CS7.405 Responsible & Safe AI Systems

Ponnurangam Kumaraguru ("PK")
#ProfGiri @ IIIT Hyderabad

pk.profgiri /in/ponguru @ponguru Ponnurangam.kumaraguru

Guest lectures how?

Daniel Paleka

Adam

Andy Zou

Arun Jose

Neel Nanda

☹ Not all join the lecture ☹

Goal

Less of Blackbox and more transparent

Motivation

Transparency tools try to provide clarity about a model's inner workings

Model changes can sometimes cause the internal representations to substantially change, so we would like to understand when models process data differently

Transparency could make it easier for monitors to detect deception and other hazards

Pixel attribution methods

Sensitivity map, saliency map, pixel attribution map, gradient-based attribution methods, feature relevance, feature attribution, and feature contribution.

Feature attribution explains individual predictions by attributing each input feature according to how much it changed the prediction (negatively or positively).

Pixel attribution methods

Occlusion- or perturbation-based

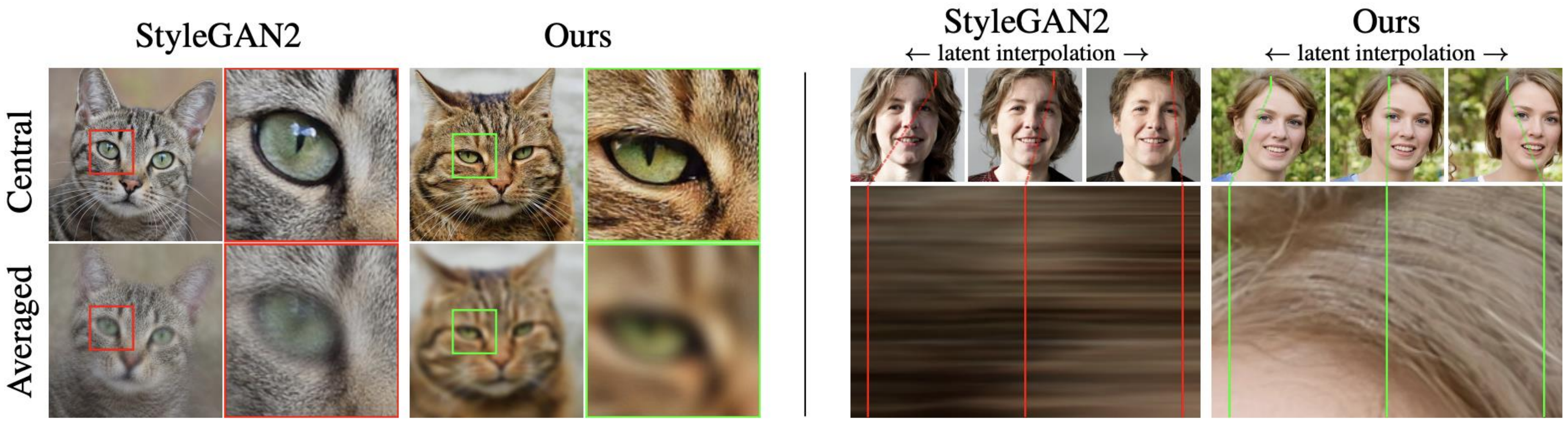
Methods like SHAP and LIME manipulate parts of the image to generate explanations (model-agnostic).

Gradient-based

Many methods compute the gradient of the prediction (or classification score) with respect to the input features.

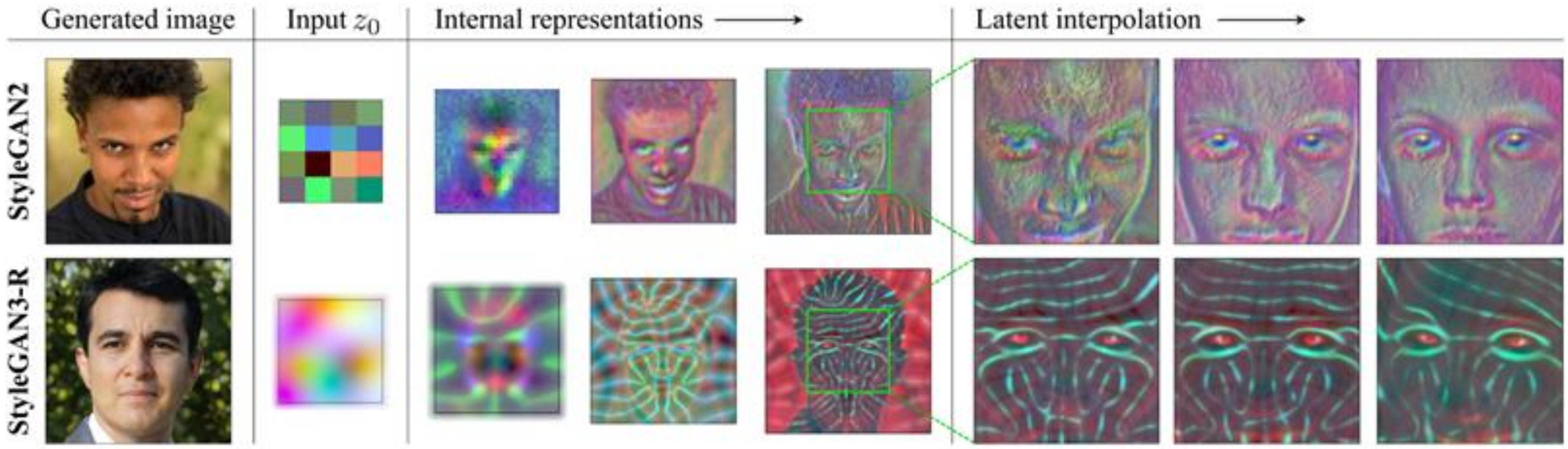
The gradient-based methods mostly differ in how the gradient is computed.

