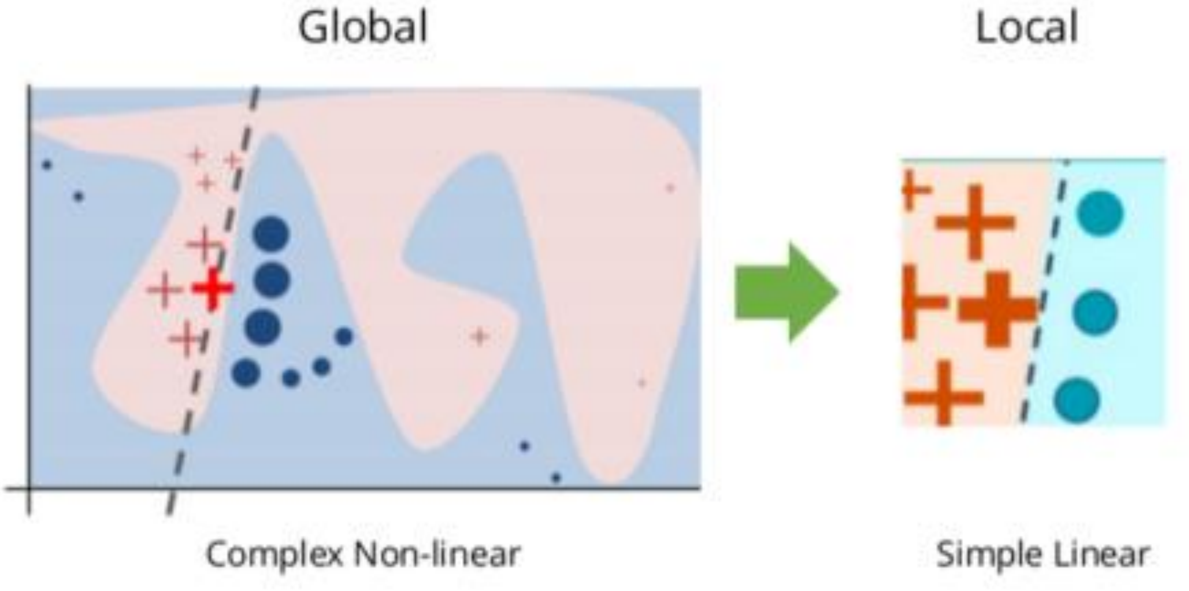
Explainability

Eric Wong

9/29/2022

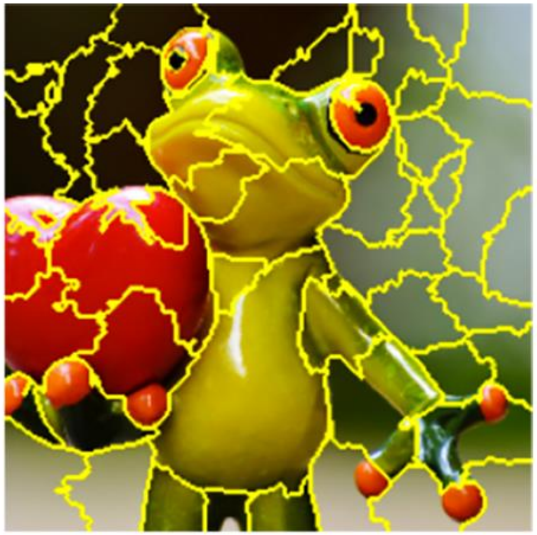
Local Linearity



Marco Tulio Ribeiro "Local Interpretable Model-Agnostic Explanations (LIME): An Introduction"

