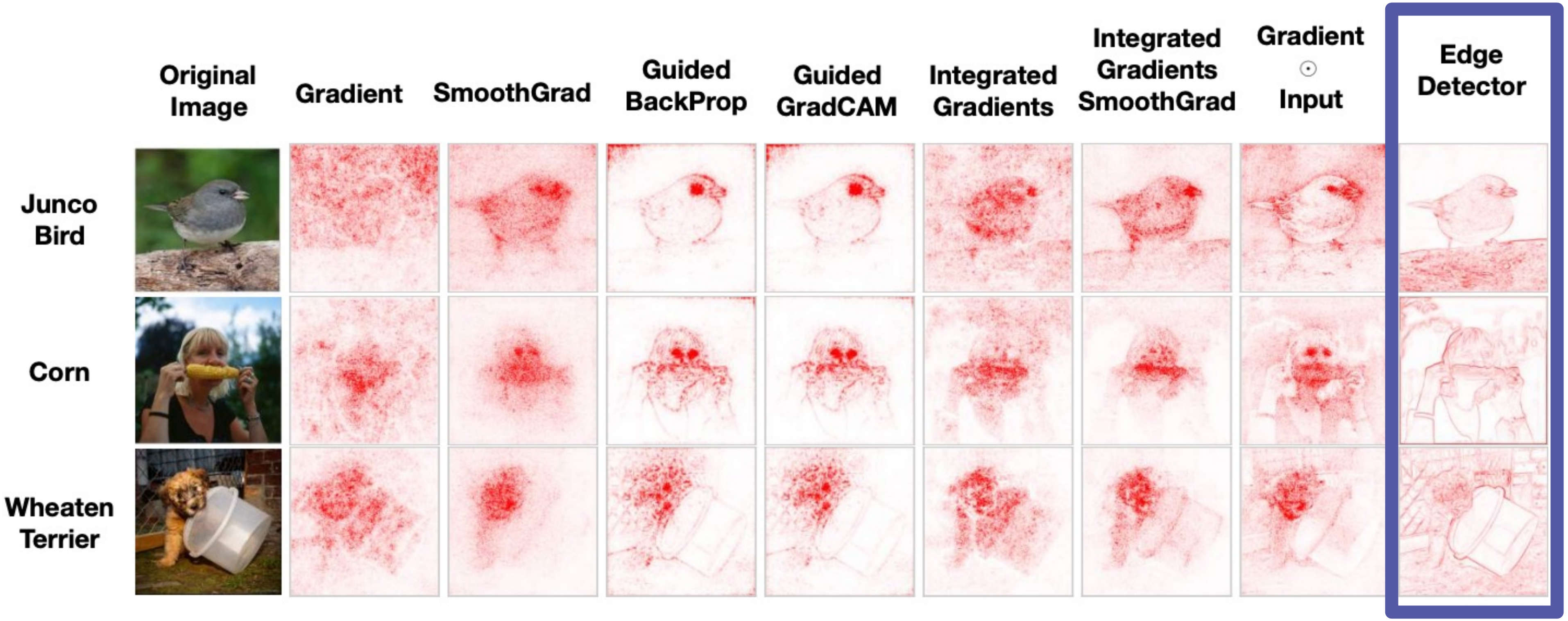
- If a constant shift is applied to all inputs, the attribution method should not change