Motivation



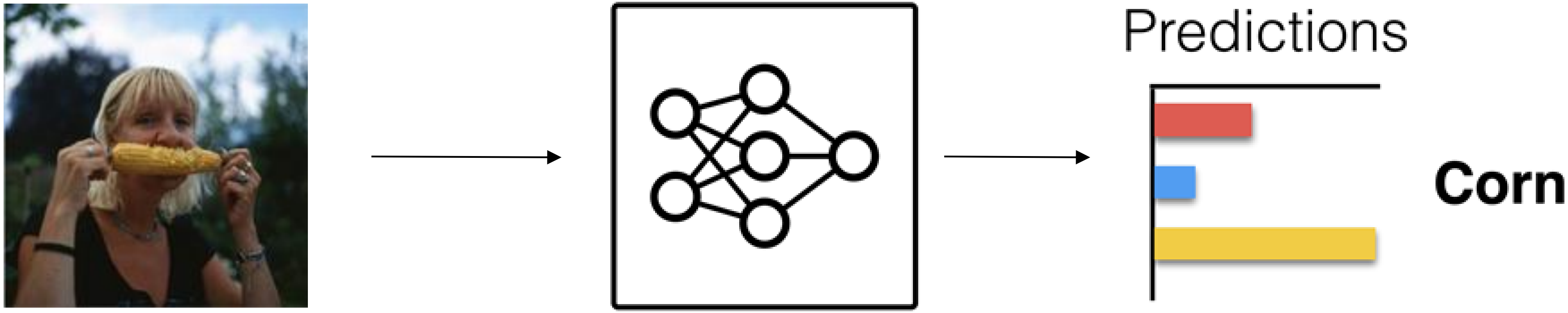
Texture sticking
Left: average of images generated from a small neighborhood around a central latent (top row)
Right: extract small vertical segment of pixels, stack horizontally
StyleGAN2, same coordinates
Hairs moving in animation

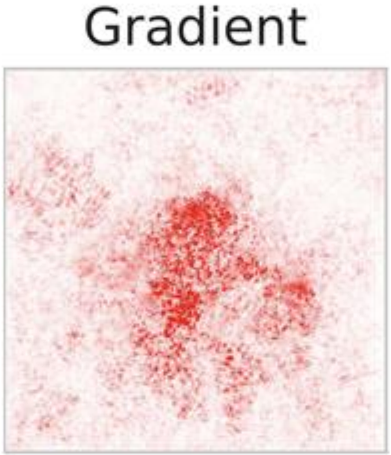
Motivation



StyleGAN2: details glued to the image vs surface; internal representations are different
StyleGAN3: fully equivariant to translation and rotation; help in identifying important properties better

Saliency Maps

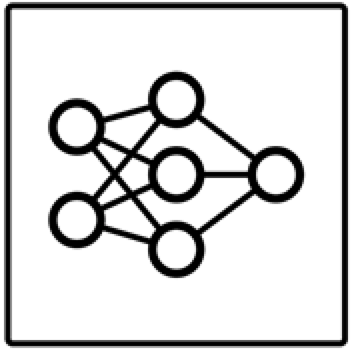


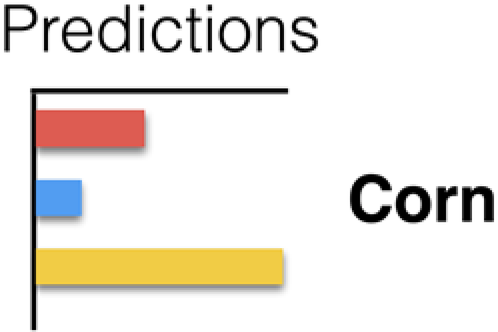


$$E_{grad}(x) = \frac{\partial S_i}{\partial x}$$

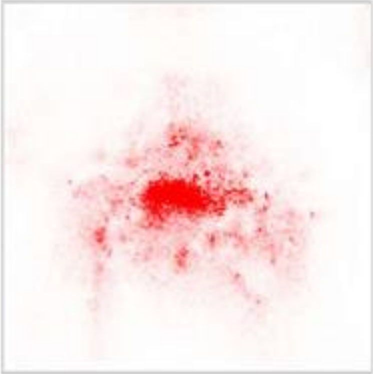
Perturbation direction of fastest ascent for the class logit

Saliency Maps





SmoothGrad

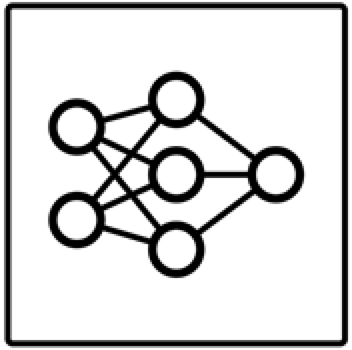


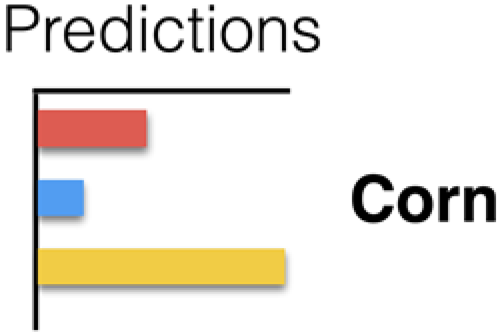
$$E_{\mathrm{sg}}(x) = \frac{1}{N}\sum_{i=1}^{N} E(x + g_i),$$

