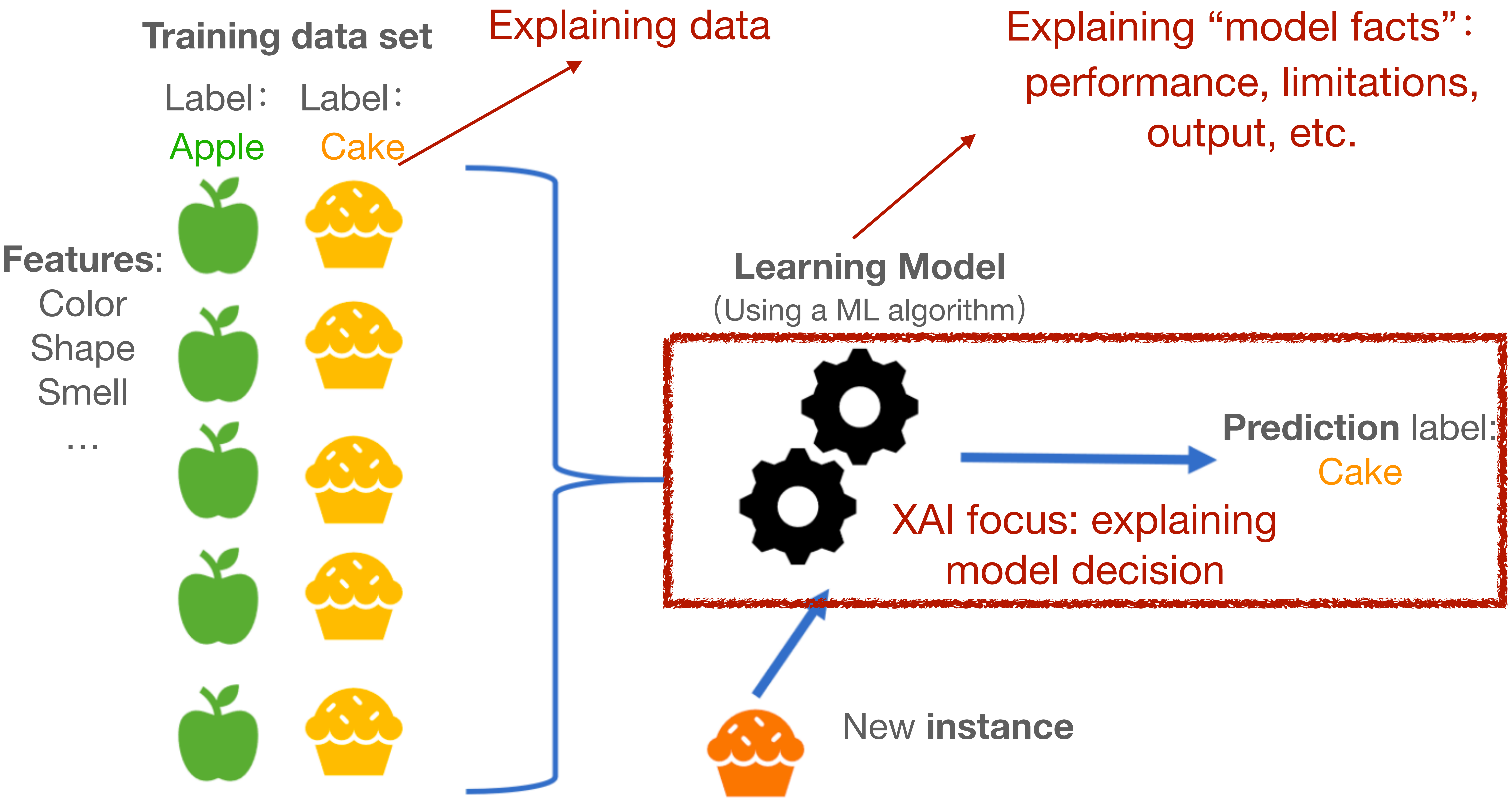
Information Visualization

eXplainable Artificial Intelligence

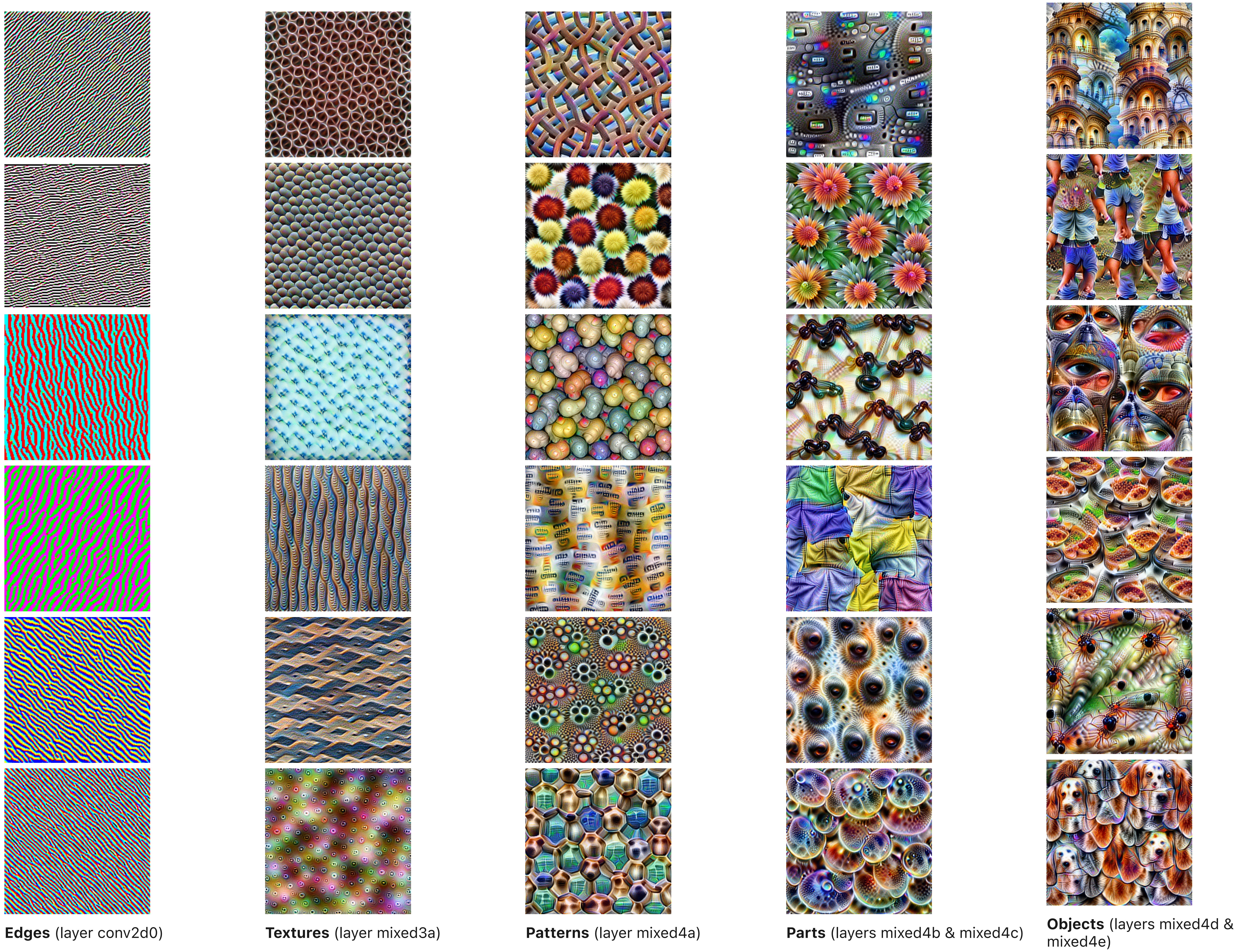