Superpixels for ”interpretable” features

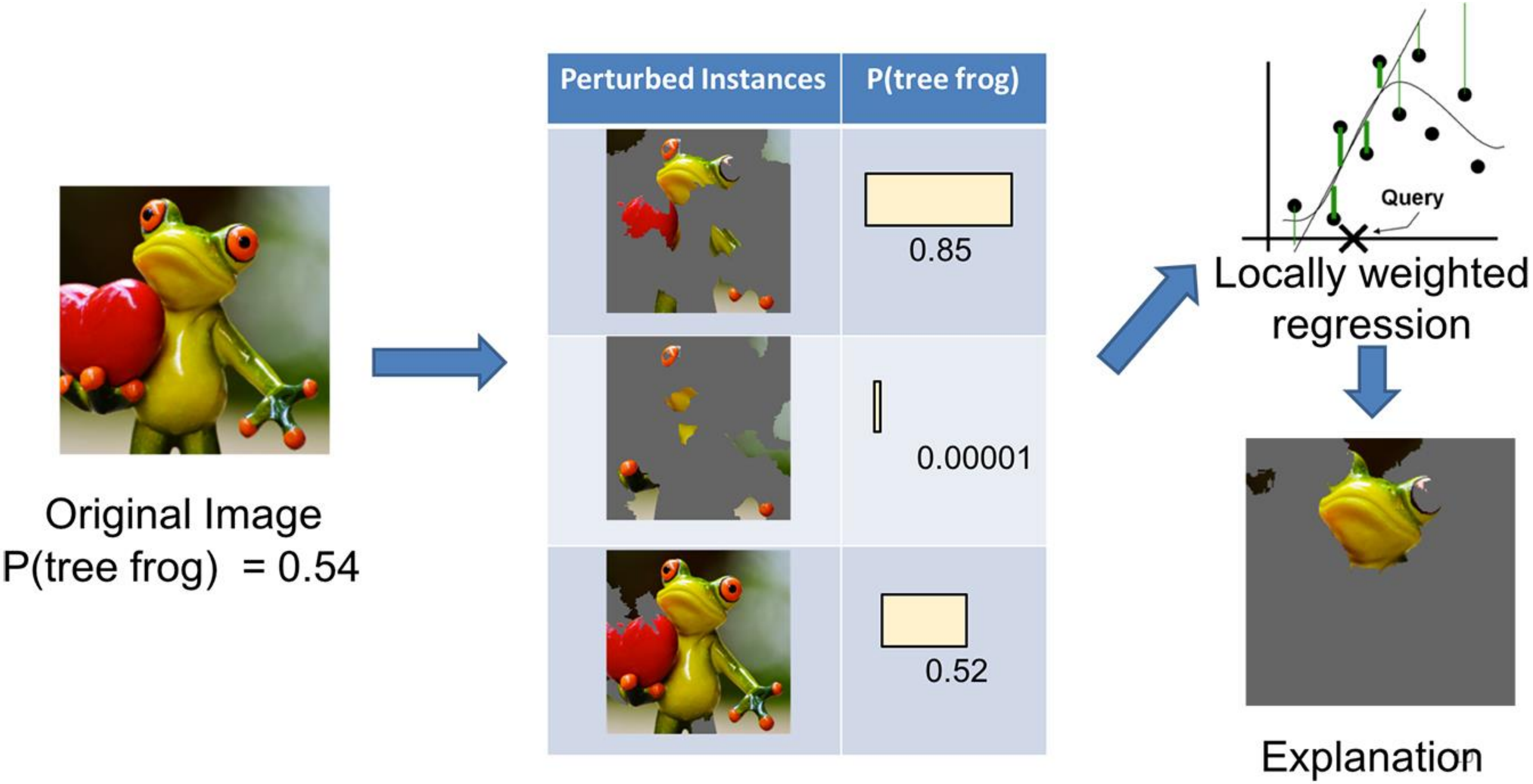
Original Image



Interpretable Components

Marco Tulio Ribeiro ”Local Interpretable Model-Agnostic Explanations (LIME): An Introduction”

Perturb superpixels



Original Image
P(tree frog) = 0.54

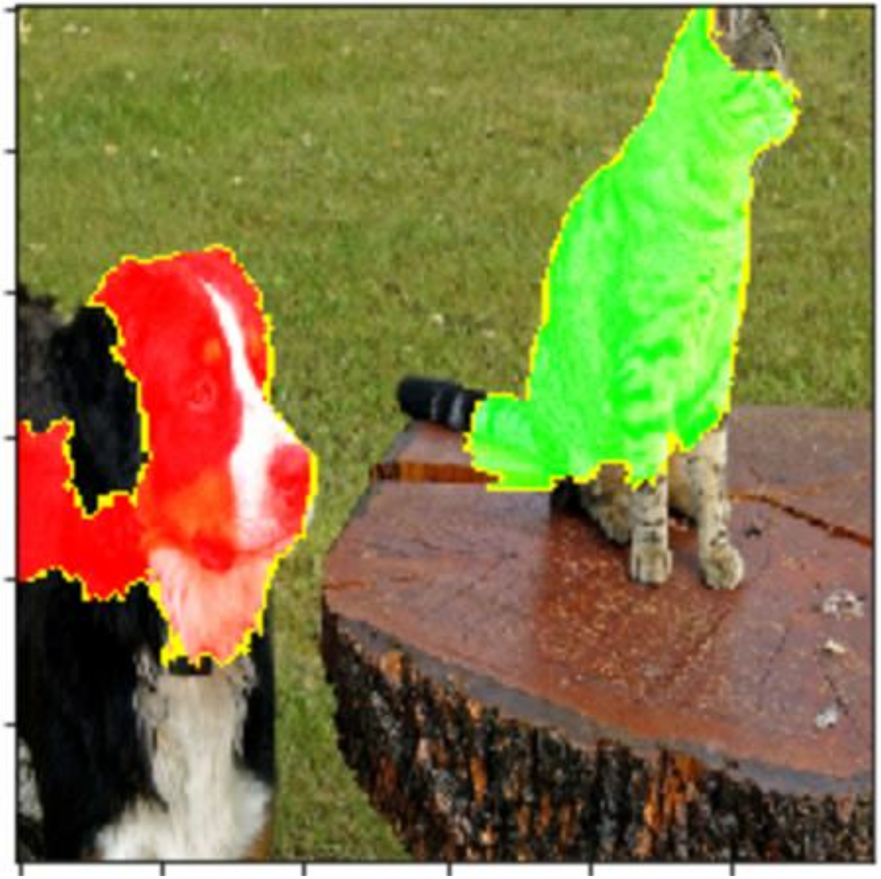
Perturbed Instances	P(tree frog)
	0.85
	0.00001
	0.52

Locally weighted regression

Explanation

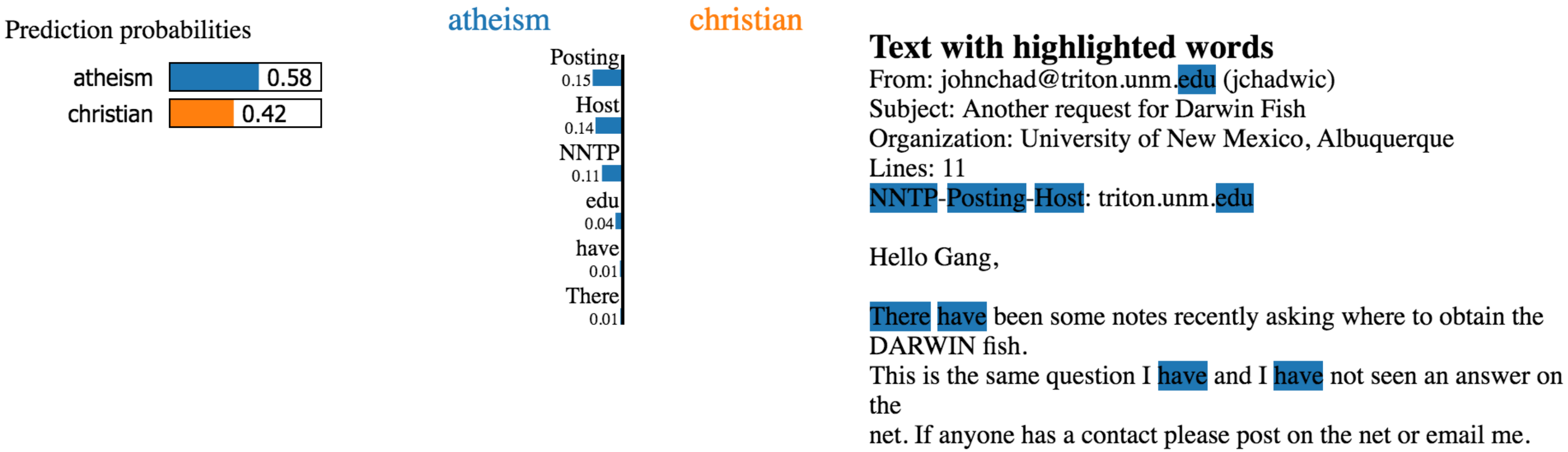
Marco Tulio Ribeiro "Local Interpretable Model-Agnostic Explanations (LIME): An Introduction"

