Explainability in Supervised Machine Learning

Forest Agostinelli
University of South Carolina

Outline

- Motivation
- Global explanations with interpretable models
- Local explanations with uninterpretable models
  - Model agnostic approaches with surrogate models
  - Model agnostic approaches without surrogate models
  - Model dependent approaches without surrogate models
- Semi-global explanations with uninterpretable models

Motivation

- We need to be able to understand how a model works and why it gives the predictions it does for safety, transparency, and knowledge discovery
- Explainability is not an exact term; however, anything that helps us with the aforementioned goal can be considered a part of explainability

Safety

- Self-driving cars use supervised learning for object detection (e.g. cars, pedestrians, stop signs, etc.)
- Some medical practitioners use supervised learning for object detection (e.g. polyps), prognosis, and image segmentation
- There is research on using supervised learning for elderly care, such as for fall detection

Transparency

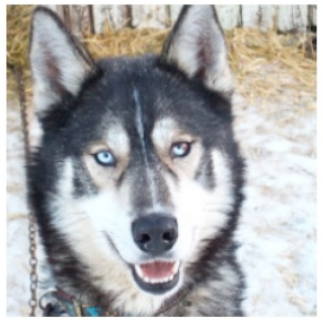
- Supervised learning has been proposed for predicting recidivism
- Some banks may use supervised learning for loans
- Some companies may use supervised learning to filtering resumes

Knowledge Discovery

- There may be supervised learning tasks that humans cannot accurately label
- Labels may be obtained from computationally expensive simulations
  - Predicting binding affinity can be done using simulations at the atomic level
- Labels may also be obtained from reinforcement learning
  - How many steps to solve the Rubik's cube?
  - How costly would it be to synthesize a given chemical?
- After training a model on such tasks, we can use explainability methods to understand how these predictions are made to obtain new knowledge

Practical Concerns

- Data leakage: creation of unexpected additional information in the training data, allowing the algorithm to make unrealistically good predictions.
  - All examples of wolves are in snow. The algorithm then detects snow instead of the actual wolf.
  - Detection of prostate cancer. The algorithm used a variable that represented whether or not a patient had prostate surgery.



(a) Husky classified as wolf

(b) Explanation

- Data shift: The data in the test set is different than the data in the training set.

Outline

- Motivation
- Global explanations with interpretable models
- Local explanations with uninterpretable models
  - Model agnostic approaches with surrogate models
  - Model agnostic approaches without surrogate models
  - Model dependent approaches without surrogate models
- Semi-global explanations with uninterpretable models

Global Explanations

- Explain the entire model
- Given a model, f, we may ask
  - What are the inputs, $\boldsymbol{x}$, that will give an output of y?
  - For what inputs is the model most reliable?

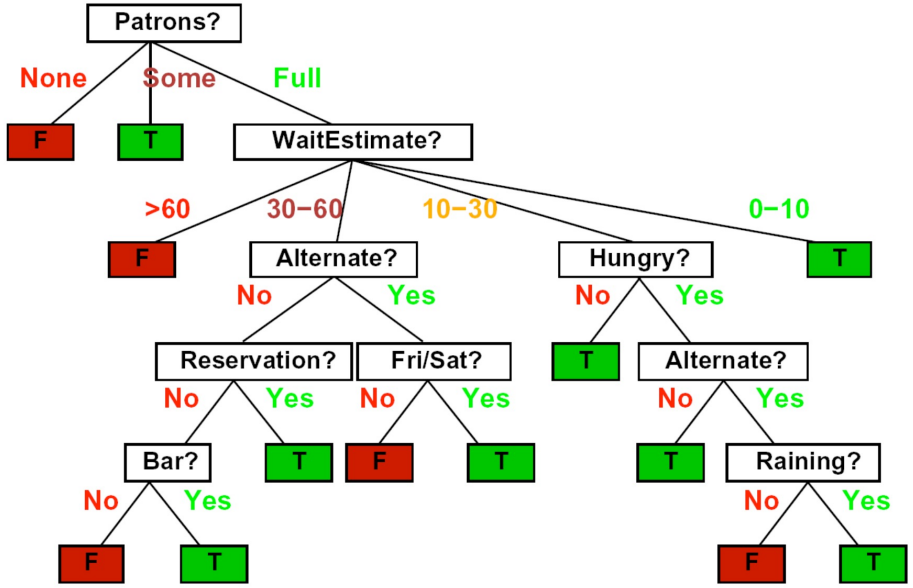
Decision Trees

- Maps a vector of attributes to a single output value by performing a sequence of tests

Data

Example	Attributes										Target
	Alt	*Bar*	*Fri*	*Hun*	*Pat*	*Price*	*Rain*	*Res*	*Type*	*Est*	*WillWait*
X_1	*T*	*F*	*F*	*T*	*Some*	*$$$*	*F*	*T*	*French*	*0–10*	*T*
X_2	*T*	*F*	*F*	*T*	*Full*	*$*	*F*	*F*	*Thai*	*30–60*	*F*
X_3	*F*	*T*	*F*	*F*	*Some*	*$*	*F*	*F*	*Burger*	*0–10*	*T*
X_4	*T*	*F*	*T*	*T*	*Full*	*$*	*F*	*F*	*Thai*	*10–30*	*T*
X_5	*T*	*F*	*T*	*F*	*Full*	*$$$*	*F*	*T*	*French*	*>60*	*F*
X_6	*F*	*T*	*F*	*T*	*Some*	*$$*	*T*	*T*	*Italian*	*0–10*	*T*
X_7	*F*	*T*	*F*	*F*	*None*	*$*	*T*	*F*	*Burger*	*0–10*	*F*
X_8	*F*	*F*	*F*	*T*	*Some*	*$$*	*T*	*T*	*Thai*	*0–10*	*T*
X_9	*F*	*T*	*T*	*F*	*Full*	*$*	*T*	*F*	*Burger*	*>60*	*F*
X_{10}	*T*	*T*	*T*	*T*	*Full*	*$$$*	*F*	*T*	*Italian*	*10–30*	*F*
X_{11}	*F*	*F*	*F*	*F*	*None*	*$*	*F*	*F*	*Thai*	*0–10*	*F*
X_{12}	*T*	*T*	*T*	*T*	*Full*	*$*	*F*	*F*	*Burger*	*30–60*	*T*