Dr. David Koop

eXplainable Artificial Intelligence



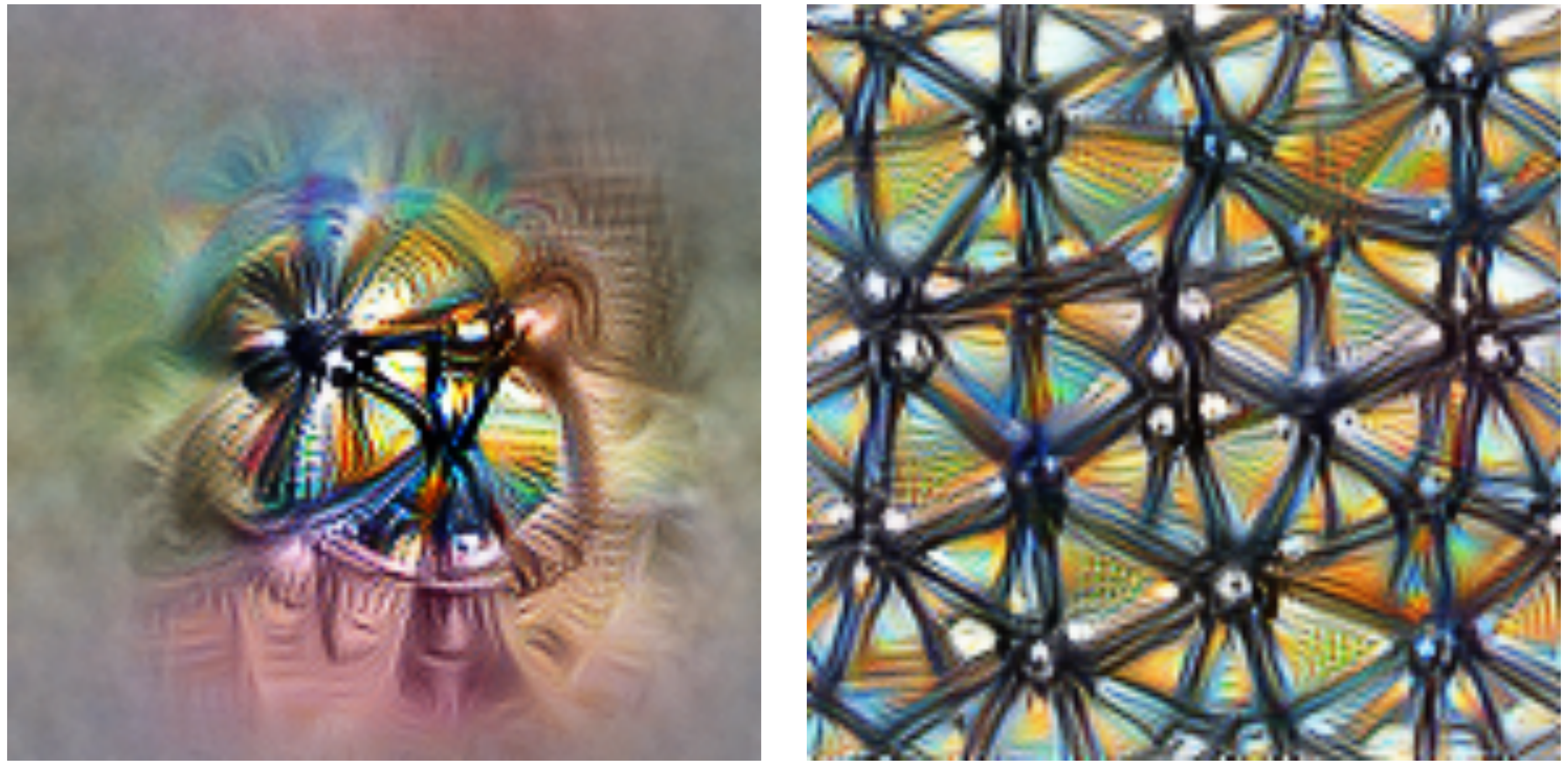
[Liao et al.]

