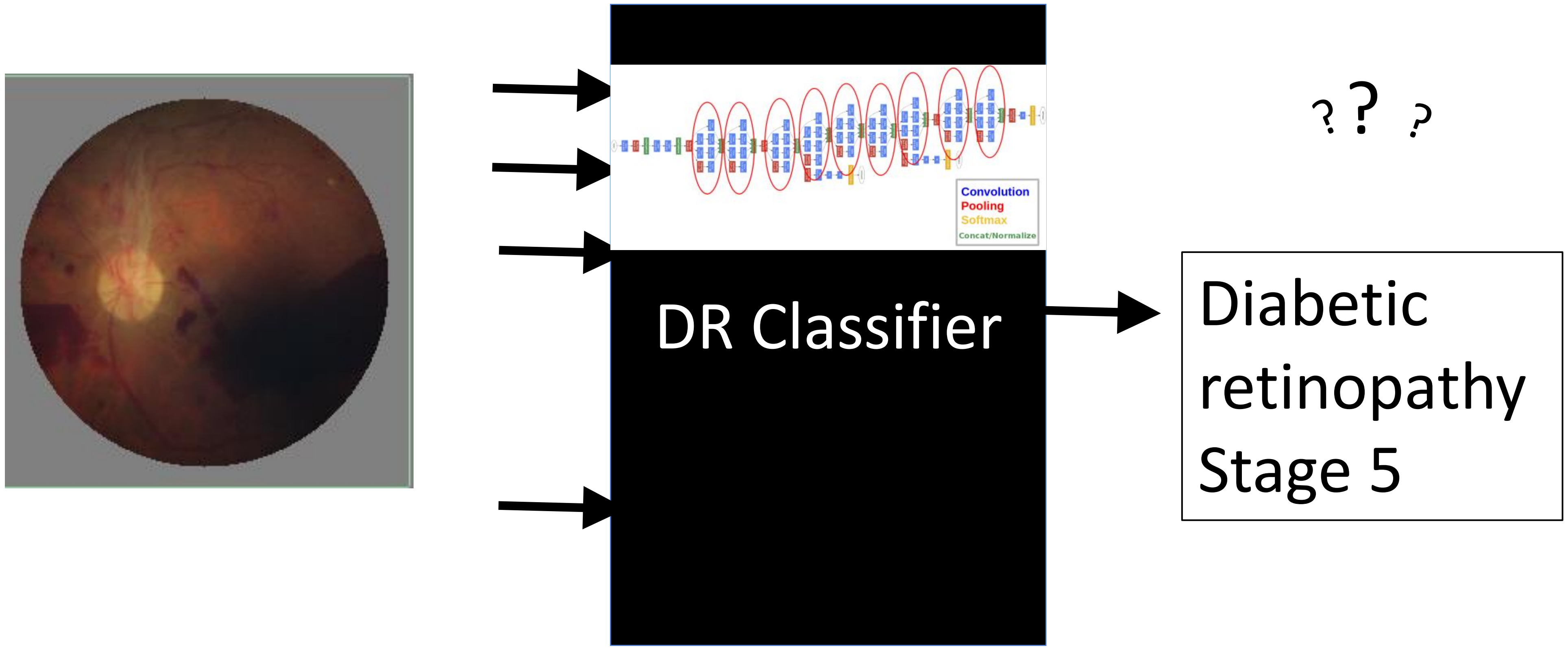
Influence-directed explanations for deep networks

Anupam Datta

John Mitchell

Stanford CS 329T, Spring 2021

Deep Learning Systems are Opaque



Why this diagnosis from the GoogleNet neural network?

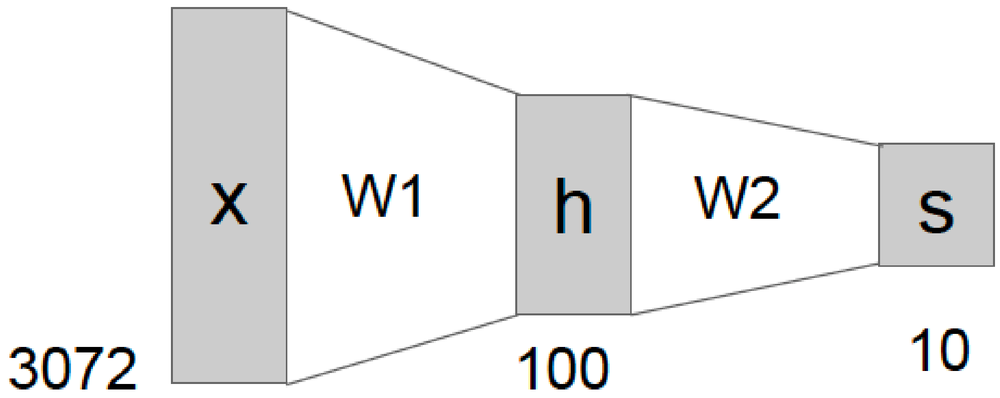
Vision: Explainable Deep Learning Systems