Decision Tree

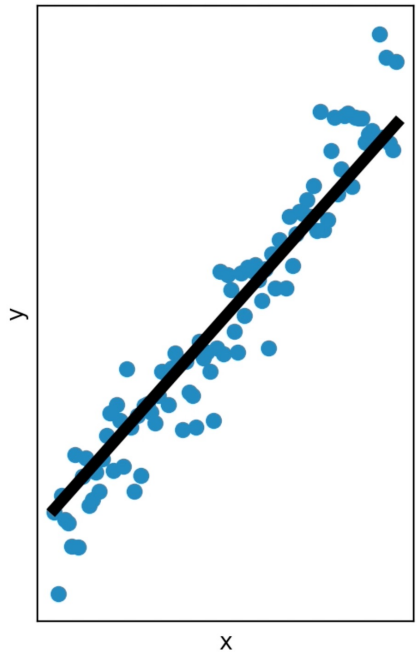


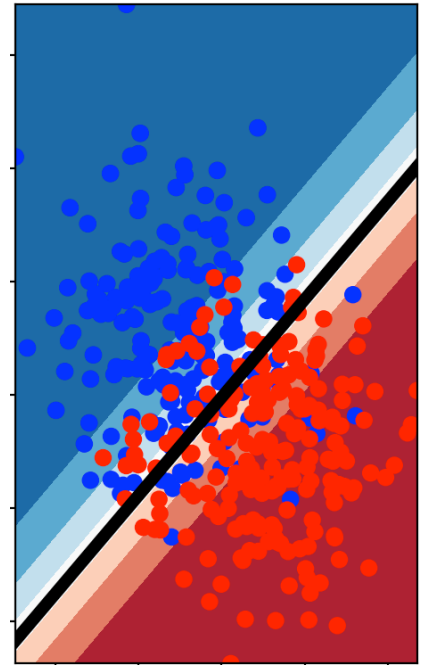
Interpretable Models

- There are models that are considered “interpretable” which can make explainability very straightforward
- A model is considered interpretable if the relationship between its inputs and outputs can easily be obtained
  - Decision trees
  - Linear models
- It is commonly considered that there is a tradeoff between interpretability and performance
- There is debate on what interpretability really means

Linear Models

- Each output, y^o, has a linear relationship with the input
  - $y^o = \sum_i w_i^o x_i + b$
- Therefore, we can easily see how the output changes as a function of any given input

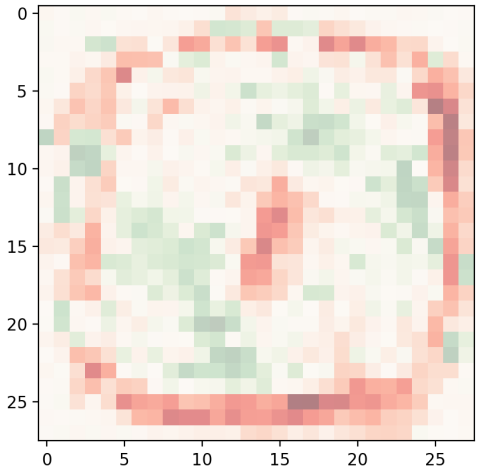




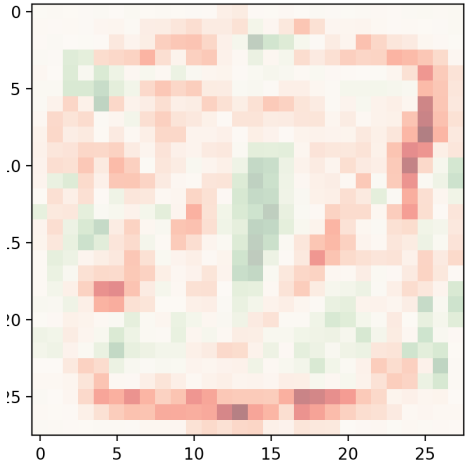
Linear Models

- For MNIST, for a given output, we can see how a pixel increases or decreases the given output
  - Green increases
  - Red: decreases
  - Darker colors mean larger weights

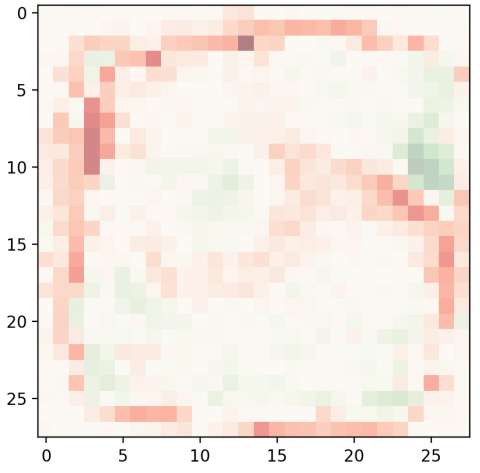
Zero



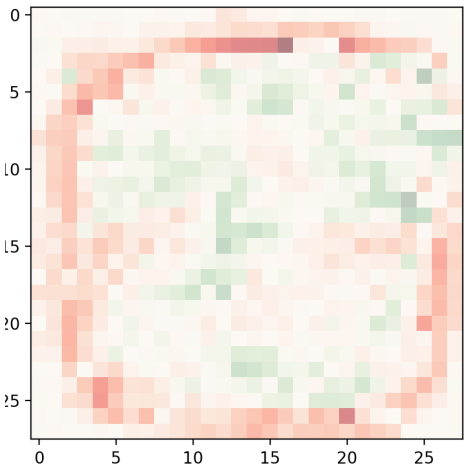
One



Five



Eight



Linear Models

- How else might we obtain explainability?
- Are there any limitations?
  - I.e. what explanations might you want to obtain that are difficult to obtain with linear models?