Feature Visualization

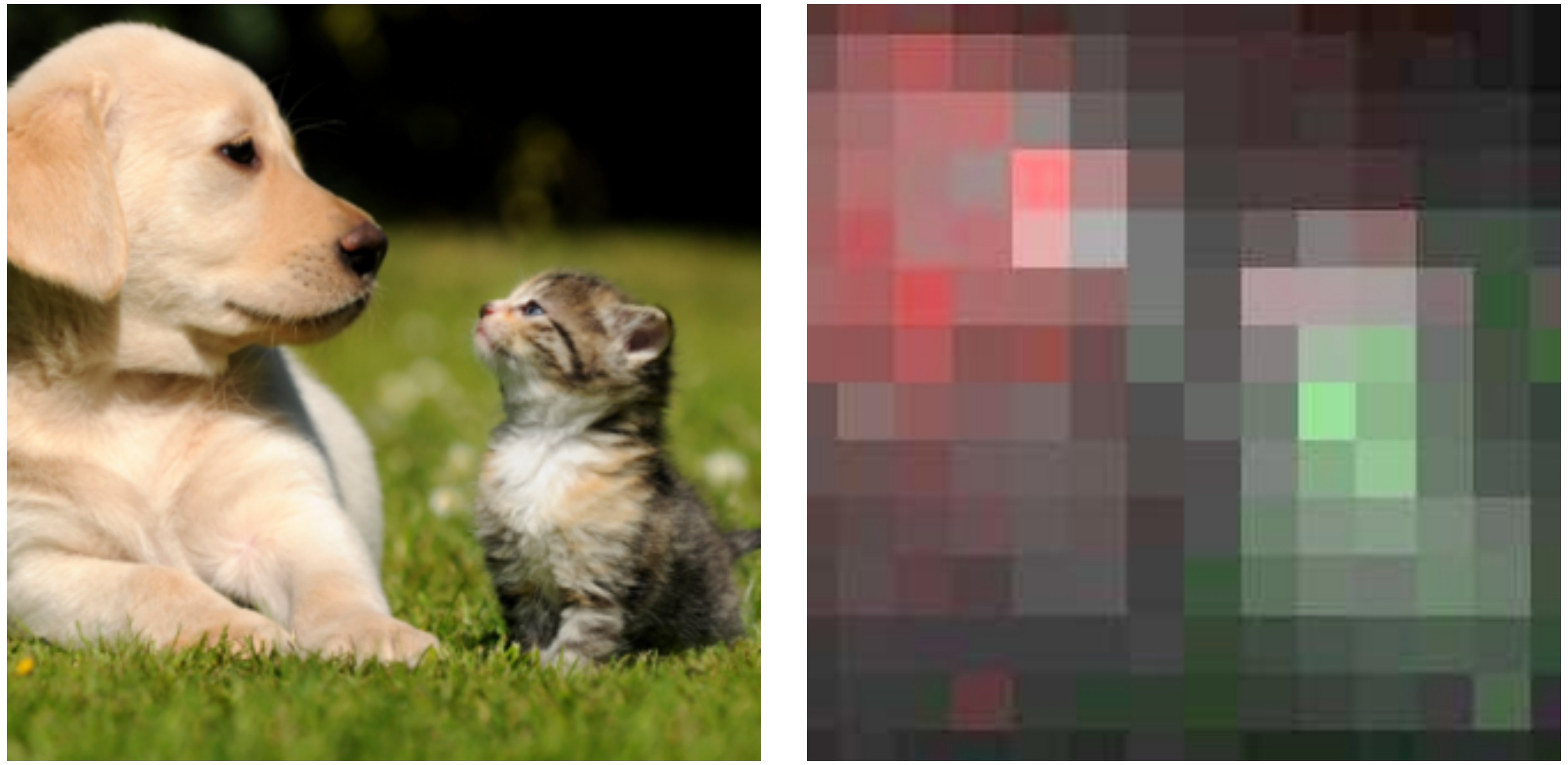


[C. Olah et al., 2017]

Feature Visualization vs. Attribution



Feature visualization answers questions about what a network — or parts of a network — are looking for by generating examples.

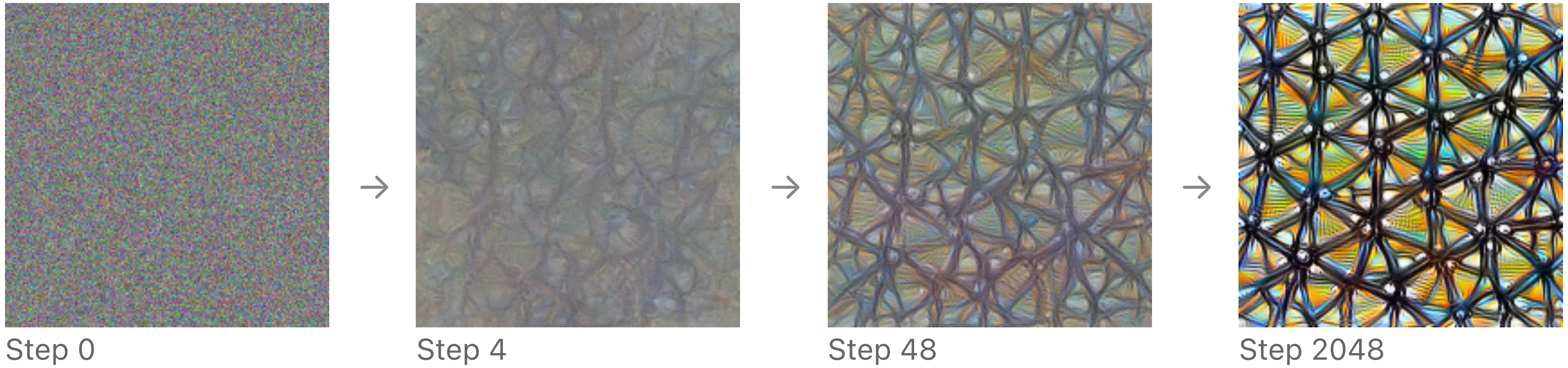


Attribution [1] studies what part of an example is responsible for the network activating a particular way.

[C. Olah et al., 2017]

Feature Vis by Optimization

- "[W]hat kind of input would cause a certain behavior"
- Start from random noise and iteratively tweak (using derivatives)