Reveal meaningful information about the logic of the machine learnt prediction/decision model

- Enable humans + machines to make better decisions together
- Build trust in and debug models
- Protect societal values (fairness, privacy)
- Applications: Finance, healthcare, …

2-Layer neural network

$$s = W_2 \max(0, W_1 x)$$



- Iterated construction: linear function followed by non-linear function
- A “deep network” has many such layers
- Difficult for humans to understand network behavior

Goals

1. Design mechanism for explaining behavior of deep neural networks by examining inner workings
   - What concept did the network use to classify an image into class A?
   - 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 instead of class B?

2. Evaluate explanation mechanism
   - Empirically and analytically

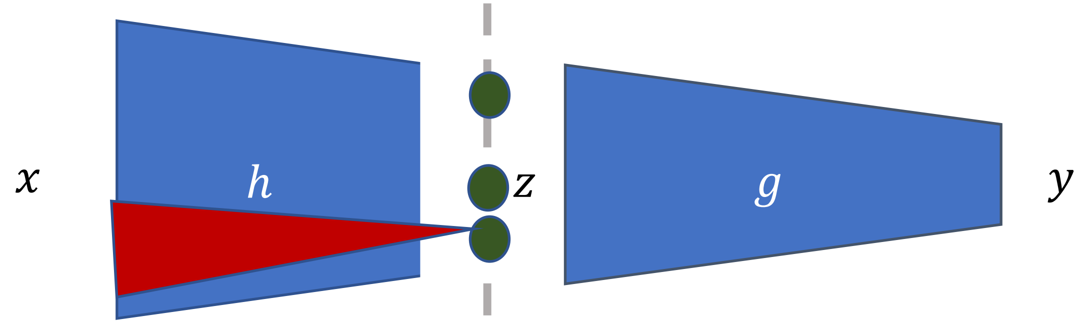
Influence-directed explanations [Leino, Sen, Datta, Fredrikson, Li 2018]

Explaining property of a ML system =
identify influential factors +
make them human interpretable

- Influence: What are important factors causing this model property?
- Interpretation: What do these factors mean?

Influence-directed explanations for deep networks

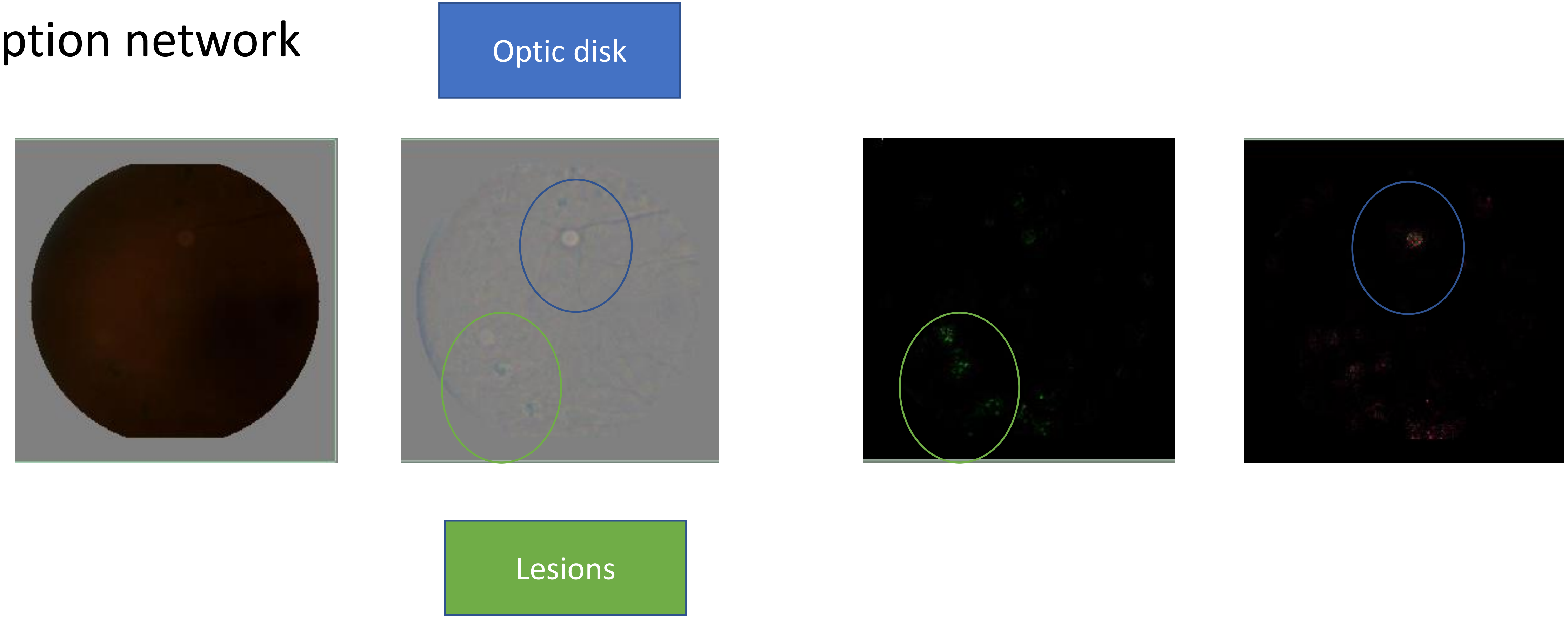
- Rank causally influential neurons in internal layers (novel!)
- Give them interpretation using visualization techniques (prior work)