Outline

- Motivation
- Global explanations with interpretable models
- Local explanations with uninterpretable models
  - Model agnostic approaches with surrogate models
  - Model agnostic approaches without surrogate models
  - Model dependent approaches without surrogate models
- Semi-global explanations with uninterpretable models

Local Explanations

- Given model predictions, $y = f(\boldsymbol{x})$, we may ask
  - What features of $\boldsymbol{x}$ are the most important for giving the label of y?
  - How would $\boldsymbol{x}$ have to be modified to change the label y to a different label?
  - How do features of $\boldsymbol{x}$ depend on one another?
  - How confident is the model about the given label, y?
  - What examples are similar to $\boldsymbol{x}$, but have a different label? Why?

Outline

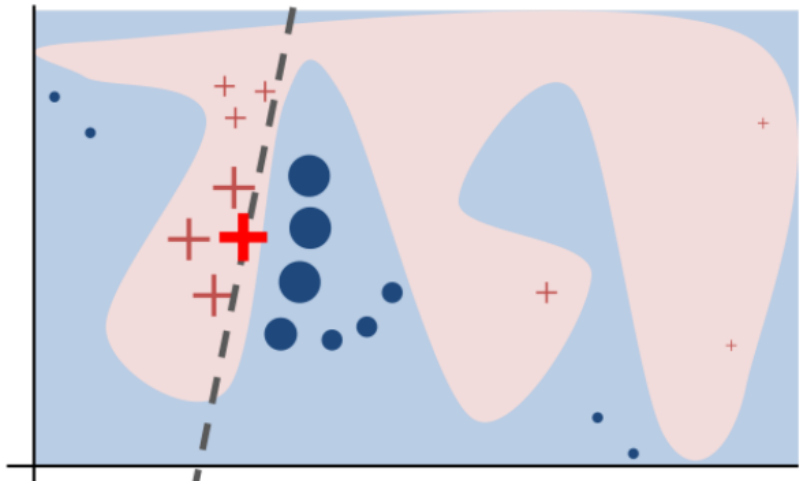
- Motivation
- Global explanations with interpretable models
- Local explanations with uninterpretable models
  - Model agnostic approaches with surrogate models
  - Model agnostic approaches without surrogate models
  - Model dependent approaches without surrogate models
- Semi-global explanations with uninterpretable models

Surrogate Models

- A surrogate model is any model that attempts to replicate the predictions of another, usually more expensive, model
- In the context of explainability, we can train an interpretable surrogate model to match the predictions of an uninterpretable model
- After the surrogate model is trained, we can then leverage the explainability approaches we already have with interpretable models
- Since uninterpretable models usually have more expressive power than interpretable ones, often times, this is done “locally” (for only a single example)

LIME

- A method that can give Locally-interpretable, model-agnostic, explanations (LIME).
- Locally-interpretable: A human should be able to interpret the results. The results are faithful in the local vicinity of the example.
- Model agnostic: This method should be able to be used on any model.



Ribeiro, Marco Tulio, Sameer Singh, and Carlos Guestrin. "" Why should I trust you?" Explaining the predictions of any classifier." Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining. 2016.

Obtaining an Interpretable Model from any Model

- We have a prediction model f, and we want an explanation for the output of f give input x.
- x may not always be interpretable, so x’ is created to be an interpretable version of x.
  - For example, individual pixels vs super-pixels
- An interpretable model g is created. g could be, for example, a linear model $g(x') = w^T x'$.
- Perturbed samples of x and x’, z and z’, are created.
  - For example, removing super-pixels in an image.
- f(z) should be equal to g(z’).

Obtaining an Interpretable Model from any Model

How close does the explaining model, g, match the model being explained, f, for example x.

$$\xi(x) = \underset{g \in G}{\operatorname{argmin}} \; \mathcal{L}(f, g, \pi_x) + \Omega(g)$$

Model complexity penalty.

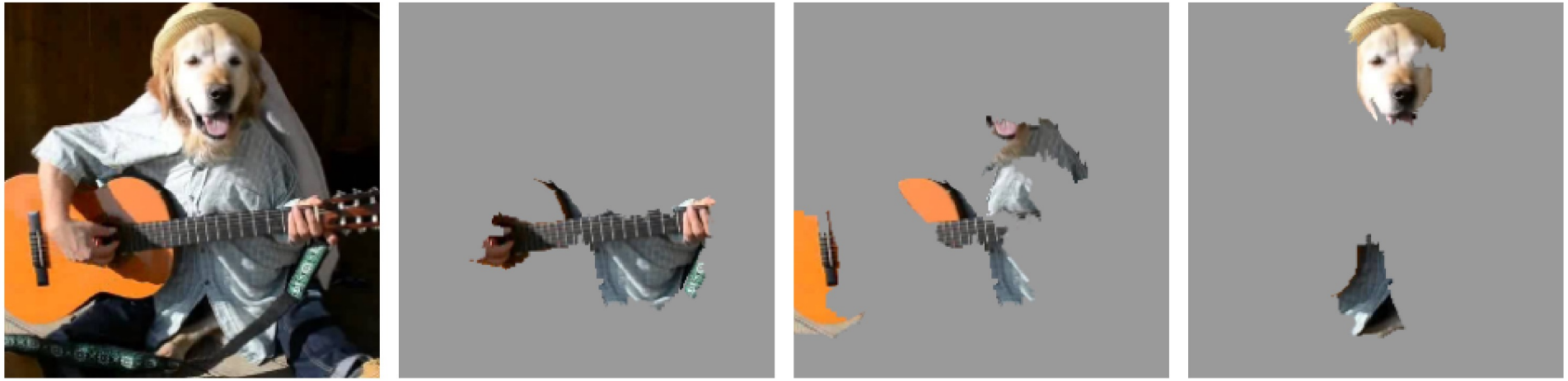
Model being explained

$$\mathcal{L}(f, g, \pi_x) = \sum_{z, z' \in \mathcal{Z}} \pi_x(z) \left(f(z) - g(z')\right)^2$$

Weights examples based on similarity to original input x.