Explaining images



Marco Tulio Ribeiro "Local Interpretable Model-Agnostic Explanations (LIME): An Introduction"

Explaining words



Text with highlighted words

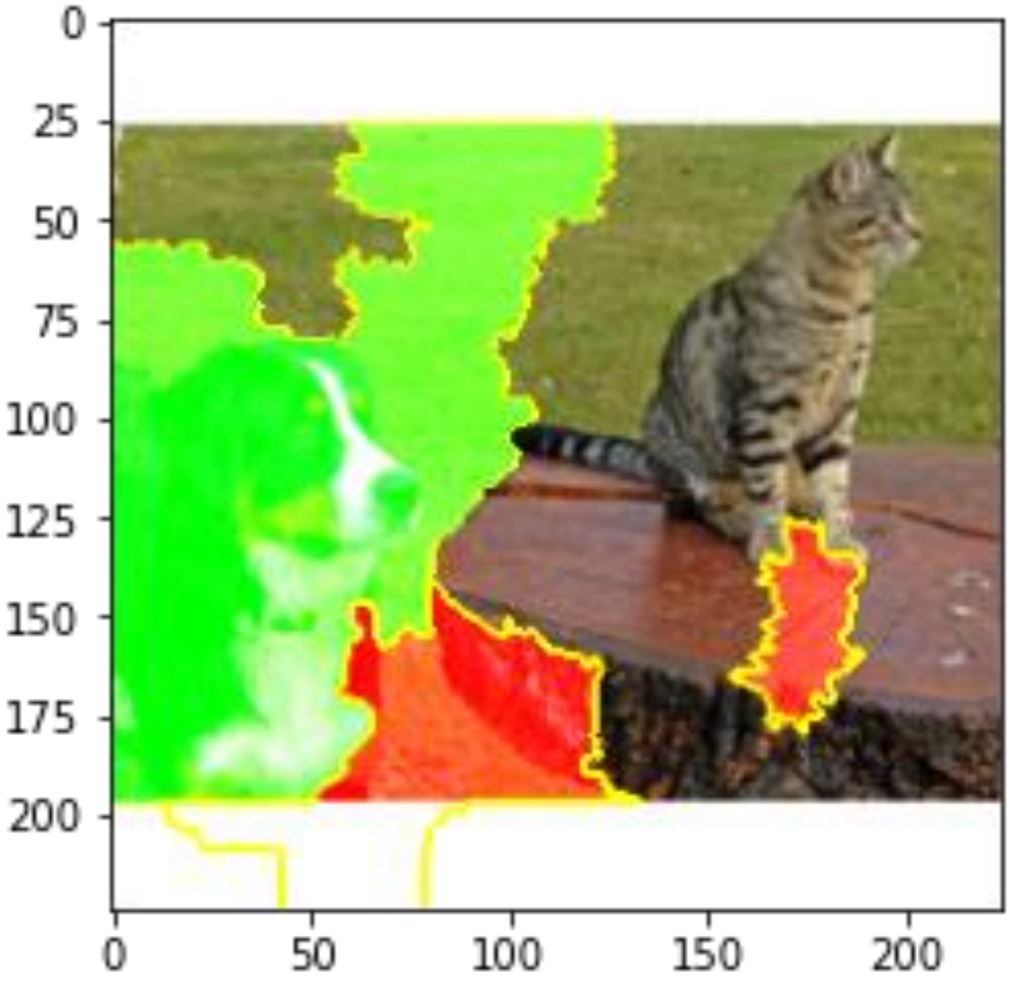
From: johnchad@triton.unm.edu (jchadwic)
Subject: Another request for Darwin Fish
Organization: University of New Mexico, Albuquerque
Lines: 11
NNTP-Posting-Host: triton.unm.edu

Hello Gang,

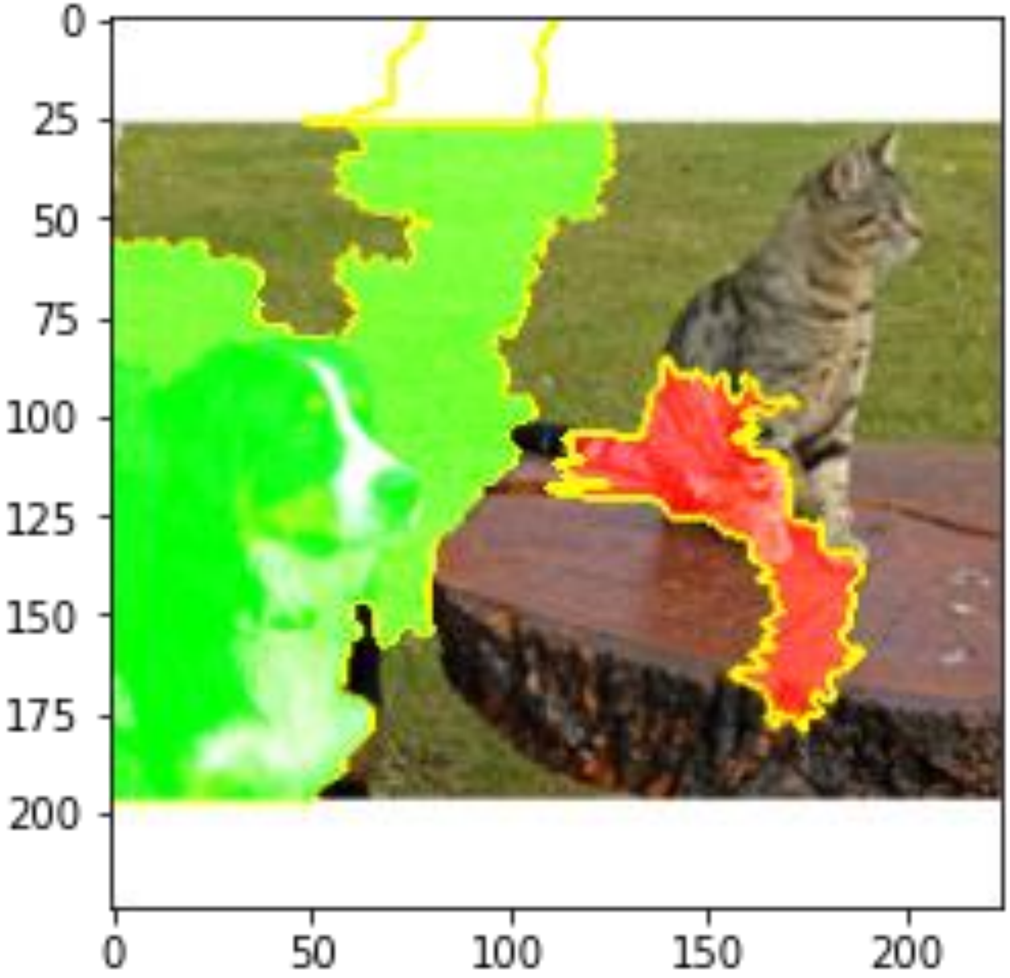
There have been some notes recently asking where to obtain the
DARWIN fish.
This is the same question I have and I have not seen an answer on
the
net. If anyone has a contact please post on the net or email me.