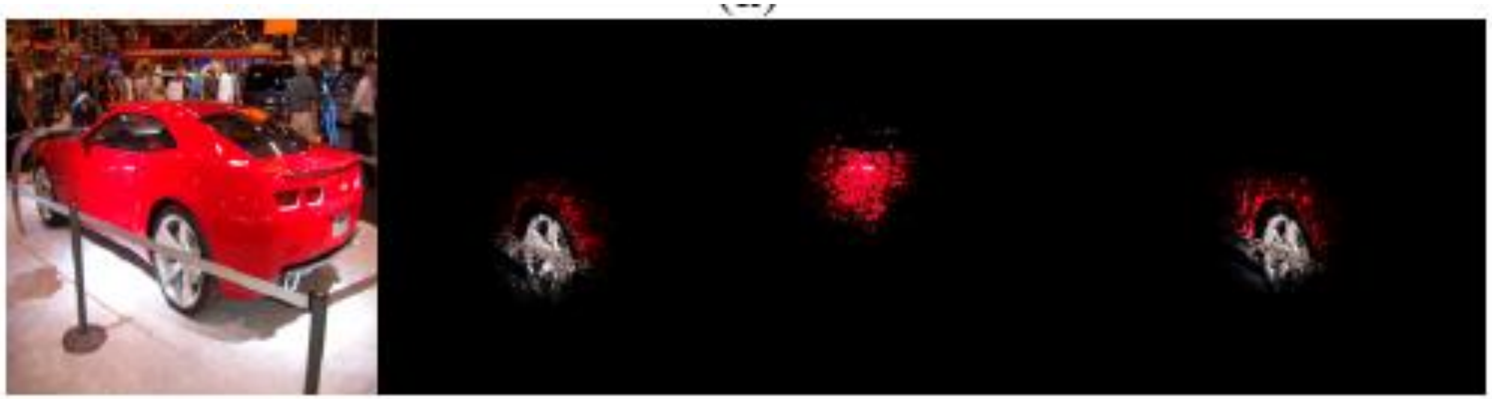
First result with internal influence measure for deep networks

Why classified as diabetic retinopathy stage 5?

Inception network



Why did the network classify input as sports car?



Input image Influence-directed Explanation

Why sports car instead of convertible?