Kindermans et al., 2019

Many saliency methods produce maps which are very similar to edge detectors



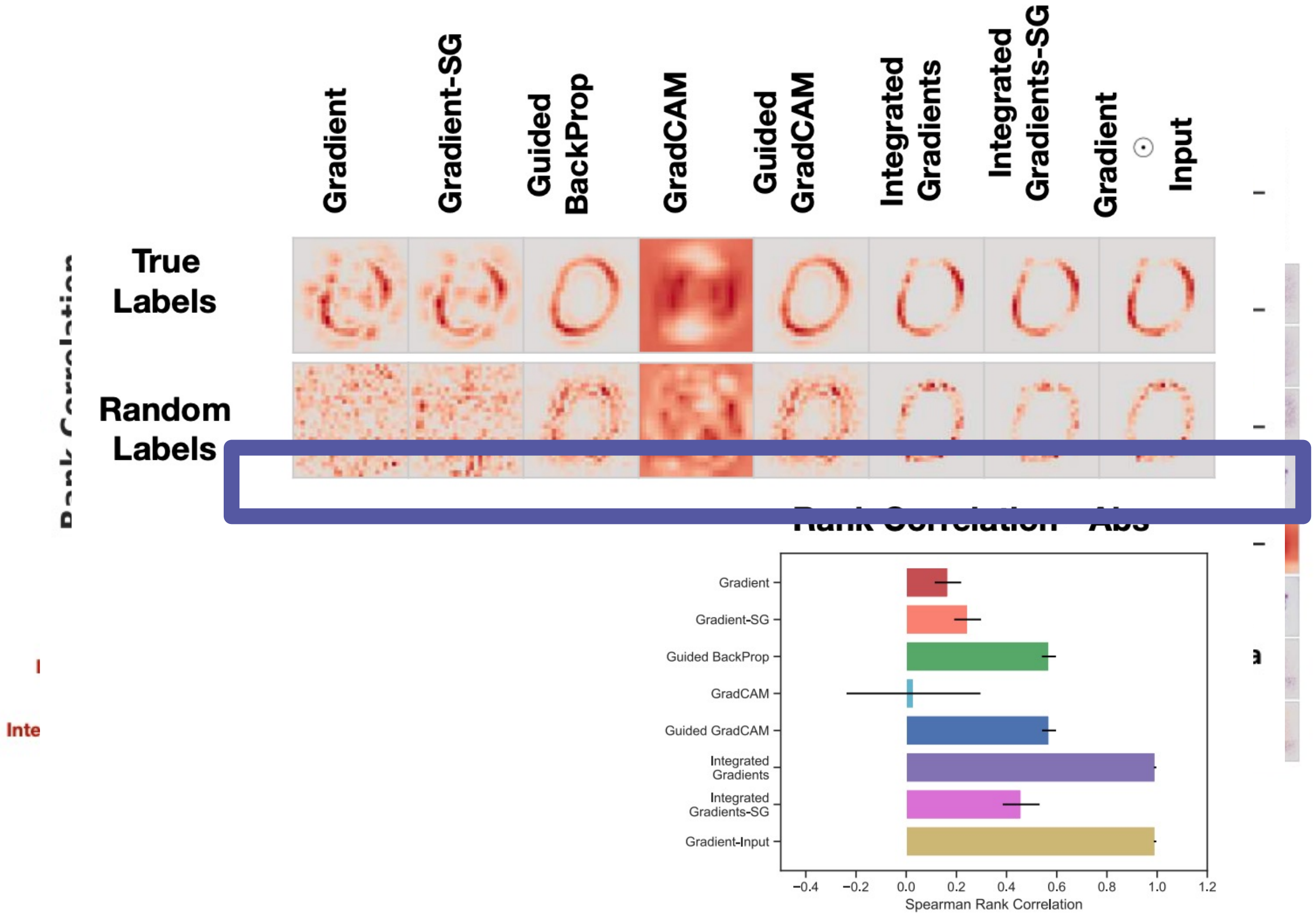
Adebayo et al., *NeurIPS* 2018

Many saliency methods fail randomization sanity checks

Two randomization tests

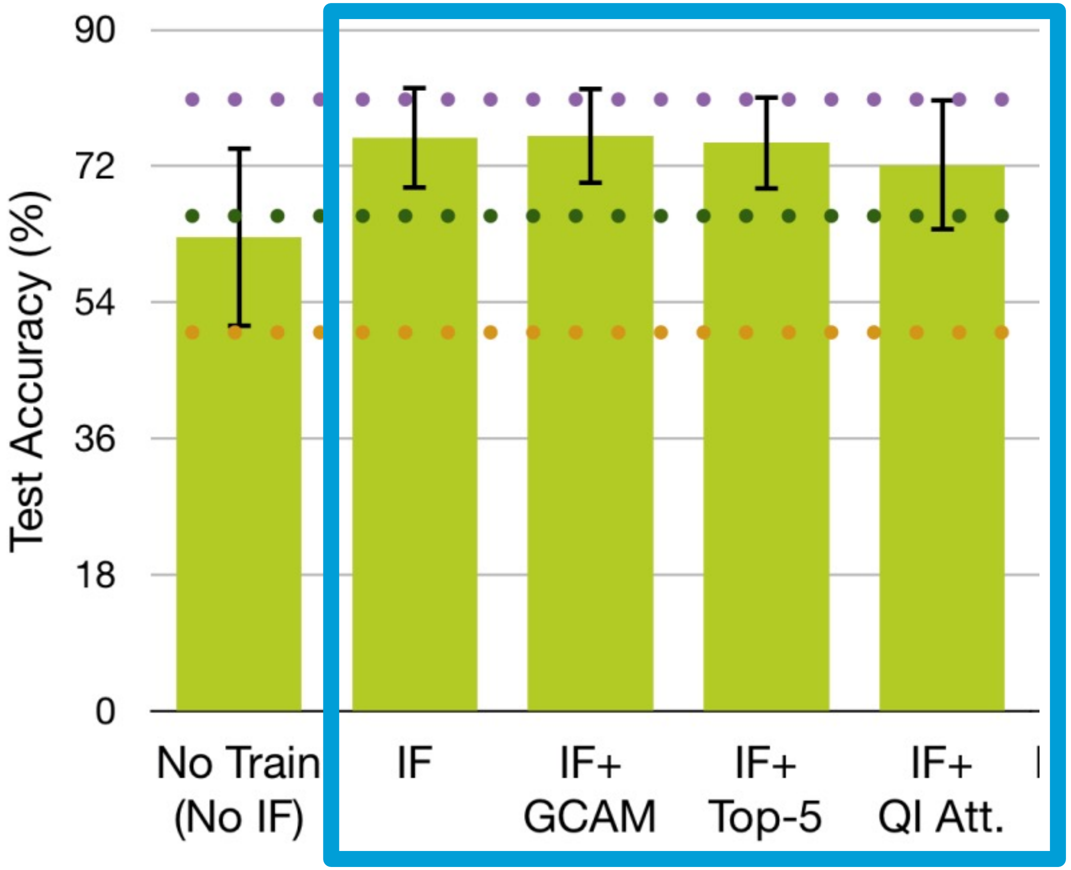
- **Model parameter randomization:** If a saliency method depends on the learned parameters of the model, the saliency maps for a randomly initialized, untrained network and a trained network should be very different

- **Data randomization:** If a saliency method depends on the data labels (i.e., $p(y|x)$), then the saliency maps for a model trained on data with randomly permuted labels should be different from the maps generated against the uncorrupted dataset