Marco Tulio Ribeiro "Local Interpretable Model-Agnostic Explanations (LIME): An Introduction"

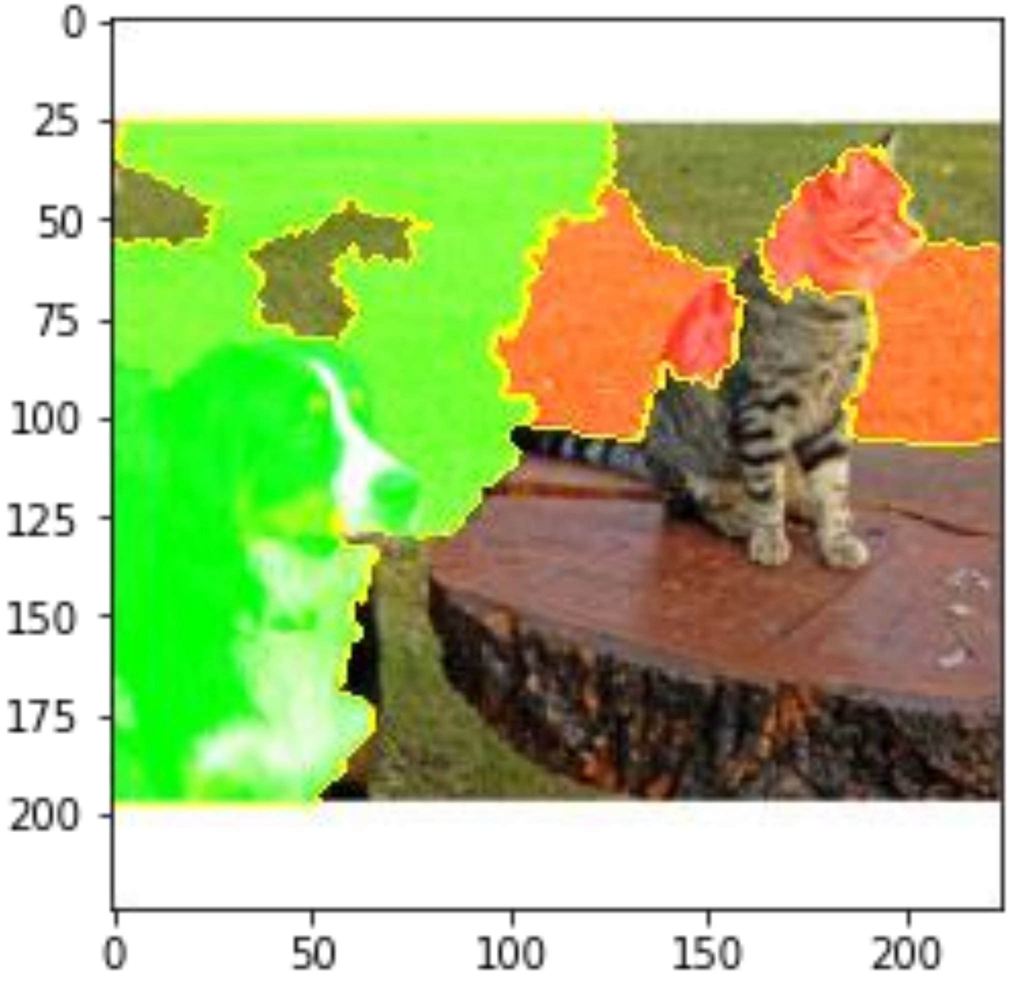
A closer look at dog



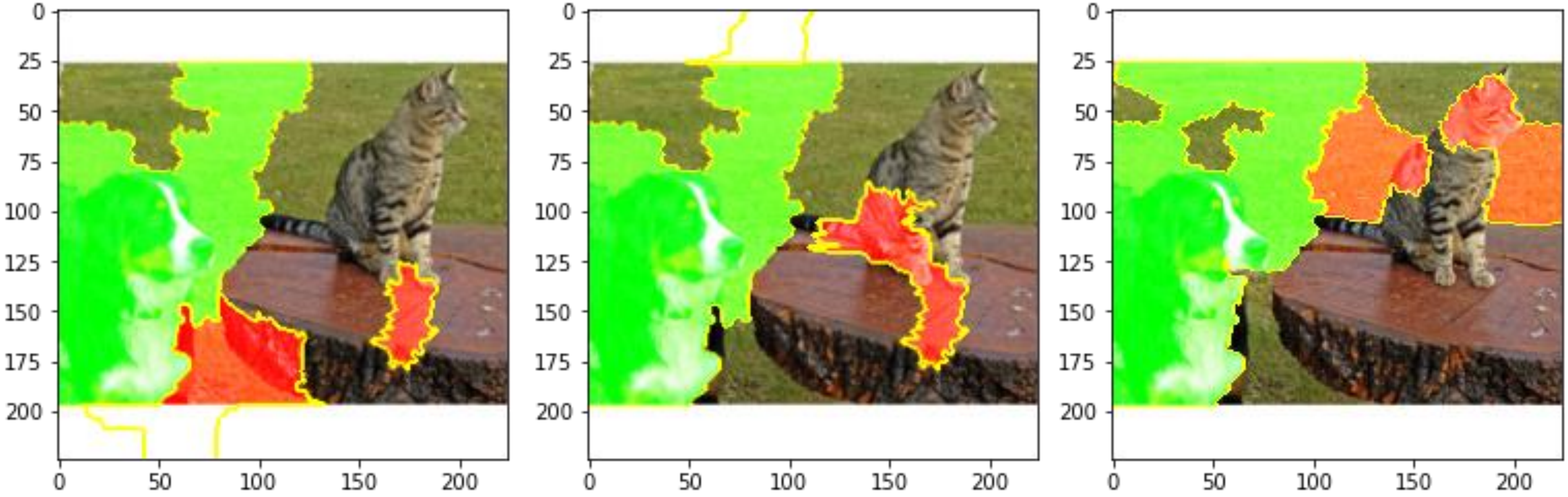
A closer look at dog



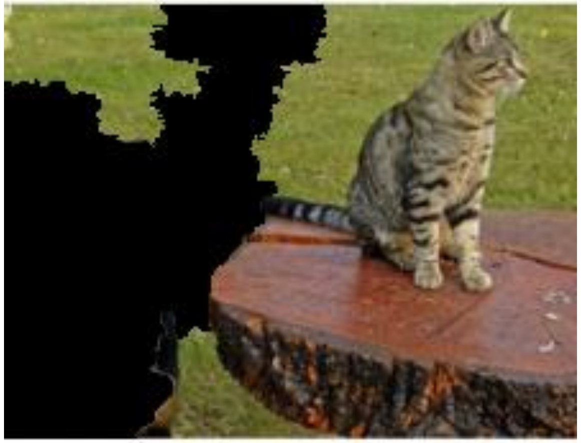
A closer look at dog



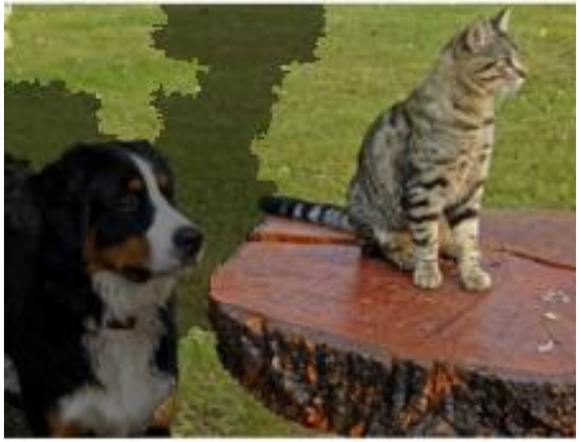
A closer look at dog



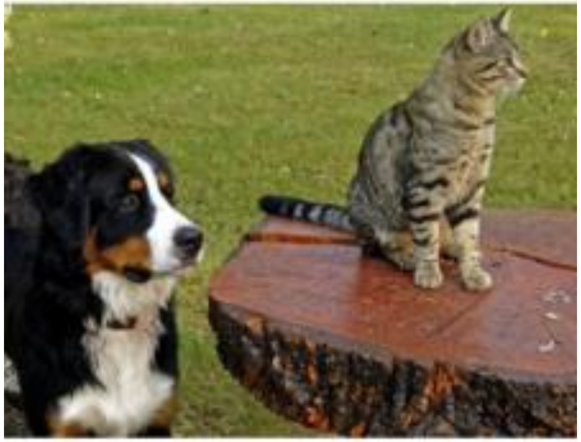
Local linearity?