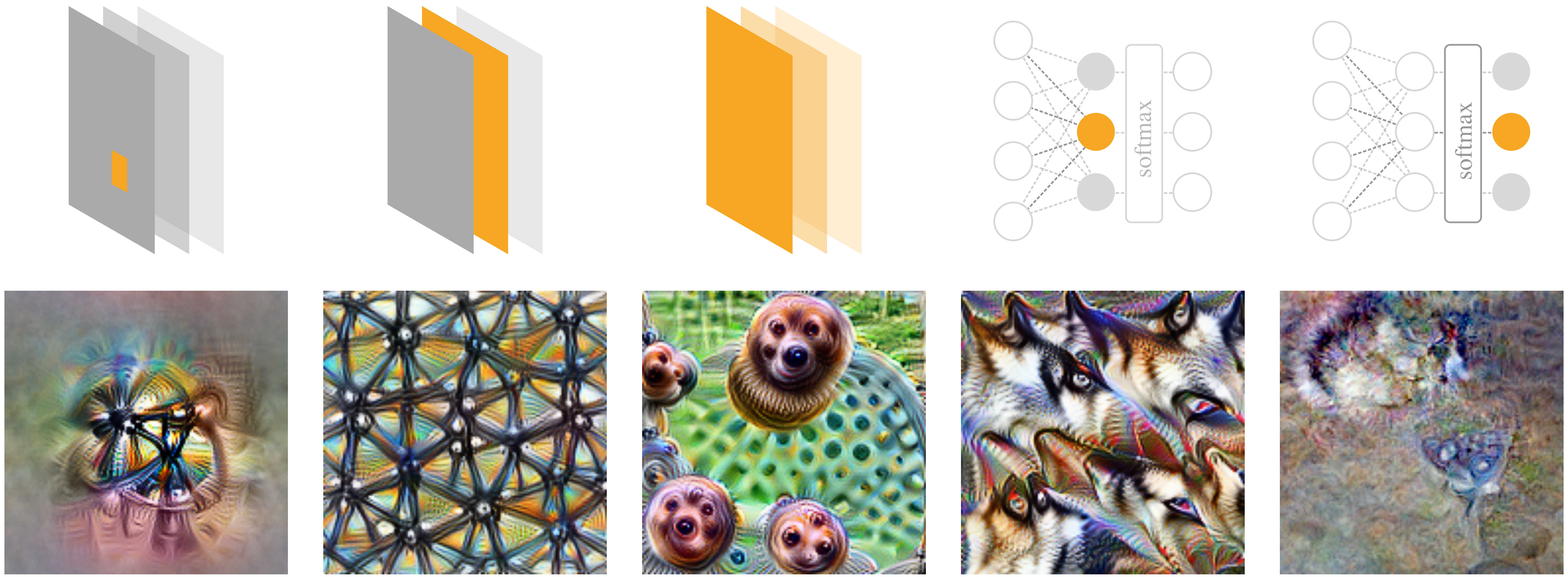
- What are the objectives? (Where are we going?)
  - Neuron, channel, layer (has DeepDream "interesting" objective

[C. Olah et al., 2017]

Optimization Objectives

Different **optimization objectives** show what different parts of a network are looking for.

`n` layer index
`x,y` spatial position
`z` channel index
`k` class index

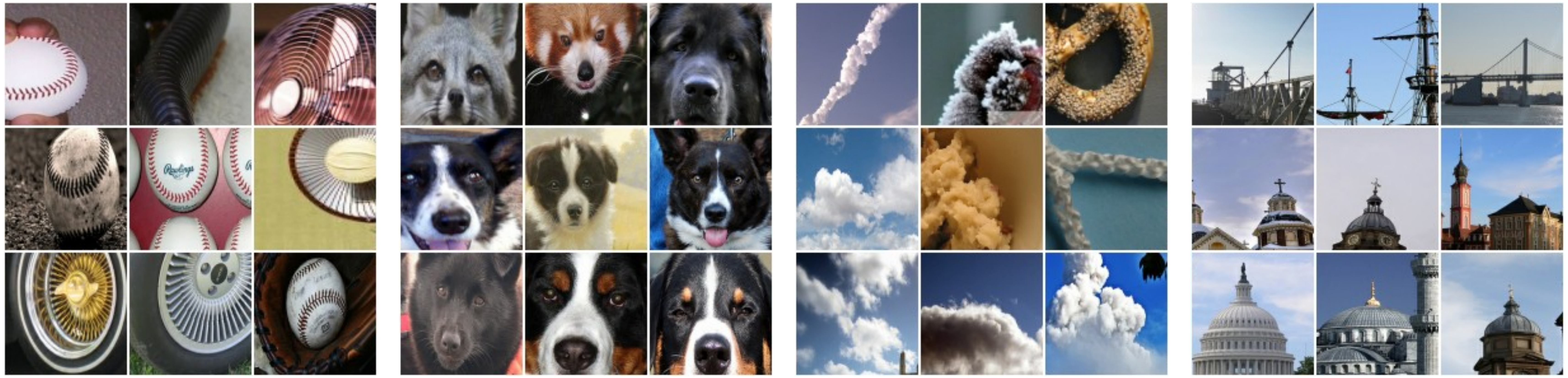


Neuron	**Channel**	**Layer**/DeepDream	**Class Logits**	**Class Probability**
$\texttt{layer}_n\texttt{[x,y,z]}$	$\texttt{layer}_n\texttt{[:,:,z]}$	$\texttt{layer}_n\texttt{[:,:,:]}^2$	`pre_softmax[k]`	`softmax[k]`

[C. Olah et al., 2017]

Why not Examples?

Dataset Examples show us what neurons respond to in practice



Optimization isolates the causes of behavior from mere correlations. A

neuron may not be detecting what you initially thought.

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

[C. Olah et al., 2017]

Diversity

- Examples can be diverse
- Optimization may give us one very positive takeaway
- Add a diversity term!

Simple Optimization

Optimization with diversity reveals multiple types of balls. *Layer mixed5a, Unit 9*

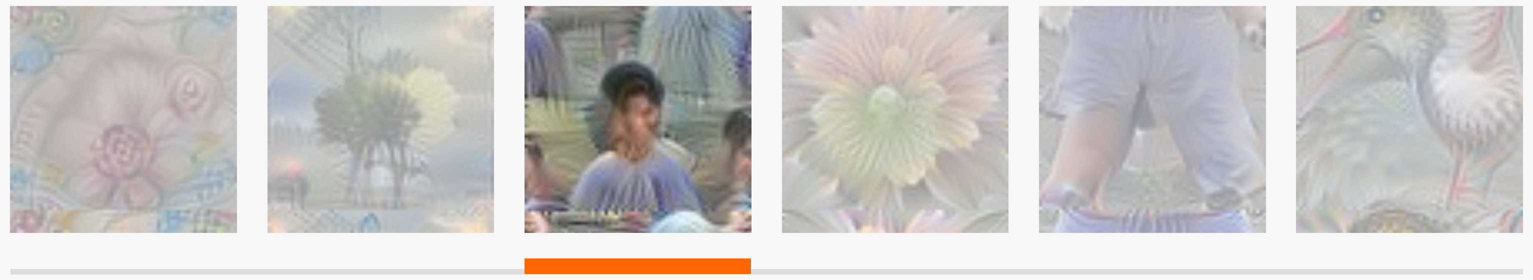
Dataset examples

[C. Olah et al., 2017]

Naive Optimization Doesn't Work

Even if you carefully tune learning rate, you'll get noise.

Optimization results are enlarged to show detail and artifacts.