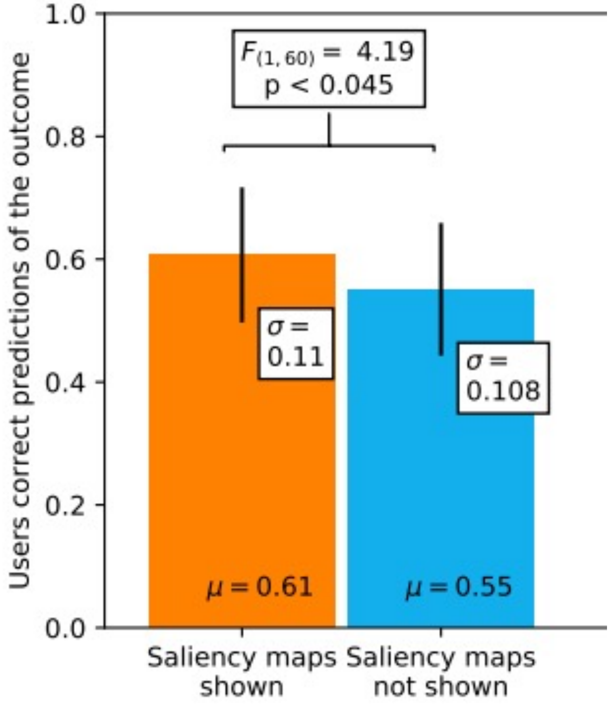


Adebayo et al., *NeurIPS* 2018

Do saliency maps help humans better predict DNN outputs?



Chandrasekaran et al., *CVPR* 2017



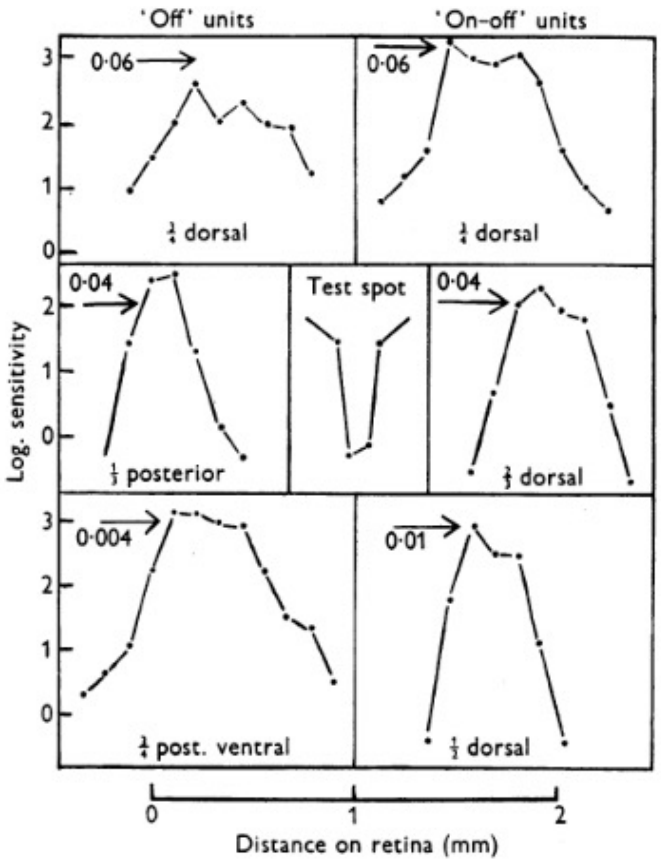
Alqaraawi et al., *IUI* 2020

Takeaways

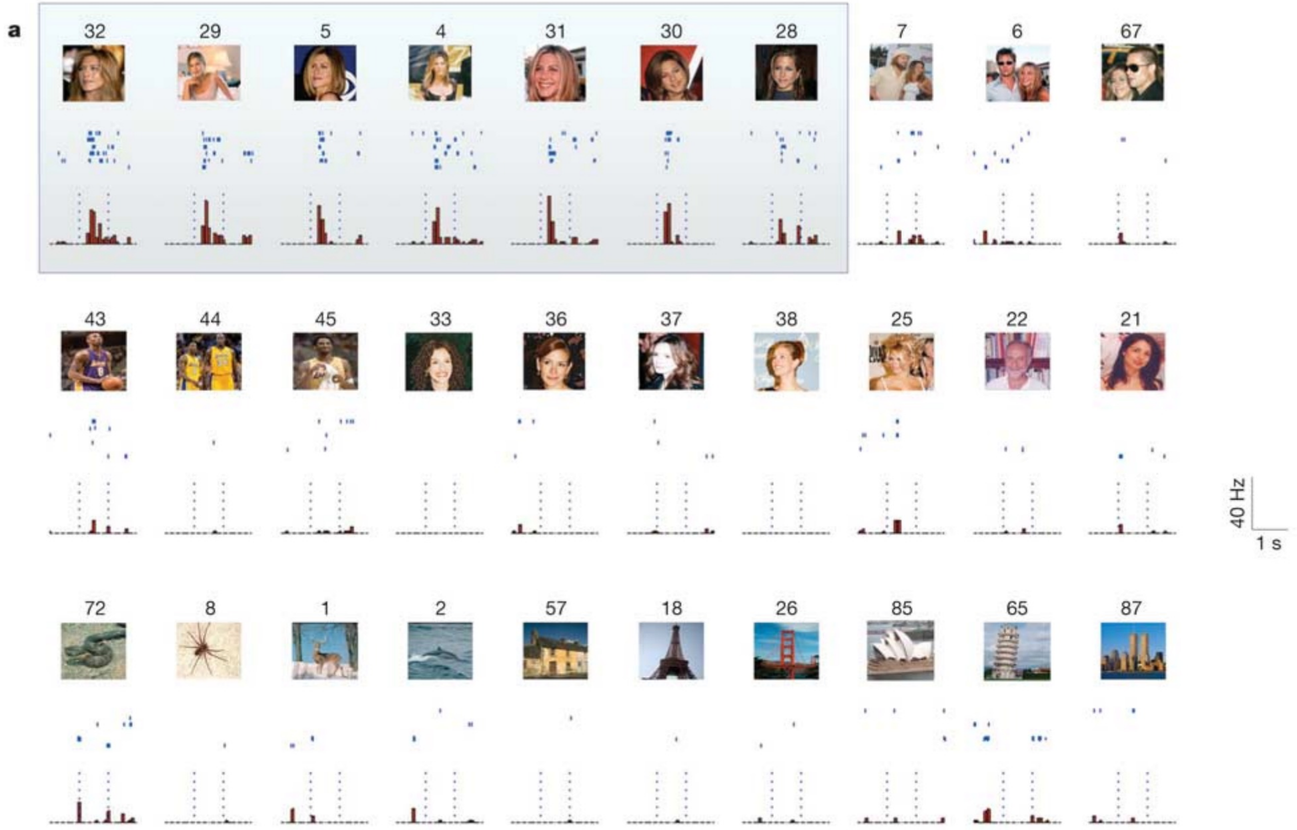
- Saliency/attribution maps generate intuitive and visually-appealing annotations which at first seem to explain why a network made a particular decision
- However, many saliency methods fail critical sanity checks – most notably, many methods do not depend on the learned function, $p(\hat{y}|x)$, but rather depend on the data distribution alone, $p(x)$
- The flaws of saliency methods may be particularly concerning when they're used in safety-critical settings, such as medical imaging where saliency has been widely used (Arun et al., 2020; Saporta et al., 2021)

