AIMLAI @ ECML PKDD 2026
September 7th 2026

Early Interpretability Methods

José Oramas

VUB ARTIFICIAL INTELLIGENCE RESEARCH GROUP
Universiteit Antwerpen
fwo Flanders AI Research Program

Coming up...

Early Interpretability Methods

José Oramas

From Classical to Mechanistic Interpretability

Ward Gauderis, José Oramas & Thomas Dooms

From Mechanistic to Compositional Interpretability

Ward Gauderis & Thomas Dooms

This chapter...

Post-hoc semantic mapping of neuron activations

What does a neuron do?

Inherently interpretable methods

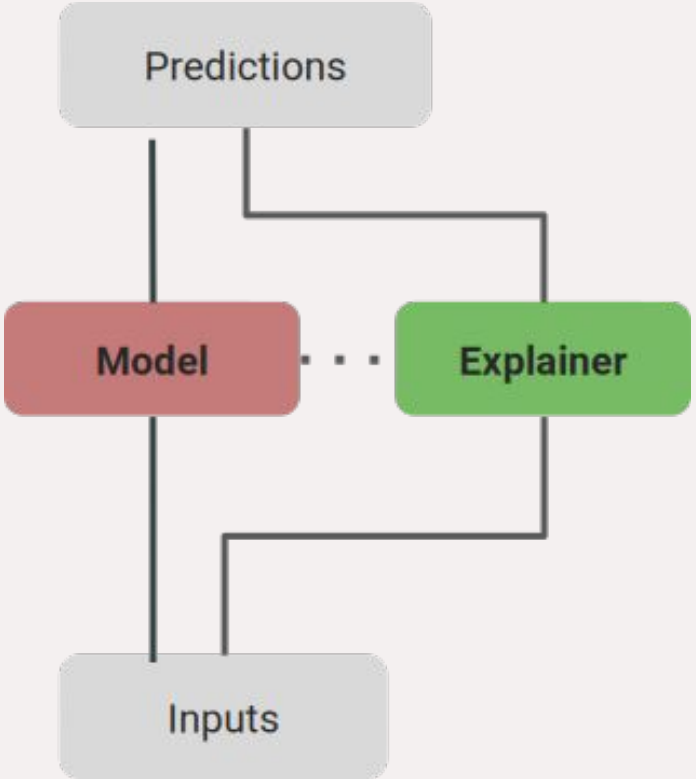
Can we make neurons understandable?

Evaluation protocols

When do we understand neurons?

Post-hoc semantic mapping of neuron activations.

What does a neuron do?



Post-hoc approaches

Feature visualization

Idea: Dense visual inspection of projected features across the model.

How?

- Feed-forward until a neuron of interest.

- Project the activations back to the input space.

- Collect visualizations of top activations.

(Zeiler & Fergus. 2014)

Feature visualization

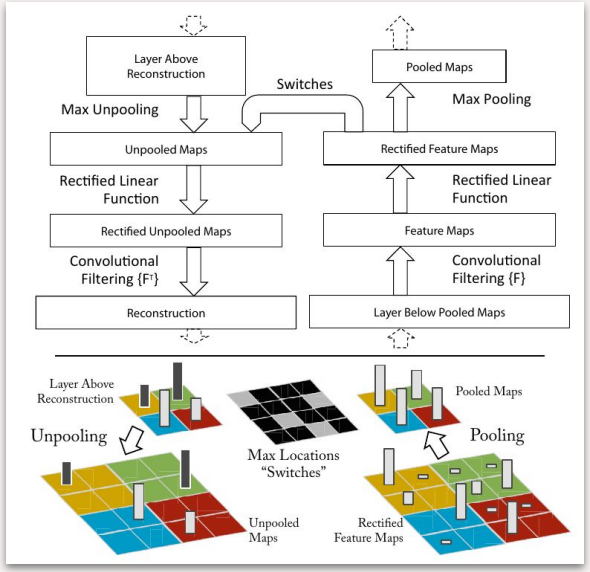
Idea: Dense visual inspection of projected features across the model.

How?

- Feed-forward until a neuron of interest.

- Project the activations back to the input space.

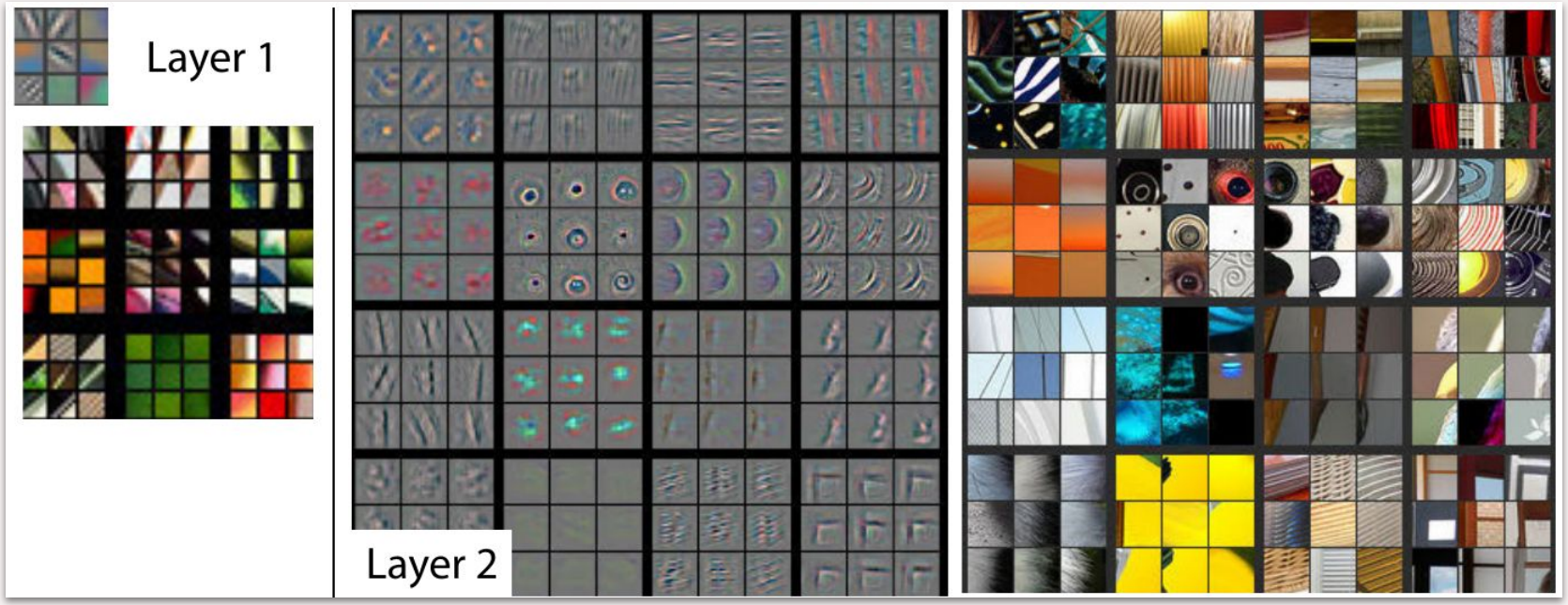
- Collect visualizations of top activations.



(Zeiler & Fergus. 2014)

Feature visualization

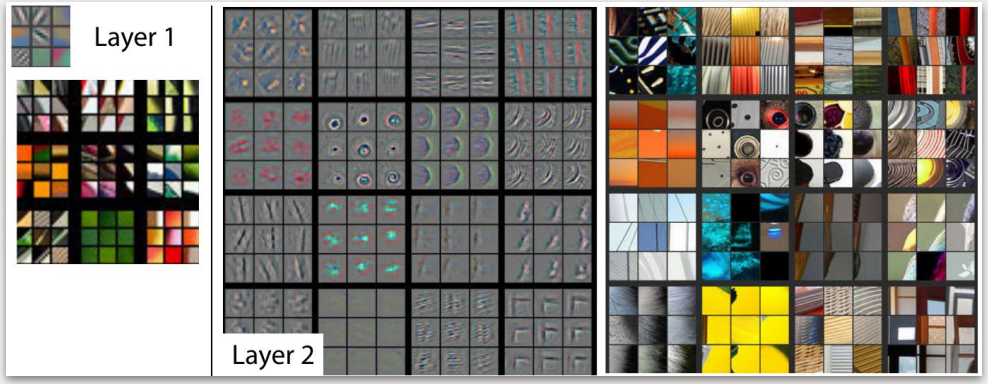
Idea: Dense visual inspection of projected features across the model.