Explaining model

LIME Explanations

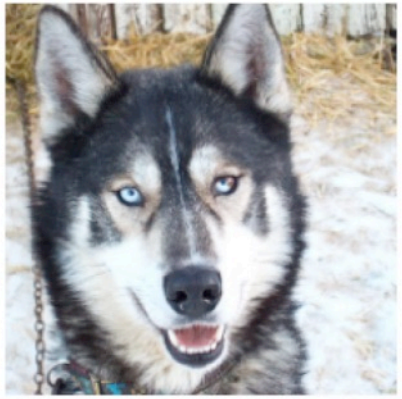


(a) Original Image (b) Explaining *Electric guitar* (c) Explaining *Acoustic guitar* (d) Explaining *Labrador*

Figure 4: Explaining an image classification prediction made by Google's Inception neural network. The top 3 classes predicted are "Electric Guitar" ($p = 0.32$), "Acoustic guitar" ($p = 0.24$) and "Labrador" ($p = 0.21$)

Purposefully Bad Classifier

- A classifier only trained on wolves in snow and huskies not in snow, leading it to rely heavily on the presence of snow when detecting wolves.
- Users were much less likely to trust this model after seeing the explanations for the predictions.



(a) Husky classified as wolf

(b) Explanation

	Before	After
Trusted the bad model	10 out of 27	3 out of 27
Snow as a potential feature	12 out of 27	25 out of 27

Local Explanations

- How can we obtain local explanations of uninterpretable models without surrogate models?
- Use MNIST as an example domain

Outline

- Motivation
- Global explanations with interpretable models
- Local explanations with uninterpretable models
  - Model agnostic approaches with surrogate models
  - Model agnostic approaches without surrogate models
  - Model dependent approaches without surrogate models
- Semi-global explanations with uninterpretable models

Feature Importance

- What features must be present in order for the model to give its prediction?
  - Remove features from the input until the prediction changes
  - Add features to the input until the prediction changes
  - Perturb the input until the prediction changes

Outline

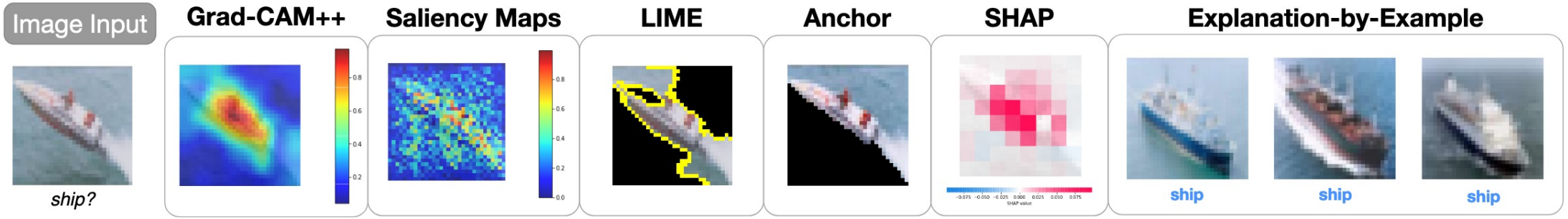
- Motivation
- Global explanations with interpretable models
- Local explanations with uninterpretable models
  - Model agnostic approaches with surrogate models
  - Model agnostic approaches without surrogate models
  - Model dependent approaches without surrogate models
- Semi-global explanations with uninterpretable models

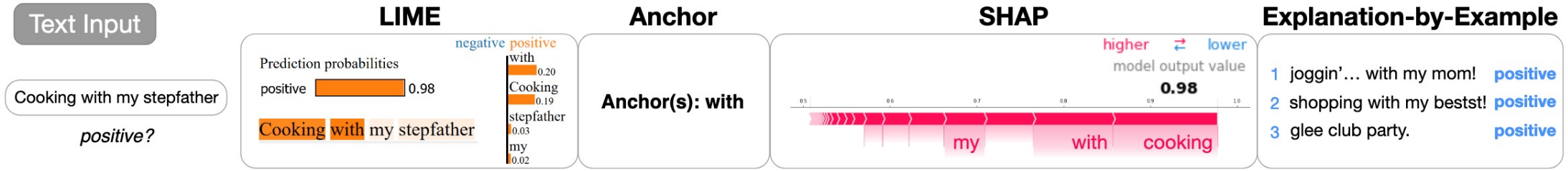
Explanation by Example

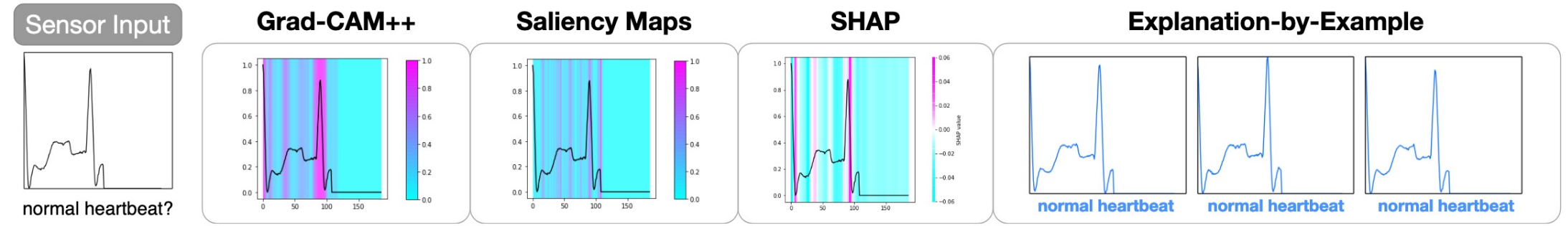
- Instead of examining features for a given example, one can find similar examples to the given example
- To compute similarity one use the distance between two examples in the input space
  - However, examples may be quantitatively distant in the input space but qualitatively similar
- Comparing examples based on the hidden activations often gives better results