Mask=0

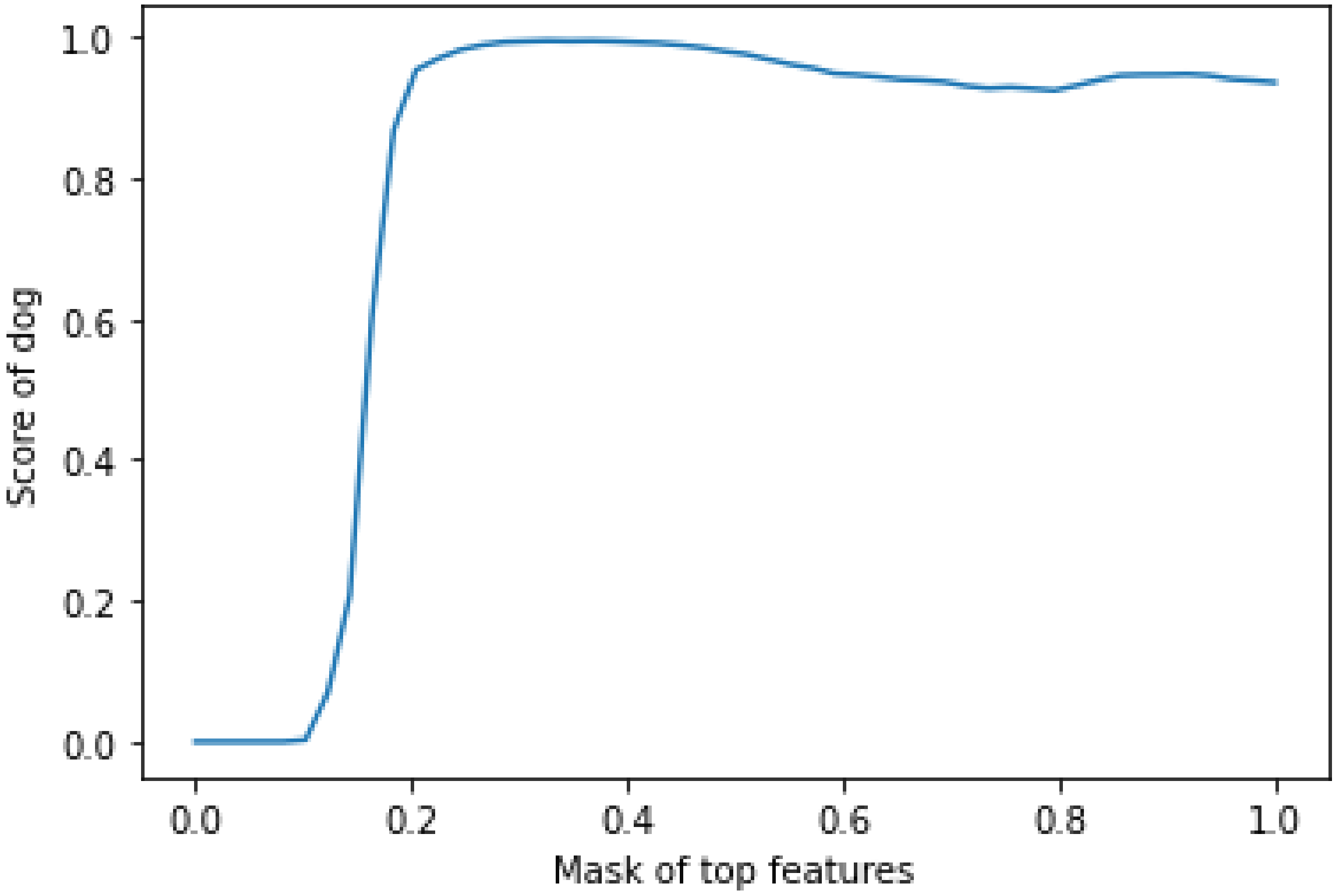


Mask=0.5



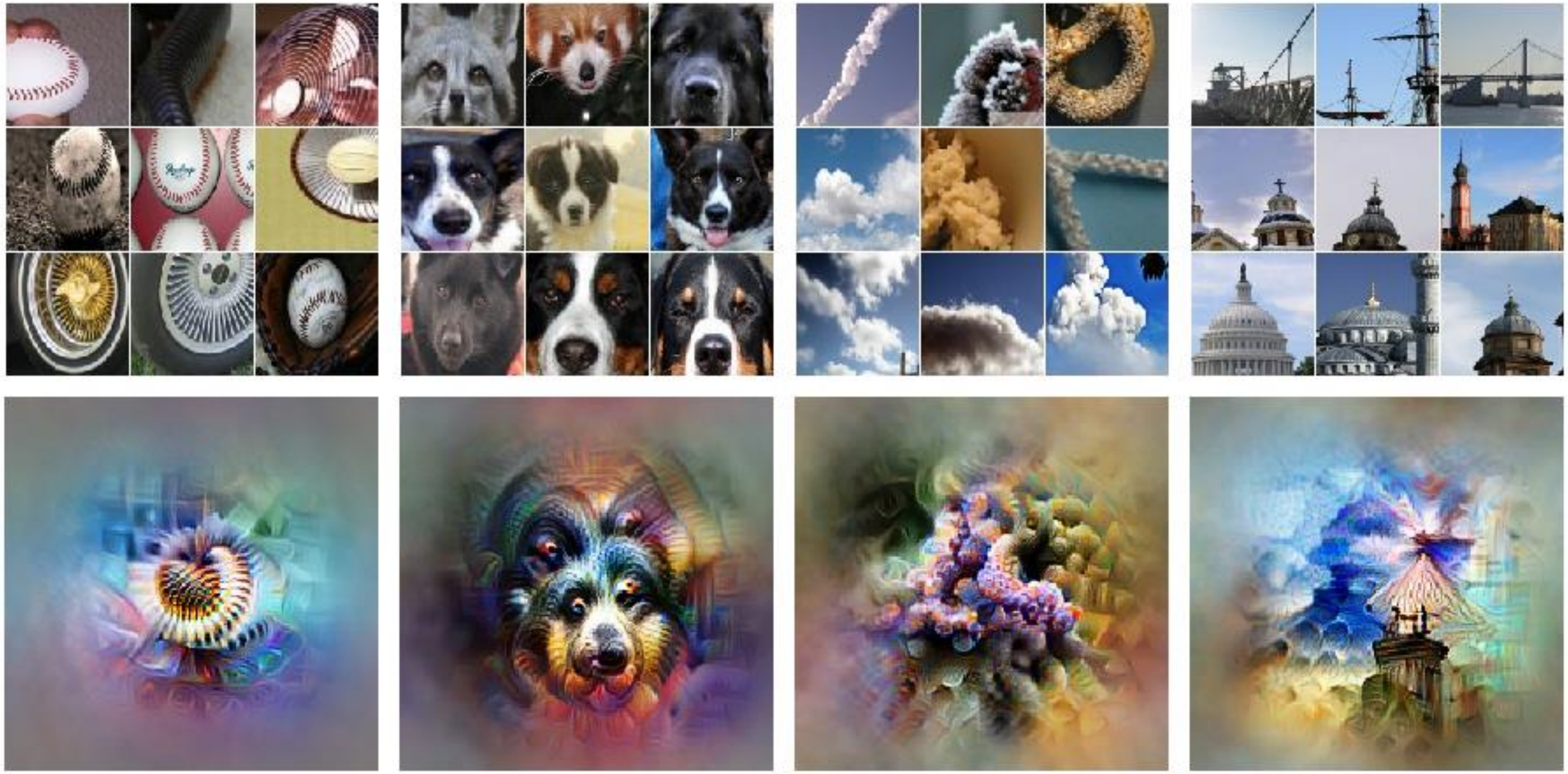
Mask=1

Local linearity?



Feature viz

