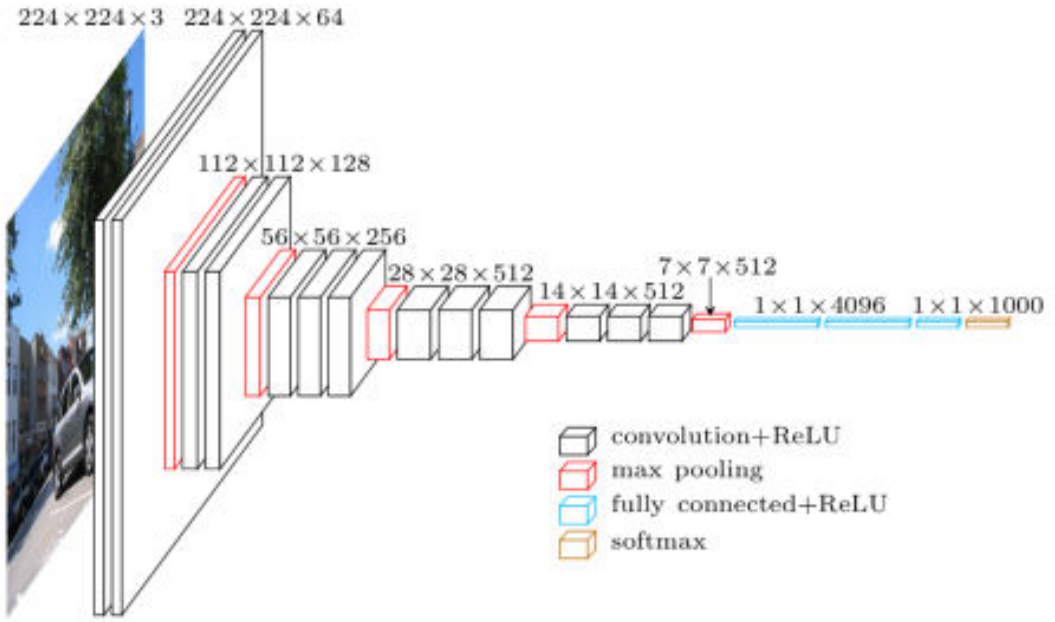
Influence-directed explanations for deep networks

Anupam Datta

CMU

Spring 2018

CNNs are complex



VGGNet (2014 ImageNet ILSVRC challenge runner-up)

- Depth of network critical for good performance (16 CONV/FC layers)
- More expensive to evaluate; many parameters (140M)

Goal

Explain behavior of deep neural networks by examining inner workings

Example questions

- What is the essence of a class from the network's point of view?
- What concept did the network use to classify an image into class A?
- What concept did the network use to classify an image into class A instead of class B?