Case study 2:
Understanding networks through easy-to-interpret neurons

Selectivity for understanding the nervous system



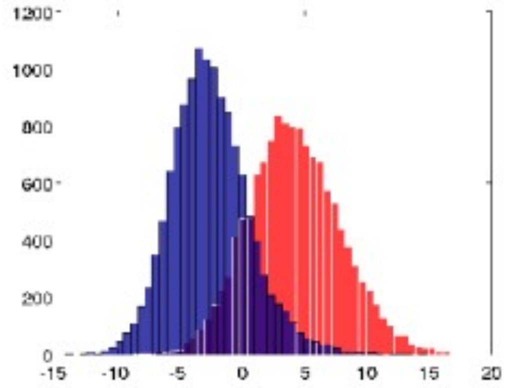
Barlow, *J. Physiol.* 1953



Quian Quiroga et al., *Nature* 2005

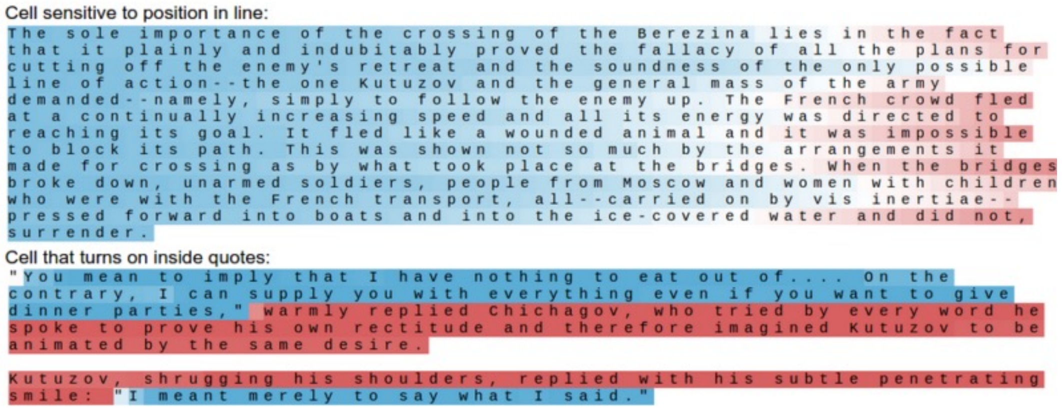
Selectivity for understanding deep networks