Exemplars vs Optimization



Baseball—or stripes?
mixed4a, Unit 6

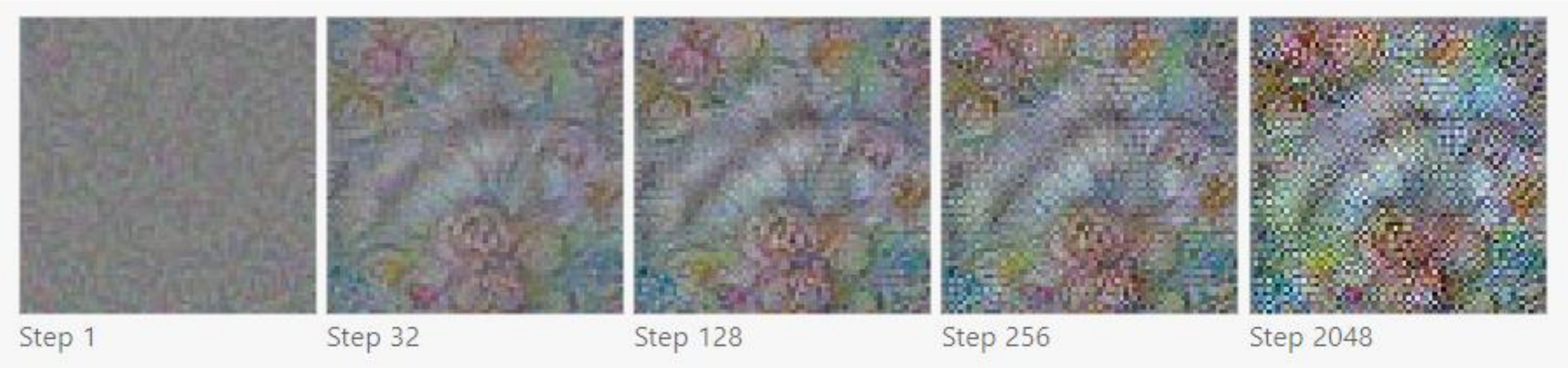
Animal faces—or snouts?
mixed4a, Unit 240