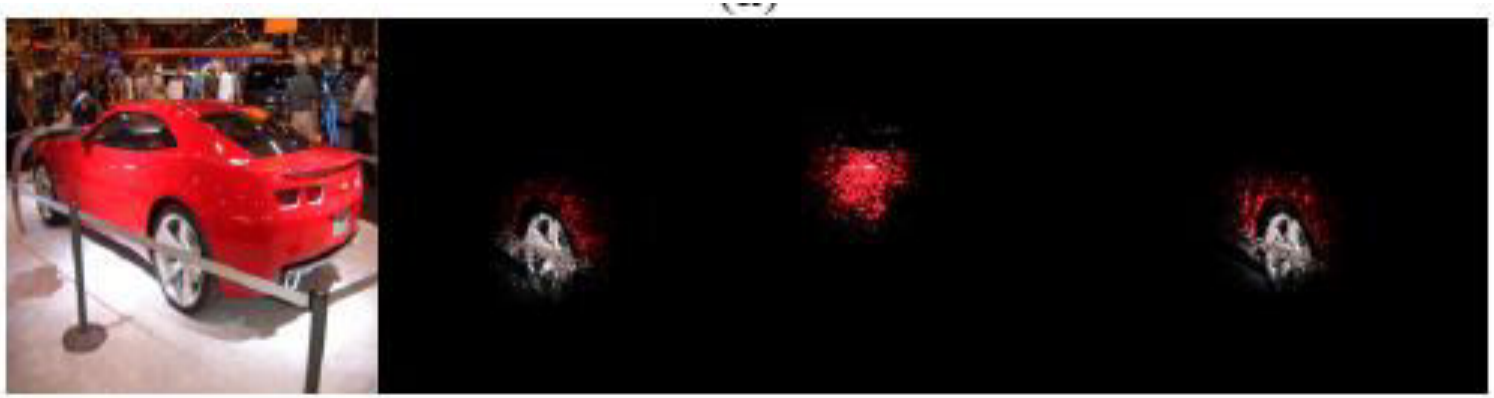
Influence-directed explanations

[Leino, Li, Sen, Datta, Fredrikson 2018]

Key idea

- Identify causally influential neurons in internal layers
- Give them interpretation using visualization techniques

Why did the network classify input as sports car?



Input image

Influence-directed Explanation

Why did the network classify input as sports car instead of convertible?

Input image

Influence-directed Explanation

Beyond input influence

- Uncovers high-level concepts learned by internal neurons that cause model's behavior
- These concepts generalize across input instances

An organizing principle

Accounting for a behavioral property of a ML system involves **interpretation of causally influential factors**

- Causation: What are important factors causing the property?
- Interpretation: What do these factors mean?

Influence-directed explanations

[Leino, Li, Sen, Datta, Fredrikson 2018]

Key idea

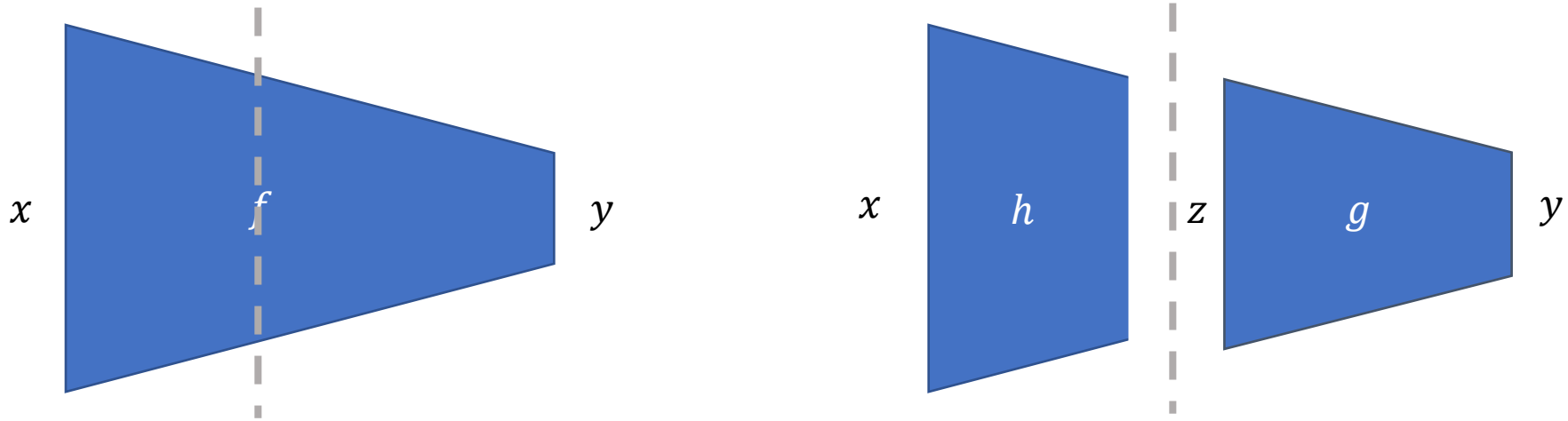
- Identify causally influential neurons in internal layers (how?)
- Give them interpretation using visualization techniques (how?)