Each input perturbed by different Gaussian noise and then averaged

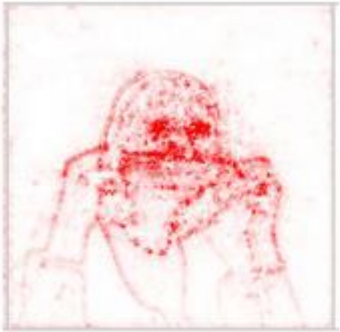
Smoother & less noisy

Saliency Maps





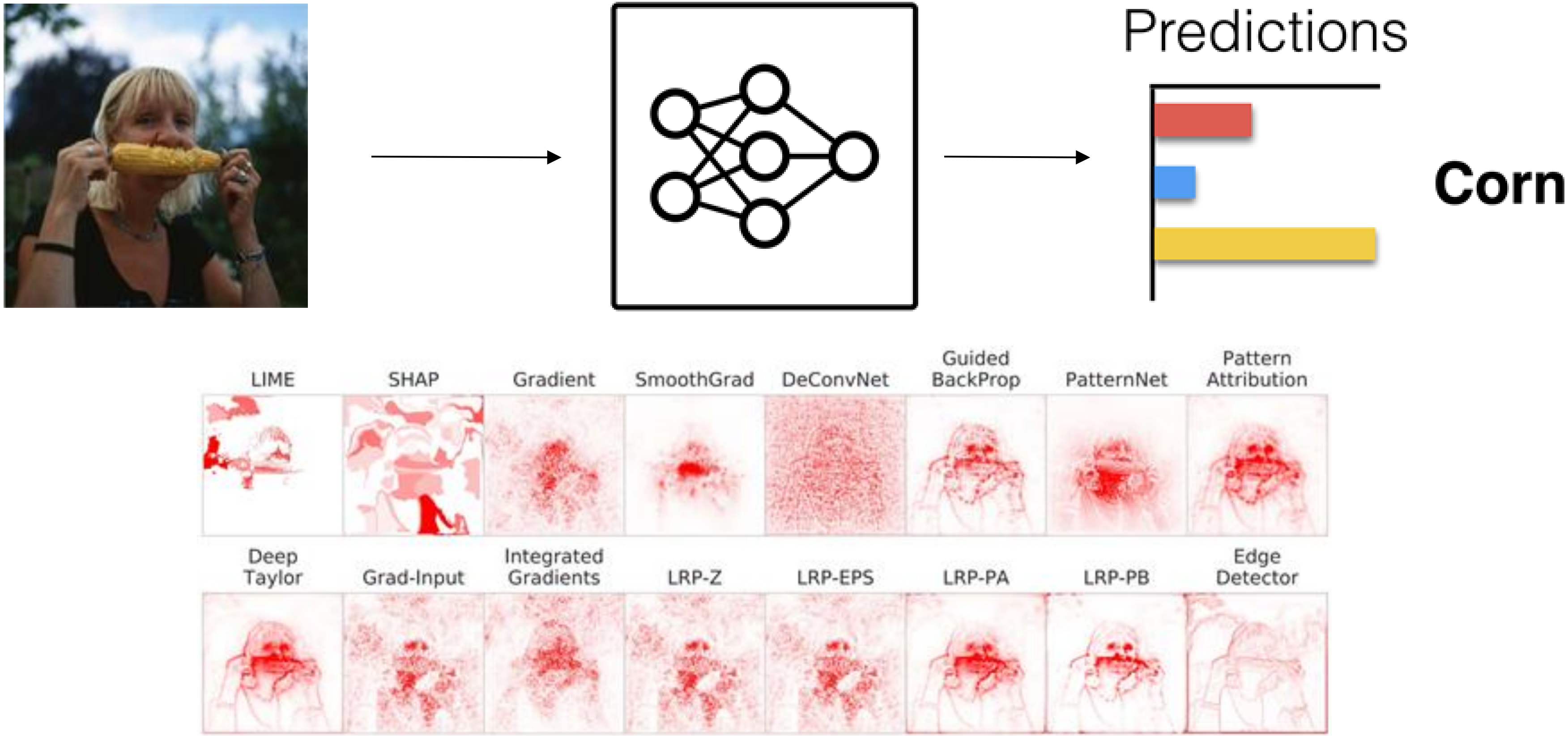
Guided BackProp



$$R^{l} = 1_{R^{l+1}>0} 1_{f^{l}>0} R^{l+1}$$

Backprop with intermediate negative activations and gradients zeroed out

Saliency Maps



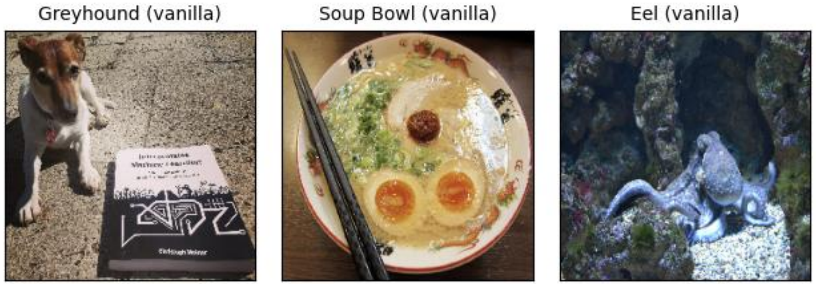
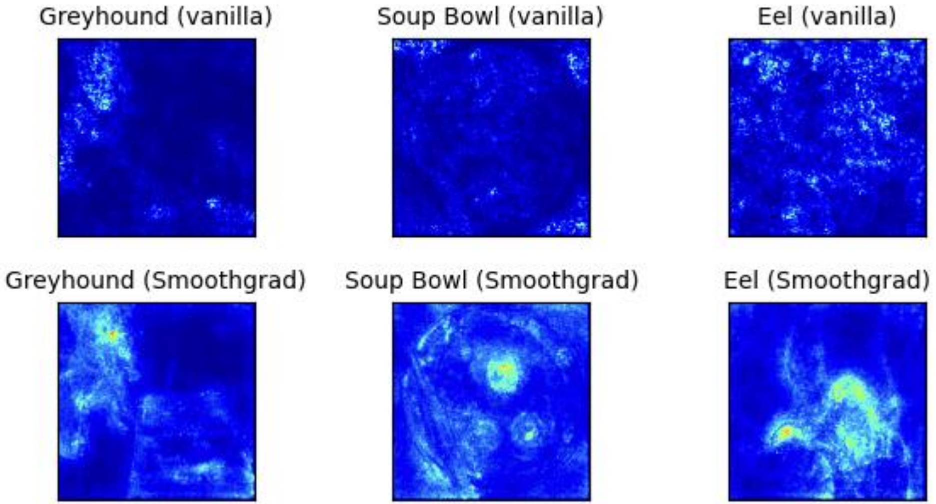
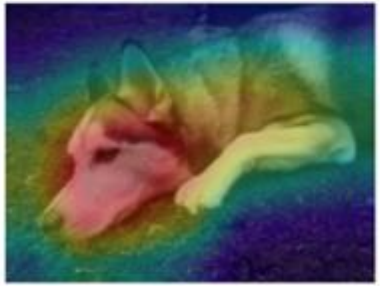
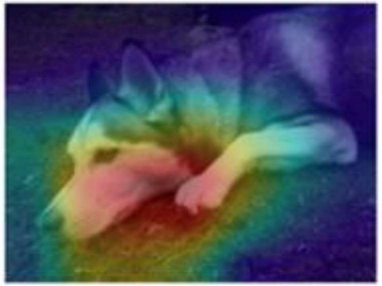


FIGURE 10.9: Images of a dog classified as greyhound, a ramen soup classified as soup bowl, and an octopus classified as eel.



