

Approaches for Explainability of AI-Enabled Systems in Medical Imaging

Berkman Sahiner, Ph.D.

Leader, Image Analysis Laboratory

Division of Imaging, Diagnostics, and Software Reliability

Office of Science and Engineering Laboratories

Center for Devices and Radiological Health

U.S. Food and Drug Administration

New Technology Necessitates New Vocabulary

Trustworthiness

Explainability

Fairness PERFORMANCE Generalizability

Robustness

Interpretability

Transparency Accountability

Augmented Intelligence

LOCKED ALGORITHM *Continuous Learning* Causality

Autonomous System **BLACK BOX**

Explainability and Interpretability

- In this presentation, I will use interpretability and explainability interchangeably, although some in the field distinguish between them
 - Roughly, AI revealing underlying causes to its decision making
- However, I will distinguish between interpretability/explainability and explanation

What is an Explanation?

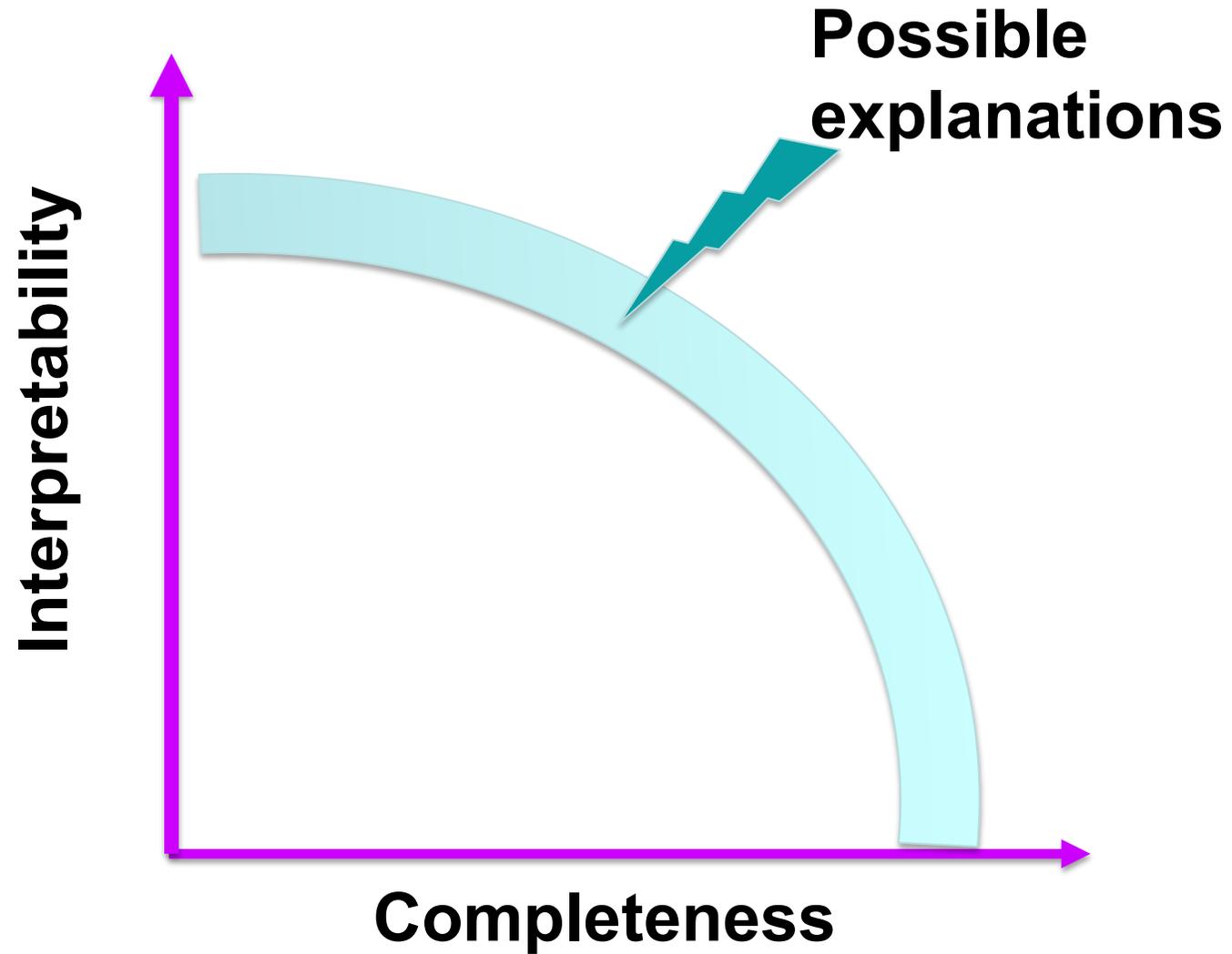
- An explanation is *typically* an answer to a why-question*
 - Why did the AI system diagnose this case with pneumonia?
- Why-questions are typically contrastive, even when not explicitly so
 - Why did the AI system diagnose this case with pneumonia **and not pleural effusion**?
 - Why did the AI system diagnose this case with pneumonia **but not the previous case**?

