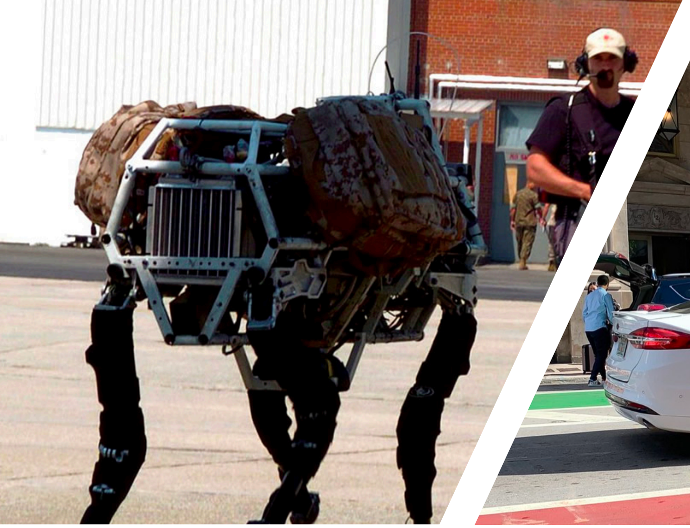
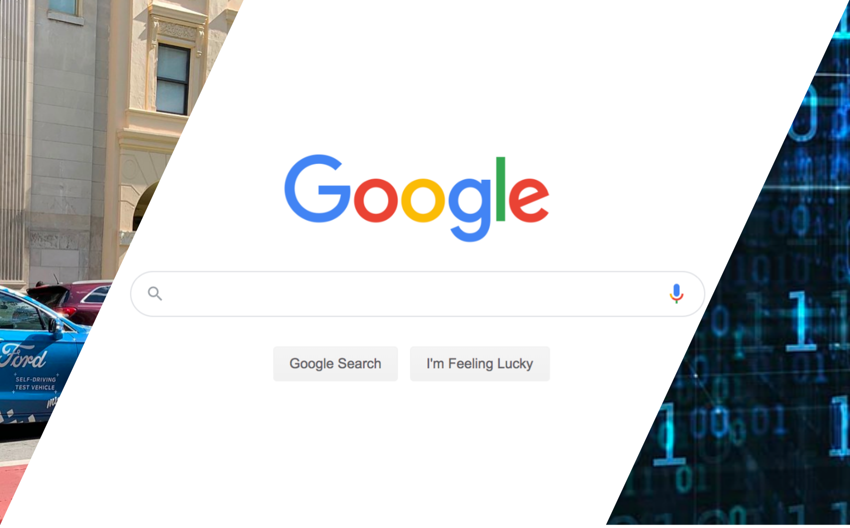
Influence-Directed Explanations for CNNs

Klas Leino

Overview

- Background on Interpretability
- Input Influence
- Internal Influence
  - Slices
  - Distributions of Interest
  - Quantities of Interest
  - Axioms
- Interpretation of Internal Features