VGG16 ImageNet model

Input image

Influence-directed Explanation

Uncovers high-level concepts that generalize across input instances

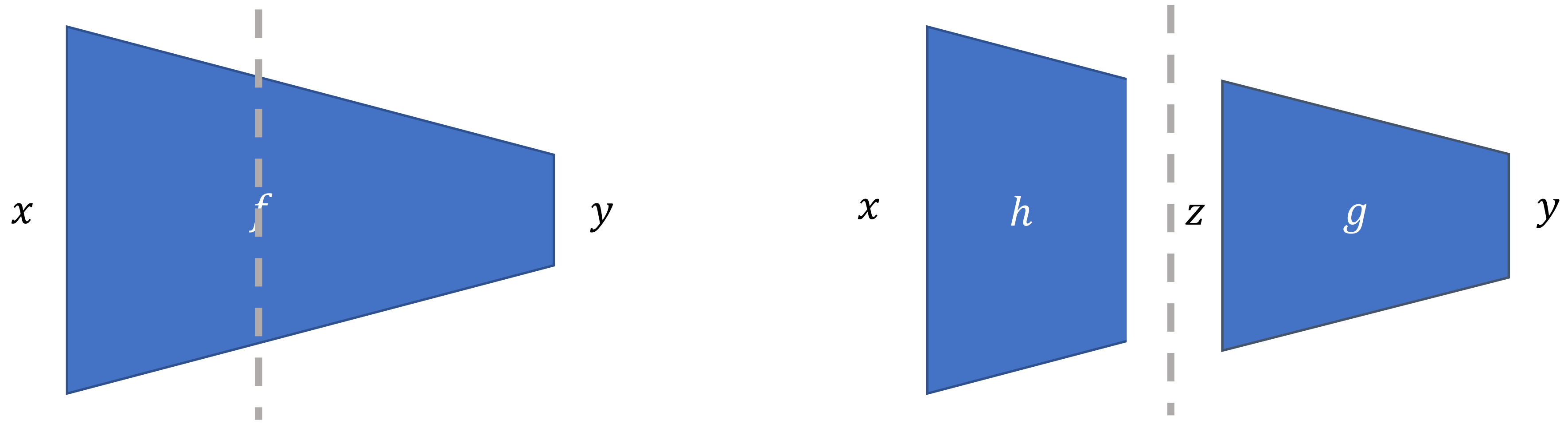

BMW

Outline

- Design of explanation mechanism
  - Distributional influence
  - Interpretation with visualization

- Evaluation of explanation mechanism
  - Explaining instances
  - Identifying influential concepts
  - Analytical justification

Decomposing network

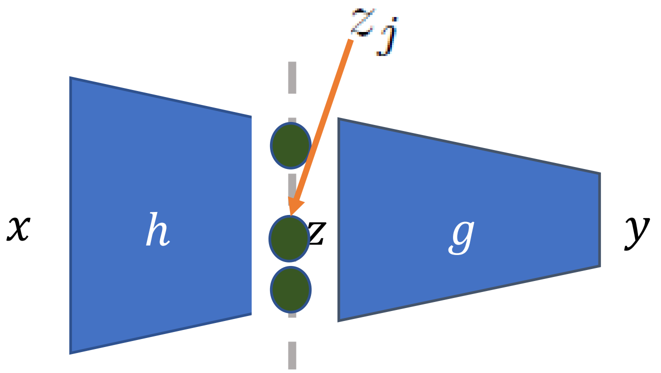


$$y = f(x) = g(h(x))$$

- 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

Influence = average gradient over distribution of interest



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

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

Gradient

Weighted by probability of input x

For input x [note z = h(x)]

Theorem: Unique measure that satisfies a set of natural properties

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”

Nearest neighbors

- Integrated gradients [Sundarajan et al., ICML 2017]
  - Input influence not internal influence
  - Analytically justified measure but different axioms

- Quantitative input influence [Datta et al., S&P 2016, Datta et al. IJCAI 2015]
  - Input influence not internal influence
  - Analytically justified measure but different axioms
  - Suited for non-differentiable model

Inspired by work in co-operative game theory

Related work

	Explanation framework properties		
	Quantity	**Distribution**	**Internal**
Influence-Directed	$\checkmark$	$\checkmark$	$\checkmark$
Integrated Gradients [3]		$\checkmark^{-}$	
Simple Taylor [4]		$\checkmark^{-}$	
Sensitivity Analysis [2]			
Deconvolution [5]			$\checkmark^{\dagger}$
Guided Backpropagation [6]			$\checkmark^{\dagger}$
Relevance Propagation [4]		$\checkmark^{-}$	$\checkmark^{\dagger}$

Only explain individual predictions

means to that end)

Outline

- Design of explanation mechanism
  - Distributional influence
  - Interpretation with visualization
- Evaluation of explanation mechanism
  - Identifying influential concepts
  - Analytical justification

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 [Simonyan et al. ICLR 2014]