Learning Rate (0.05)

REPRODUCE IN A CO NOTEBOOK

Step 1 Step 32 Step 128 Step 256 Step 2048

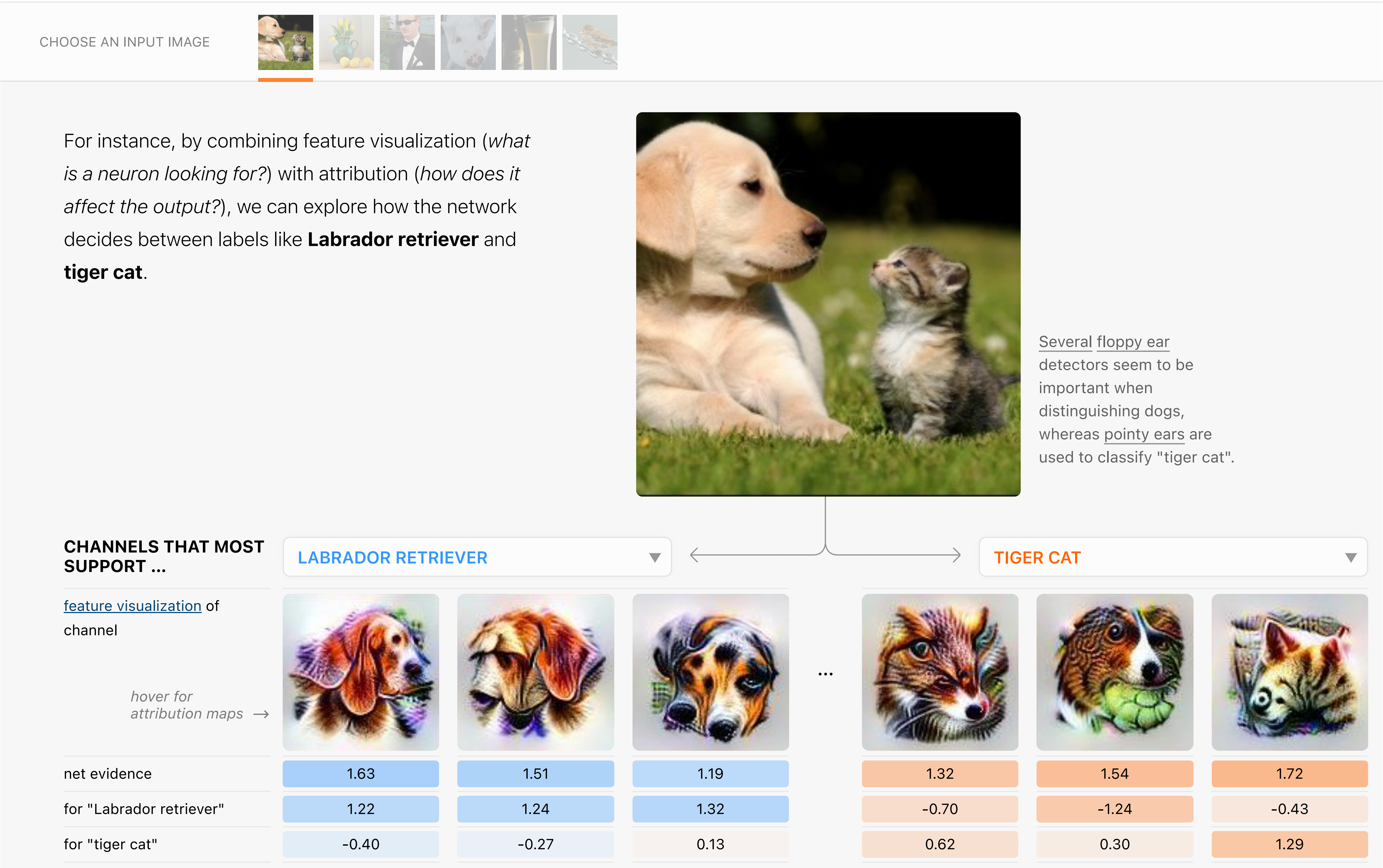
[C. Olah et al., 2017]

Regularization to Avoid Noise

- Frequency penalization: penalize high-frequency noise
- Transformation robustness: jitter/rotate/scale images and still activate
- Learned priors: learn a prior and try to enforce it

[C. Olah et al., 2017]

The Building Blocks of Interpretability



The Building Blocks of Interpretability

Interpretability techniques are normally studied in isolation.
We explore the powerful interfaces that arise when you combine them—
and the rich structure of this combinatorial space.

CHOOSE AN INPUT IMAGE

For instance, by combining feature visualization (*what is a neuron looking for?*) with attribution (*how does it affect the output?*), we can explore how the network decides between labels like **Labrador retriever** and **tiger cat**.

Several floppy ear detectors seem to be important when distinguishing dogs, whereas pointy ears are used to classify "tiger cat".

CHANNELS THAT MOST SUPPORT ... LABRADOR RETRIEVER / TIGER CAT

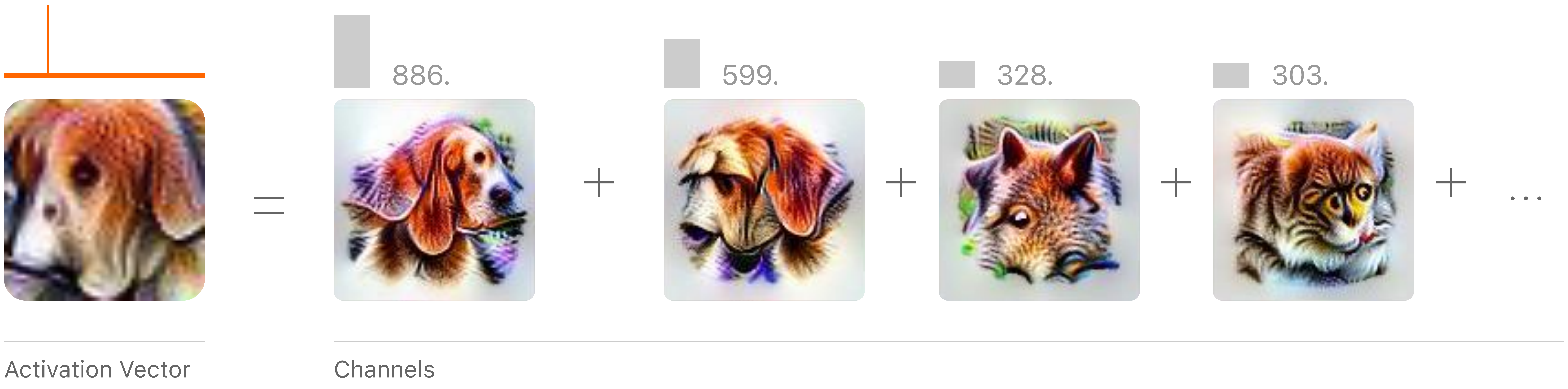
feature visualization of channel

hover for attribution maps →

				...			
net evidence	1.63	1.51	1.19		1.32	1.54	1.72
for "Labrador retriever"	1.22	1.24	1.32		-0.70	-1.24	-0.43
for "tiger cat"	-0.40	-0.27	0.13		0.62	0.30	1.29