*Miller T, “Explanation in AI: Insights from the Social Sciences,” arXiv 1706.07269

Interpretability and Completeness

- An explanation can be evaluated in two ways: according to its interpretability, and according to its completeness*
- Interpretability:
 - Describe the internals of the system in a way that is understandable to humans
 - Tied to the cognition, knowledge, and biases of the user
- Completeness:
 - Describe the operation of a system in an accurate way
 - A perfectly complete explanation can always be given by revealing all the mathematical operations and parameters in the system

Levels of Explanation



Levels of Explanation

- When asked to provide an explanation, people provide different levels of explanation depending on
 - Who is asking
 - Is it my 5-yr old niece or a graduate student?
 - What they believe are the most relevant causes
 - The chain of causes needs to be truncated at some level: Depth of explanation
- In health AI, we have different “users” and different needs for “depth of explanation”
 - Patients, caregivers, insurance, regulators, scientific community

Types of Explanations for Deep Networks*

- 1) Explaining the processing of the data
- 2) Explaining the representation of data
- 3) Creating explanation-producing systems

Will illustrate with examples from deep networks in imaging

Explaining the Processing of the Data

- Most common
- Includes
 - Proxy models
 - An interpretable model that behaves similarly to the original in a local neighborhood
 - Linear proxy models, rule extraction, proxy decision trees, etc.
 - Occlusion studies, saliency maps, class activation maps
 - Highlight which portion of the computation is most relevant

Linear Proxy Model: LIME (Local Interpretable Model-Agnostic Explanations)