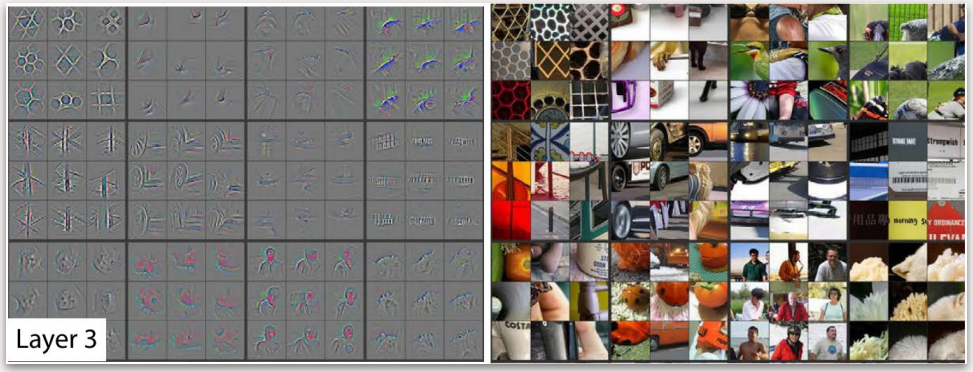


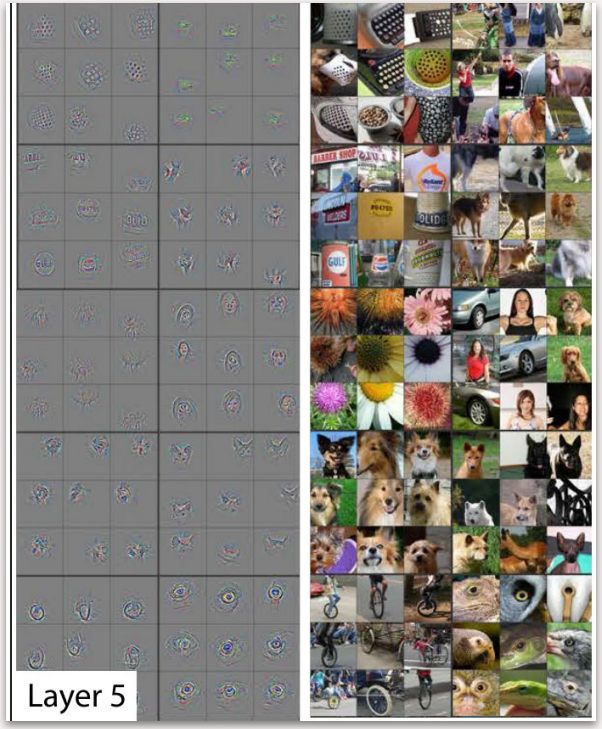
(Zeiler & Fergus. 2014)

Feature visualization

Idea: Dense visual inspection of projected features across the model.



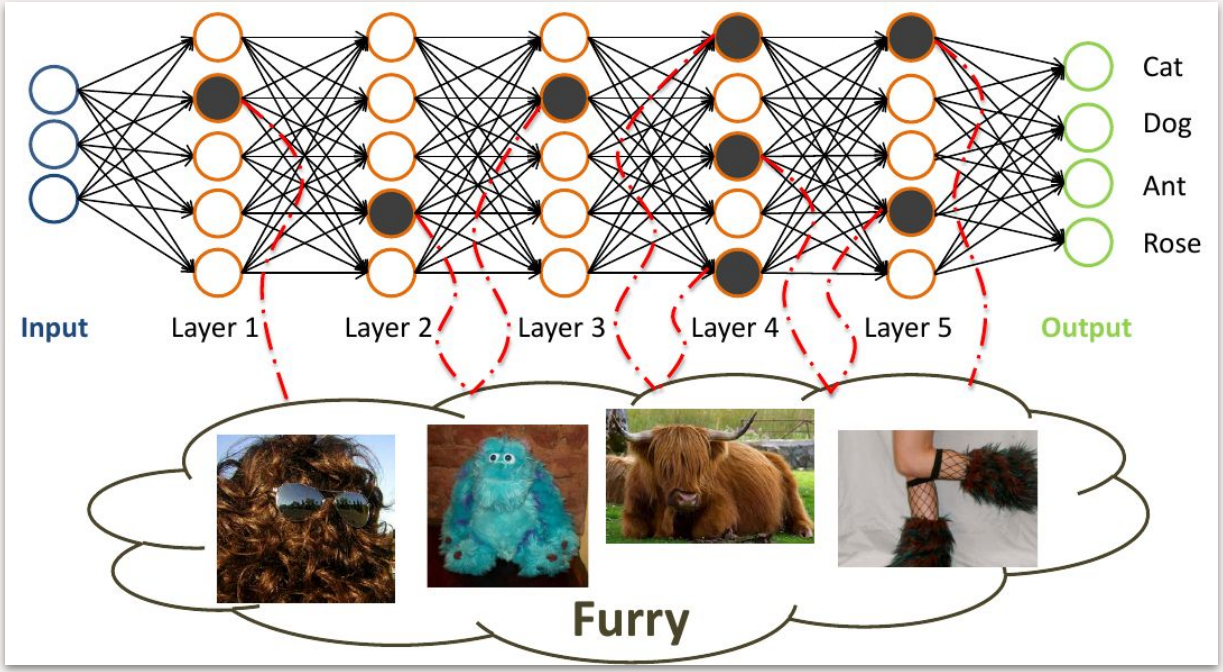




(Zeiler & Fergus. 2014)

Linking neuron activations with visual attributes.

Idea: Identify relationships between activations and attributes.

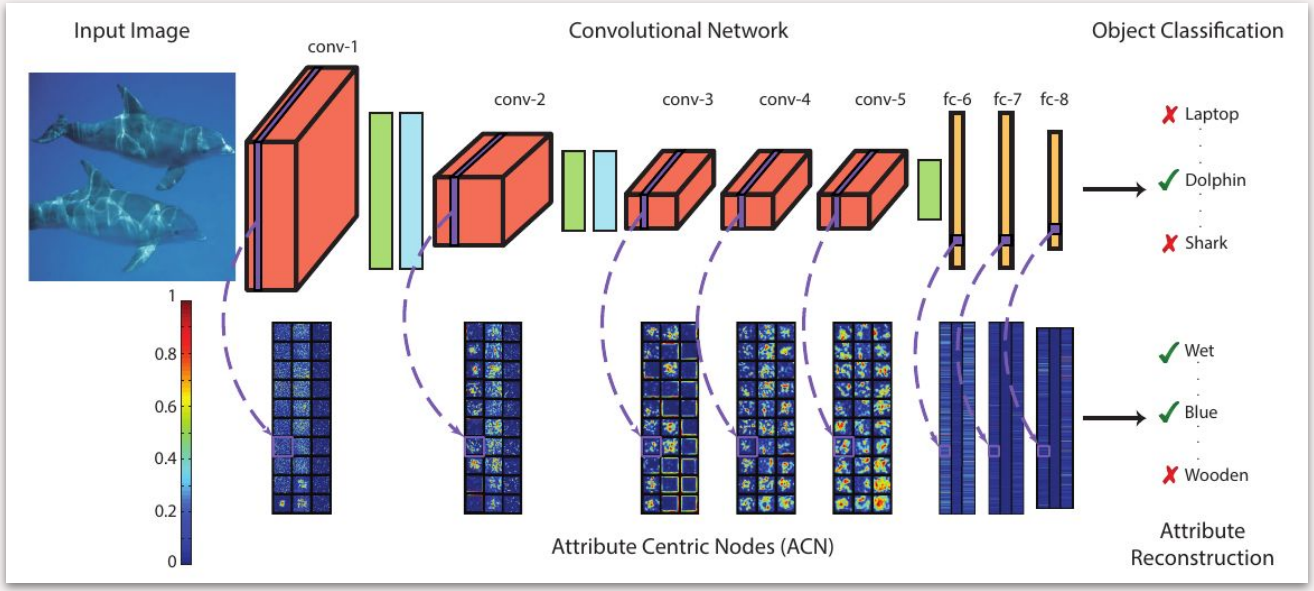


Assumption: there is a set of units responsible of encoding certain information

(Escorcia et al. 2015)

Linking neuron activations with visual attributes.

Idea: Identify relationships between activations and attributes.