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

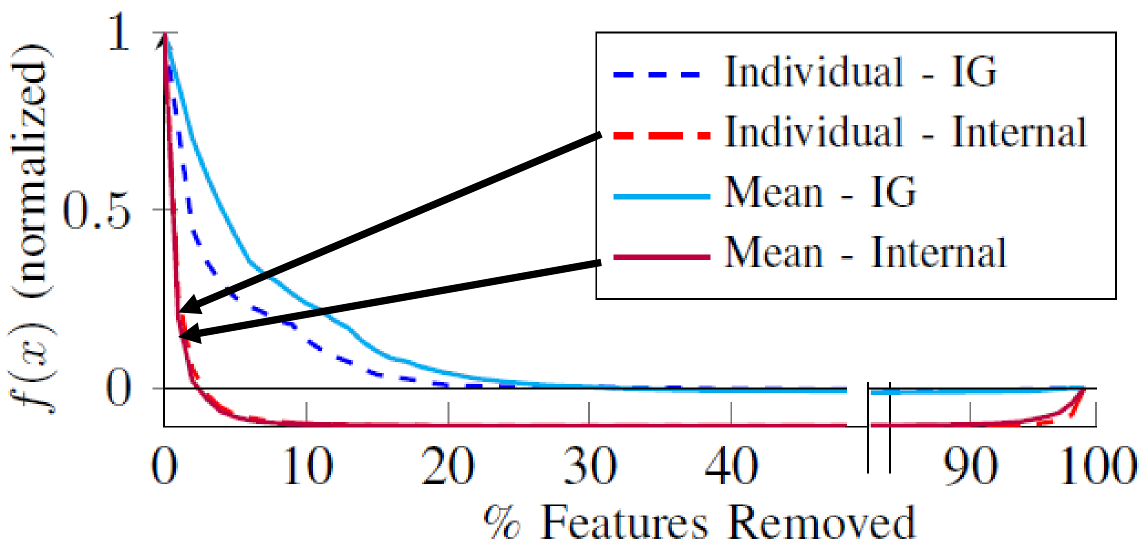
Outline

- Design of explanation mechanism
  - Distributional influence
  - Interpretation with visualization

- Evaluation of explanation mechanism
  - Identifying influential concepts
  - Analytical justification

Distributional influence captures general concepts

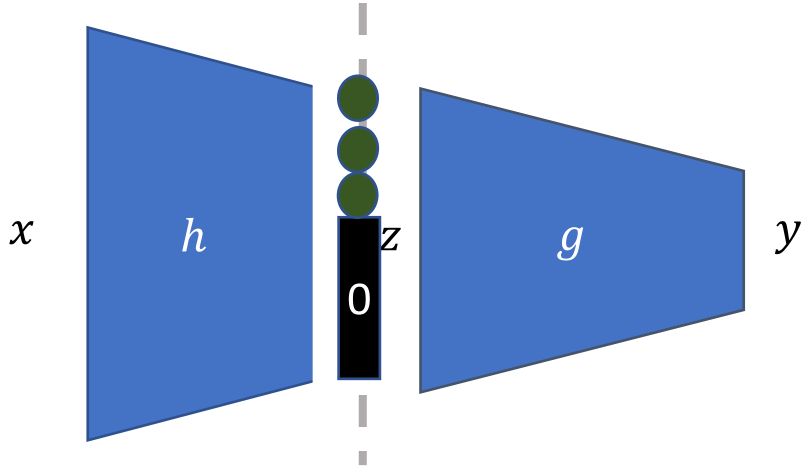
- Neurons influential for class on-average also influential for individual instances of class
- Not so for input influence (Integrated Gradients)



Score for correct class drops rapidly as most influential neurons are turned off

Validating the essence of a class

- Produce compressed model by keeping only most influential neurons for class *i*
- Convert to binary class predictor that distinguishes class *i* from all others



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

Validating the essence of a class

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

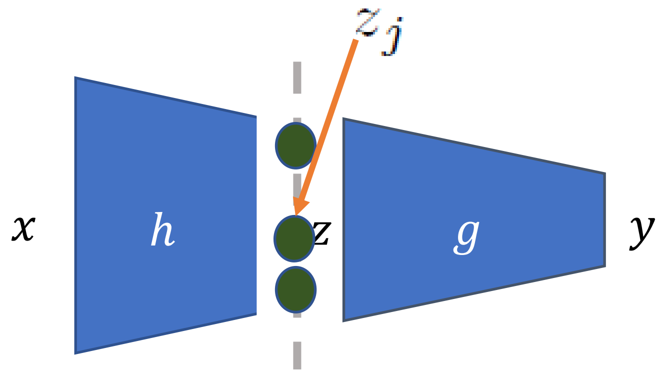
Compressed model with ~ top 1% influential neurons has comparable recall

Outline

- Design of explanation mechanism
  - Distributional influence
  - Interpretation with visualization
- Evaluation of explanation mechanism
  - Identifying influential concepts
  - Analytical justification

Unique measure theorem

Influence = average gradient over distribution of interest



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

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

Theorem: Unique measure that satisfies a set of natural properties

What are these “natural properties”?

1. Linear agreement
   - For linear models, the influence of an input variable is its coefficient

Novel ideas here!