Outline

- Distributional influence
- Interpretation with visualization
- Identifying influential concepts
- Explaining instances
- Justifying influence measure

Distributional influence

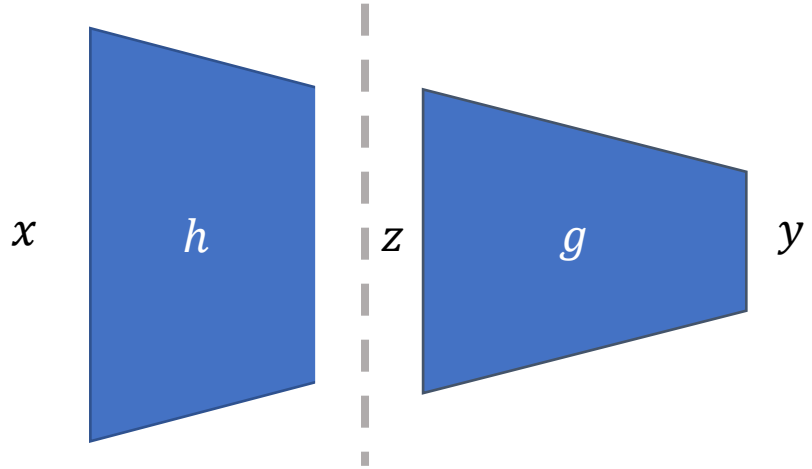
Decomposing network



$$y = f(x) = g(h(x))$$

- Slice of network $s = \langle g, h \rangle$ identifies layer whose neurons are examined
- Inputs drawn from distribution of interest P
- Quantity of interest f identifies network behavior to be explained

Distributional influence



$$y = f(x) = g(h(x))$$

Definition 1. *The influence of an element j in the internal representation defined by $s = \langle g, h \rangle$ is given by*

$$\chi_j^s(f, P) = \int_{\mathcal{X}} \left. \frac{\partial g}{\partial z_j} \right|_{h(\mathbf{x})} P(\mathbf{x}) d\mathbf{x} \qquad (1)$$

VGG16 model trained on ImageNet

Input image

Influence-directed Explanation

- Slice of network identifies layer whose neurons are examined: conv4_1
- Inputs drawn from distribution of interest P: training distribution
- Quantity of interest f identifies network behavior to be explained: difference in class scores of “sports car” and “convertible”

Outline

- Distributional influence
- Interpretation with visualization
- Identifying influential concepts
- Explaining instances
- Justifying influence measure

Interpretation with visualization

Interpreting influential neurons

Depicts interpretation (visualization) of 3 most influential neurons

- Slice of VGG16 network: conv4_1
- Inputs drawn from distribution of interest: delta distribution
- Quantity of interest: class score for correct class

Interpreting influential neurons

Visualization method: Saliency maps

- Compute gradient of neuron activation wrt input pixels
- Scale pixels of original image accordingly

Outline

- Distributional influence
- Interpretation with visualization
- Identifying influential concepts
- Explaining instances
- Justifying influence measure