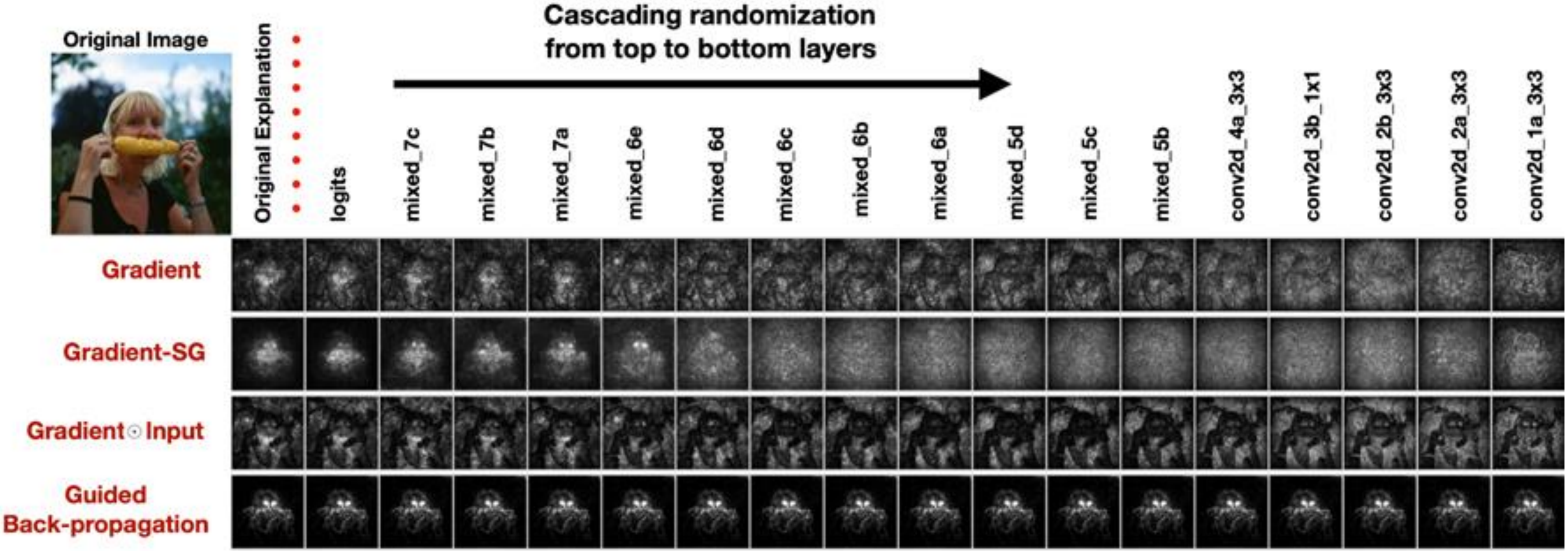
Saliency Maps Can Be Deceptive

Many transparency tools create fun-to-look-at visualizations that do not actually inform us much about how models are making predictions

	Test Image	Evidence for Animal Being a Siberian Husky	Evidence for Animal Being a Transverse Flute
Explanations Using Attention Maps			

Sanity Checks for Saliency Maps

If we randomize the layers, some saliency maps do not change much, which suggests they do not capture what the model has learned