Original Image



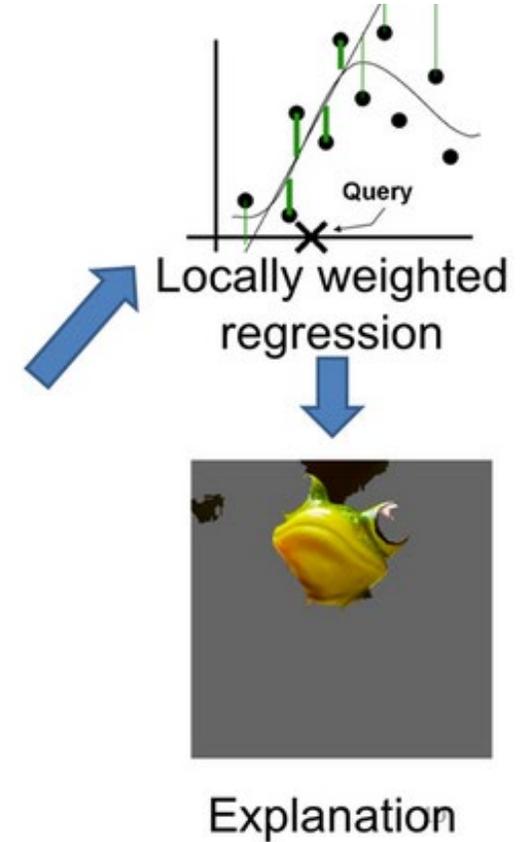
Interpretable Components



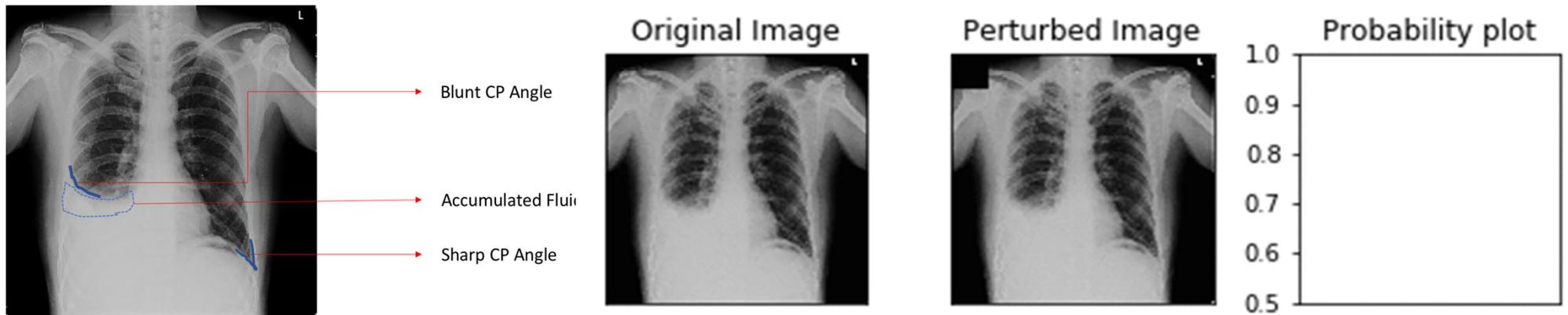
Original Image
P(tree frog) = 0.54



Perturbed Instances	P(tree frog)
	0.85
	0.00001
	0.52



Occlusion Visualization



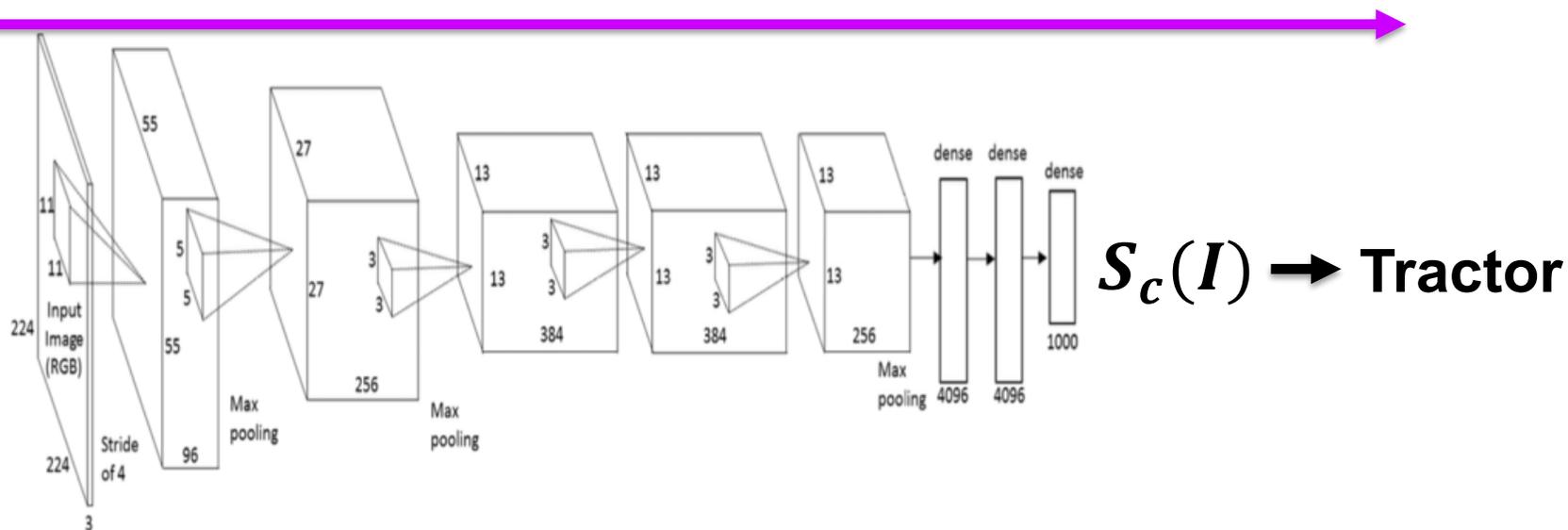
When the right costophrenic angle and accumulated fluid region is occluded to the network, the probability of the pleural effusion drops