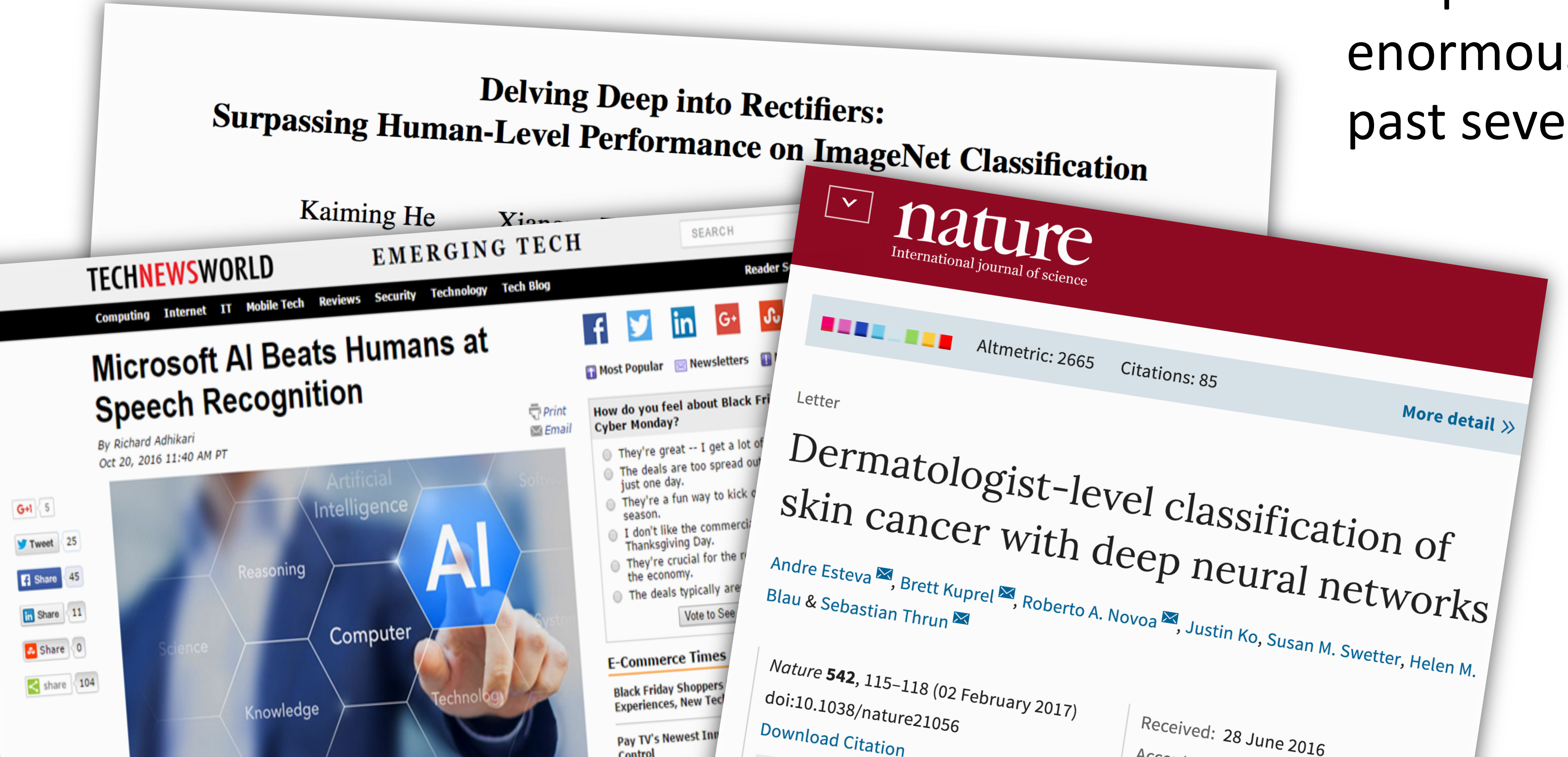




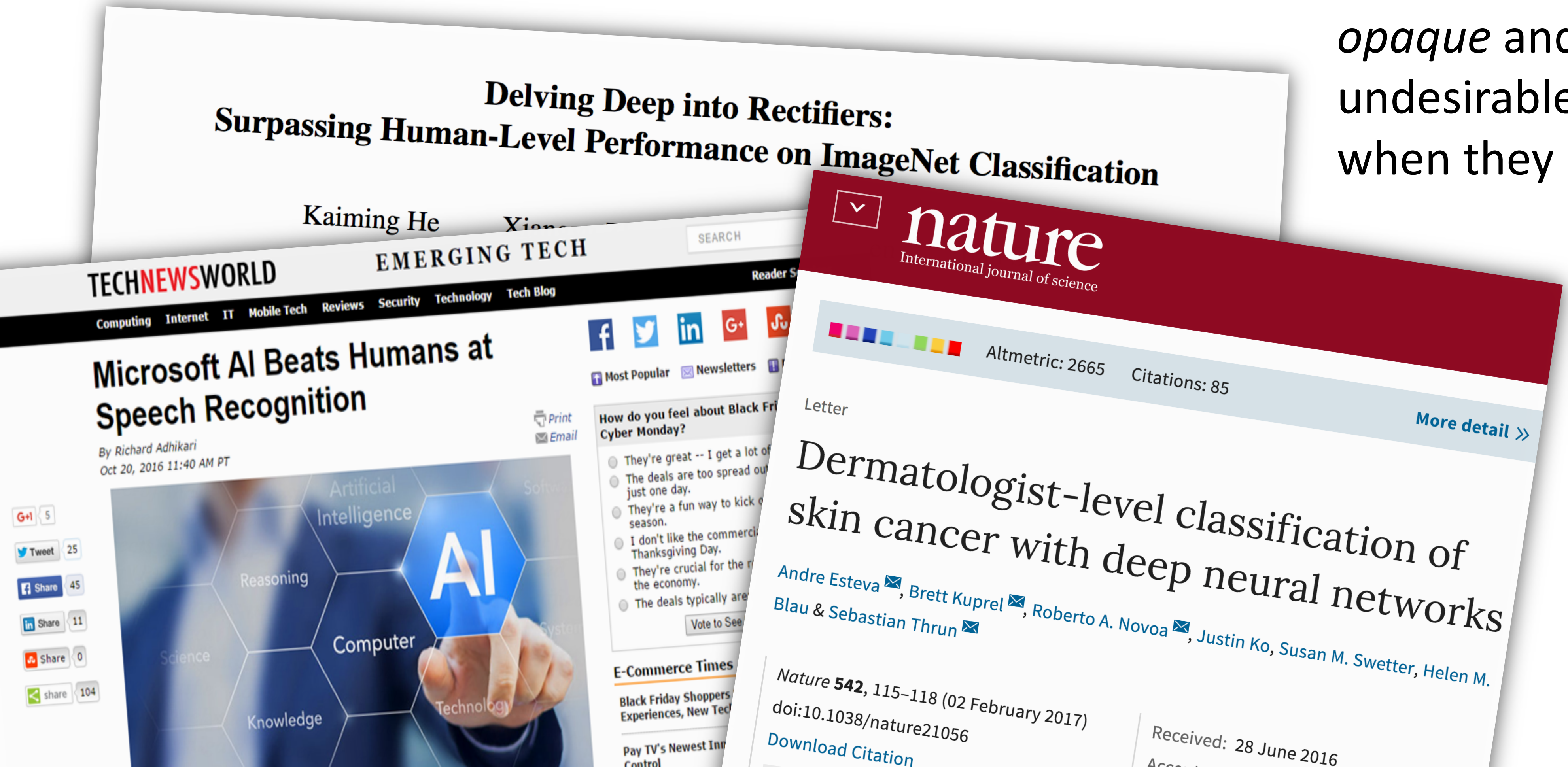
Machine Learning is Everywhere

How Much can We Trust DNN Predictions?

Deep learning has seen enormous success in the past several years