"Face" neurons



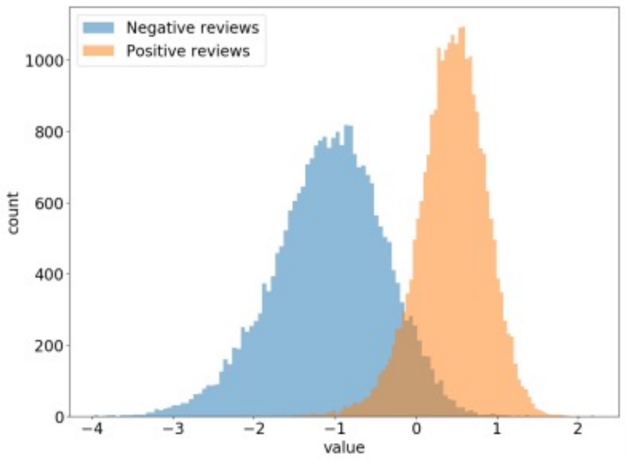
Le et al., *ICML* 2011

Text-selective neurons



Karpathy et al., 2016

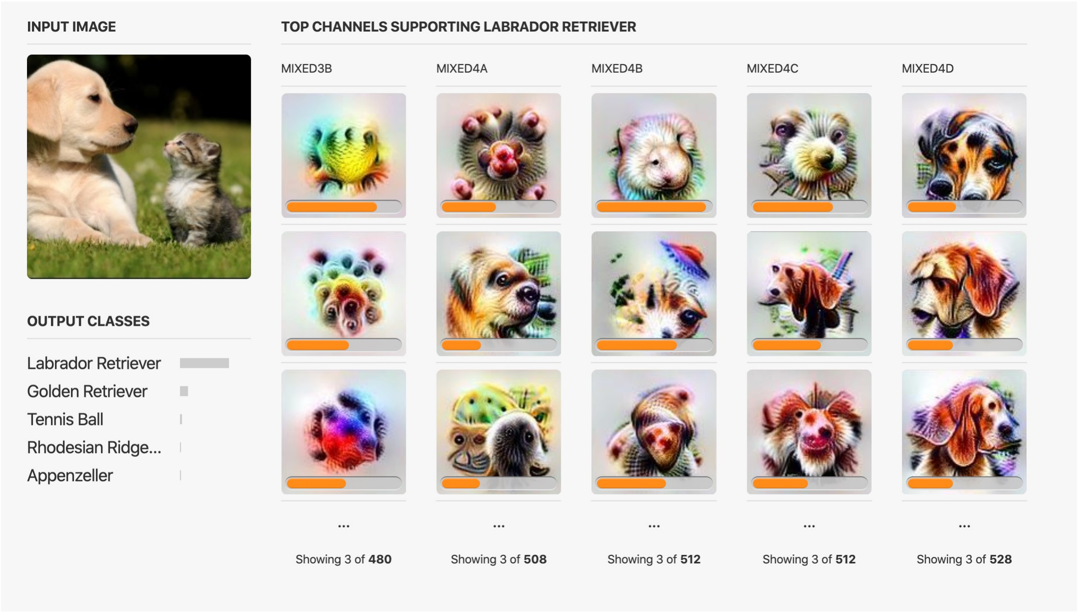
Sentiment neuron



Radford et al., 2017

Activation maximization for understanding deep learning

Channel attribution

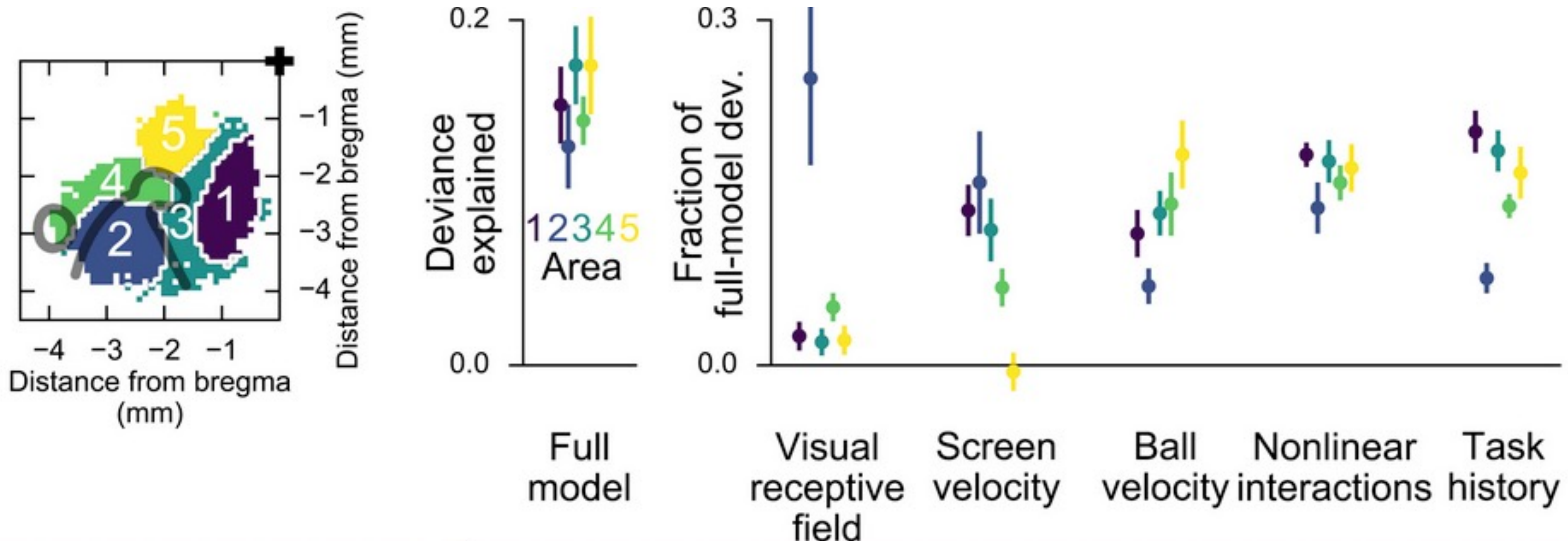


Olah et al., *Distill* 2018

Optimized for CaffeNet output neurons

Nguyen et al., *NIPS* 2016

A shift in neuroscience from single units to populations

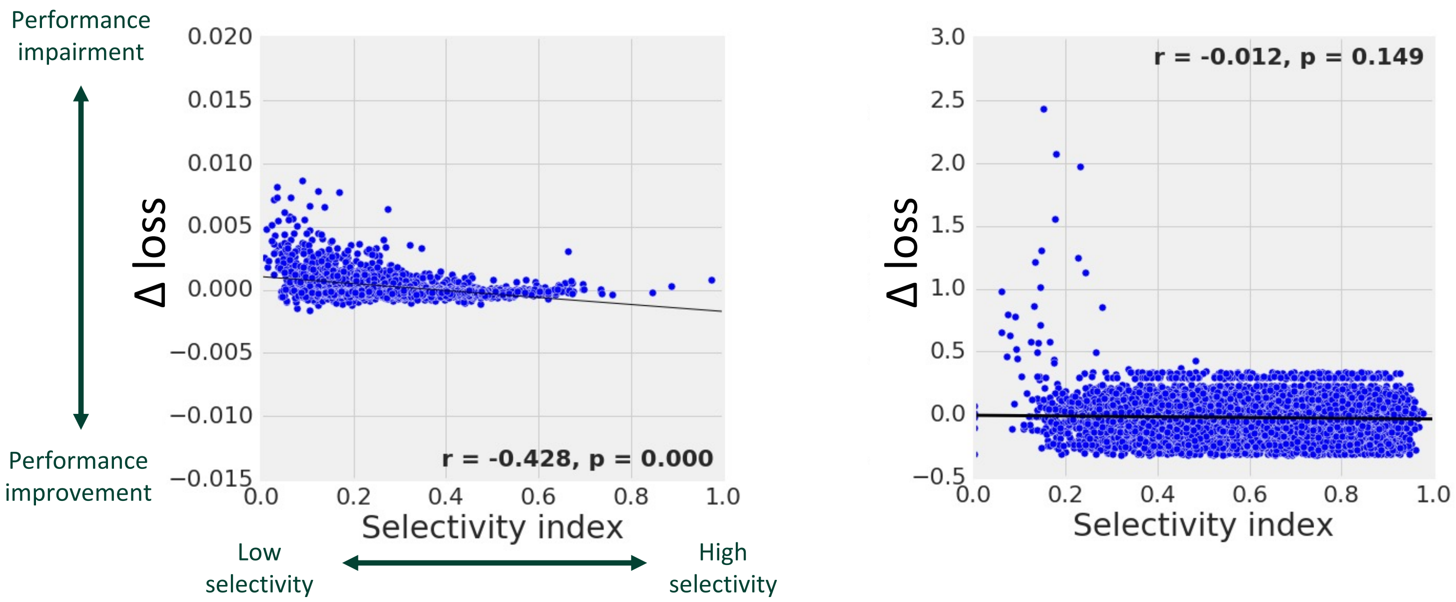


Minderer et al., *Neuron* 2019

- Non-selective neurons can contribute to population coding
  - Macaque prefrontal cortex (Leavitt et al., 2017); mouse visual cortex (Zylberberg, 2017); rat auditory and frontal cortex (Insanally et al., 2019)