Saliency Maps

How to tell which pixels matter for classification



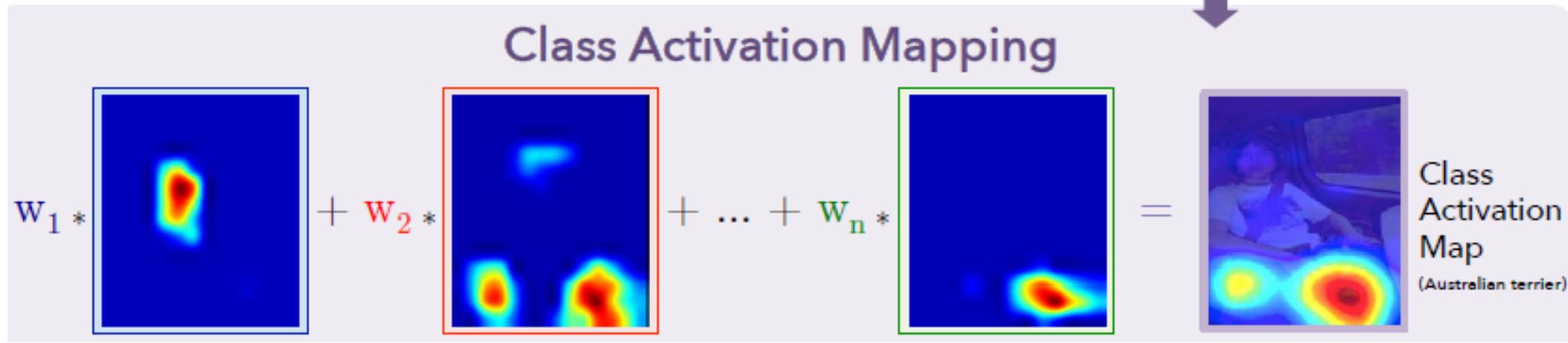
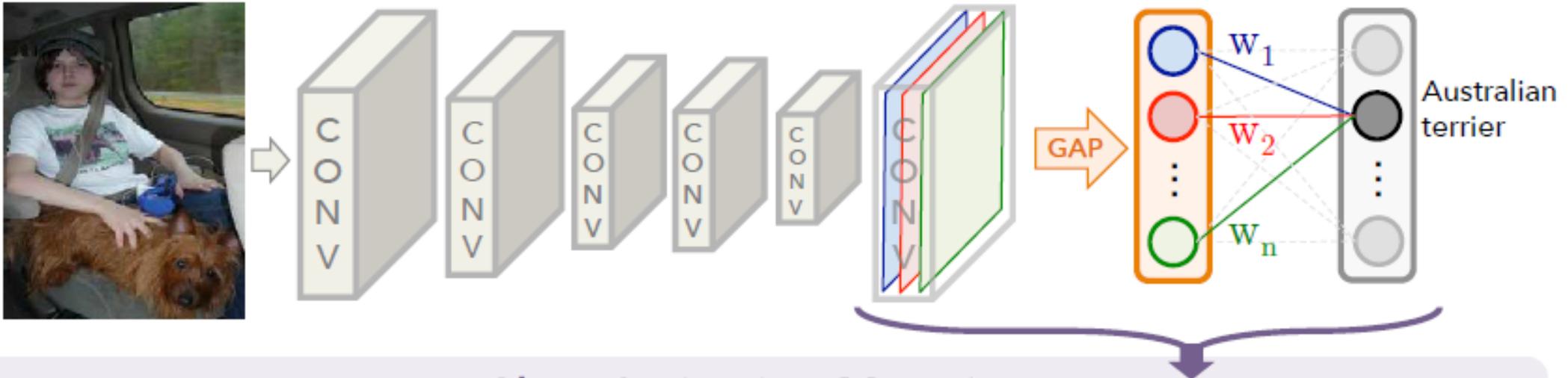
=



Compute gradient of $S_c(I)$ with respect to image pixels, take absolute value and max over RGB channels

Simonyan K et al, "Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps," arXiv 1312.6034