How:

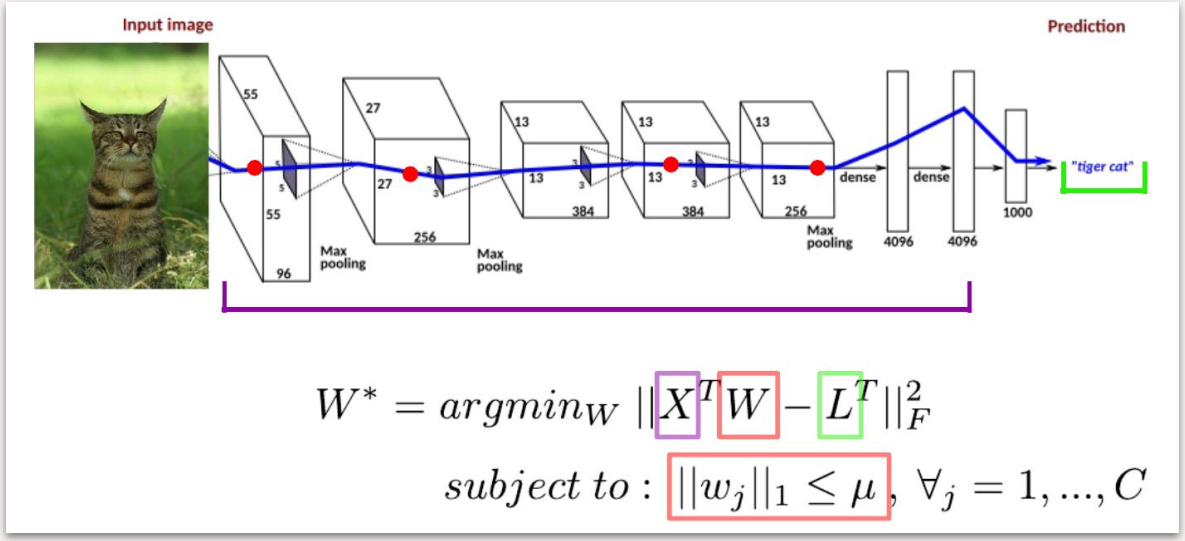
$$\mathbf{W}^* = \underset{\mathbf{W}}{\operatorname{argmin}} \ \|\mathbf{X}^\top \mathbf{W} - \mathbf{L}^\top\|_F^2$$
$$\text{subject to: } \|\mathbf{w}_j\|_1 \leq \mu_j \ \ \forall j = 1, \cdots, d$$

(Escorcia et al. 2015)

Linking neuron activations wrt. the original task.

Idea: Identify relationships between activations and the classes of interest.

Advantage: No need for additional data.



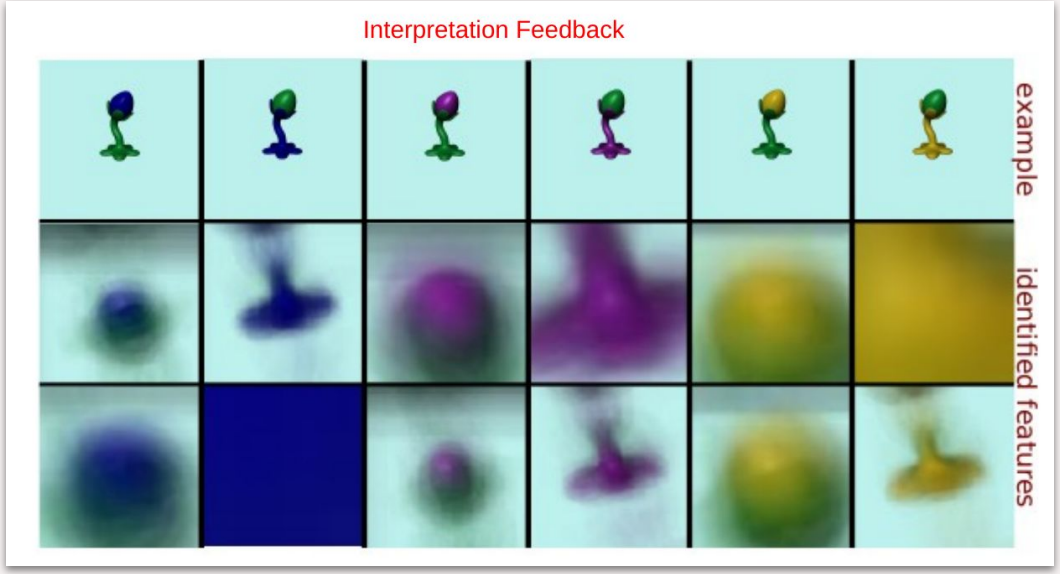
Assumption: there is a set of units responsible of encoding certain information

(Oramas et al. 2017)

Linking neuron activations wrt. the original task.

Idea: Identify relationships between activations and the classes of interest.

Advantage: No need for additional data.

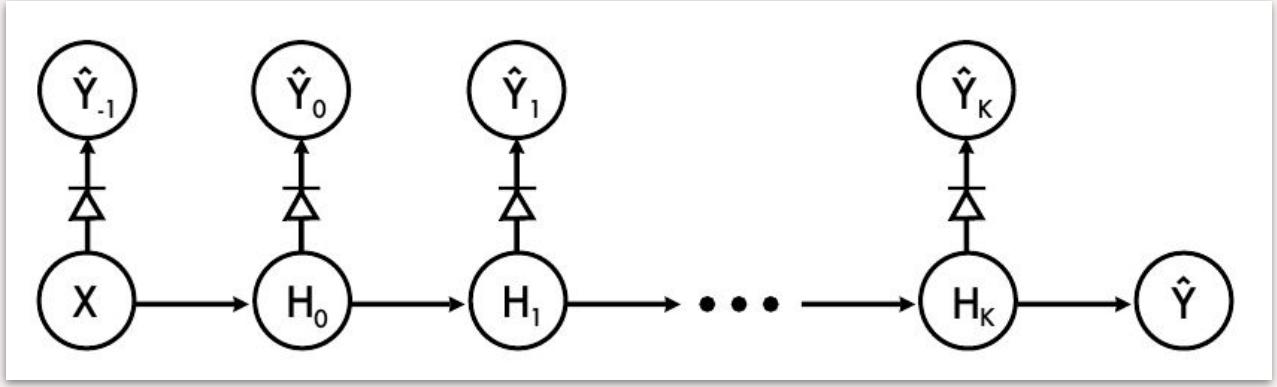


Assumption: there is a set of units responsible of encoding certain information

(Oramas et al. 2017)

Analyzing the predictive signal of neuron activations.

Idea: Use simple classifier (probes) to inspect the representation.

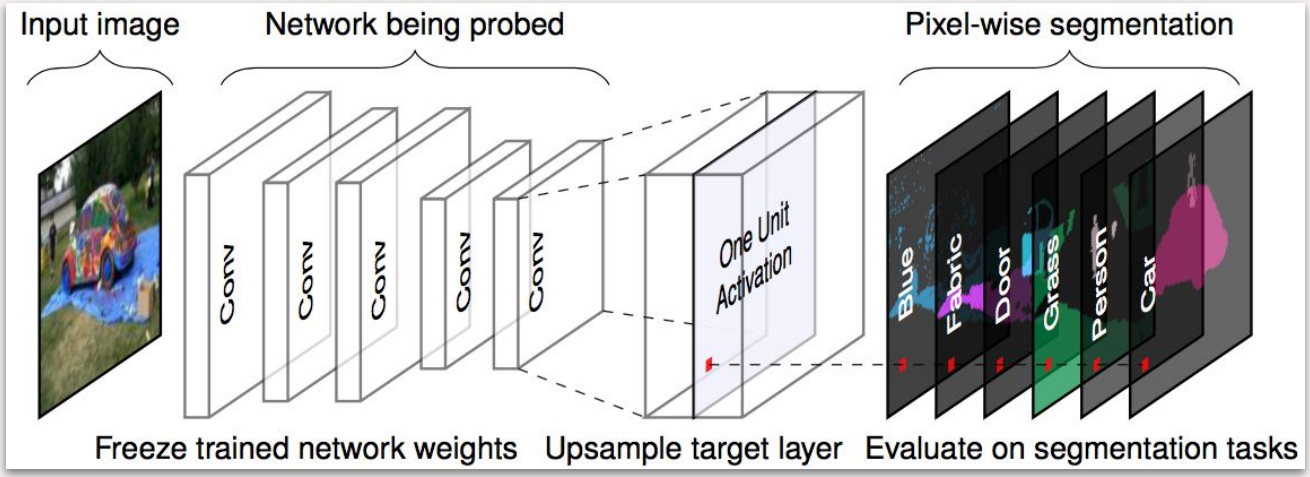


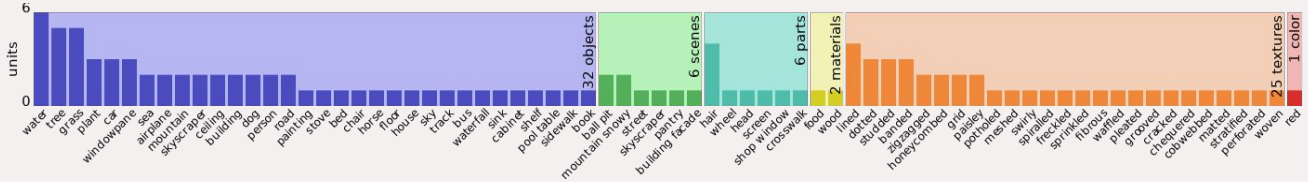
- Lightweight probes -> low capacity & expressivity
- Measure predictive performance across layers (and throughout training)
- Performance on the original task vs. on a new task

(Alain & Bengio, 2017)

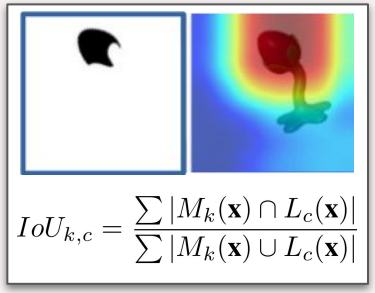
Linking neuron activations with semantic concepts.