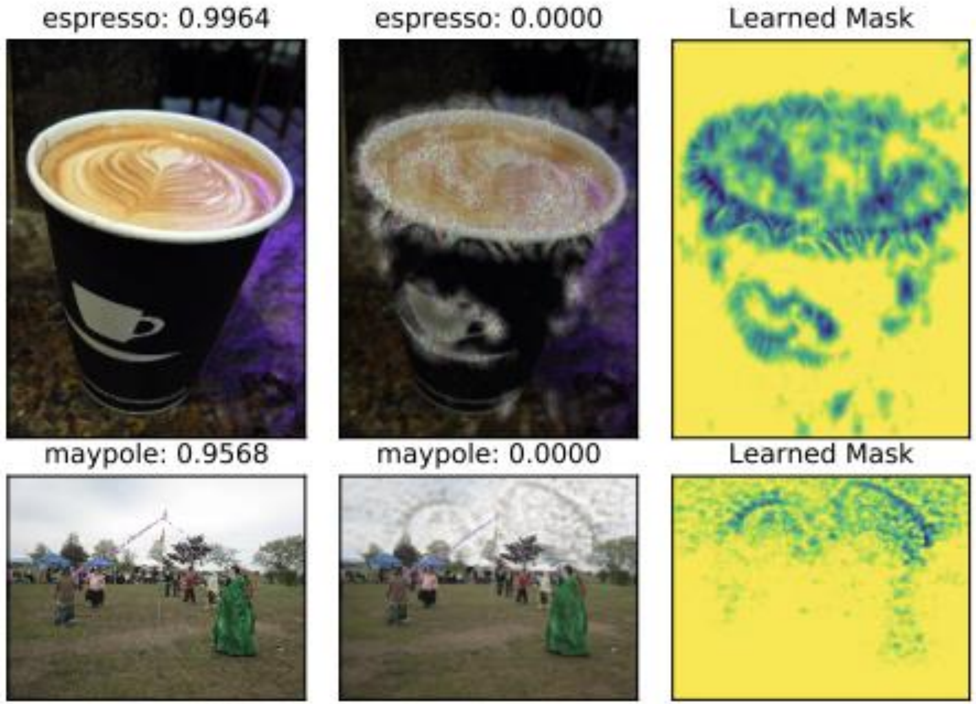


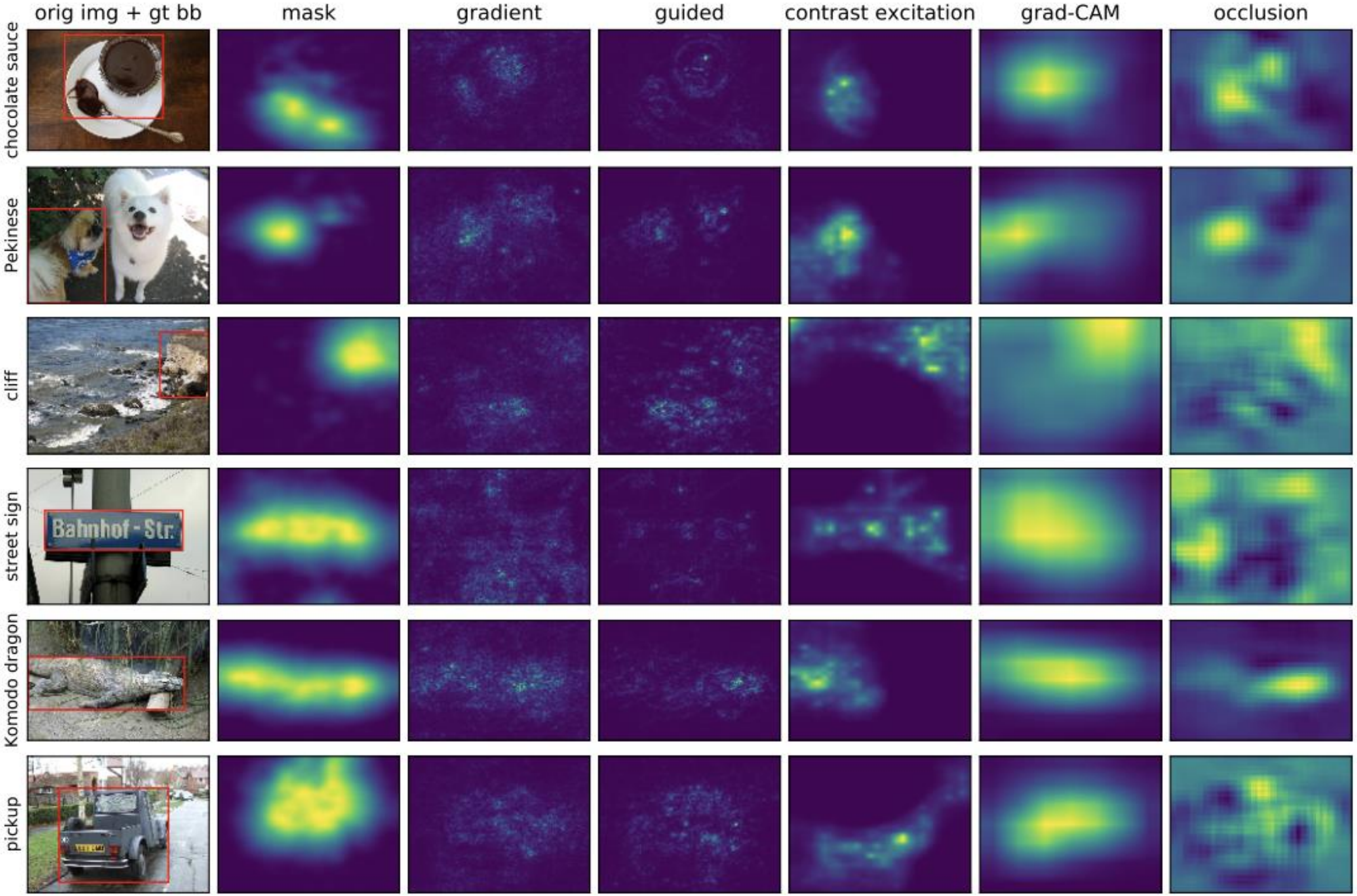
Optimized Masks for Saliency

Some saliency maps optimize a mask to locate and blur salient regions

Figure 1. An example of a mask learned (right) by blurring an image (middle) to suppress the softmax probability of its target class (left: original image; softmax scores above images).

This is highly sensitive to hyperparameters and mask initialization


espresso: 0.9964
espresso: 0.0000
Learned Mask
maypole: 0.9568
maypole: 0.0000
Learned Mask


orig img + gt bb
mask
gradient
guided
contrast excitation
grad-CAM
occlusion
chocolate sauce
Pekinese
cliff
street sign
Bahnhof - Str.
Komodo dragon
pickup

Pros & Cons Gradient based

Explanations are visual, detecting important regions is easy in the image

Faster to compute than model-agnostic methods

LIME & SHAP are very expensive

Difficult to know whether an explanation is correct

Very fragile - adversarial perturbations produce same prediction

Saliency Maps for Text

Saliency maps can be used for text models too

	$p(y \mid \mathbf{x}; \theta)$	y	c
the year 's best and most unpredictable comedy	0.91	pos	pos
we never feel anything for these characters	0.95	neg	neg
handsome but unfulfilling suspense drama	0.18	neg	pos

y = gold, c = predicted

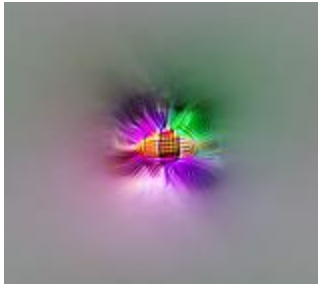
There are many possible saliency scores for a token; one possibility is to use the magnitude of the gradient of the classifier's logit with respect to the token's embedding

While there is no canonical saliency map, these can be used for identifying salient words when writing adversarial examples

Feature Visualization

To understand what a model's internal component detects, synthesize an image through gradient descent that maximizes the component