Clouds—or fluffiness?
mixed4a, Unit 453

Buildings—or sky?
mixed4a, Unit 492

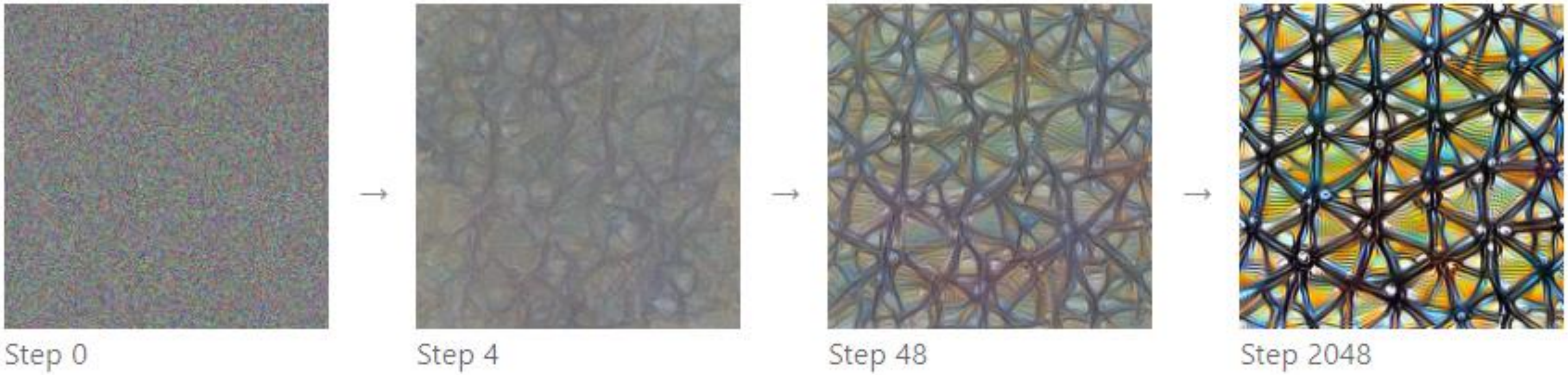
Olah et al. 2017 “Feature Visualization”

Standard gradient ascent is not useful



Olah et al. 2017 “Feature Visualization”