Idea: Quantify model interpretability with segmentation overlap.





Measuring overlap

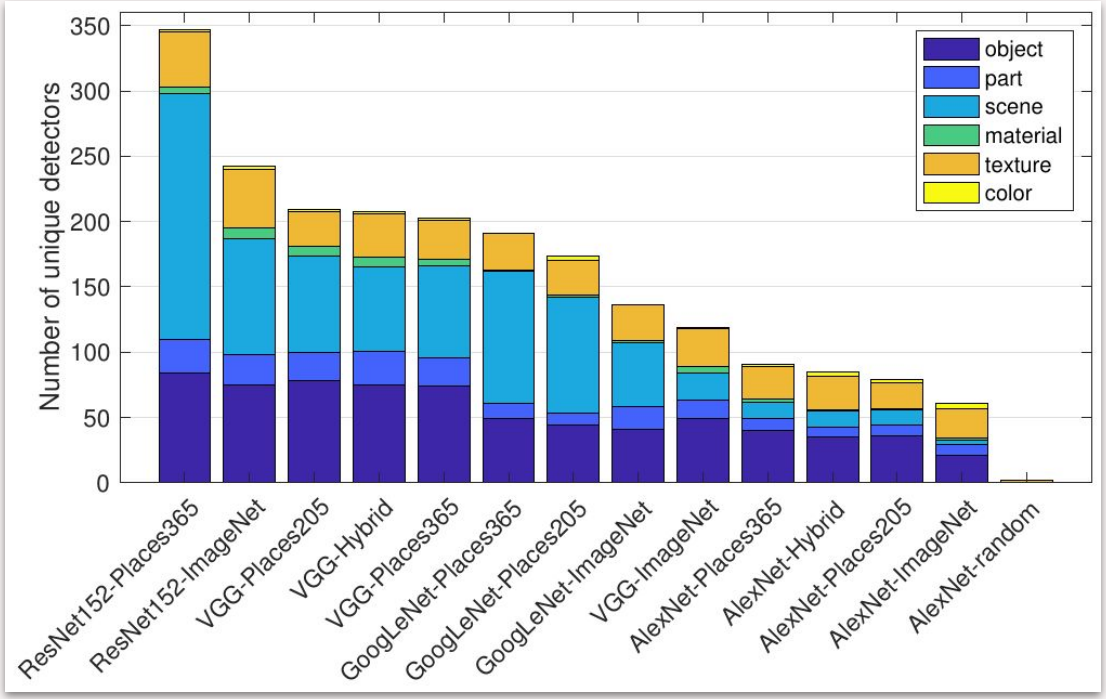


$$IoU_{k,c} = \frac{\sum |M_k(\mathbf{x}) \cap L_c(\mathbf{x})|}{\sum |M_k(\mathbf{x}) \cup L_c(\mathbf{x})|}$$

(Bau et al. 2017)

Linking neuron activations with semantic concepts.

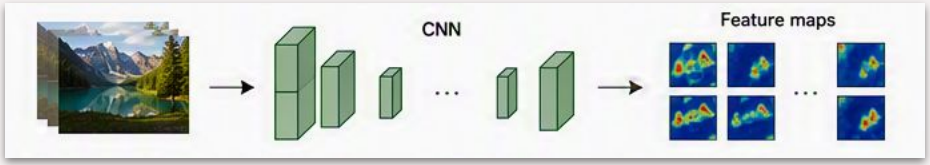
Idea: Quantify model interpretability with concept histograms.

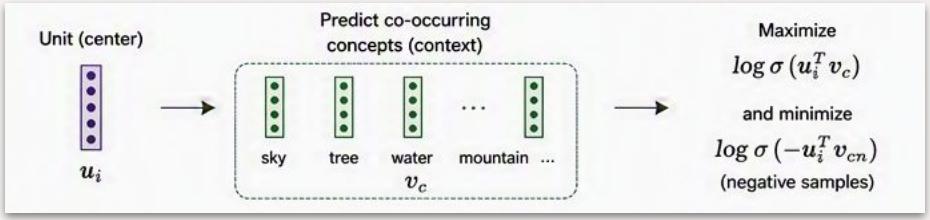


(Bau et al. 2017)

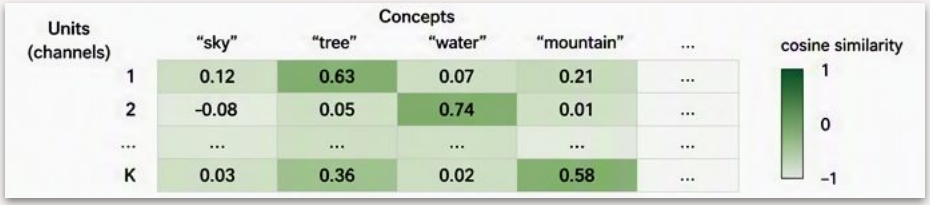
Linking sets of neuron activations with concepts.

Assumption: semantic representations are distributed





Concept-embedding Analysis



Units must be studied in conjunction

- Collect unit activations for inputs from a specific probe

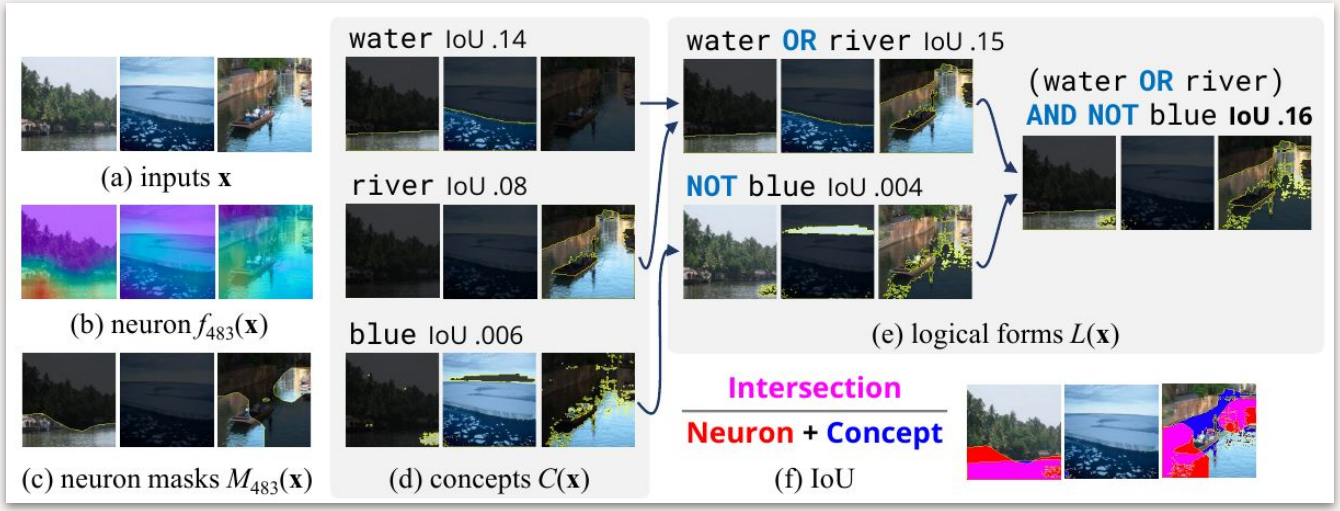
- Learn how to weight the collected activations to perform semantic tasks

Learned weights -> Embedding

(Fong & Vedaldi. 2018)

Linking neuron activations with concept compositions.

Observation: some abstract concepts can't be covered by a single neuron



Proposal

- Expand the search space to include compositional logical forms.

- For a neuron, find a logical formula that best match its activation pattern.

How: $$\delta(n, C) \triangleq \text{IoU}(n, C) = \left[\sum_{\mathbf{x}} \mathbb{1}(M_n(\mathbf{x}) \wedge C(\mathbf{x}))\right] / \left[\sum_{\mathbf{x}} \mathbb{1}(M_n(\mathbf{x}) \vee C(\mathbf{x}))\right]$$

FInding: concept compositions were better approximators of neuron behavior

(Mu & Andreas. 2020, Oikarine & Weng 2024)

This chapter...