How Much can We Trust DNN Predictions?

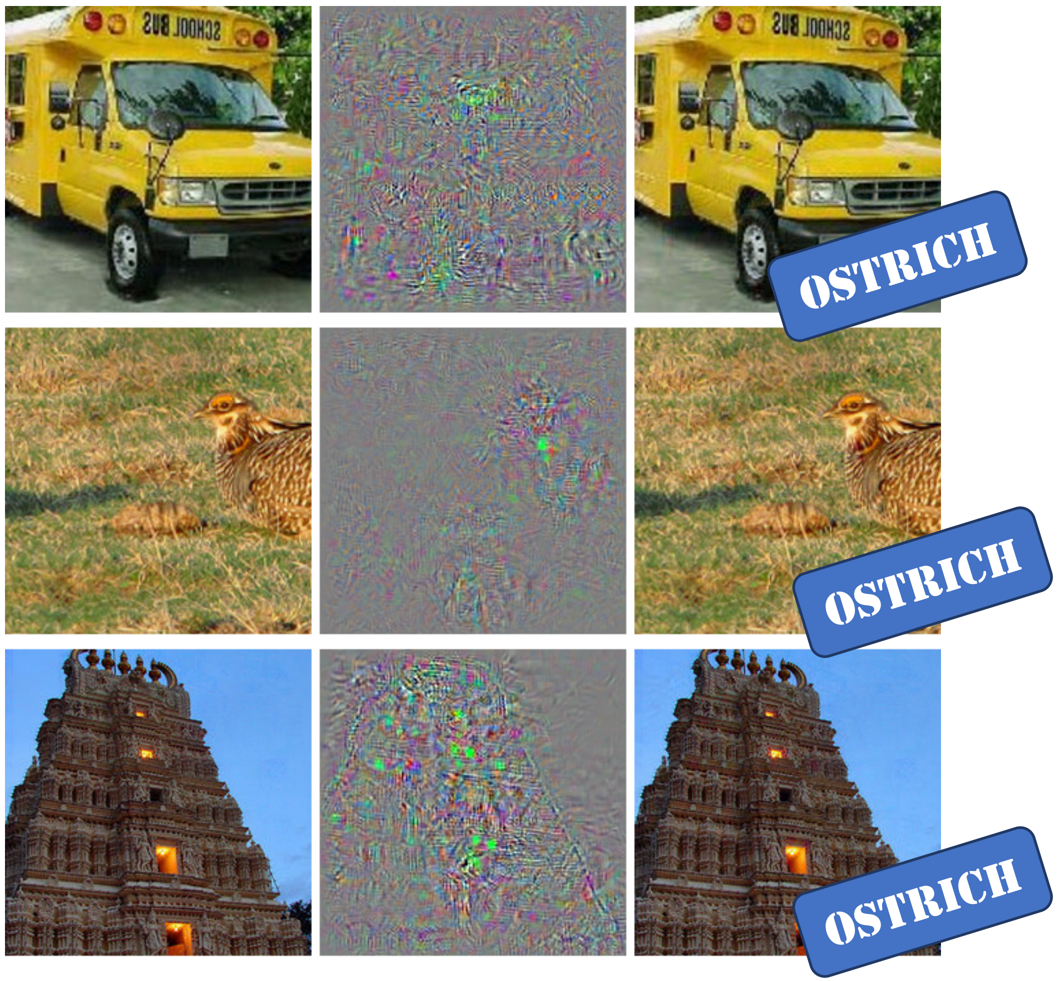


But deep networks remain *opaque* and often exhibit undesirable behavior even when they appear to work well

Example: Adversarial Attacks

what is this a picture of?