- Movement towards population-level phenomena for characterizing neural systems (Shenoy et al.,2013; Raposo et al., 2014; Fusi et al., 2016; Morcos and Harvey, 2016; Pruszynski and Zylberberg,2019; Heeger and Mackey, 2019; Saxena and Cunningham, 2019)

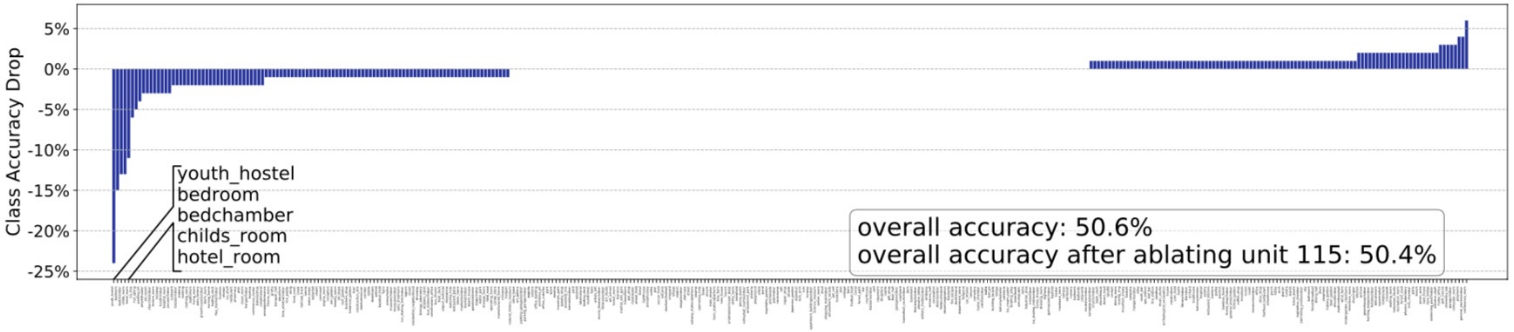
Class selectivity is a poor predictor of unit importance

The previous studies showed the existence of individually interpretable single units, but did not test whether these units causally lead to better performance



Morcos et al., *ICLR* 2018

Ablating selective units causes class-specific deficits



Zhou et al., 2018

Ablating the sentiment neuron can improve performance

Features	SST	MR	CR	IMDB
All features	91.76	87.52	91.38	92.28
SN deleted	91.87	86.96	90.72	91.77