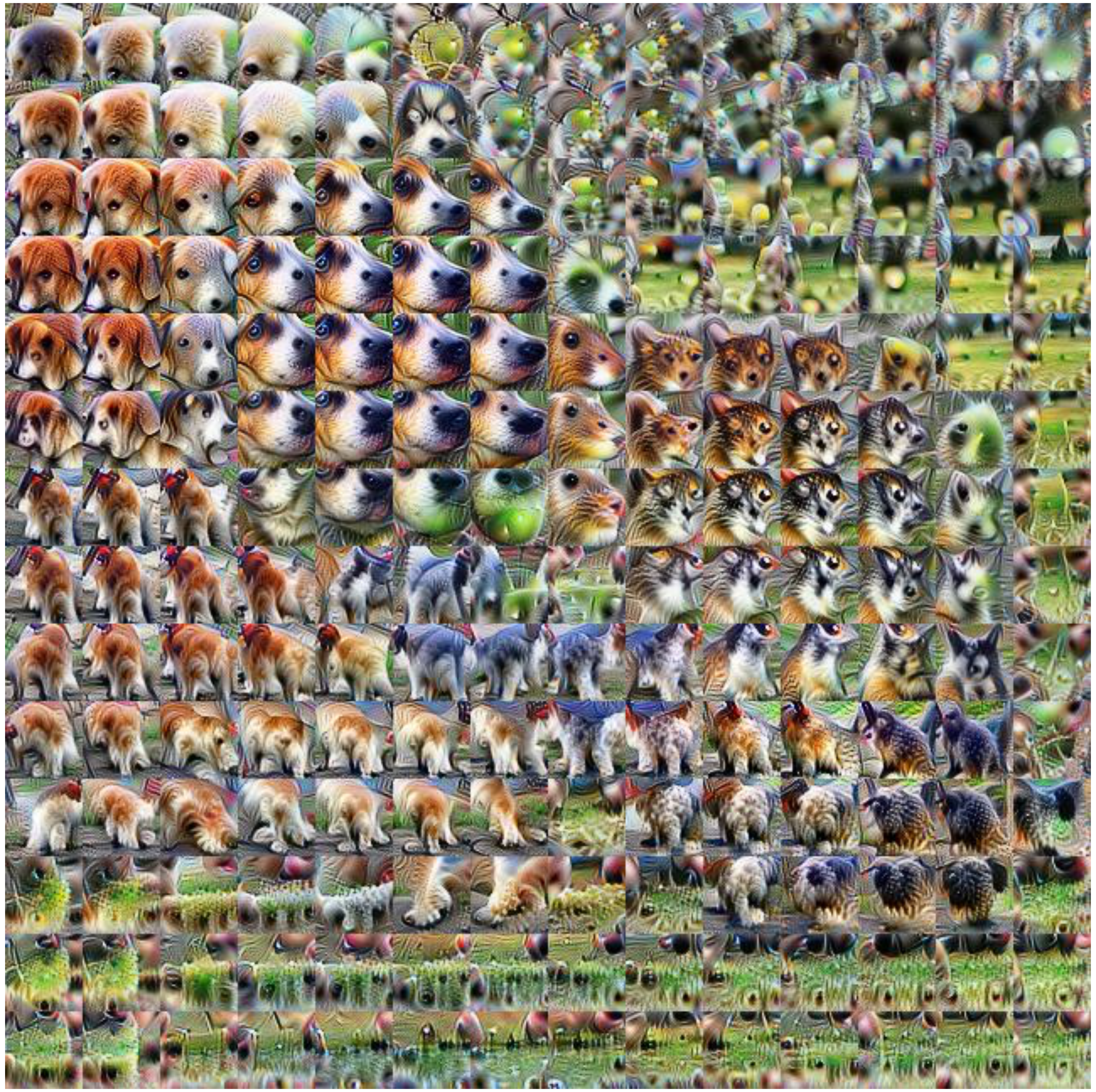
[C. Olah et al.]

Combine Feature Vis and Activation



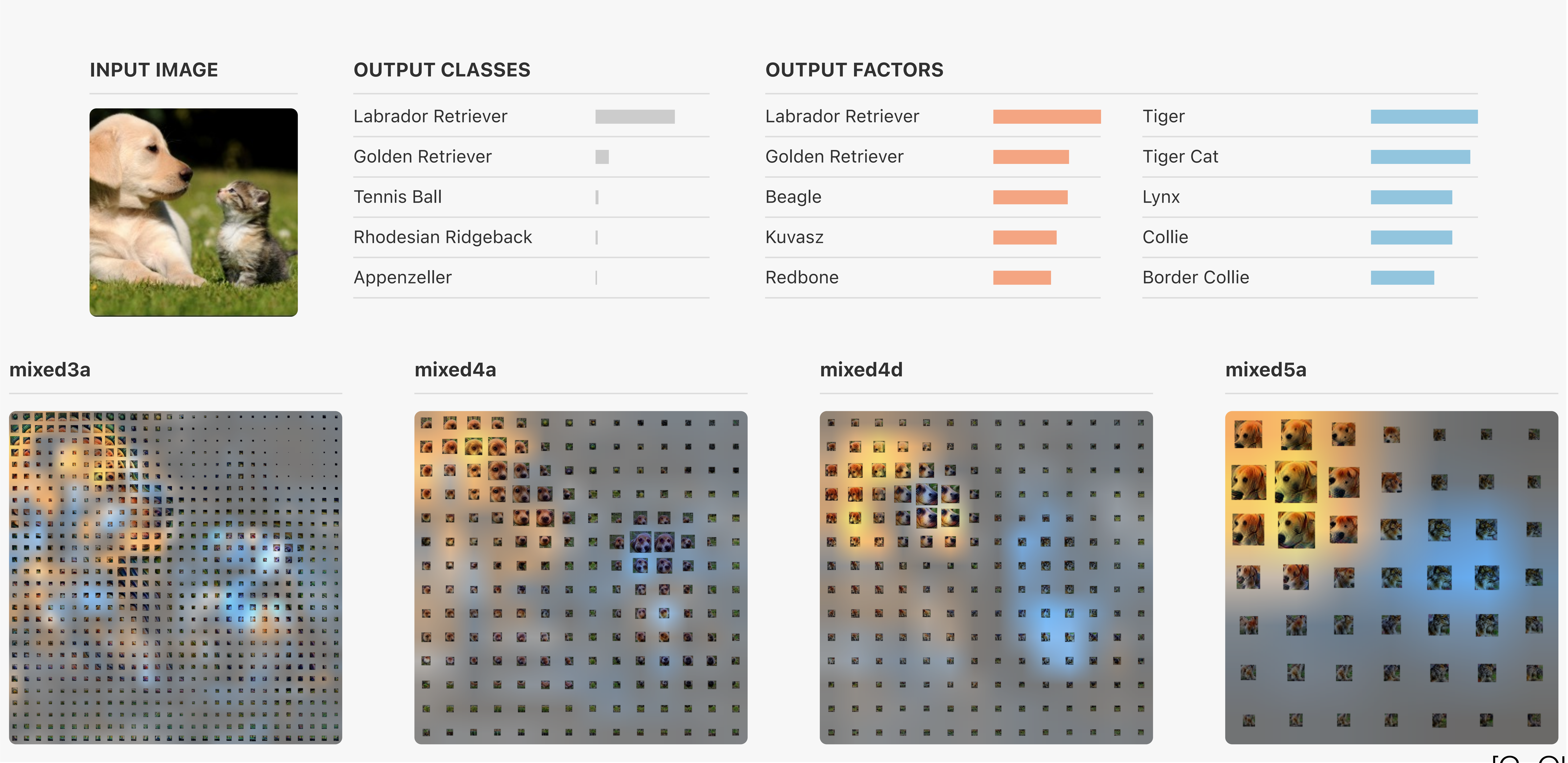
[C. Olah et al.]