2. Distributional faithfulness
   - Incorporate information about training distribution in influence measure

3. Internal influence invariances
   - Make influence measures depend only on the computed functions (ignoring differences in implementations)

Distributional marginality property

marginality *If*

$$\left(\frac{\partial f_1}{\partial x_i}\bigg|_X = \frac{\partial f_2}{\partial x_i}\bigg|_X \right)$$

then

$$\chi_i(f_1, P) = \chi_i(f_2, P).$$

- Marginality principle well known in co-operative game theory (e.g., Integrated Gradients)
- Restriction to distribution important for deep networks since network behavior unpredictable outside manifold

Summary

1. Design mechanism for explaining behavior of deep neural networks by examining inner workings
   - Distributional influence

2. Evaluate explanation mechanism
   - Empirically: explaining instances, identifying general concepts
   - Analytically: Unique influence measure that satisfies natural properties

Connections

- Explanations for other kinds of models
  - Shapley Values -- Datta et al. S&P 2016, Lundberg, Lee NIPS 2017
- Explanations to improve privacy and fairness
  - Part II, III of course
- Explanations that span the training process
  - Koh, Liang 2017, …
- Adversarial training, robustness and its interaction with explanations
  - Part IV of course

Thanks! Questions?

Additional slides

Formal properties

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

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

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*

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

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

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

Proof. Choose any function f and $P_{\mathbf{a}}(\mathrm{x}) = \delta(\mathrm{x}-\mathrm{a})$, where δ is the Dirac delta function on $\mathcal{X}$. Now, choose $f'(\mathrm{x}) = \frac{\partial f}{\partial \mathbf{x}_i}|_{\mathbf{a}} x_i$. By linearity agreement, it must be the case that, $\chi(f', P_{\mathbf{a}}(\mathrm{x})) = \frac{\partial f}{\partial x_i}|_{\mathbf{a}}$. By distributional marginality, we therefore have that $\chi_i(f, P_{\mathbf{a}}) = \chi_i(f', P_{\mathbf{a}}) = \frac{\partial f}{\partial x_i}|_a$. Any distribution P can be written as $P(\mathrm{x}) = \int_{\mathcal{X}} P(\mathrm{a})P_{\mathbf{a}}(\mathrm{x})d\mathrm{a}$. Therefore, by the distribution linearity axiom, we have that $\chi(f,P) = \int_X P(\mathrm{a})\chi(f, P_a)da = \int_{\mathcal{X}} P(\mathrm{a})\frac{\partial f}{\partial x_i}|_{\mathbf{a}} d\mathrm{a}$. $\square$

Related work

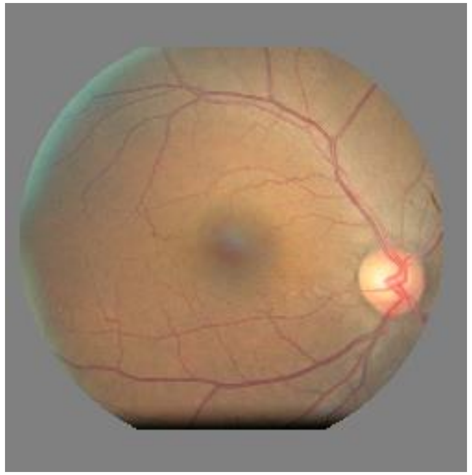
	Explanation framework properties			*Influence properties*	
	Quantity	**Distribution**	**Internal**	**Faithfulness**	**Sensitivity**
influence-directed	✓	✓	✓	✓*	✓
integrated gradients		✓*		✓*	✓
simple Taylor		✓*		✓*	✓
sensitivity analysis				✓	
deconvolution			✓†	✓	
guided backpropagation			✓†	✓	
relevance propagation			✓†	✓	✓*

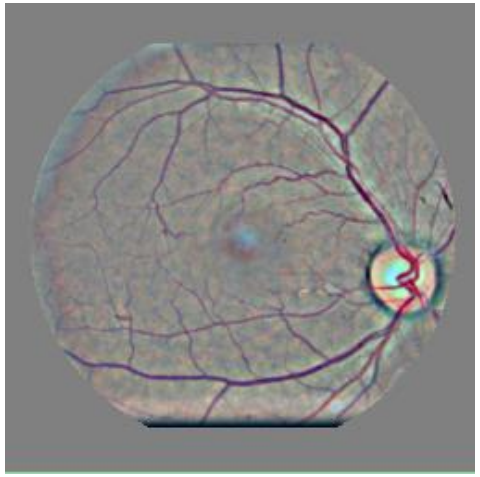
Diabetic retinopathy

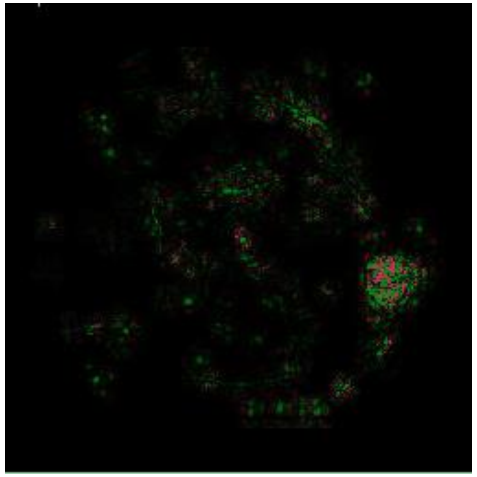
Source: American Academy of Ophthalmology