Identifying influential concepts

Distributional influence captures general concepts

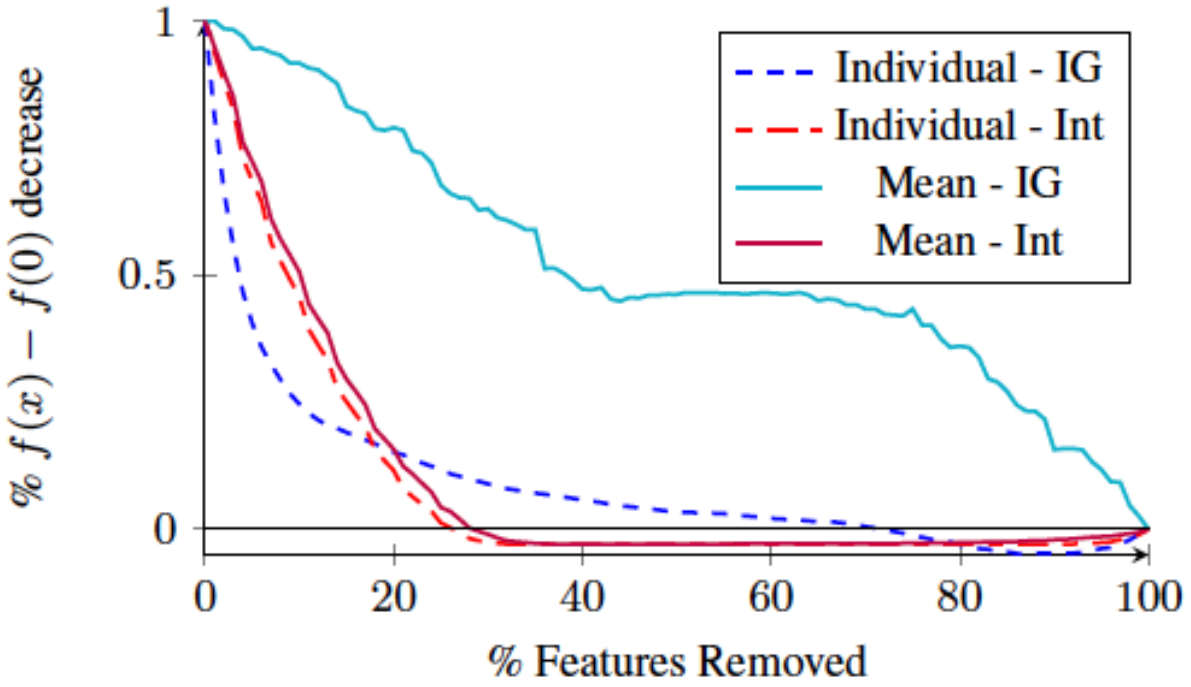
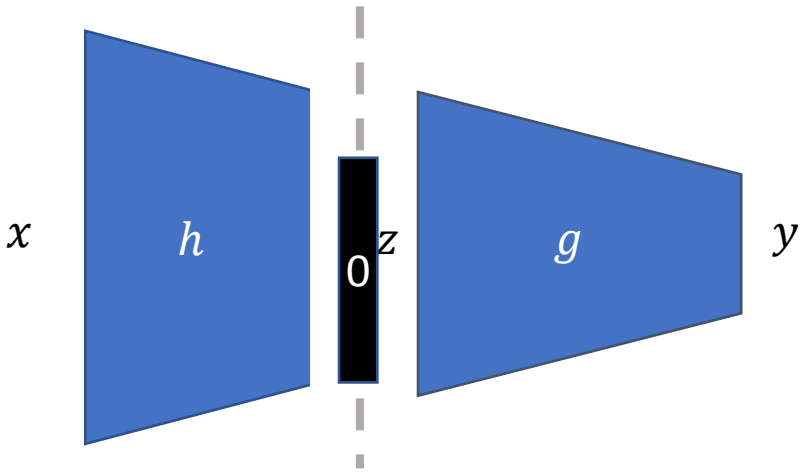


Figure 2. Plot of the decrease in the function value (before the softmax), $f(x)$, as features are removed in order of influence. The model is based on LeNet (LeCun et al., 1998) trained on the MNIST dataset. In each case, the most influential feature or hidden unit is incrementally removed, and the resulting value of $f(x)$ is depicted on the vertical axis. The vertical axis was normalized so that the average value of $f(x)$ on class "3" is 1, and 0 is the value of $f(0)$. The dashed curves depict the quantity when influence is measured for each instance individually (averaged across all instances of class "3"), and the solid curves when influence is with the empirical distribution of the entire class as the distribution of interest. Plots are shown for both integrated gradients (**IG**), as well as for internal influence (**Int**) for a slice at the first fully-connected layer of the network.

- Neurons influential for class on-average also influential for individual instances of class
- Not so for Integrated Gradients

Validating the essence of a class

- Produce compressed model + convert to binary class predictor



$$f_i = \left(f\Big|^{i}, \sum_{j \neq i} f\Big|^{j} \right)$$

Validating the essence of a class

Class	Orig.	Act.	Infl.
Chainsaw (491)	.14	0.	.71
Bonnet (452)	.62	0.	.92
Park Bench (703)	.52	0.	.71
Sloth Bear (297)	.36	0.	.75
Pelican (144)	.65	0.	.95