Jeyakumar, Jeya Vikranth, et al. "How can i explain this to you? an empirical study of deep neural network explanation methods." Advances in neural information processing systems 33 (2020): 4211-4222.

Explanation Visualizations

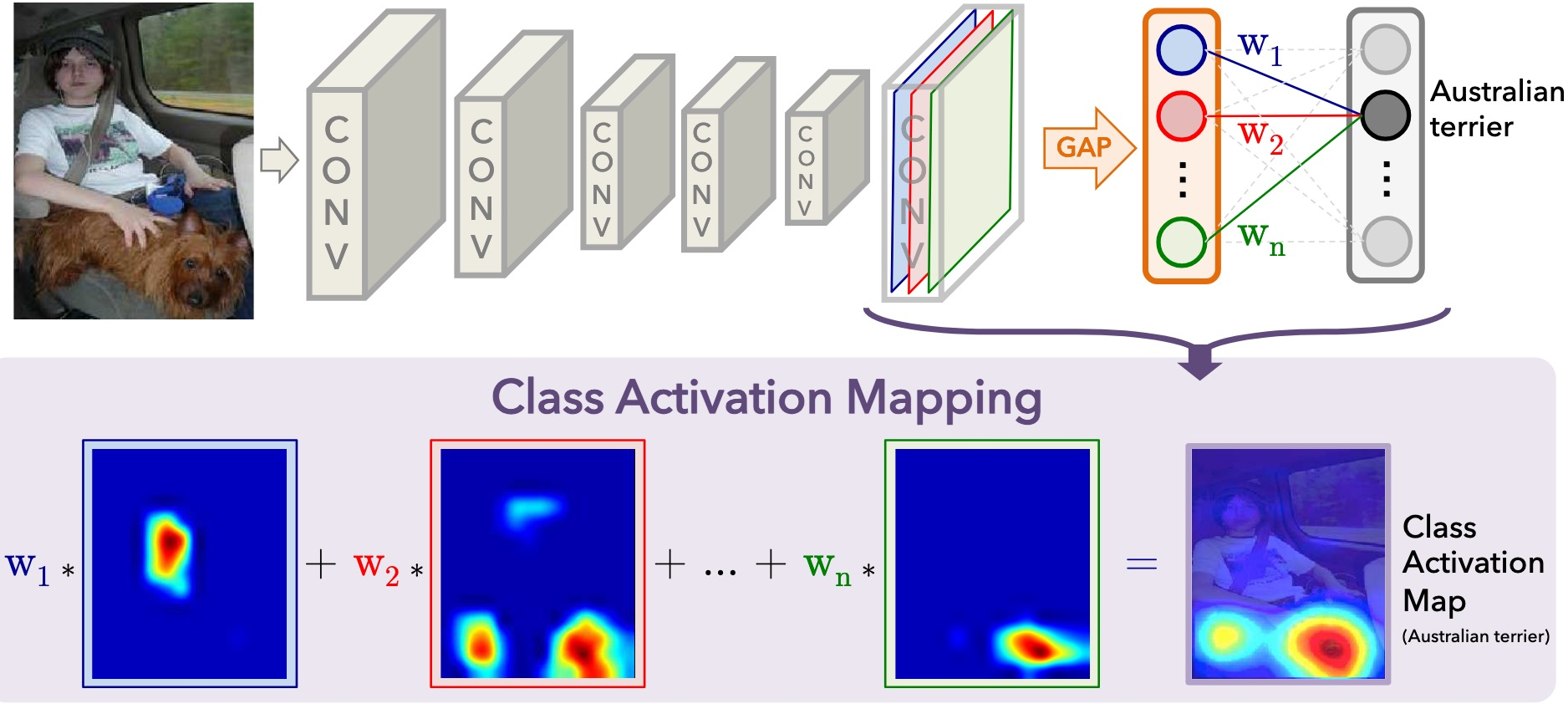






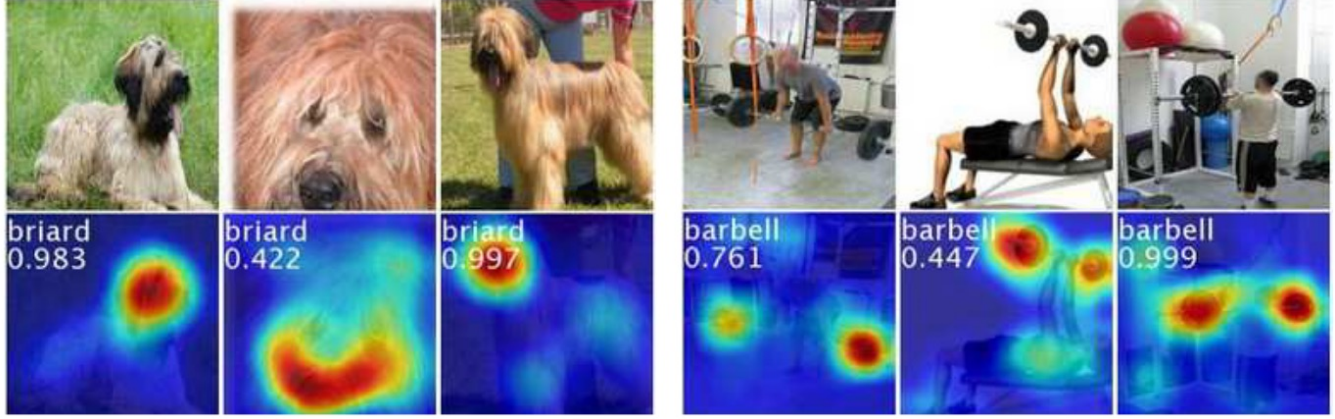
Class Activation Mapping (CAM)

- Find out what regions of the region are most relevant to the given class using the known architecture



Zhou, Bolei, et al. "Learning deep features for discriminative localization." Proceedings of the IEEE conference on computer vision and pattern recognition. 2016.