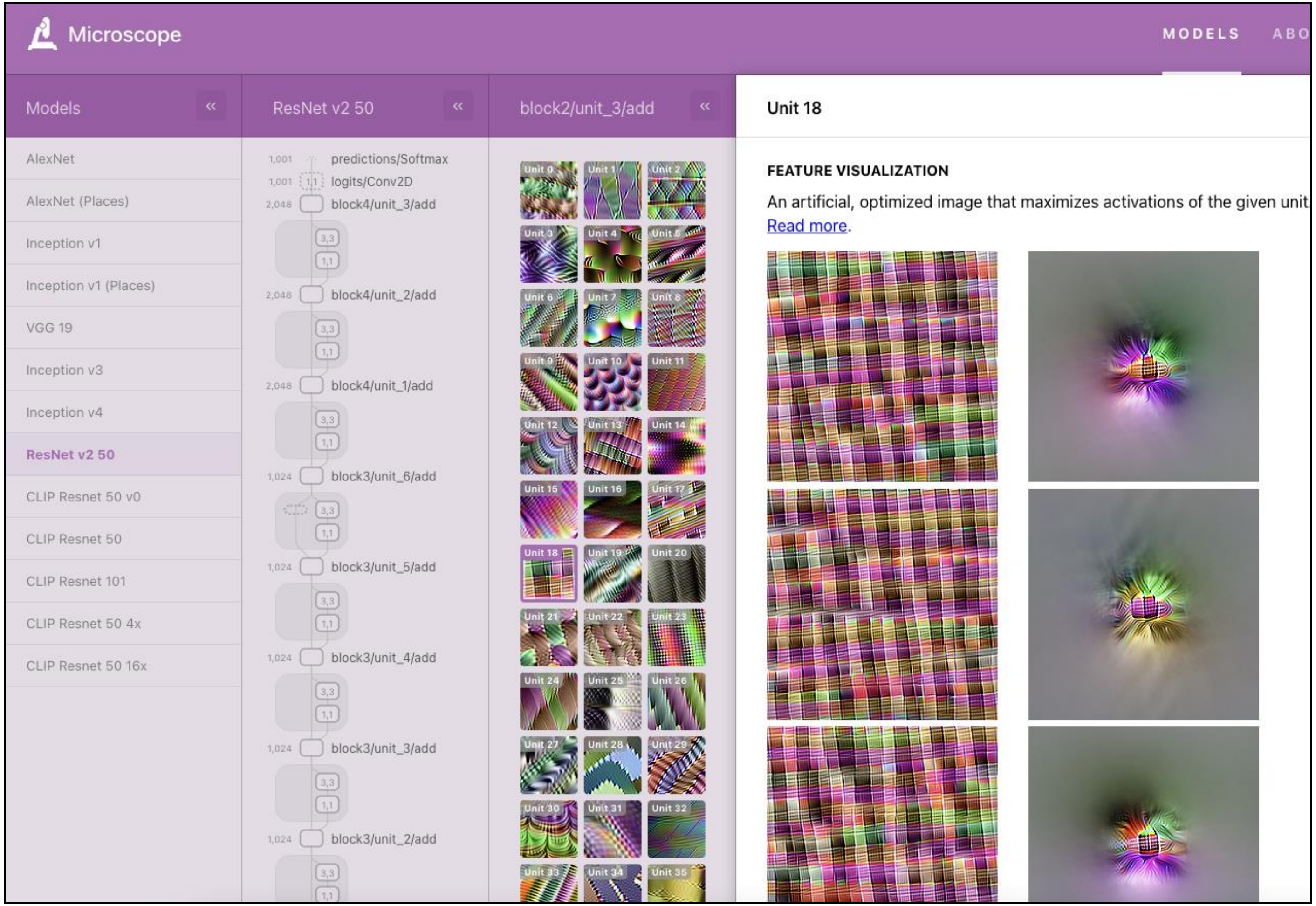
Neuron Visualization



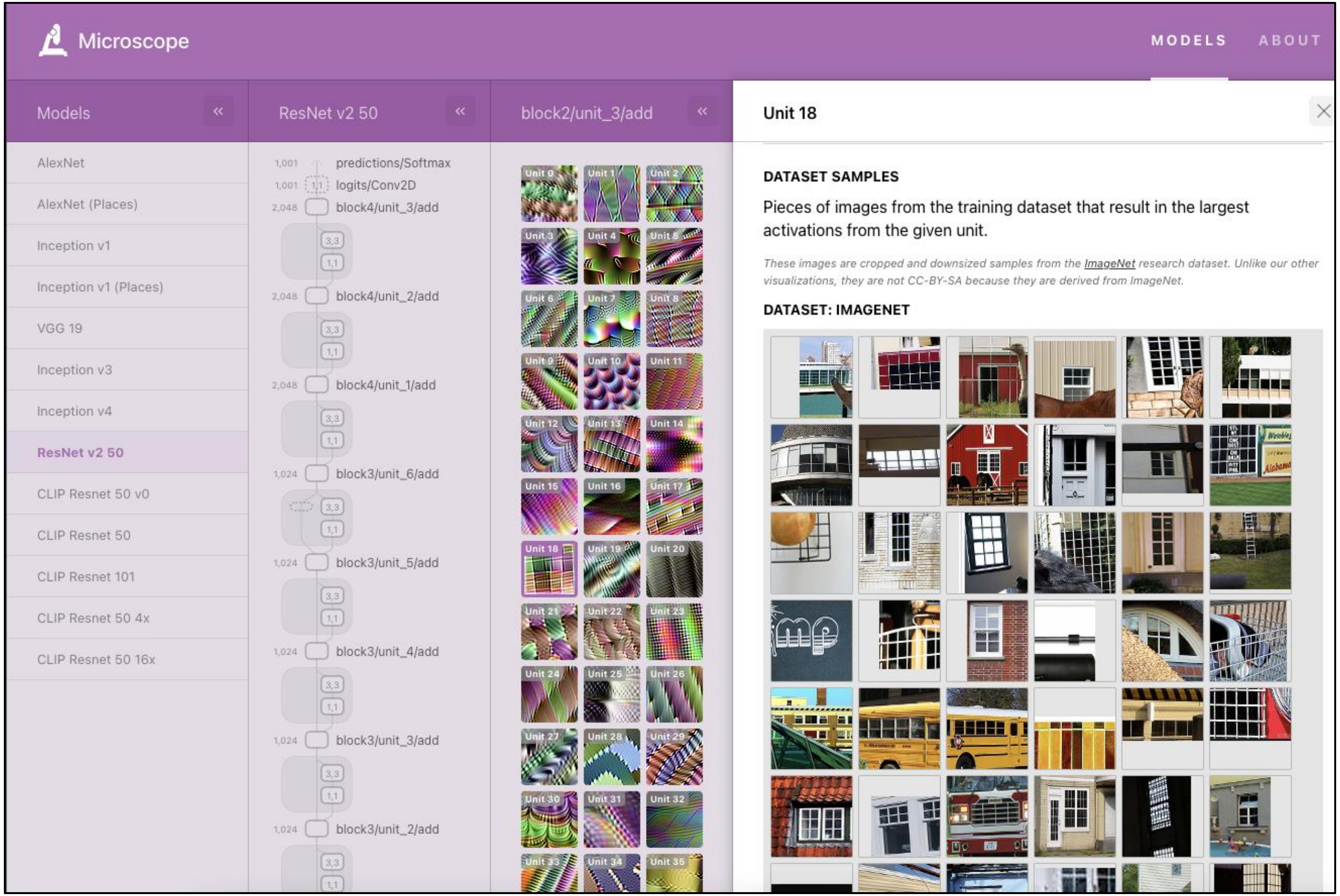
Channel Visualization

Maximally Activating Natural Images

NV: Component = Neuron, optimize the image to maximally activate the neuron, repeated round of GD optimize the noise image
CV: Like Neuron Viz, both gradient descent, Loss of channel visualization might be sum of the squares of all neurons in the channel, lot of squares



https://microscope.openai.com/models/resnetv2_50_slim/resnet_v2_50_block2_unit_3_bottleneck_v2_add_0/18



https://microscope.openai.com/models/resnetv2_50_slim/resnet_v2_50_block2_unit_3_bottleneck_v2_add_0/18

The OpenAI Microscope is a collection of visualizations of every significant layer and neuron of 13important vision models. LEARN MORE▷

AlexNet

A landmark in computer vision, this 2012 winner of ImageNet has over 50,000 citations.

26 nodes