But can work with lots of tricks



Step 0

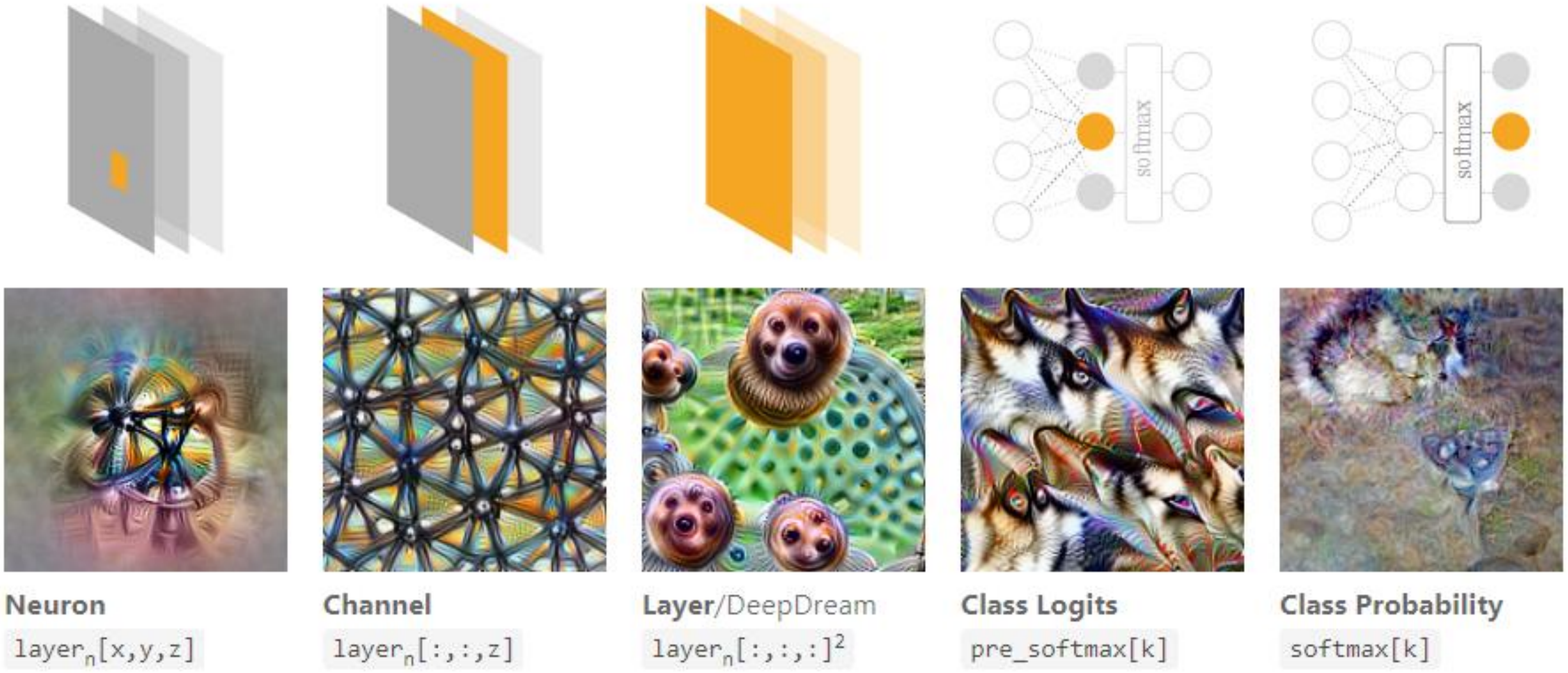
Step 4

Step 48

Step 2048

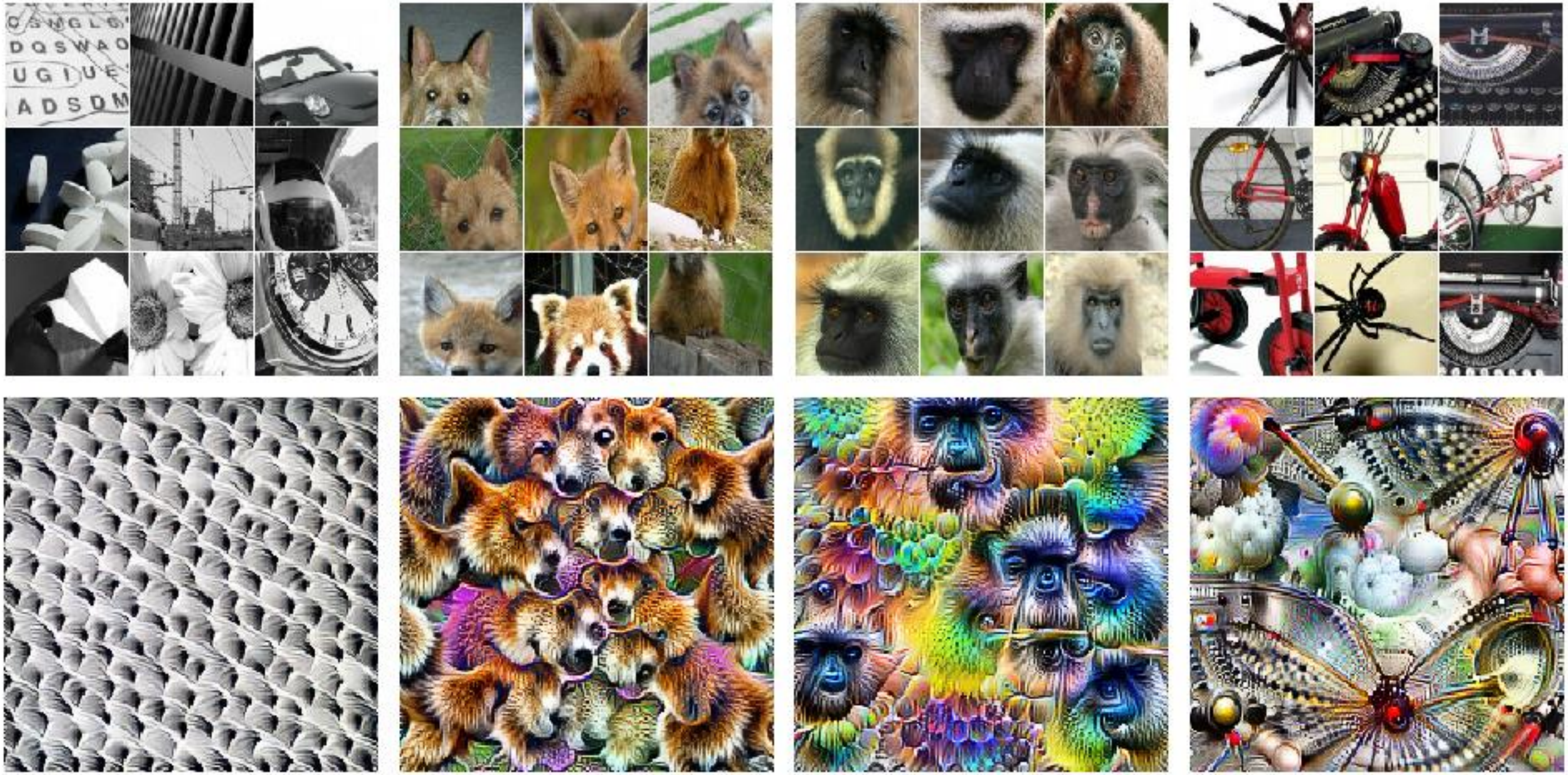
Olah et al. 2017 “Feature Visualization”

Objectives



Olah et al. 2017 “Feature Visualization”

What direction?



mixed3a, random direction

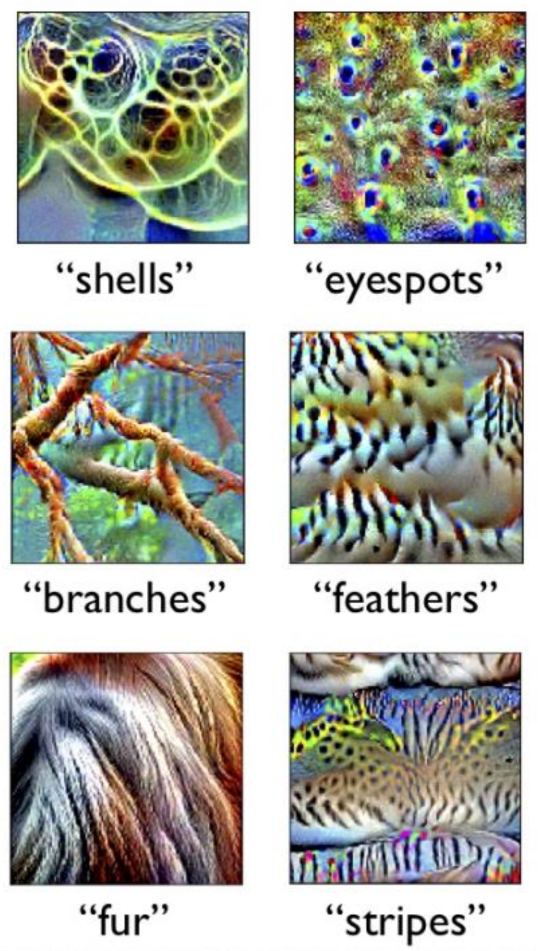
mixed4c, random direction

mixed4d, random direction

mixed5a, random direction

Olah et al. 2017 “Feature Visualization”

Robust models



Engstrom et al. 2019 "Adversarial Robustness as a Prior for Learned Representations"