CAM



Mushroom

Penguin

Teapot

Cleaning the floor

Cooking

Fixing a car

Dining room

Frequent object:
wall:0.99
chair:0.98
floor:0.98
table:0.98
ceiling:0.75
window:73

Informative object:
table:0.96
chair:0.85
chandelier:0.80
plate:0.73
vase:0.69
flowers:0.63

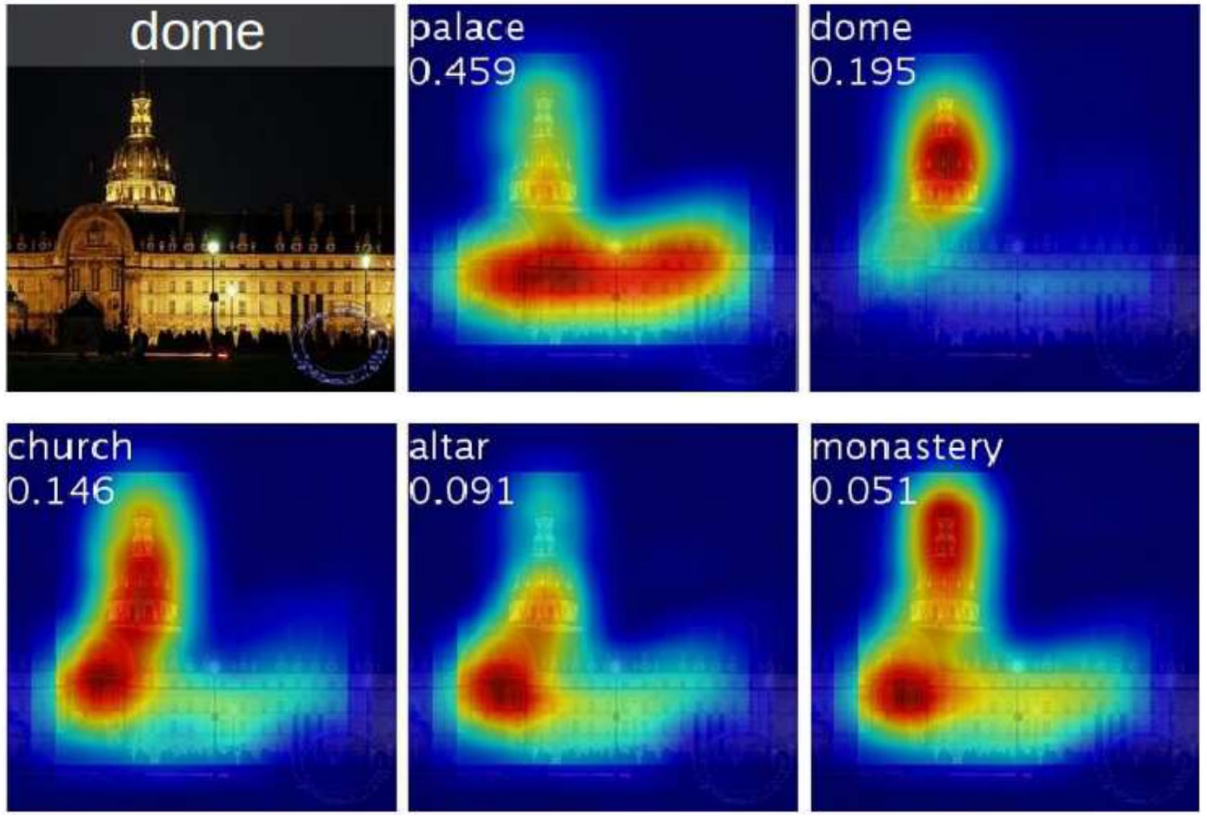
Bathroom

Frequent object:
wall: 1
floor:0.85
sink: 0.77
faucet:0.74
mirror:0.62
bathtub:0.56

Informative object:
sink:0.84
faucet:0.80
countertop:0.80
toilet:0.72
bathtub:0.70
towel:0.54

CAM: Understanding Misclassifications

- With CAM, we can see what regions led to other classes having high probabilities