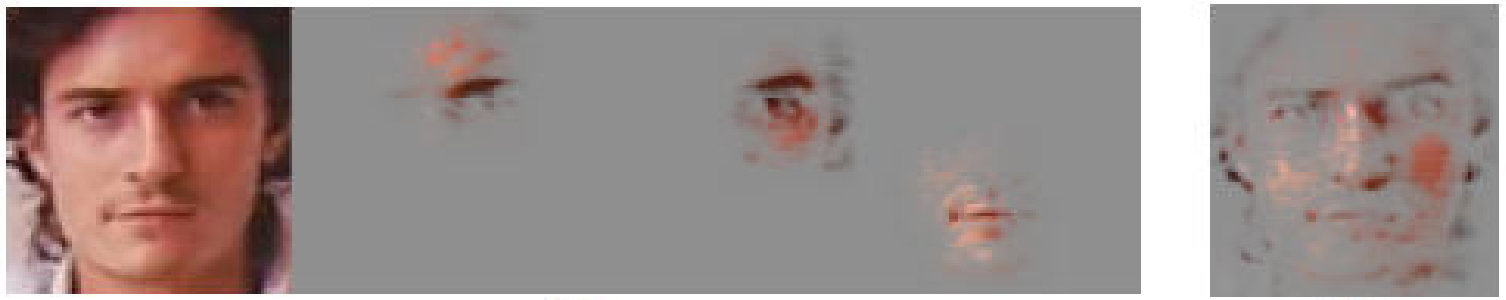
Table 1. Model compression recall for five randomly-selected ImageNet classes. Columns marked Orig. correspond to the original model, Act. to experts computed using activation levels, and Infl. to experts computed using influence measures. Precision in all cases was 1.0.

Outline

- Distributional influence
- Interpretation with visualization
- Identifying influential concepts
- Explaining instances
- Justifying influence measure

Explaining instances

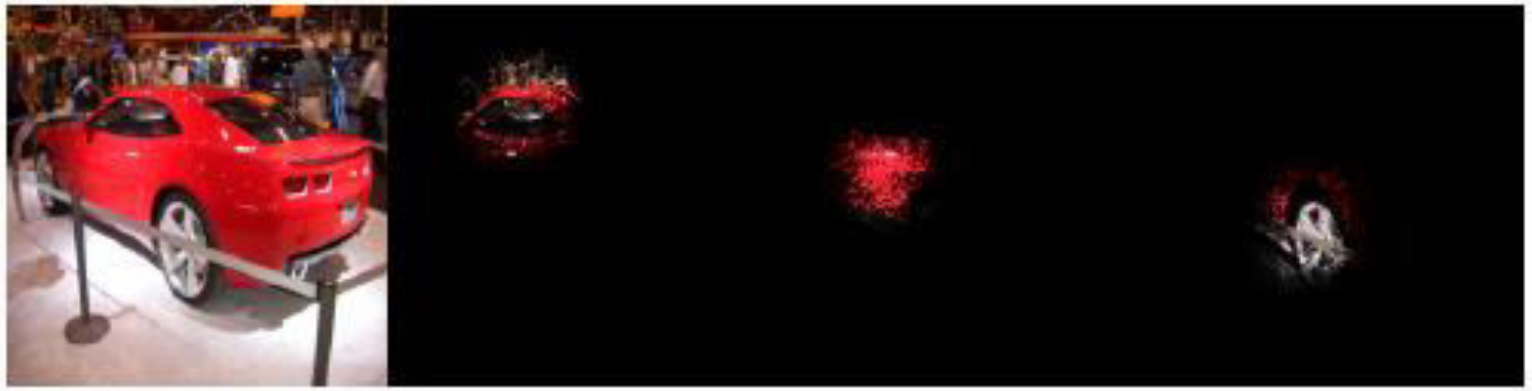
Focused explanations from slices



(a) (b)