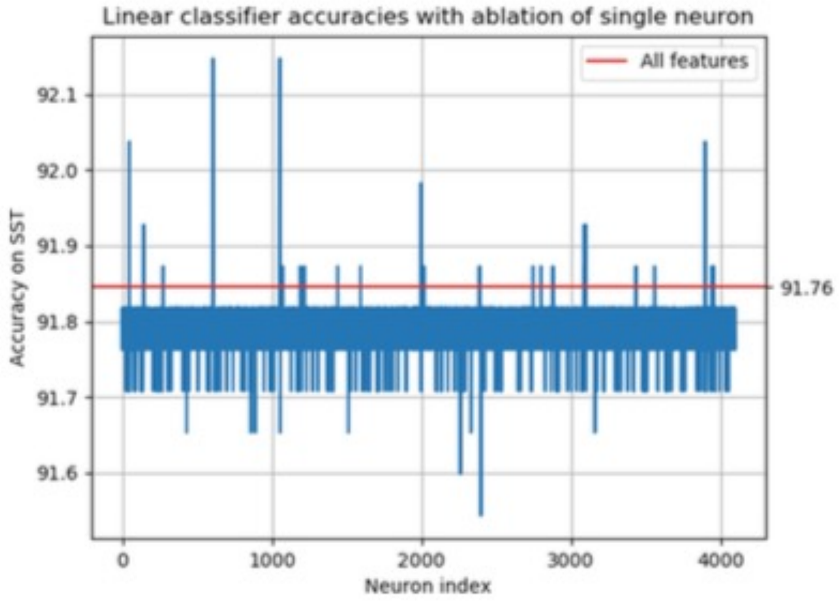
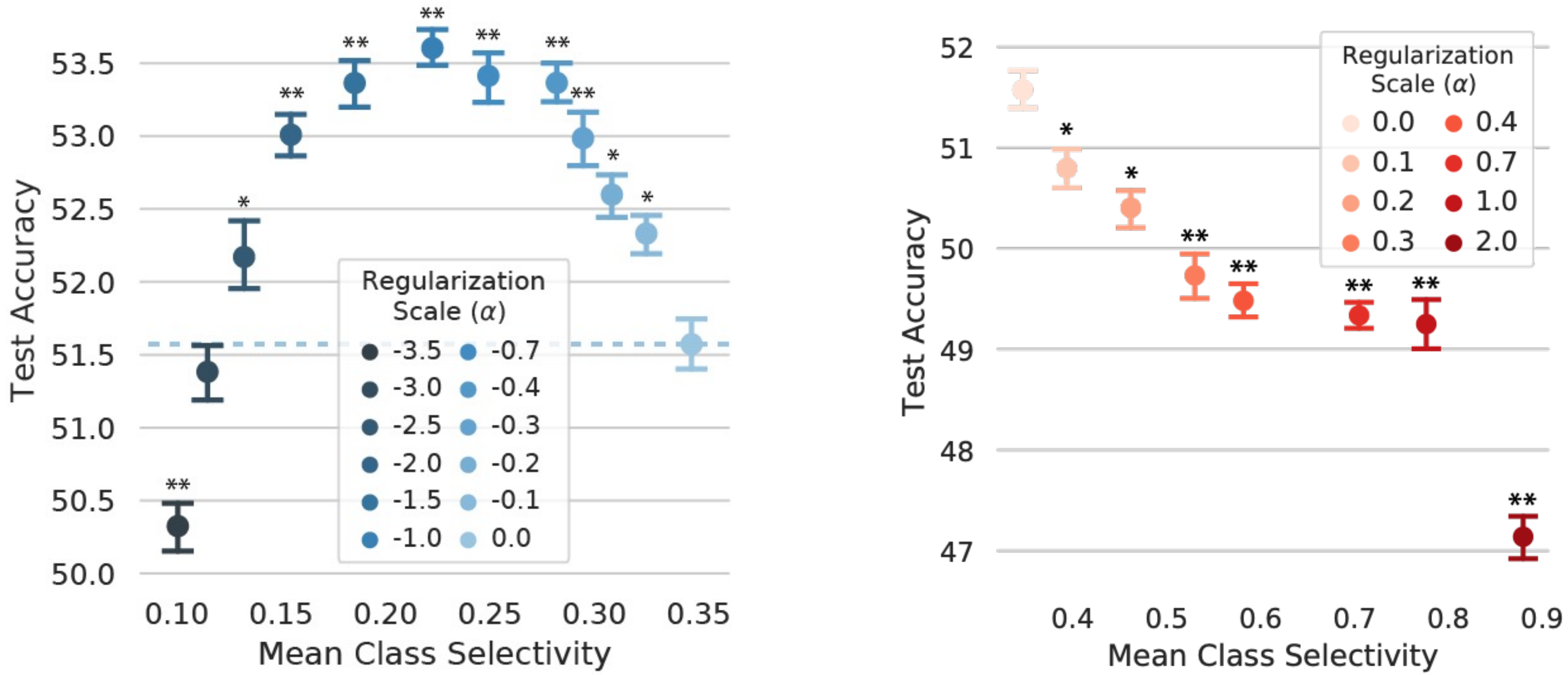


Fig. 2. Accuracy scores of classifier with each neuron ablated individually.