Outline

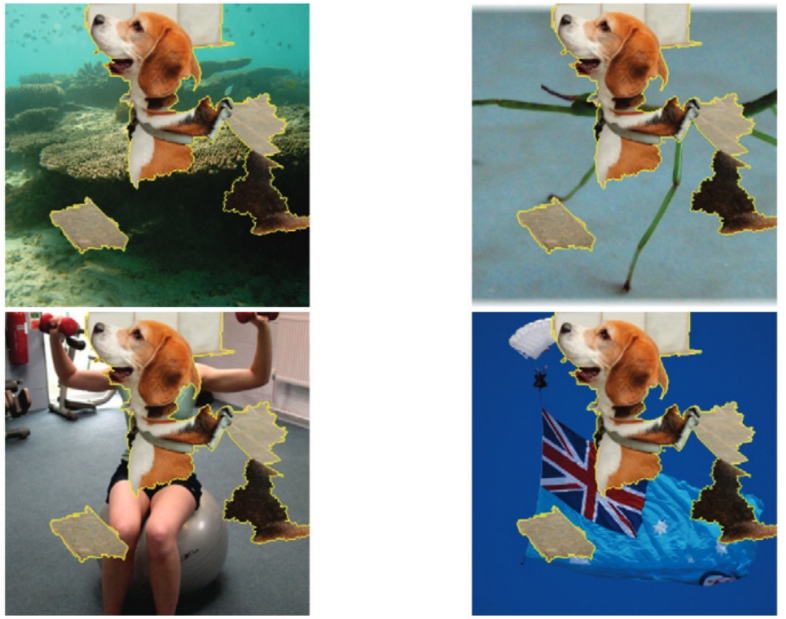
- Motivation
- Global explanations with interpretable models
- Local explanations with uninterpretable models
  - Model agnostic approaches with surrogate models
  - Model agnostic approaches without surrogate models
  - Model dependent approaches without surrogate models
- Semi-global explanations with uninterpretable models

Anchors

- Attempt to find features in an input that will anchor the output to a given prediction
  - Other features may change, but this does not change the ouput

(a) Original image

(b) Anchor for “beagle”



(c) Images where Inception predicts $P(\text{beagle}) > 90\%$

Ribeiro, Marco Tulio, Sameer Singh, and Carlos Guestrin. "Anchors: High-precision model-agnostic explanations." Proceedings of the AAAI conference on artificial intelligence. Vol. 32. No. 1. 2018.

Anchors: Catching Bad Predictions

(a) Instances

(b) LIME explanations

{"not", "bad"} → Positive {"not", "good"} → Negative

(c) Anchor explanations

Deconvolution and Unpooling

- Determine which neurons had the highest activation for a given example
- Project the activation back to the input space

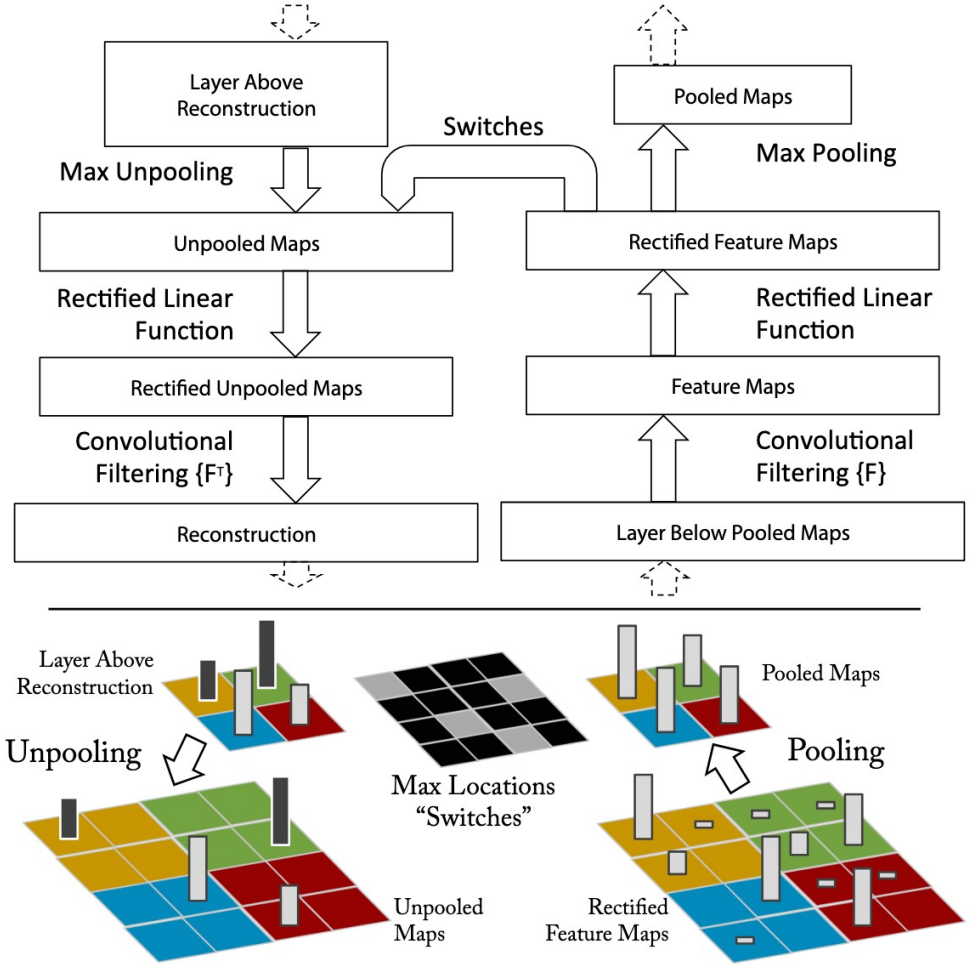
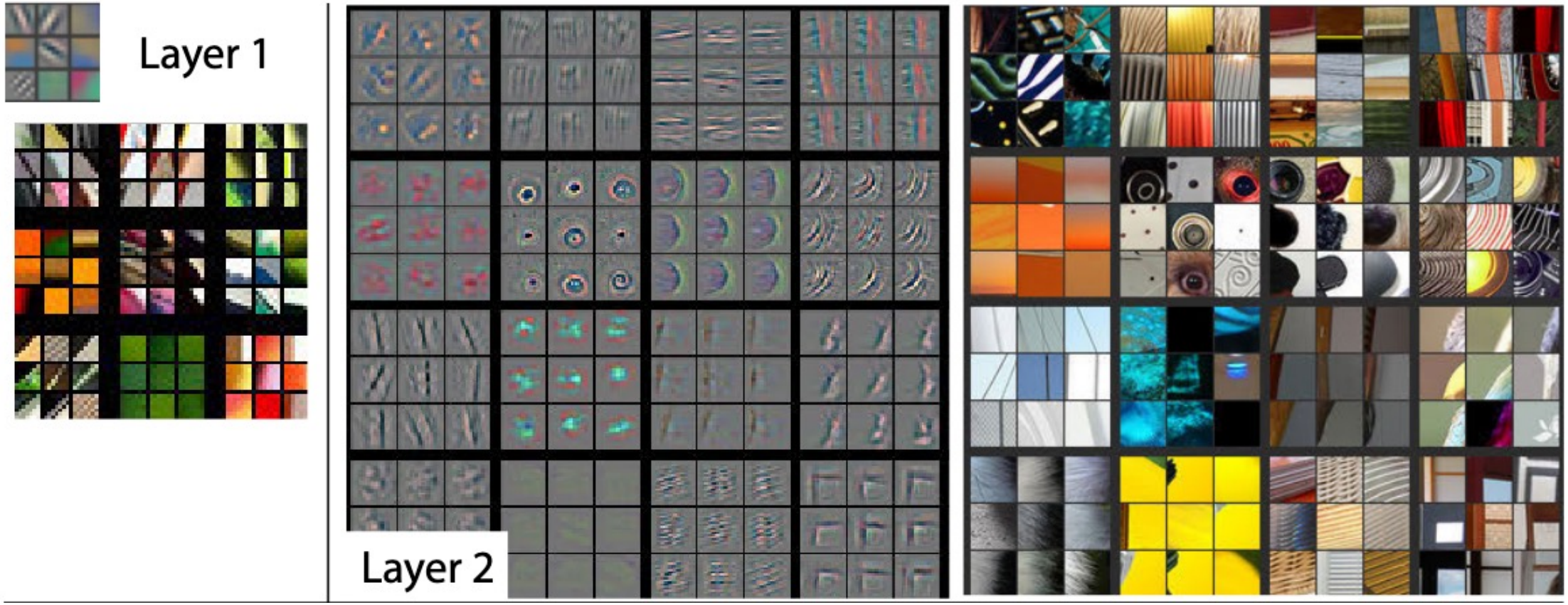