Post-hoc semantic mapping of neuron activations

What does a neuron do?

Inherently interpretable methods

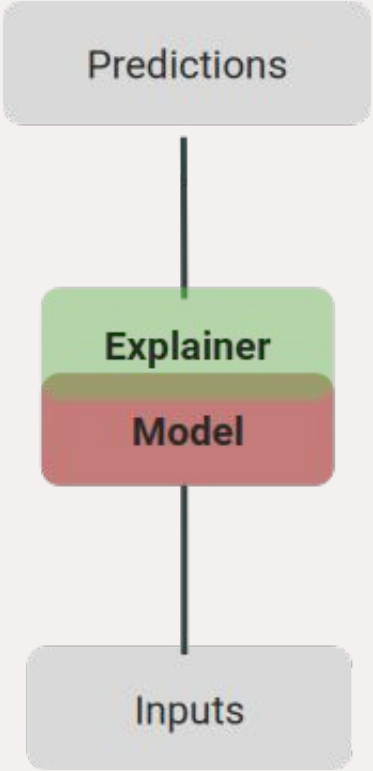
Can we make neurons understandable?

Evaluation protocols

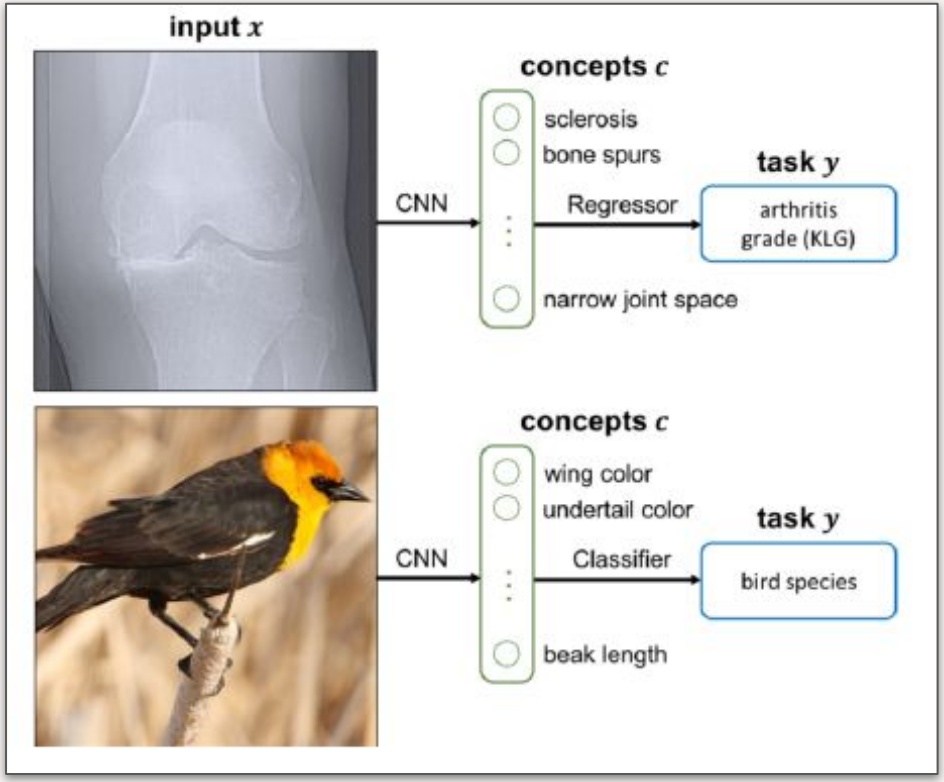
When do we understand neurons?

Inherently interpretable methods

How can we make neurons understandable?



Injecting interpretable units via concept bottlenecks.

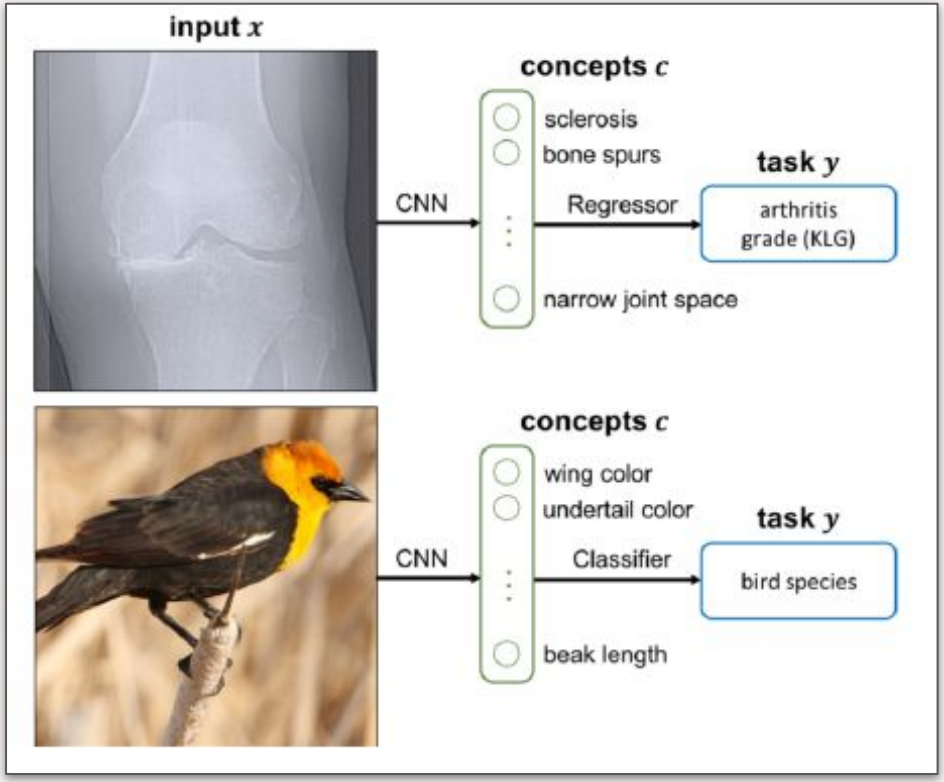


- Enforce the learning of pre-defined intermediate semantic concepts

- Make a prediction out of these concepts

(Koh et al. 2020)

Injecting interpretable units via concept bottlenecks.



- Enforce the learning of pre-defined intermediate semantic concepts

- Make a prediction out of these concepts

How:

$\{(x^{(i)}, y^{(i)}, c^{(i)})\}_{i=1}^{n}$ Data

$f(g(x))$ Target model

where

$\hat{c} = g(x)$ Concept prediction

$\hat{y} = f(g(x))$ Downstream task

(Koh et al. 2020)

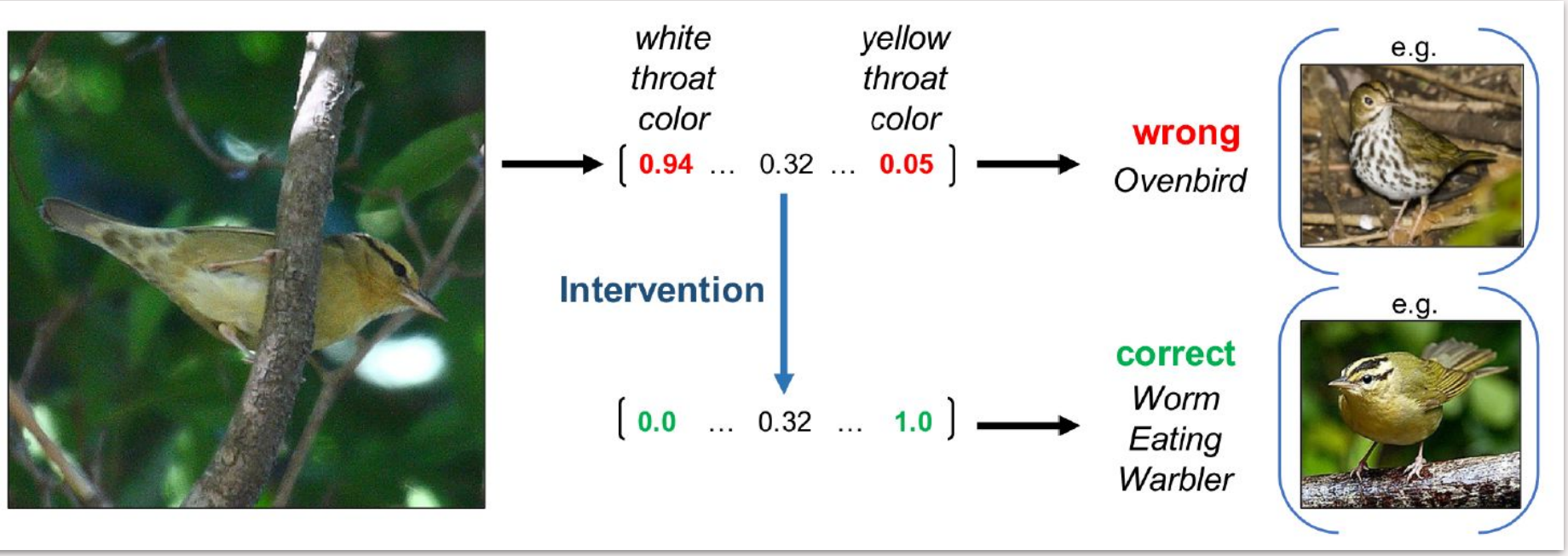
Injecting interpretable units via concept bottlenecks.