AlexNet (Places)

The same architecture as the classic AlexNet model, but trained on the Places365 dataset.

26 nodes

Inception v1

Also known as GoogLeNet, this network set the state of the art in ImageNet classification in 2014.

84 nodes

Inception v1 (Places)

The same architecture as the classic Inception v1 model, but trained on the Places365 dataset.

VGG 19

Introduced in 2014, this network is simpler than Inception variants, using only 3x3 convolutions and no

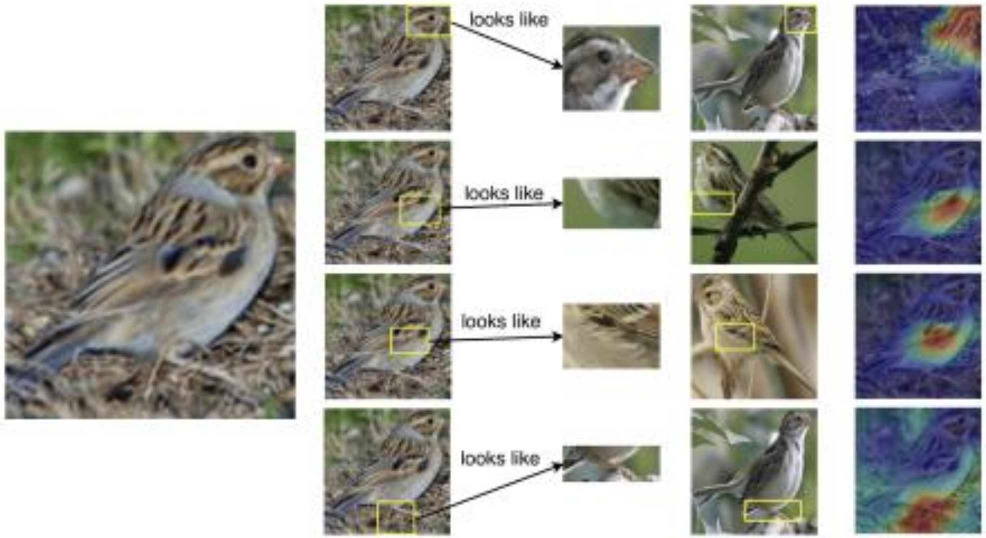
Inception v3

Released in 2015, this iteration of the Inception architecture improved performance and efficiency.

https://microscope.openai.com/models

ProtoPNet (“This Looks Like That”)

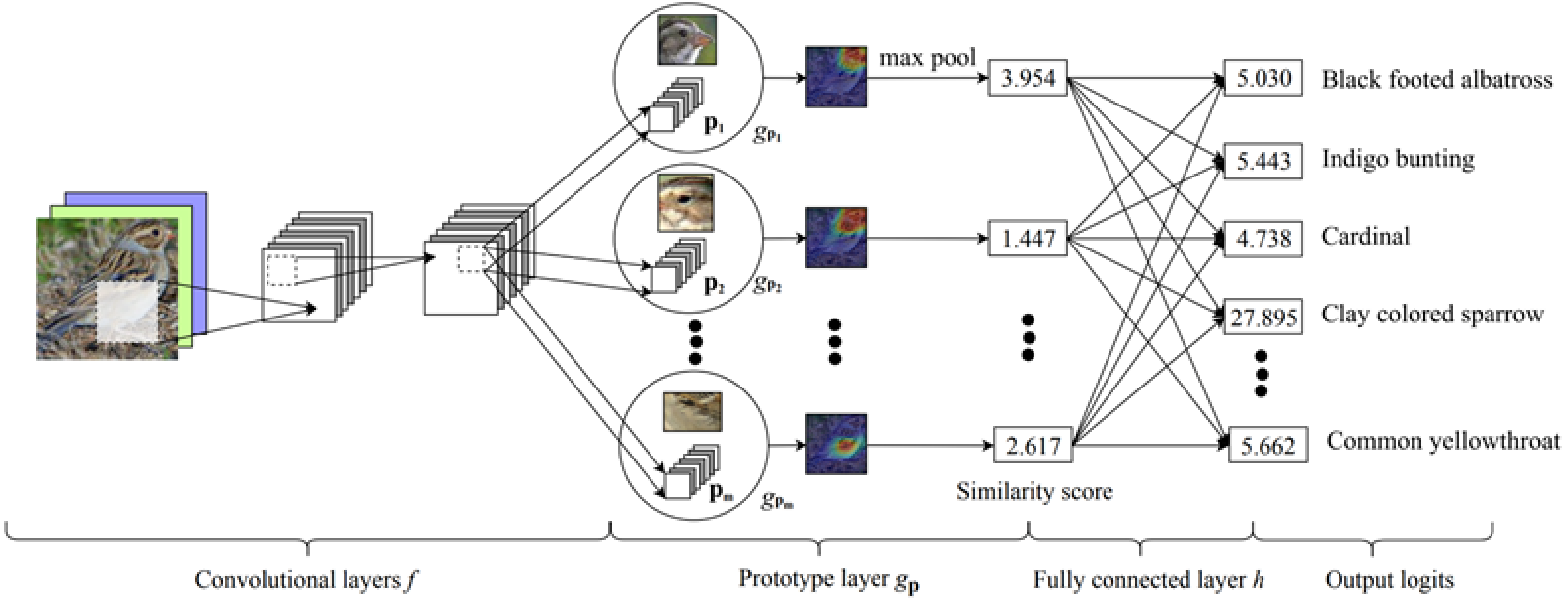
These models perform classifications based on the most important patches of training images, using patches that are prototypical of the class