Feature Vis + Attribution



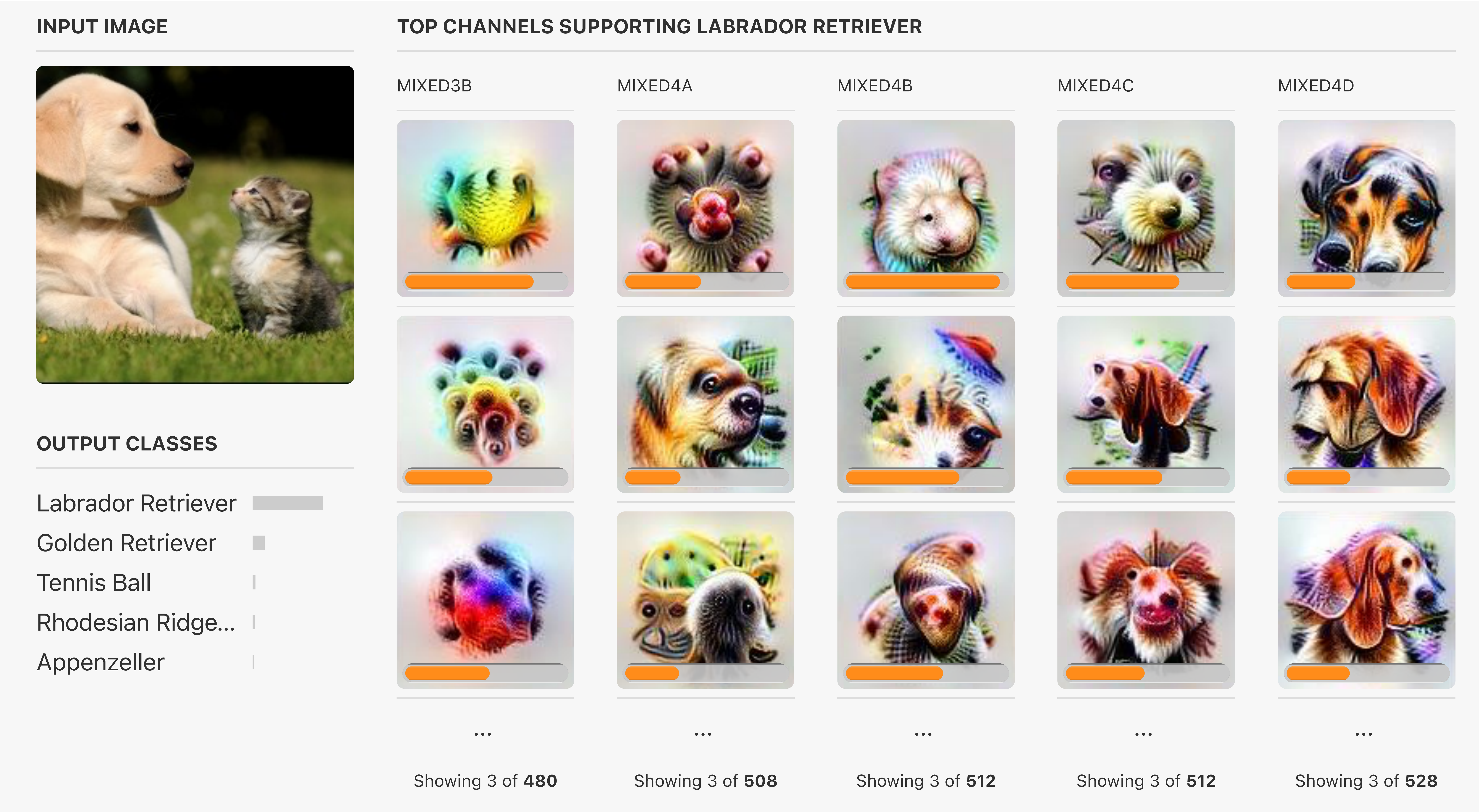
[C. Olah et al.]

Spatial Attribution with Saliency Maps



[C. Olah et al.]

Channel Attribution



[C. Olah et al.]