Class Activation Maps

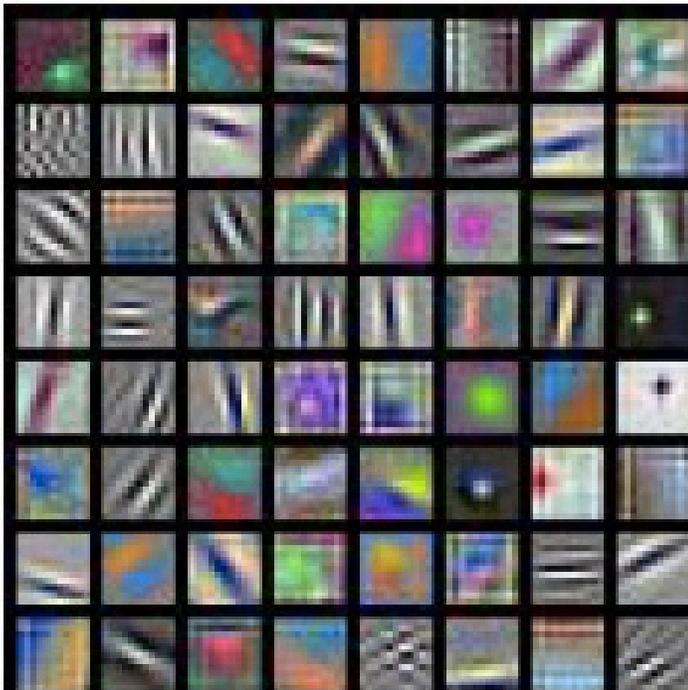


Zhou K et al, "Learning Deep Features for Discriminative Localization," arXiv 1512.04150

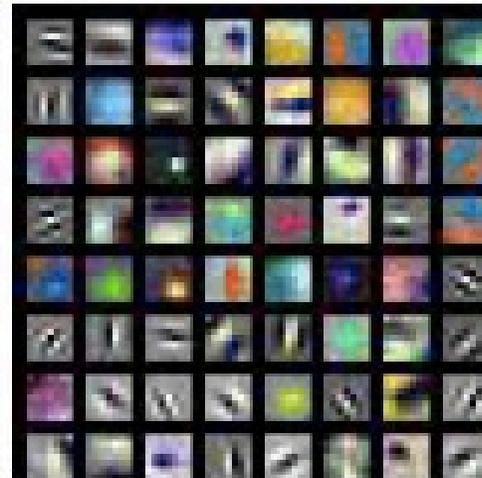
Explaining the Representation of Data

- Understand the role and structure of the deep network subcomponents
 - By layer
 - By unit
 - By vector