Original Image | Adversarial Perturbation | Perturbed Image



[Szegedy et al. 2014]

Increasing Model Trust

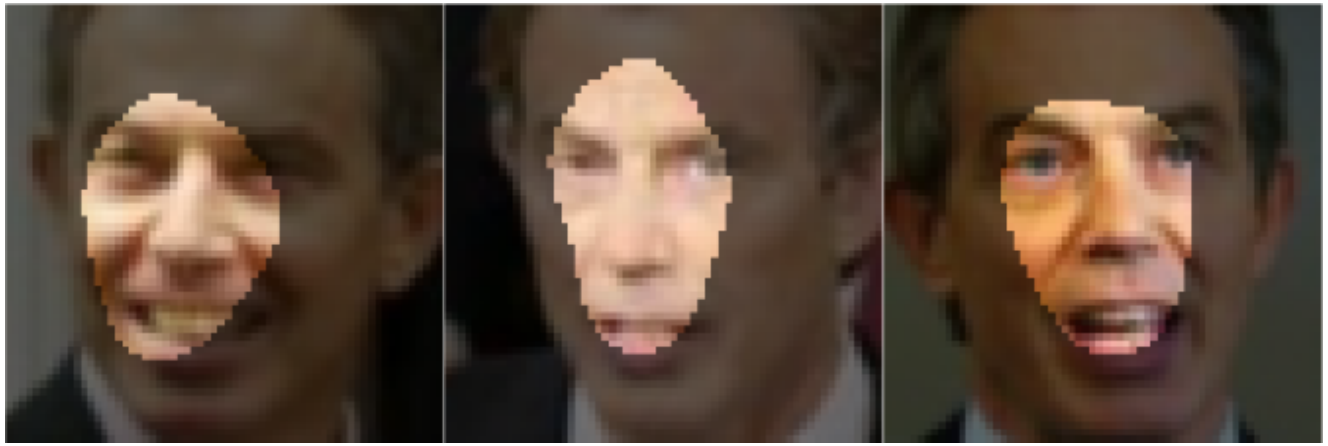
- Generalization error might not be sufficient to instill model trust
- Question: when a model makes a decision, did it make it for the right reason?

- By examining the inner workings of a network, we may be able to address these types of questions

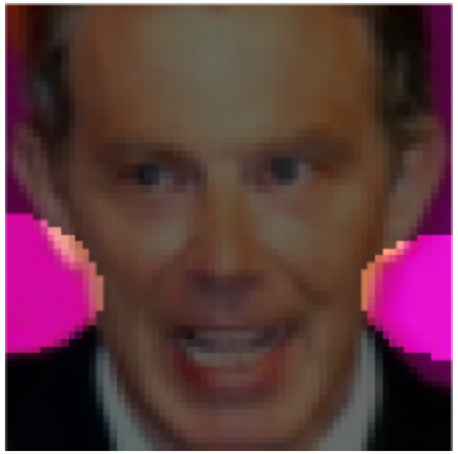
Example: Overfitting

notice the distinctive pink background

Sample of LFW training instances



Typical explanations on test instances of Tony Blair



Explanation [Leino et al. 2018] on training instance of Tony Blair with distinctive pink background. The model uses the background to classify the instance as Tony Blair.

[Leino and Fredrikson, 2019]

What Else Might We Want to Understand?

- Explaining mistakes
  - Question: when a model makes a mistake, why?
- Uncovering new knowledge
  - Question: did the model learn a pattern that we overlooked but might find useful?

Purpose of an Explanation Framework

- Answer *queries* like the questions posed in previous slides
- Goal: provide a framework for rigorously formulating and answering as broad a set of specific queries as possible

Overview

- Background on Interpretability
- Input Influence
- Internal Influence
  - Slices
  - Distributions of Interest
  - Quantities of Interest
  - Axioms
- Interpretation of Internal Features

Notation

- We will take a functional view of a neural network:

  A model is a function, $f: \mathbb{R}^n \rightarrow \mathbb{R}^m$, where n is the number of input features and m is the number of classes

  - Let $x \in \mathbb{R}^n$ be an input to the model
    - We say x_j for $j \in [n]$ is a *feature* or *variable*
  - Let $f_c(x)$ be the model's output for class c on input x

Influence Measures

- An (input) *influence measure*, χ, for a model, f, assigns a value to each of the input features, x_i, specifying how important x_i was in determining the model's output, $f(x)$

Saliency Maps

- Informally, for an influence measure to be *causal* (with respect to the model), a feature should be considered important if changing it slightly* would change the output of the model
- Gradient w.r.t. features captures this intuition precisely
- Simple influence definition [Simonyan et al. 2014]