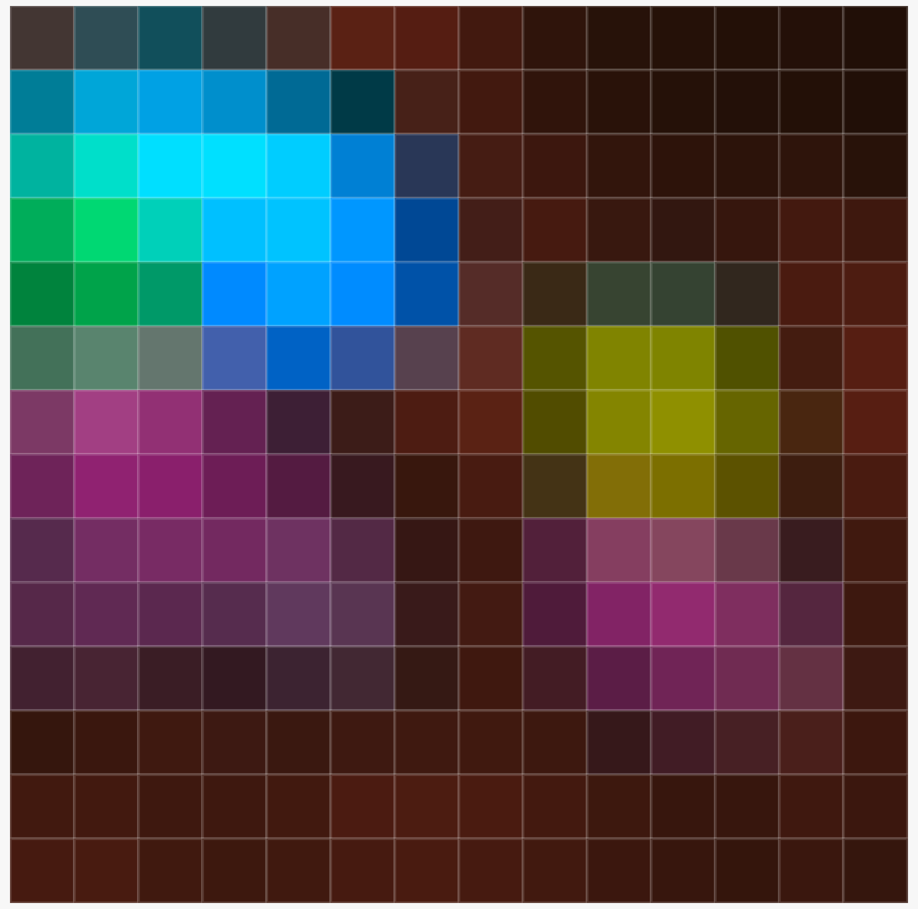
Factoring into Neuron Groups

By using non-negative matrix factorization we can reduce the large number of neurons to a small set of groups that concisely summarize the story of the network.

REPRODUCE IN A NOTEBOOK

INPUT IMAGE

ACTIVATIONS of neuron groups



NEURON GROUPS based on matrix factorization of mixed4d layer

6 groups

color key

feature visualization of each group

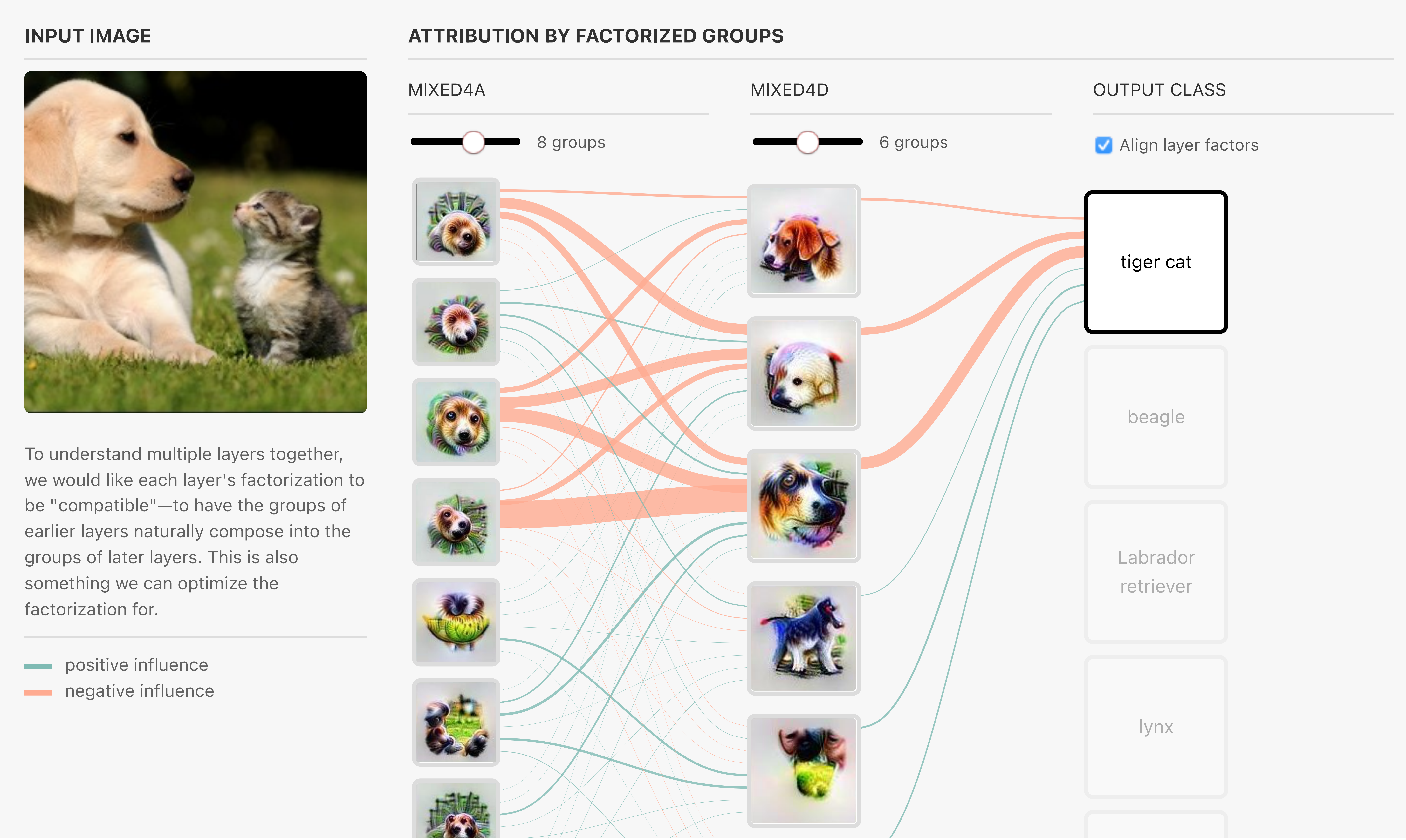
hover to isolate →

EFFECT of neuron groups on output classes

Labrador retriever	2.249	3.755	-1.193	-1.141	1.117	-1.892
beagle	3.298	0.599	-0.110	-0.356	-0.133	-2.618
tiger cat	-0.350	-0.994	-1.607	0.116	0.248	0.205
lynx	0.111	-0.642	-0.057	0.117	1.120	0.152
tennis ball	0.920	1.336	0.152	-0.885	1.227	-0.480

[C. Olah et al.]

Adding InfoVis



[C. Olah et al.]

Progress Reports