Advantage: enables manual intervention.

(Koh et al. 2020)

Injecting interpretable units via concept bottlenecks.

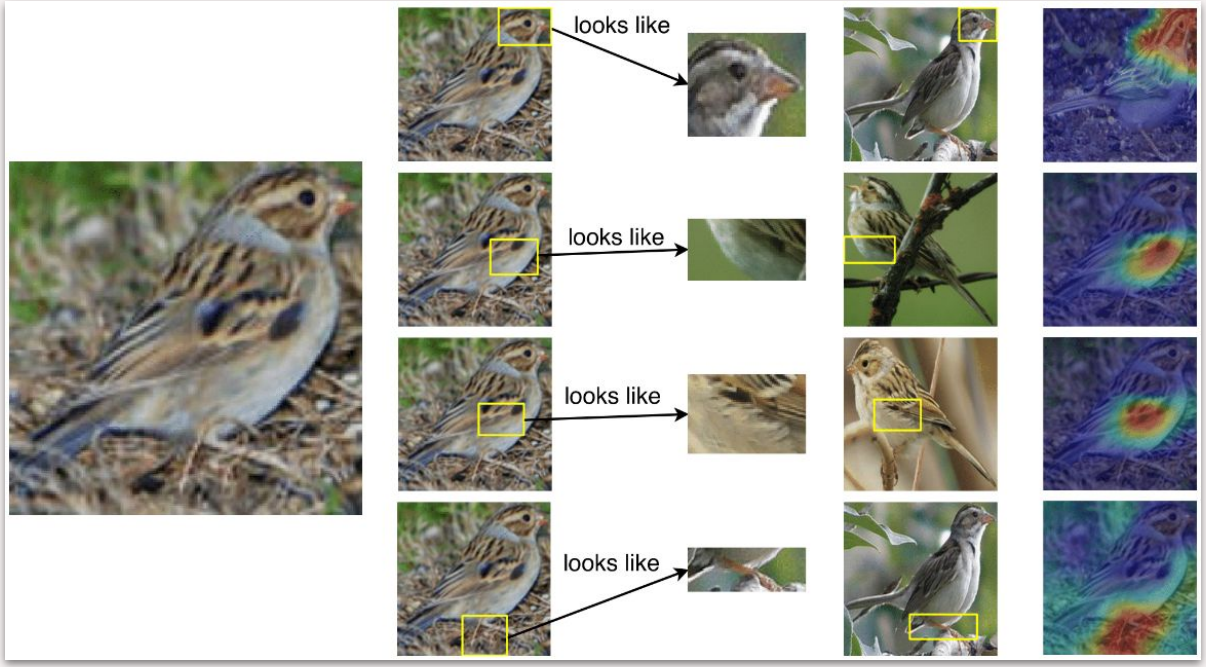
Advantage: enables manual intervention.



(Koh et al. 2020)

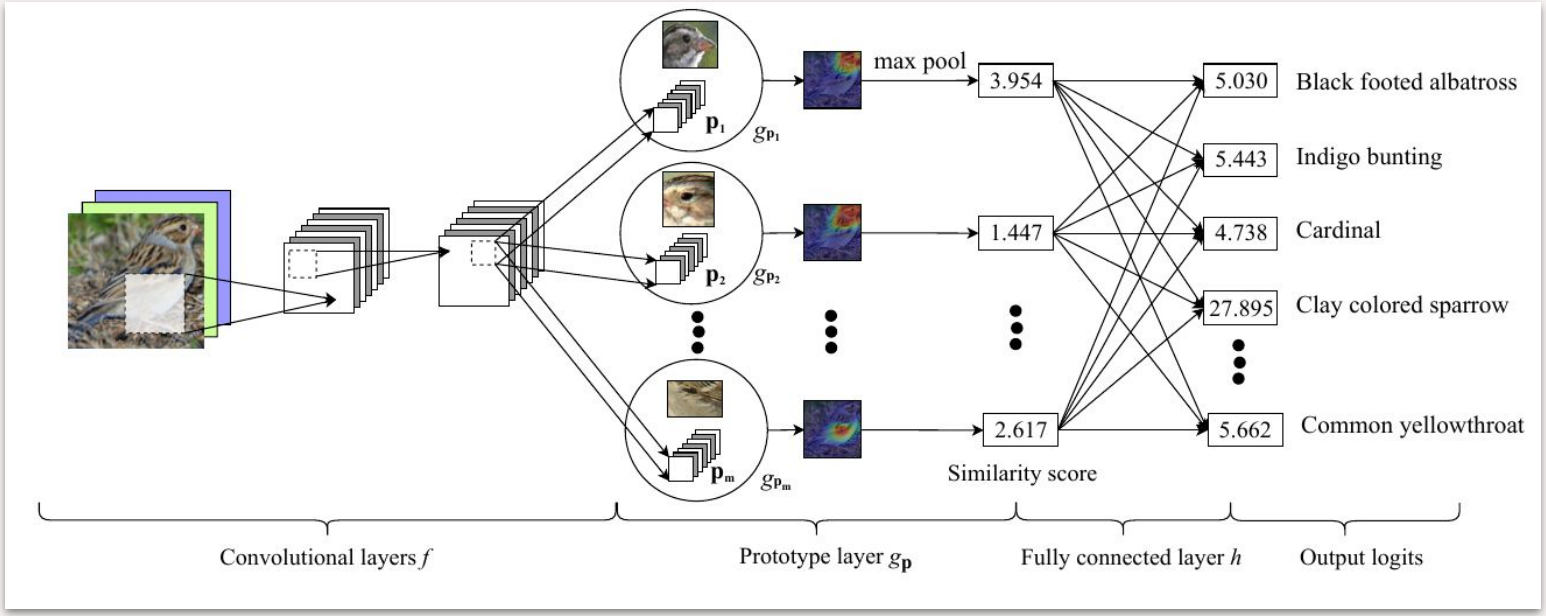
Case-based reasoning with prototype networks.

Idea: find prototypical parts and combine evidence into output.



(Chen et al. 2019)

Case-based reasoning with prototype networks.

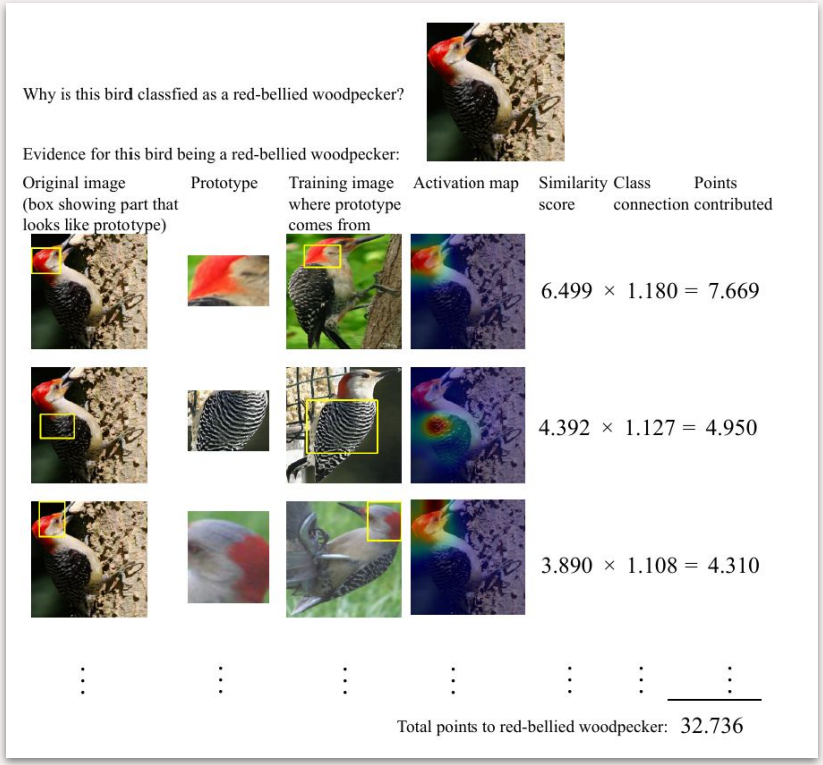


- Learn embeddings that link the representation with the task of interest (i.e. the predictions).
- Associate the embedding with intermediate concepts of interest (e.g. object parts).
- Produced explanations reflect the decision-making process

(Chen et al. 2019)

Case-based reasoning with prototype networks.