Debugging misclassification of stage 2 image

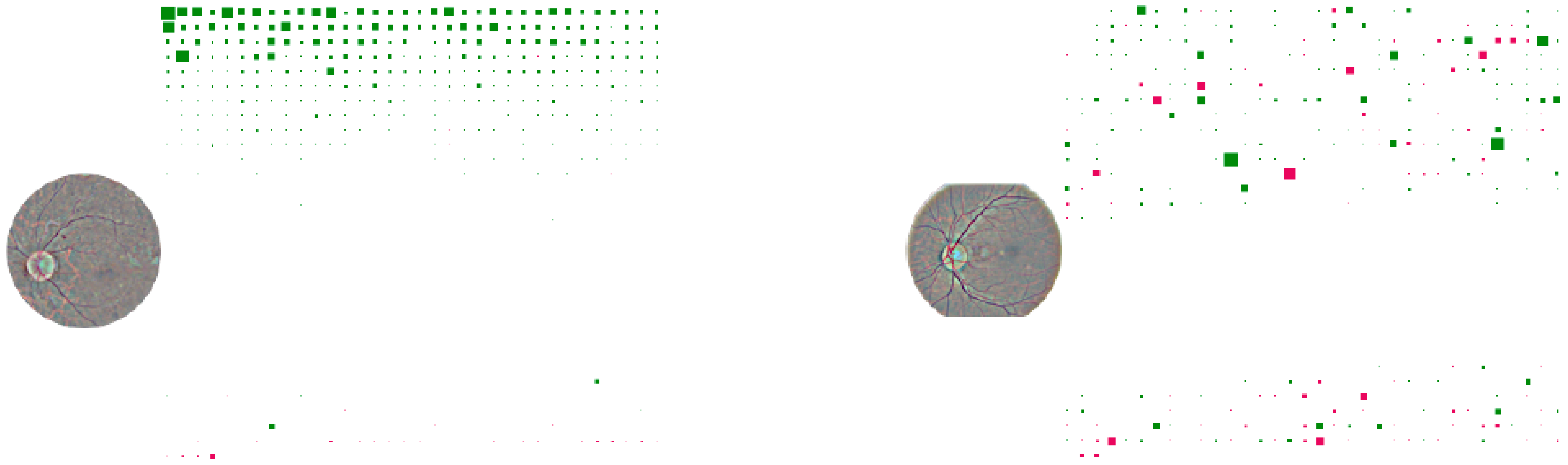
Inception network







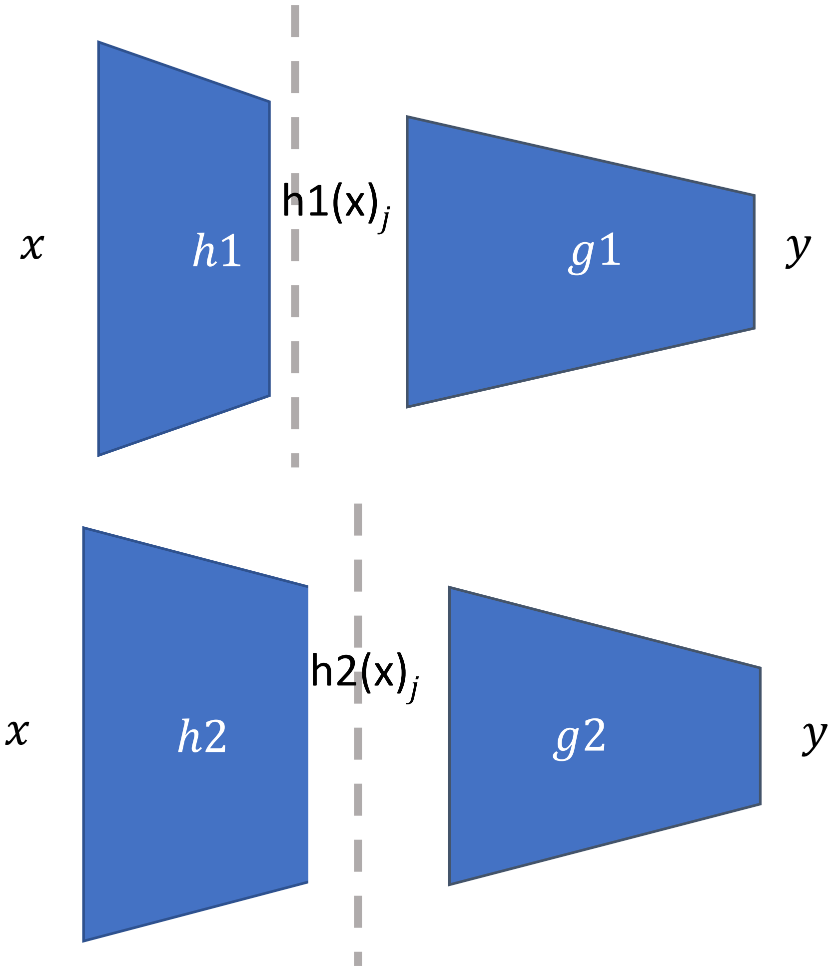
Misclassification as deviations from class influence profiles



(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.