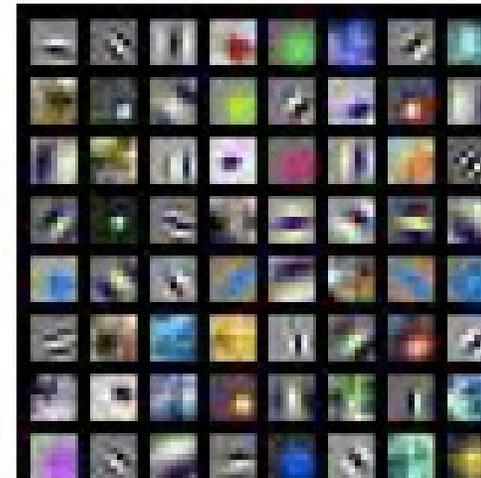
Visualization of First Layer Filters



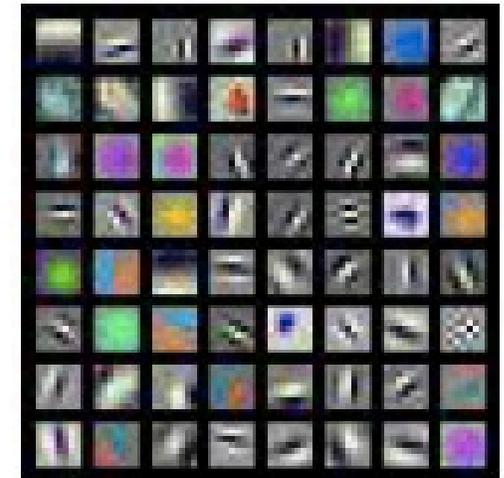
AlexNet:
64 x 3 x 11 x 11



ResNet-18:
64 x 3 x 7 x 7



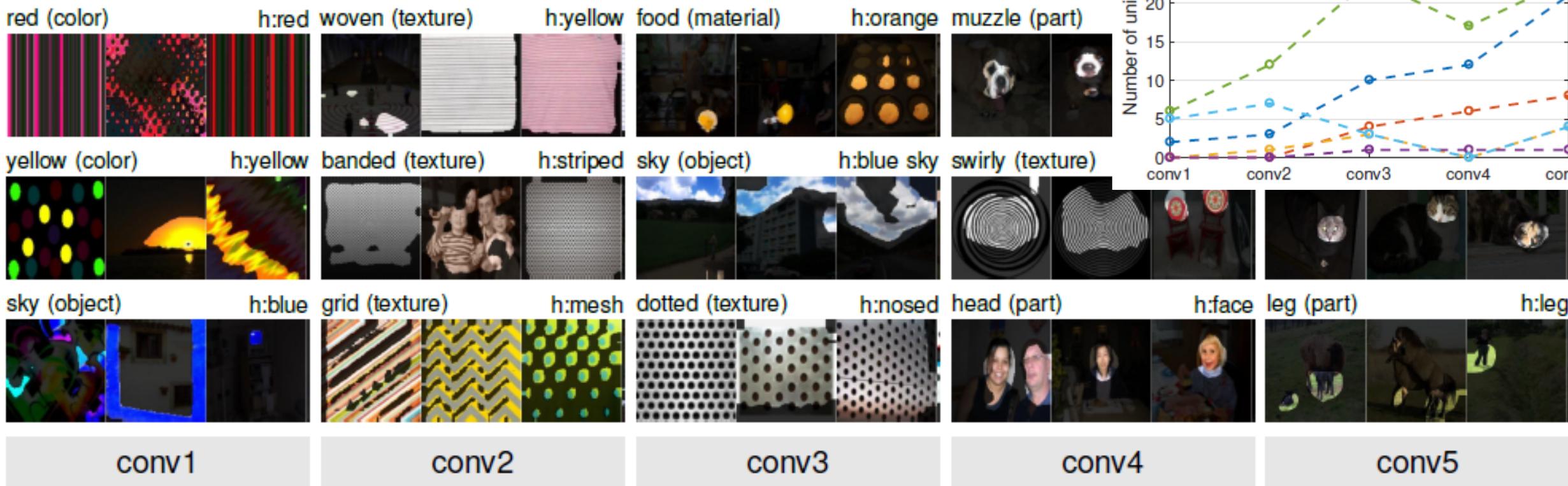
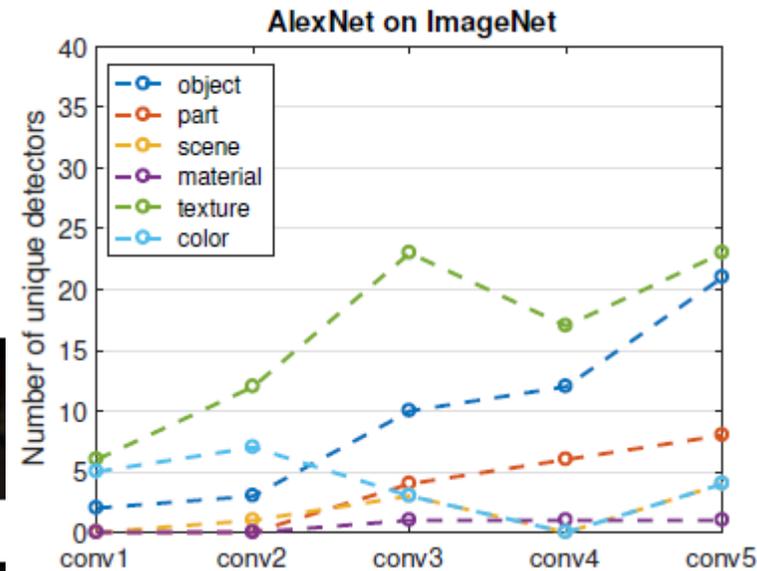
ResNet-101:
64 x 3 x 7 x 7



DenseNet-121:
64 x 3 x 7 x 7

Weights are useful to visualize because well-trained networks usually display smooth filters without noisy patterns.