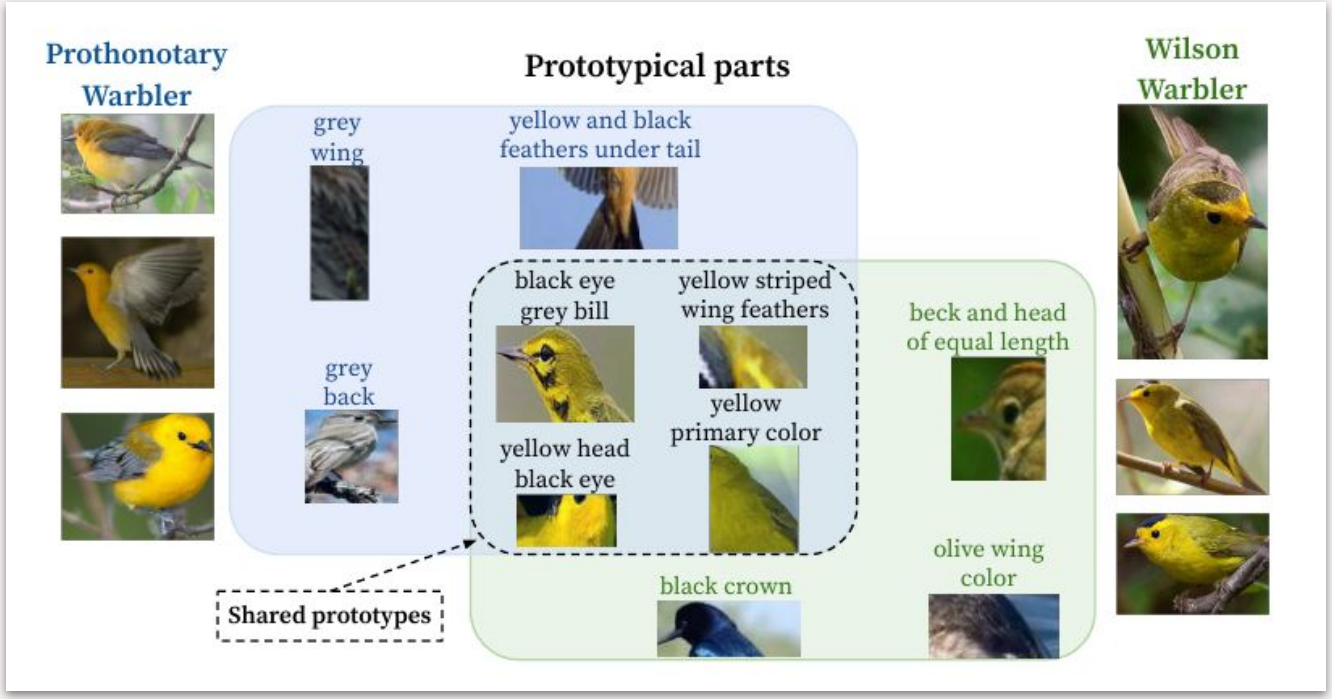
Goal: produced explanations reflect the decision-making process.



(Chen et al. 2019)

Case-based reasoning with prototype networks.

Goal: Addressing scalability by sharing prototypes across classes.



(Rymarczyk et al., 2020, 2022)

This chapter...

Post-hoc semantic mapping of neuron activations

What does a neuron do?

Inherently interpretable methods

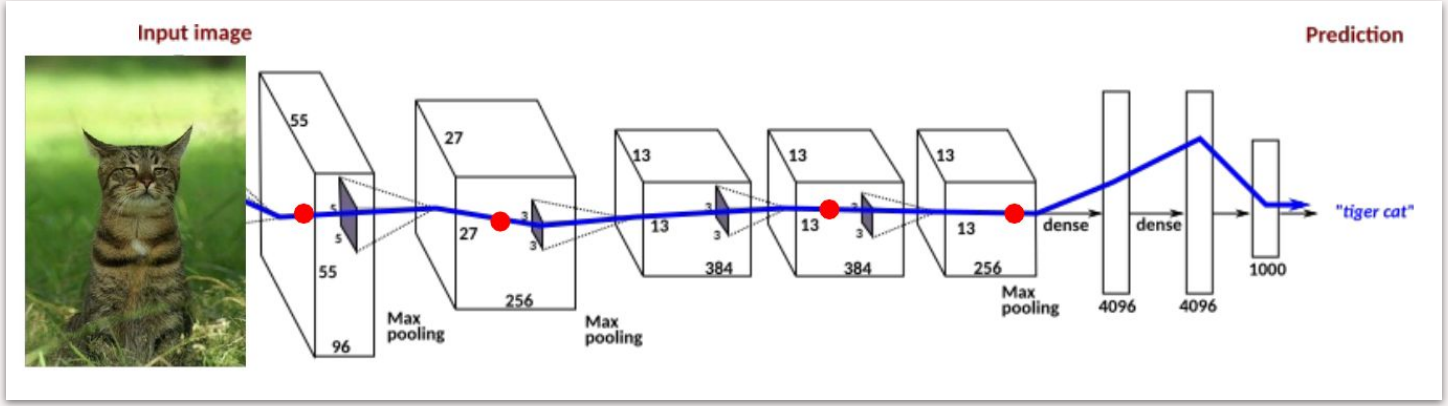
Can we make neurons understandable?

Evaluation protocols

When do we understand neurons?

Ablating the effect of relevant units/neurons in the model.

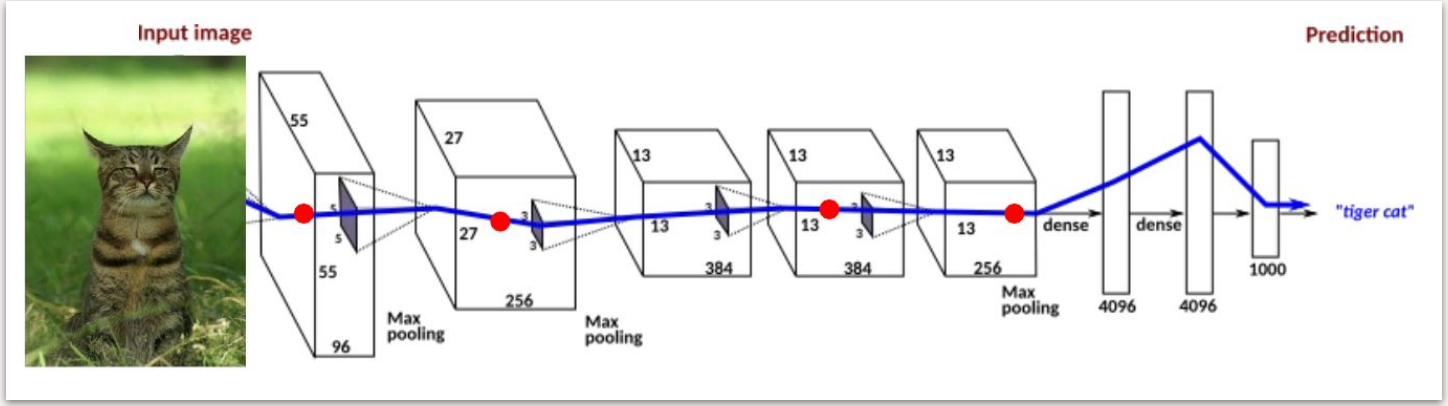
Goal: Assessing the fidelity of the units on the decision-making process.



- **Define a mask**
- **Define a baseline**

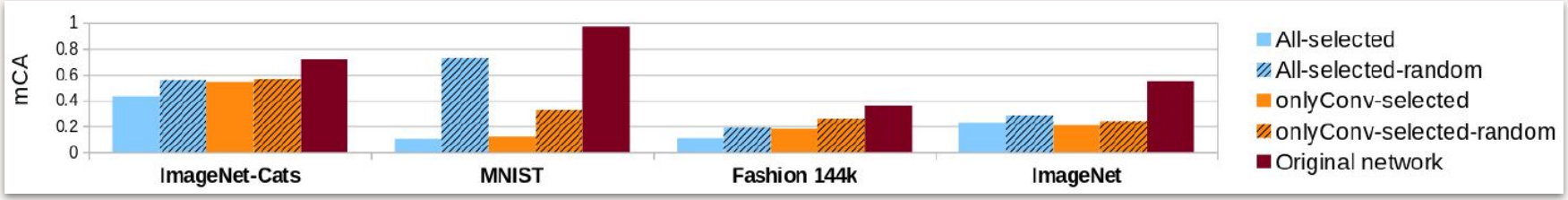
Ablating the effect of relevant units/neurons in the model.