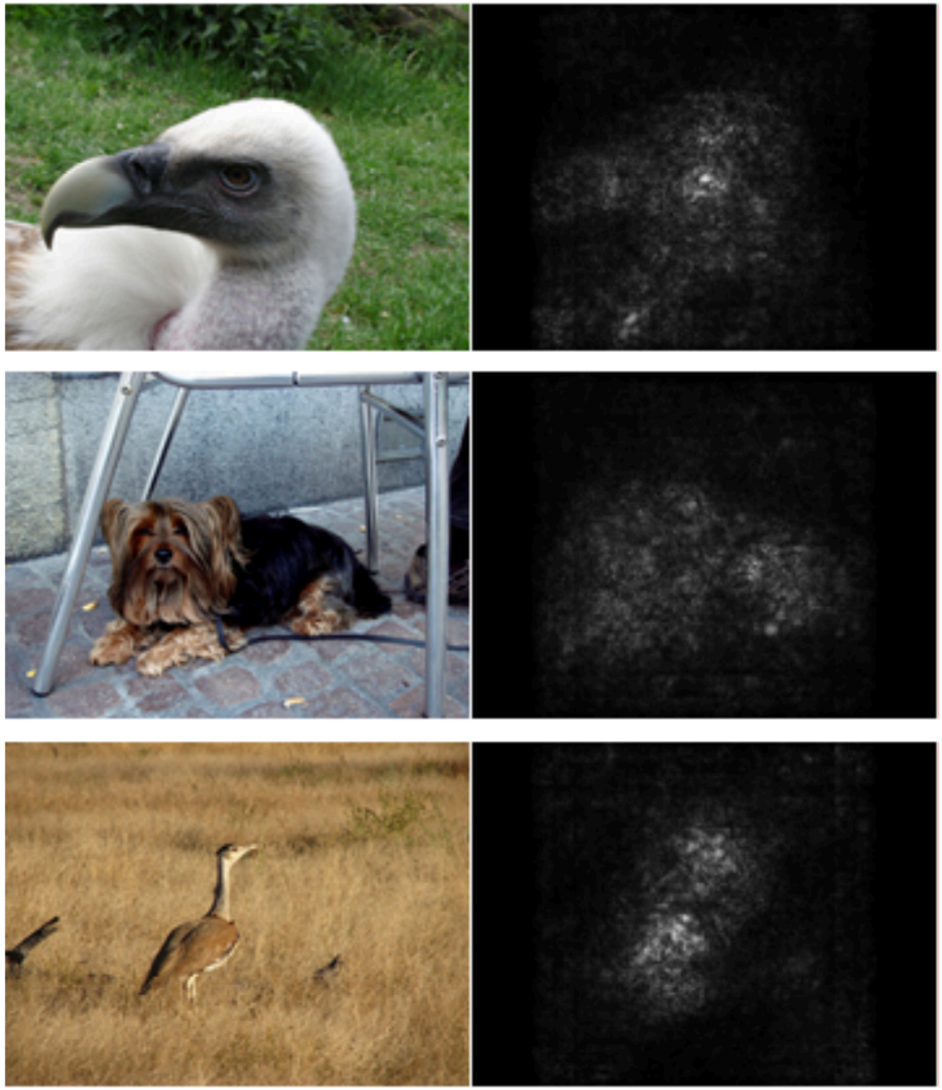
c′ is the predicted class

$$\chi_{saliency}(f, \boldsymbol{x}) = \frac{\partial f_{c'}}{\partial x}[\boldsymbol{x}]$$

take the gradient w.r.t. the input

evaluate at the point we are calculating the influence for

Example: Saliency Maps



[Simonyan et al. 2014]

Integrated Gradients

- Gradient at a point may describe behavior that is too local
- Example:
  - let $f(x) = \max\{x, 1\}$ (where $x \in \mathbb{R}$, i.e., the input is 1-dimensional)
  - let $\boldsymbol{x} = 1.5$
  - Then $f(\boldsymbol{x}) = 1$, but $\frac{\partial f}{\partial x}[\boldsymbol{x}] = 0$
  - It seems natural to give some influence to $\boldsymbol{x}$, but according to a very local view, $\boldsymbol{x}$ does not change f
- Integrated gradients [Sundararajan et al. 2017] addresses this by taking the average gradient between the point, $\boldsymbol{x}$, and a *baseline* point

Integrated Gradients

- Integrated gradients [Sundararajan et al. 2017]

$$\chi_{IG}(f, \boldsymbol{x}, \boldsymbol{x_0}) = (\boldsymbol{x} - \boldsymbol{x_0}) \int_{\alpha=0}^{1} \frac{\partial f_{c'}}{\partial x}[\boldsymbol{x_0} + \alpha(\boldsymbol{x} - \boldsymbol{x_0})] d\alpha$$