Network Dissection



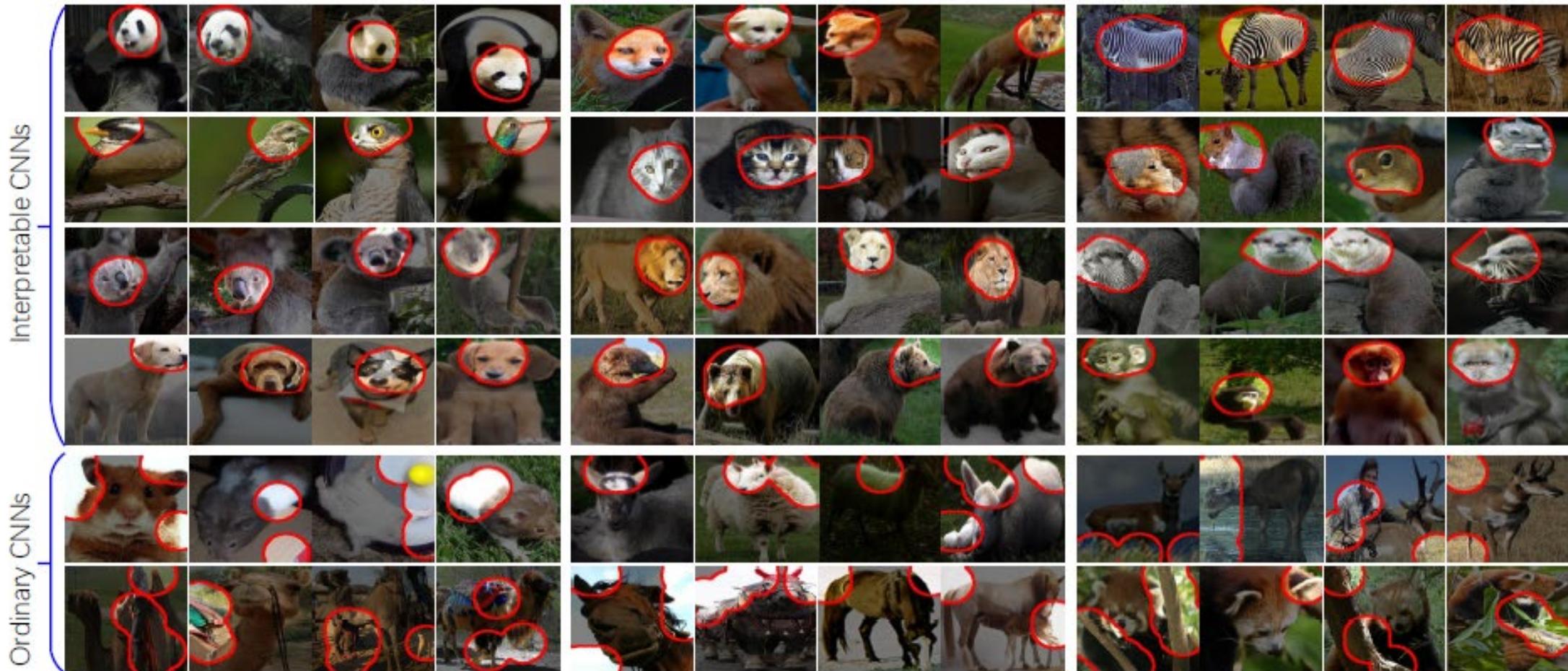
A comparison of the interpretability of all five convolutional layers of AlexNet (five layers, three units from top to bottom for each layer). For each unit, the segmentation generated by that unit is shown superimposed on the three example input images.