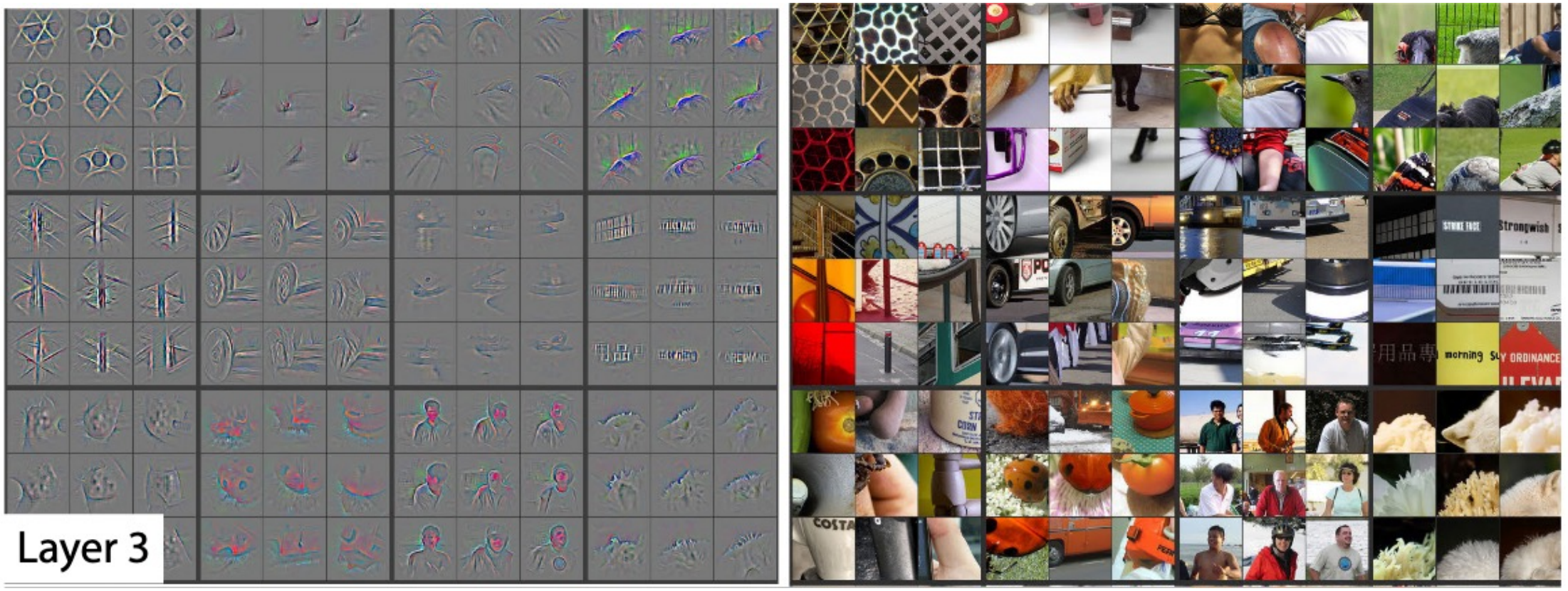
Figure 1. Top: A deconvnet layer (left) attached to a convnet layer (right). The deconvnet will reconstruct an approximate version of the convnet features from the layer beneath. Bottom: An illustration of the unpooling operation in the deconvnet, using *switches* which record the location of the local max in each pooling region (colored zones) during pooling in the convnet.

Zeiler, Matthew D., and Rob Fergus. "Visualizing and understanding convolutional networks." Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part I 13. Springer International Publishing, 2014.

Deconvolution and Unpooling



Deconvolution and Unpooling



Deconvolution and Unpooling