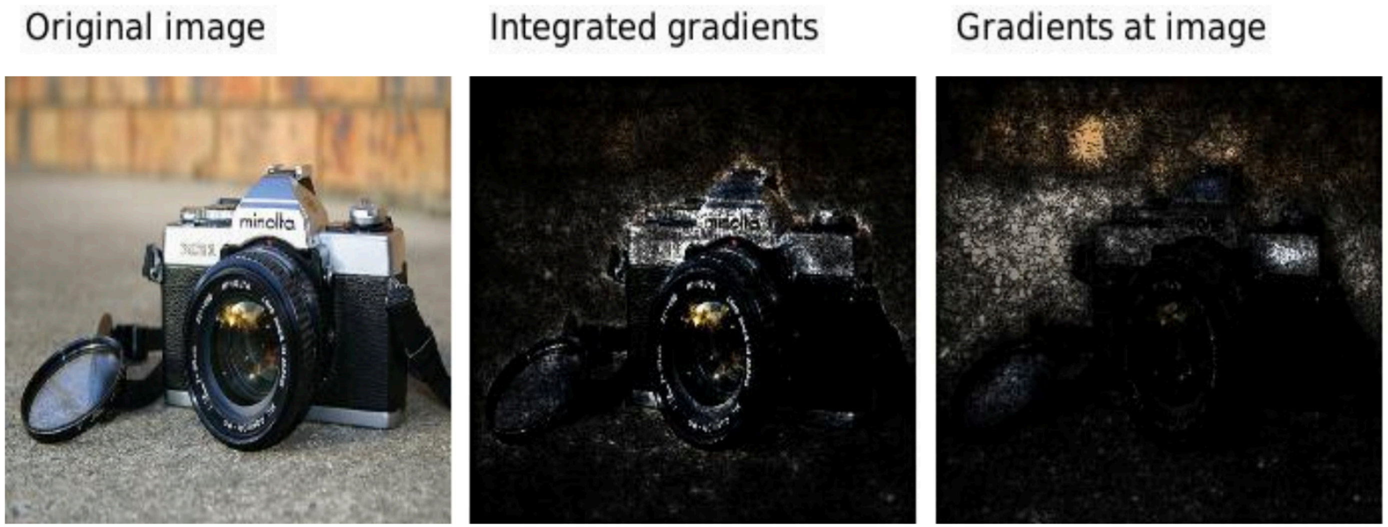
α interpolates between x_0 and x

baseline point

note: this is different from saliency maps conceptually because we multiply the gradient term by the input value (minus the baseline)

this is essentially an integral along the straight-line path from the baseline, x_0, to the point, x

Example: Integrated Gradients



[Sundararajan et al. 2017]

Selecting a Baseline

- Baseline is arbitrary, but affects how influence should be interpreted
- Commonly set to zero, i.e., a black image
  - Could be a specific point we want to compare to

Why Take a Line?

- Line between point and baseline gives rise to some natural axioms
  - **Sensitivity** | states that if the baseline differs from $\boldsymbol{x}$ in exactly one variable, and $f(\boldsymbol{x}) \neq f(\boldsymbol{x_0})$ then that variable must have non-zero influence
  - **Dummy Antisensitivity** | states that if f does not mathematically depend on a variable, that variable's influence should be zero
  - **Linear Agreement** | states that for a linear model, the influence of each feature is just the weight of that feature
  - **Efficiency** | states that the sum of the influences must be equal to the difference in output on $\boldsymbol{x}$ and on $\boldsymbol{x_0}$
  - **Symmetry Preserving** | states that symmetrical inputs to f receive equal influence

Overview

- Background on Interpretability
- Input Influence
- Internal Influence
    - Slices
    - Distributions of Interest
    - Quantities of Interest
    - Axioms
- Interpretation of Internal Features

Generalizing Input Influence

- Become *internal*
  - Assign a meaningful influence score to internal features learned by a deep network
- Become *distributional*
  - Flexibility in defining which points the influence should be supported by
- Support general quantities of interest
  - Flexibility to specify what network behavior we are trying to explain

Internal Influence

- Internal influence [Leino et al. 2018]

take gradient of QoI rather than output of f

$$\chi_{int}(f = g \circ h, D, q) = \int_{x \in \mathbb{R}^n} \frac{\partial q \circ g}{\partial h(x)} [h(x)] D(x) dx$$

weight each point according to the DoI

slice

distribution of interest (DoI)