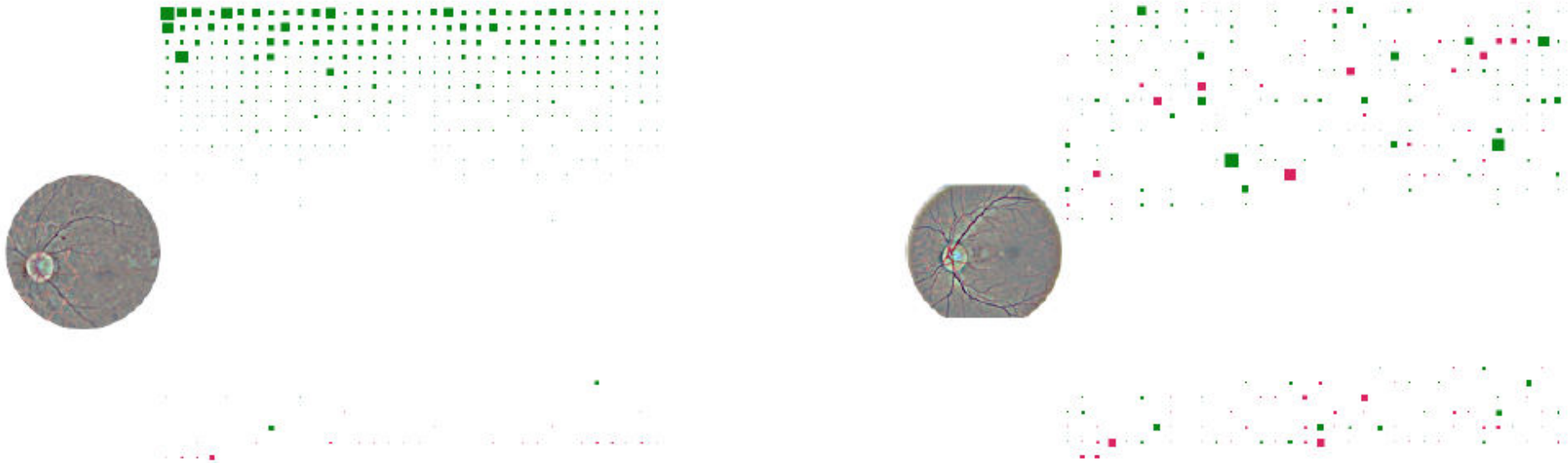
Figure 4. **(a)** Interpretation of the three most influential units from a slice corresponding to a convolutional layer (`conv4_1`), for the VGG16 (Simonyan & Zisserman, 2014) network. **(b)** Explanation based on integrated gradients (Sundararajan et al., 2017), taken on the same network and image. The interpretation in both cases was computed by scaling pixels in the original image using the results of either method.

Comparative explanations



Comparative explanation of classes "sports car" and "convertible" taken from the top-three most influential units at the `conv4_1` layer (VGG16 (Simonyan & Zisserman, 2014)).

Misclassification as deviations from class influence profiles



(a) (b)

Figure 6. Distributional influence measurements taken on DR model (Section 3.3) at bottom-most fully connected layer. To compute the grid, the distribution of influence was conditioned on class 5 (a) and class 1 (b). Figure (a) depicts an instance from class 5 that was correctly classified as such, and (b) an instance from class 5 that was incorrectly classified as class 1. In (a) the influences depicted in the grid align closely with the class-wide ordering of influences, whereas in (b) they are visibly more random. White space in the middle of the grid corresponds to units with no influence on the quantity.

Outline

- Distributional influence
- Interpretation with visualization
- Identifying influential concepts
- Explaining instances
- Justifying influence measure

Justifying influence measure

Axioms

Axiom 1 (Linear Agreement). *For linear models of the form* $f(\mathbf{x}) = \sum_i \alpha_i x_i$, $\chi_i(f, P) = \alpha_i$.

Axioms

Axiom 2 (Distributional marginality (DM)). *If,* $P(\left.\frac{\partial f_1}{\partial x_i}\right|_X = \left.\frac{\partial f_2}{\partial x_i}\right|_X) = 1$, *where* X *is the random variable over instances from* $\mathcal{X}$, *then* $\chi_i(f_1, P) = \chi_i(f_2, P)$.