Leftmost: a test image of a clay-colored sparrow
Second column: same test image, each with a bounding box generated by our model -- the content within the bounding box is considered by our model to look similar to the prototypical part (same row, third column) learned by our algorithm
Third column: prototypical parts learned by our algorithm
Fourth column: source images of the prototypical parts in the third column
Rightmost column: activation maps indicating how similar each prototypical part resembles part of the test bird

ProtoPNet ("This Looks Like That")

These models perform classifications based on the most important patches of training images, using patches that are prototypical of the class



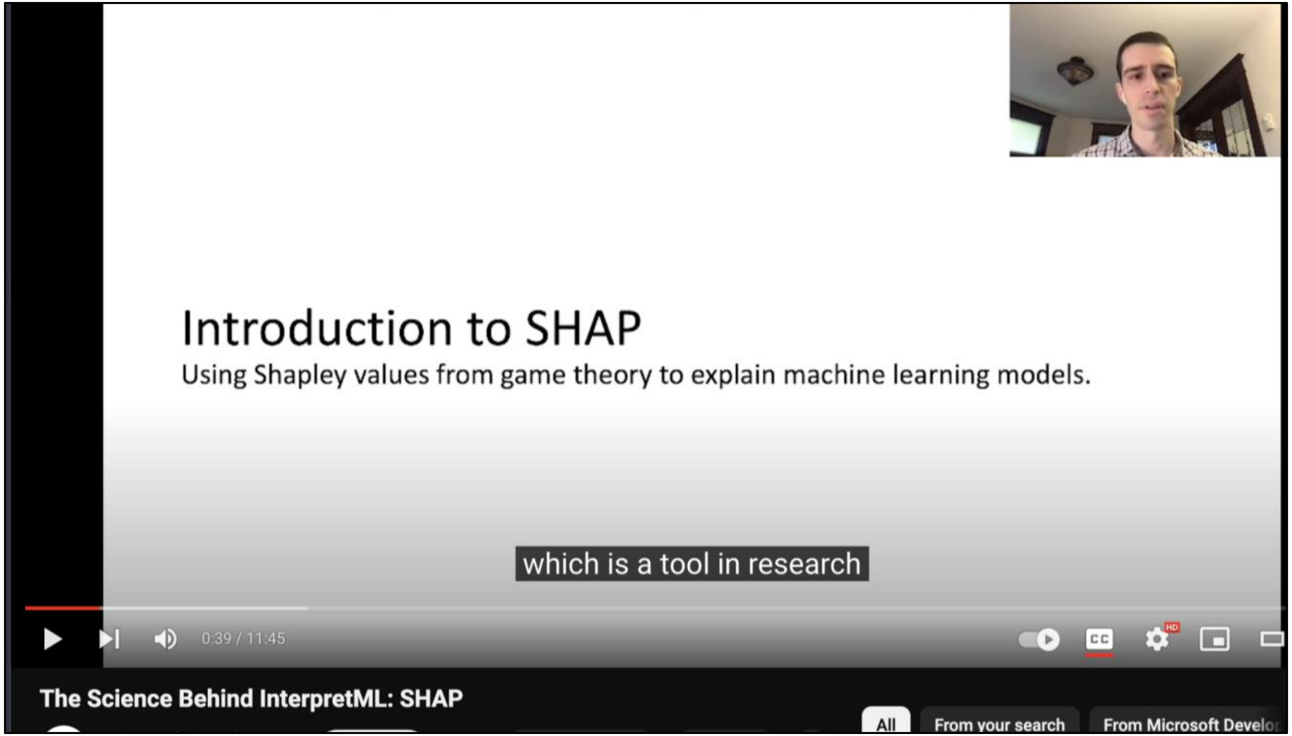
Administrativia

Project review today, hopefully all of you have made decent progress

Poster presentation

- 24 April [Wednesday] 1600 – 1800hrs?
  - Figure out a location?
- 25 April [Thursday] class timing or anytime will be hard to get audience?

SHAP



https://youtu.be/-taOhqkiuIo?si=TGDmiUD9X-kElV_j

Bibliography / Acknowledgements

https://course.mlsafety.org/

https://rdi.berkeley.edu/understanding_llms/s24

https://aisafetyfundamentals.com/

https://inst.eecs.berkeley.edu/~cs294-149/fa18/

https://aisafety.stanford.edu/

https://docs.google.com/document/d/1goyFTfu-_EB90yuepTFN2unmf-fel3TycTPYh7Kn6dI/edit

https://medium.com/@richardcngo/visualizing-the-deep-learning-revolution-722098eb9c5

pk.profgiri

Ponnurangam.kumaraguru

/in/ponguru

ponguru

pk.guru@iiit.ac.in

Thank you for attending the class!!!