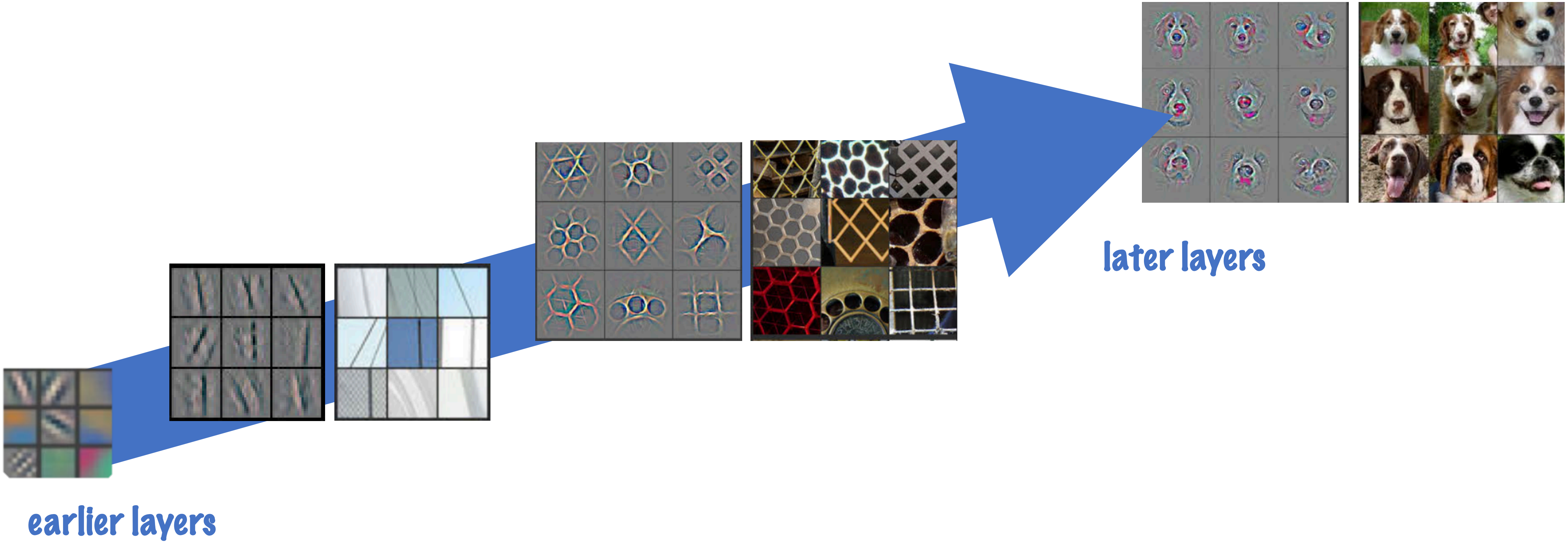
quantity of interest (QoI)

take gradient w.r.t. internal features

Overview

- Background on Interpretability
- Input Influence
- Internal Influence
  - Slices
  - Distributions of Interest
  - Quantities of Interest
  - Axioms
- Interpretation of Internal Features

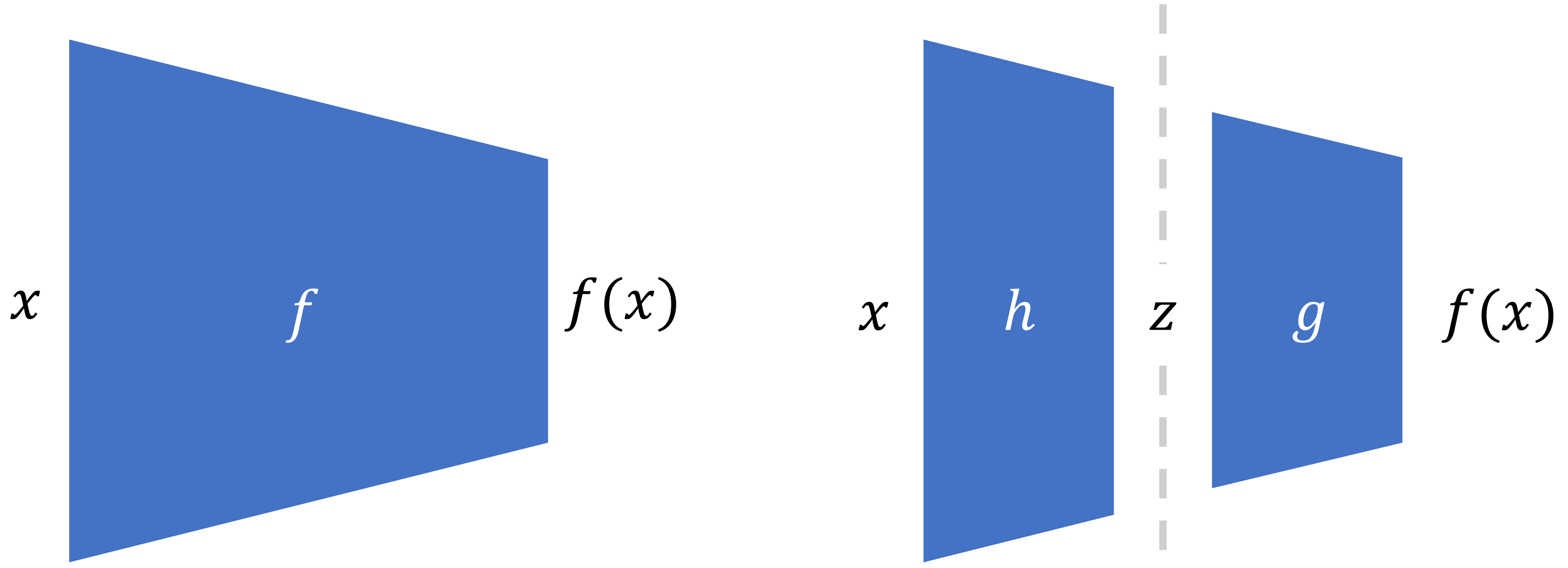
Different Layers Learn Different Abstractions