Axioms

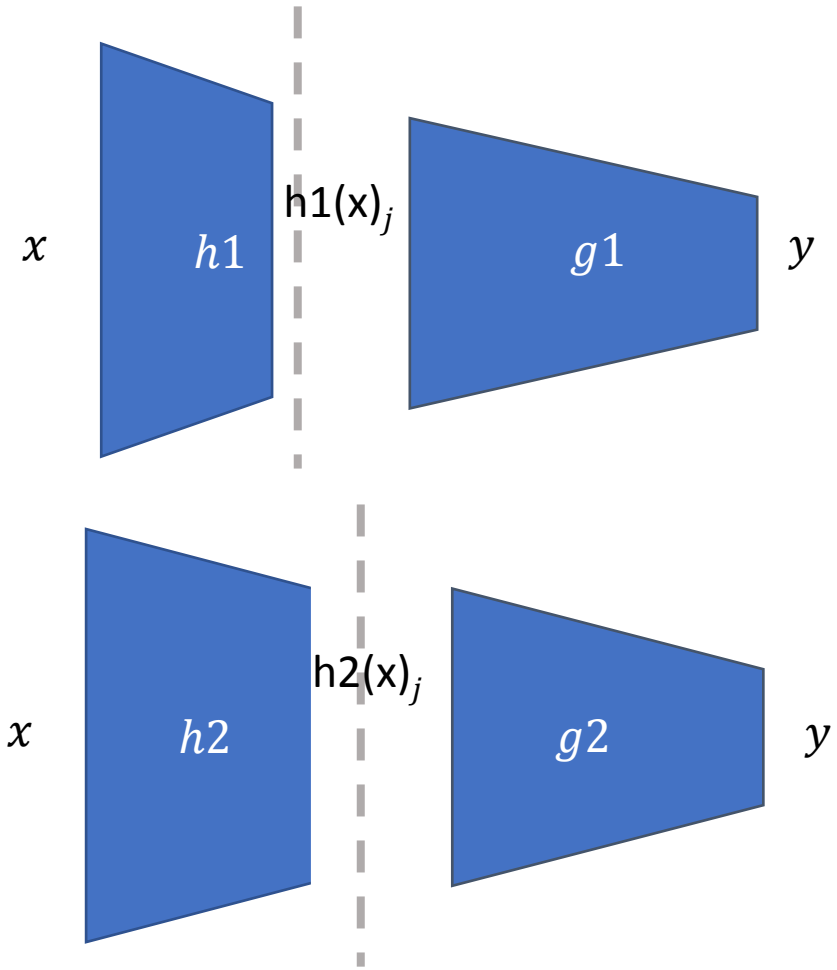
Axiom 3 (Distribution linearity (DL)). *For a family of distributions indexed by some* $a \in \mathcal{A}$, $P(x) = \int_{\mathcal{A}} g(a) P_a(x) da$, *then* $\chi_i(f, P) = \int_{\mathcal{A}} g(a) \chi_i(f, P_a) da$.

Unique input influence measure

Theorem 1. *The only measure that satisfies linear agreement, distributional marginality and distribution linearity is given by*

$$\chi_i(f, P) = \int_{\mathcal{X}} \left. \frac{\partial f}{\partial x_i} \right|_{\mathbf{x}} P(\mathbf{x}) d\mathbf{x}.$$

j-equivalent slices



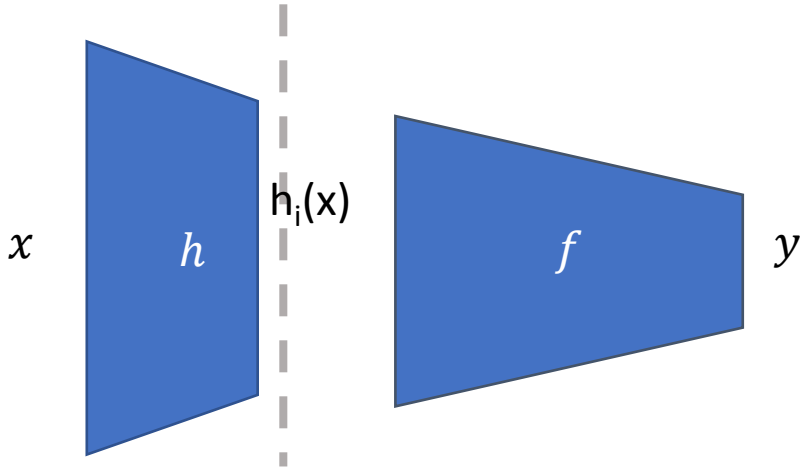
Two slices $s_1 = \langle g_1, h_1 \rangle$ and $s_2 = \langle g_2, h_2 \rangle$ are j-equivalent if for all $\mathbf{x} \in \mathcal{X}$, and $z_j \in \mathcal{Z}_j$, $h_1(\mathbf{x})_j = h_2(\mathbf{x})_j$, and $g_1(h_1(\mathbf{x})_{-j} z_j) = g_2(h_2(\mathbf{x})_{-j} z_j)$. Informally, two slices