Bau D et al, "Network Dissection: Quantifying Interpretability of Deep Visual Representations," arXiv 1704.05796

Creating Explanation-Producing Systems

- Disentangled representations
 - Describe meaningful and independent factors of variation
 - Similar to principal component analysis or independent component analysis
- Explicitly train networks to generate explanations

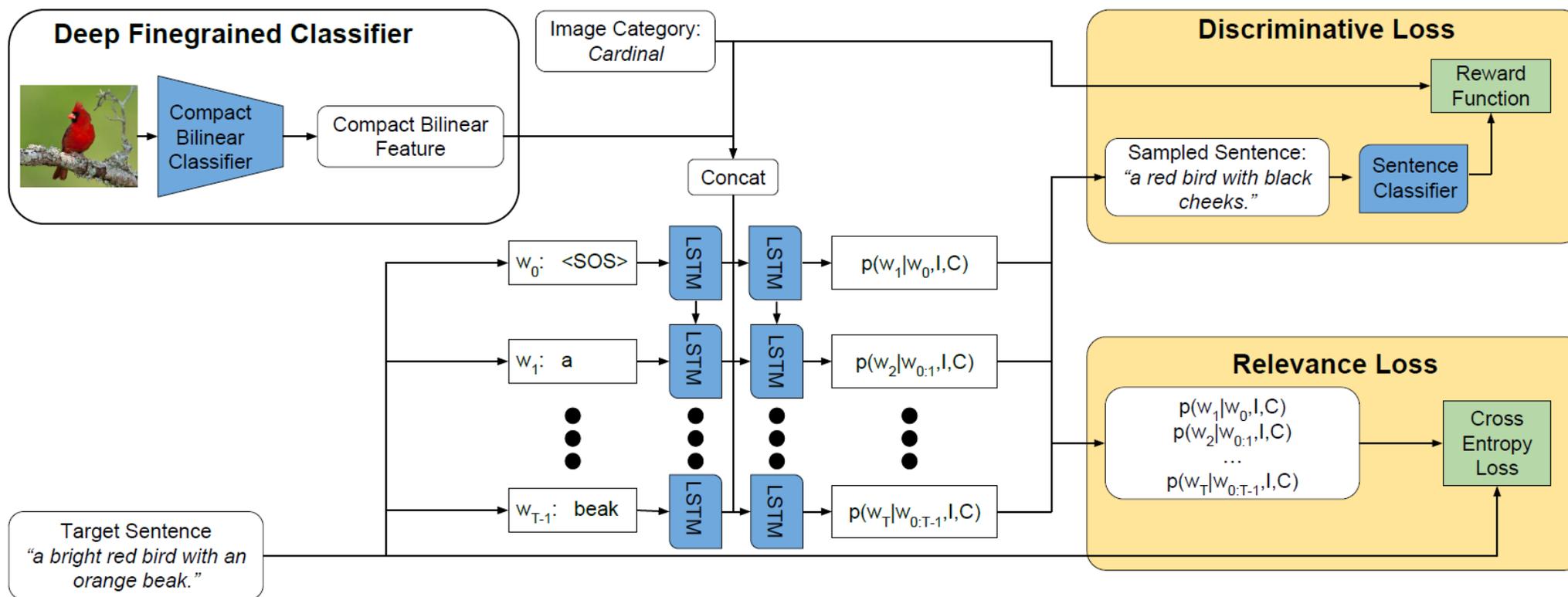
Distangled Representations



Visualization of receptive fields of filters in top conv-layers

Trained with a modified loss function that forces different units to focus on different object parts of a given class

Network Explicitly Trained to Generate Explanations



Evaluating Explanations*

- **Application-grounded: Real users, real tasks**
 - Evaluate the quality of an explanation in the context of its end-task, e.g., whether it results in better identification/reduction of errors
- **Human-grounded: Real humans, simplified tasks**
 - Example: Humans are presented with pairs of explanations, and must choose the one that they find of higher quality
- **Functionally-grounded: No humans, proxy tasks**
 - Example: Show that one's method performs better with respect to certain regularizers, e.g., is more sparse/distangled compared to other baselines

*Doshi-Velez F and Kim B, "Towards A Rigorous Science of Interpretable Machine Learning," arXiv 1702.08608

Summary

- Crisp, scientific definitions of the topics discussed in this workshop will advance the field
- Different levels of explainability may be needed depending on the user and desired depth of explanation
- A number of methods have been proposed in the literature to explain deep networks to humans
 - Interpretability and completeness are two competing factors
 - Methods to evaluate explainability need further attention