[Zeiler et al. 2013]

Slices

- A *slice* of a network, f, is a pair of functions (or sub-networks), $\langle g, h \rangle$, such that $f = g \circ h$
- Intuitively, this exposes the internals of the network at a chosen layer



Slices Help Decompose Explanations into Natural Components

Internal Influence

Input Influence

Overview

- Background on Interpretability
- Input Influence
- Internal Influence
  - Slices
  - Distributions of Interest
  - Quantities of Interest
  - Axioms
- Interpretation of Internal Features

Defining the Set of Instances to be Faithful on

- Point may describe behavior that is too local
- Alternatives:
  - Neighborhood around point (smooth gradients)
  - Line to baseline (realizes IG)
  - Entire class
  - All training points
  - Entire space

Distributions of Interest

- A *distribution of interest* (DoI) is a probability distribution over input points in $\mathbb{R}^n$, represented by its PDF, D
- E.g., to get a linear path from $\boldsymbol{x}$ to $\boldsymbol{x_0}$ (as in IG), we can define the DoI to be a uniform distribution over the points on the line segment between $\boldsymbol{x}$ and $\boldsymbol{x_0}$, i.e.,

$$D(x') = \begin{cases} \dfrac{1}{|\boldsymbol{x} - \boldsymbol{x_0}|} & \text{if } x' \text{ is on the line segment } \overrightarrow{\boldsymbol{x}\boldsymbol{x_0}} \\ 0 & \text{otherwise} \end{cases}$$

Overview

- Background on Interpretability
- Input Influence
- Internal Influence
  - Slices
  - Distributions of Interest
  - Quantities of Interest
  - Axioms
- Interpretation of Internal Features

Defining the Quantity to Explain

- We may be interested in explaining a model behavior besides its prediction, for example
  - Which features contributed to some other class that wasn't chosen by the model?
  - Why was class A chosen rather than class B?
  - Which features contributed to the activation of a particular internal neuron?

Quantities of Interest

- A *quantity of interest* (QoI) is a function, q, of the output* of f that specifies what network behavior we would like to calculate influence towards.
- E.g.,
  - to use the network's prediction as before, $q\left(f(x)\right) = \max\{f(x)\}$
  - to compare class A with class B, $q\left(f(x)\right) = f_A(x) - f_B(x)$

Example: Comparative Quantities of Interest

Top neuron for quantity $f_{sportscar}(x)$

Top neuron for (comparative) quantity $f_{sportscar}(x) - f_{convertible}(x)$

same neuron generalizes to other instances

[Leino et al. 2018]

Overview

- Background on Interpretability
- Input Influence
- Internal Influence
  - Slices
  - Distributions of Interest
  - Quantities of Interest
  - Axioms
- Interpretation of Internal Features

Justification for Internal Influence

- Internal influence follows from a few natural axioms
  - **Linear Agreement** | states that for a linear model, the influence of each feature is just the weight of that feature
  - **(distributional) Marginality** | essentially captures that the influence must be causal with respect to the model – a feature can only get influence according to its marginal contribution to the quantity of interest
  - **Distributional Linearity** | states that each point must be weighted according to its probability density given by the distribution of interest
  - **Slice Invariance** | states that the influence doesn't depend on the implementation of h and g, only on the parts of the network that are exposed
  - **Preprocessing** | states that computing internal influence for a slice should be the same as computing input influence for g, where g's inputs are preprocessed by h

Summary of Internal Influence

- Goal is to enable a broad set of queries that can be tailored to the specific application/context
  - *Slice* allows us to specify level of abstraction
    - e.g., raw inputs or high-level features
  - *Distribution* allows us to specify relevant points
    - e.g., line from baseline or entire class
  - *Quantity* allows us to specify what we are explaining
    - e.g., specific class or comparison of two classes

Overview

- Background on Interpretability
- Input Influence
- Internal Influence
  - Slices
  - Distributions of Interest
  - Quantities of Interest
  - Axioms
- Interpretation of Internal Features

How do We Interpret Influential Internal Neurons?

- Backpropagation techniques, e.g., Zeiler et al. 2013

- Use input influence with a quantity of interest that selects a particular internal neuron