Axioms

Axiom 4 (Slice Invariance). *For all j-equivalent slices s_1 and s_2, $\chi_j^{s_1}(f, P) = \chi_j^{s_2}(f, P)$.*

Consistency of input and internal influence

- Equate the input influence of an input with the internal influence of a perfect predictor of that input



Axioms

Axiom 5 (Preprocessing). *Consider* h_i *such that* $P(X_i = h_i(X_{-i})) = 1$. *Let* $s = \langle f, h \rangle$, *be such that* $h(x_{-i}) = x_{-i}h_i(x_{-i})$, *which is a slice of* $f'(\mathbf{x}_{-i}) = f(\mathbf{x}_{-i}h_i(\mathbf{x}_{-i}))$, *then* $\chi_i(f, P) = \chi_i^s(f', P)$.

Unique internal influence measure

Theorem 2. *The only measure that satisfies slice invariance and preprocessing is Equation 1.*

$$\chi_j^s(f,P) = \int_{\mathcal{X}} \left.\frac{\partial g}{\partial z_j}\right|_{h(\mathbf{x})} P(\mathbf{x})d\mathbf{x}$$

Goal

Explain behavior of deep neural networks by examining inner workings

Example questions

- What is the essence of a class from the network's point of view?
- What concept did the network use to classify an image into class A?
- What concept did the network use to classify an image into class A instead of class B?

Influence-directed explanations

[Leino, Li, Sen, Datta, Fredrikson 2018]

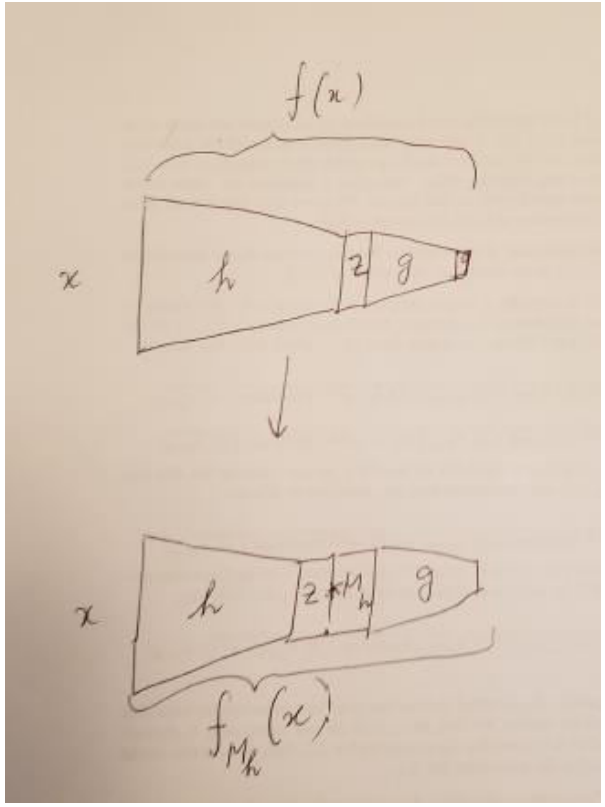
Key idea

- Identify causally influential neurons in internal layers
- Give them interpretation using visualization techniques

Thanks! Questions?

Validating the essence of a class

- Produce compressed model + convert to binary class predictor



$$f^i = \left(f|_i ; \sum_{j \neq i} f|_j \right)$$