Goal: Assessing the fidelity of the units on the decision-making process.



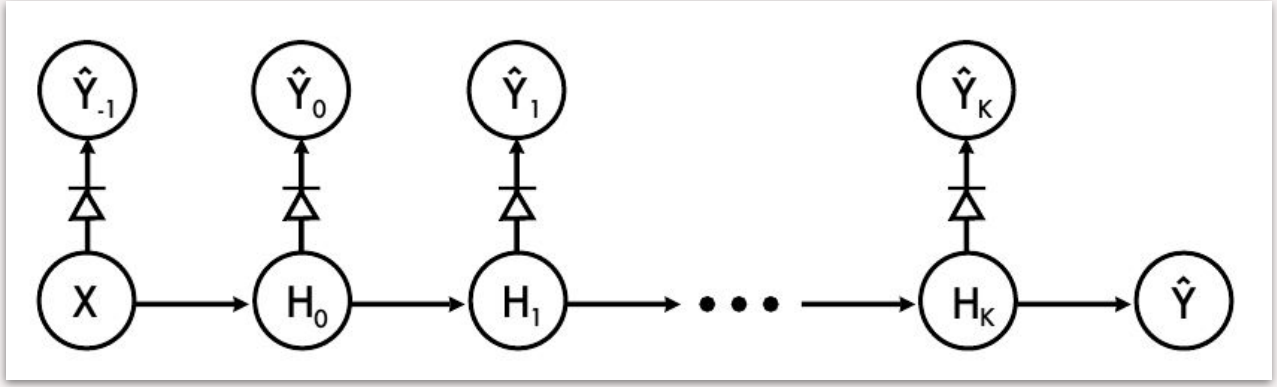
- **Define a mask**
- **Define a baseline**

... and measure the effect.



Analyzing the predictive signal of neuron activations

Idea: Use simple classifier (probes) to inspect the representation

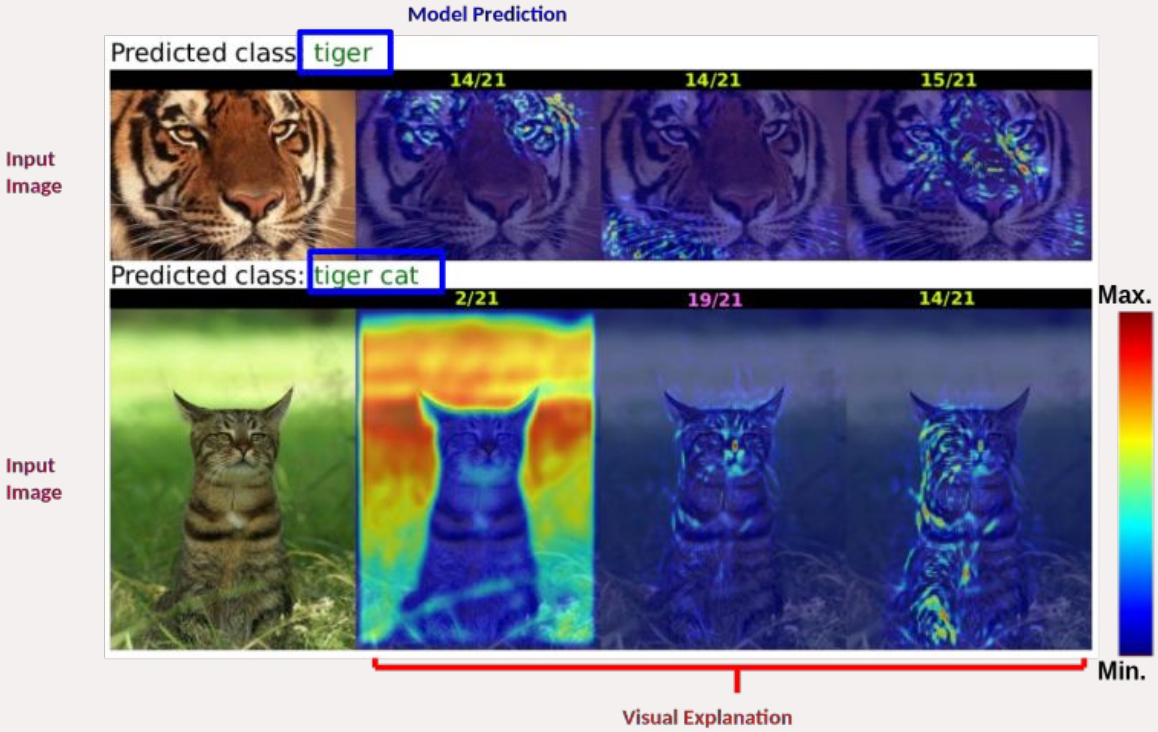


- Targeted probing procedure
- Sparser analysis

(Alain & Bengio, 2017)

(Proxy) Evaluation of derived explanations.

Idea: Evaluate the attribution maps produced from the relevant neurons.

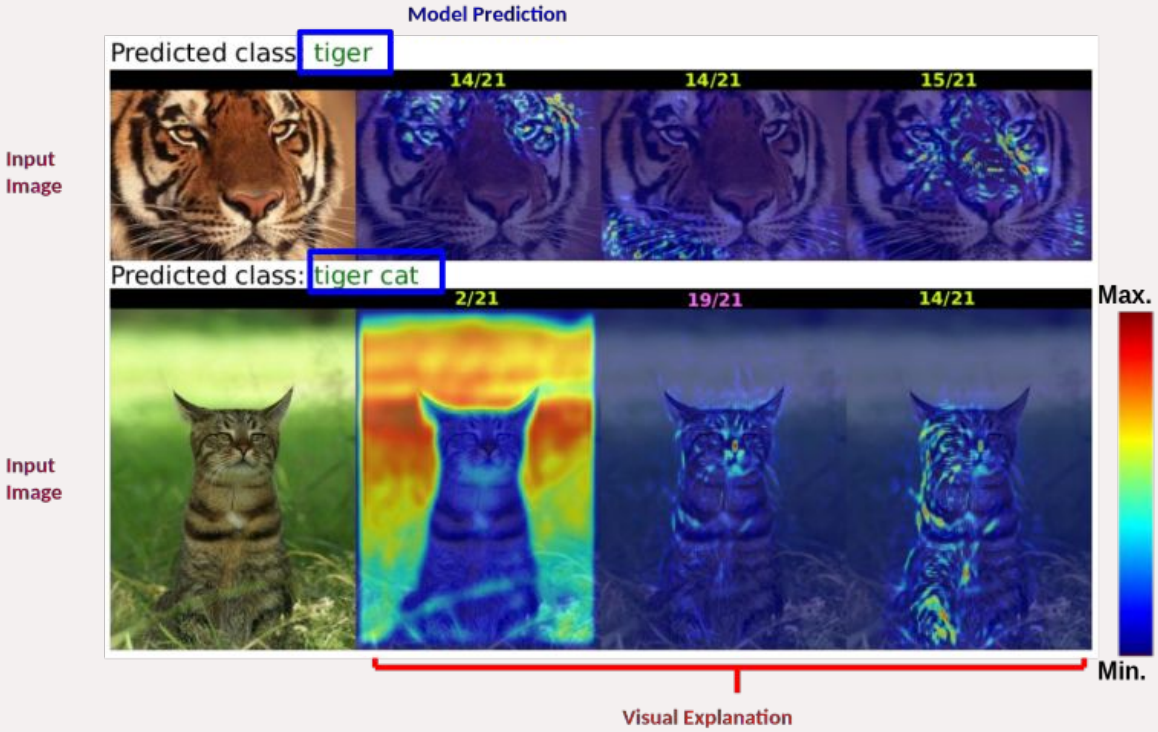


How?

- Produce an attribution map from the interpreted neurons.
- Measure performance following an attribution evaluation protocol.

(Proxy) Evaluation of derived explanations.

Idea: Evaluate the attribution maps produced from the relevant neurons.



How?

- Produce an attribution map from the interpreted neurons.
- Measure performance following an attribution evaluation protocol.

but...

- Inherit the weaknesses of those protocols

Coming up...

Early Interpretability Methods

José Oramas

From Classical to Mechanistic Interpretability

Ward Gauderis, José Oramas & Thomas Dooms

From Mechanistic to Compositional Interpretability

Ward Gauderis & Thomas Dooms