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

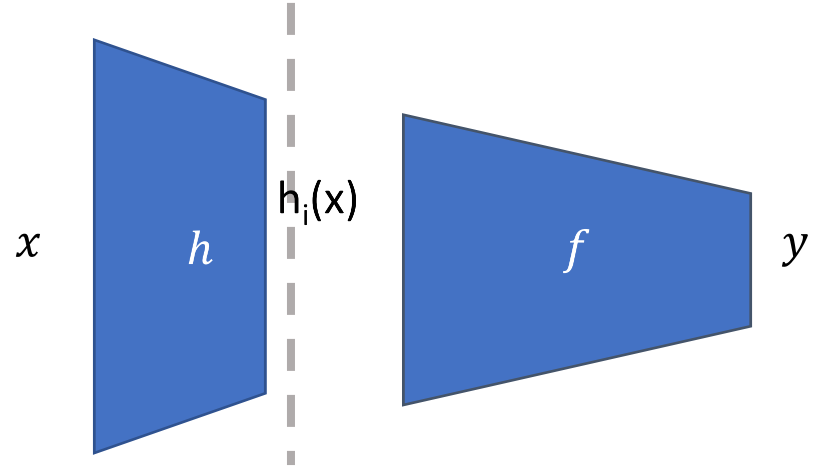
if for all $\mathbf{x} \in \mathcal{X}$, and $z_j \in \mathcal{Z}_j$, $h_1(\mathbf{x})_j = h_2(\mathbf{x})_j$, and $g_1(h_1(\mathbf{x})_{-j} z_j) = g_2(h_2(\mathbf{x})_{-j} z_j)$. Informally, two slices

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.*

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

Outline

- Design of explanation mechanism
  - Distributional influence
  - Interpretation with visualization

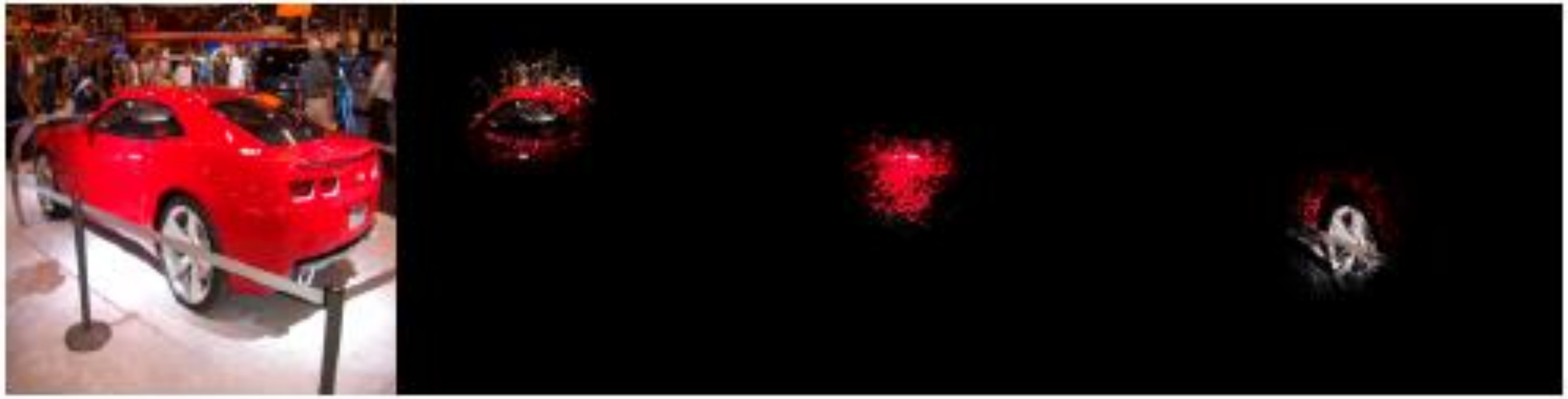
- Evaluation of explanation mechanism
  - Identifying influential concepts
  - Explaining instances
  - Analytical justification

Focused explanations from slices

Influence-directed Explanation

Integrated Gradients

Comparative explanations



Influence-directed Explanation