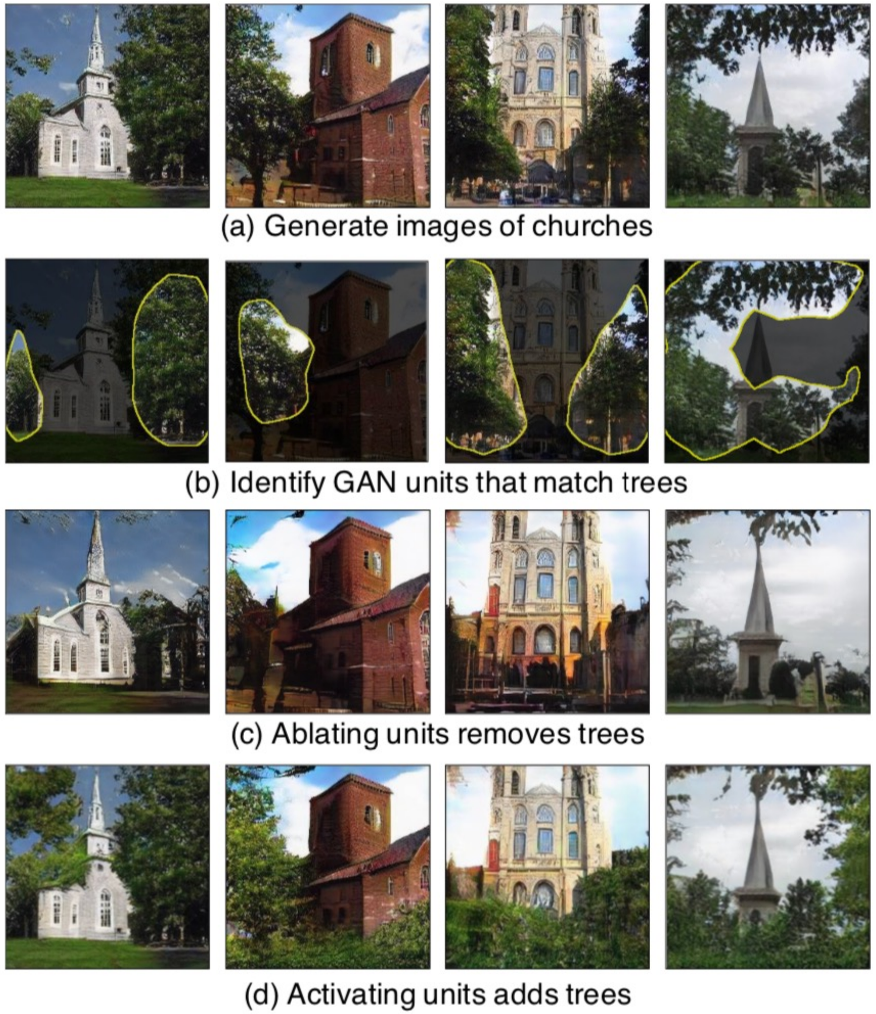
Donnelly and Roegiest, 2019

Class selectivity is neither sufficient nor strictly necessary for high accuracy



Leavitt and Morcos, *ICLR* 2021

Individual interpretable units in generative models causally change function



Bau et al., *ICLR* 2019

Takeaways

- The tractability of low-dimensional signals can encourage researchers to focus on the properties of individual units as representative of network behavior
- Studies using single unit ablation, class selectivity regularization, and generative models show that the properties of single units do not reliably extrapolate to populations
- Approaches for understanding networks should utilize functionally—ideally *causally*—relevant properties
- Researchers should focus on properties that exist *across* neurons—distributed, high-dimensional representations—and develop tools to make these properties more tractable and accessible to intuition

Building better hypotheses

Building hypotheses of increasing strength – very weak

Hypothesis: *Feature-selective neurons are the foundation of DNN function.*

Pros

Cons

- Not falsifiable
- “Foundation” is vaguely defined; how do you test whether something is the “foundation of DNN function”?
- No concept of a baseline for comparison

Building hypotheses of increasing strength – weak

Hypothesis: *If feature selectivity is important for DNN function, then we should find feature-selective neurons.*

Pros

- Is falsifiable! If there are no feature-selective neurons, this hypothesis has been proven false

Cons

- Proving the non-existence of something is often challenging. What if one just didn't use the right method to look?
- Unclear what we should expect by chance: how many feature-selective neurons would we expect to occur randomly?
- The presence of feature-selective neurons does not necessarily imply their functional importance

Building hypotheses of increasing strength – average

Hypothesis: *If feature selectivity is necessary to maximize test accuracy, ablating feature-selective single neurons should cause a decrease in test accuracy.*

Pros

- Falsifiable
- Addresses causality – "necessary" is a much more concrete statement than "important" and leads to a specific experiment to test this hypothesis

Cons

- No discussion of alternative possibilities
- No discussion of baseline. What if ablating all individual neurons causes a decrease in test accuracy? Does that satisfy this hypothesis?

Building hypotheses of increasing strength – strong

Hypothesis: *If feature selectivity in single neurons is necessary to maximize test accuracy, ablating selective neurons should cause a decrease in test accuracy proportional to the strength of the neuron's feature selectivity. Alternatively, if networks rely more on feature selectivity across neurons than on feature selectivity in individual neurons, then zeroing activity in feature-selective directions (i.e. a linear combination of units that represents curves) that are not axis-aligned should cause a decrease in test accuracy that is proportional to the strength of feature selectivity and exceeds the decrease from ablating only single units.*

Pros

- Falsifiable
- Makes clear testable predictions
- Presents multiple competing hypotheses
- Provides a baseline comparison (single units vs. non-axis-aligned linear combinations)

Cons

- Verbose

Key takeaways

Intuition is critical, but be wary of unverified intuition, especially in interpretability!

4 key recommendations:

1. Make clear, specific, testable, and falsifiable hypotheses
2. Be wary of visualization! One's skepticism should be proportional to the feeling of intuitiveness
3. Quantify wherever you can. An unquantifiable hypothesis risks being an unfalsifiable hypothesis.
4. Remember the "human" in "human explainability." If a method aims to help humans understand DNNs, test this explicitly.

Thank you!

Matthew Leavitt

Former Facebook AI Resident

For more detail and references, see our position paper:
https://arxiv.org/abs/2010.12016

FACEBOOK