6.S979

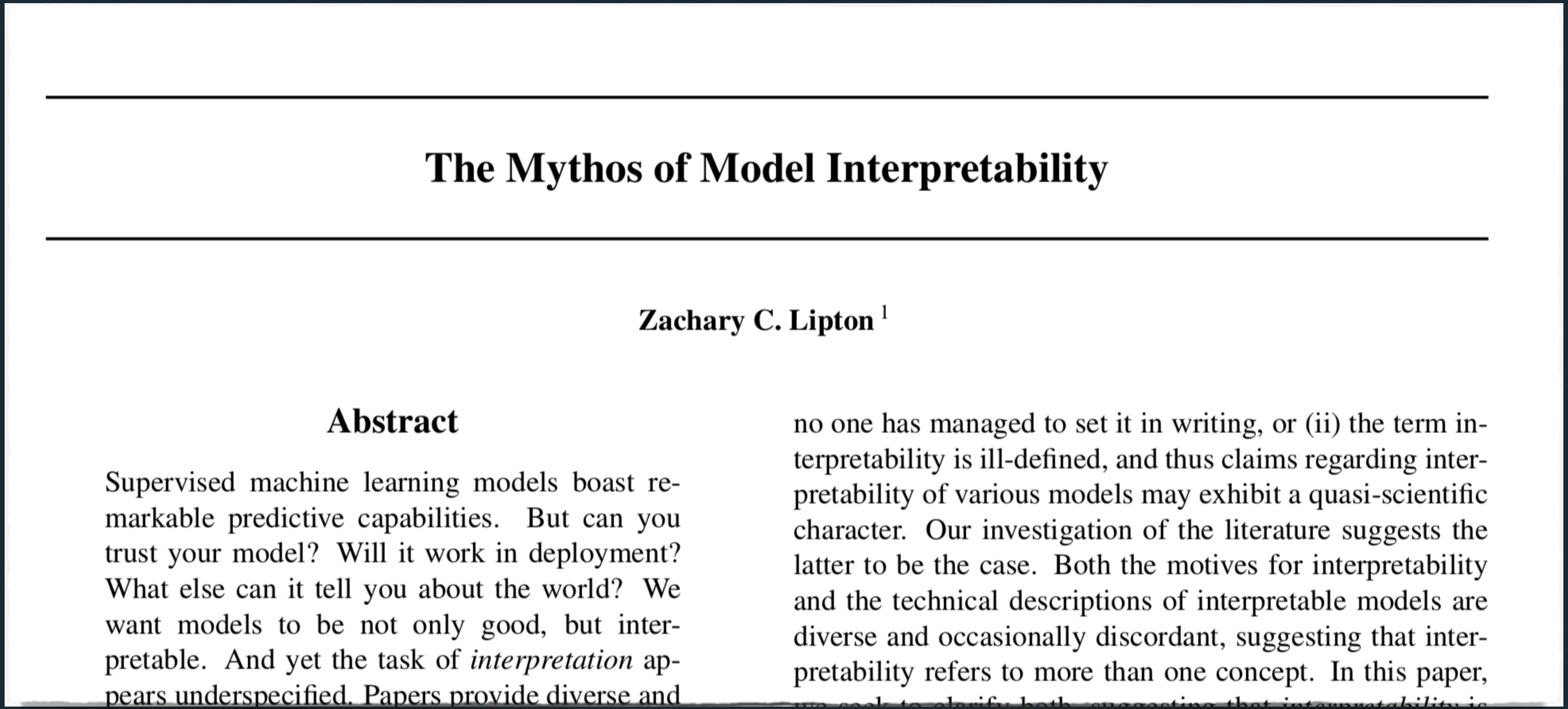
Model Interpretability

Arvind Satyanarayan

What is "Interpretability"?

Causal: the degree to which a person can understand the cause of the result.

Post-Hoc: what does the model tell me?



The Mythos of Model Interpretability

Zachary C. Lipton [1]

Abstract

Supervised machine learning models boast remarkable predictive capabilities. But can you trust your model? Will it work in deployment? What else can it tell you about the world? We want models to be not only good, but interpretable. And yet the task of *interpretation* appears underspecified. Papers provide diverse and

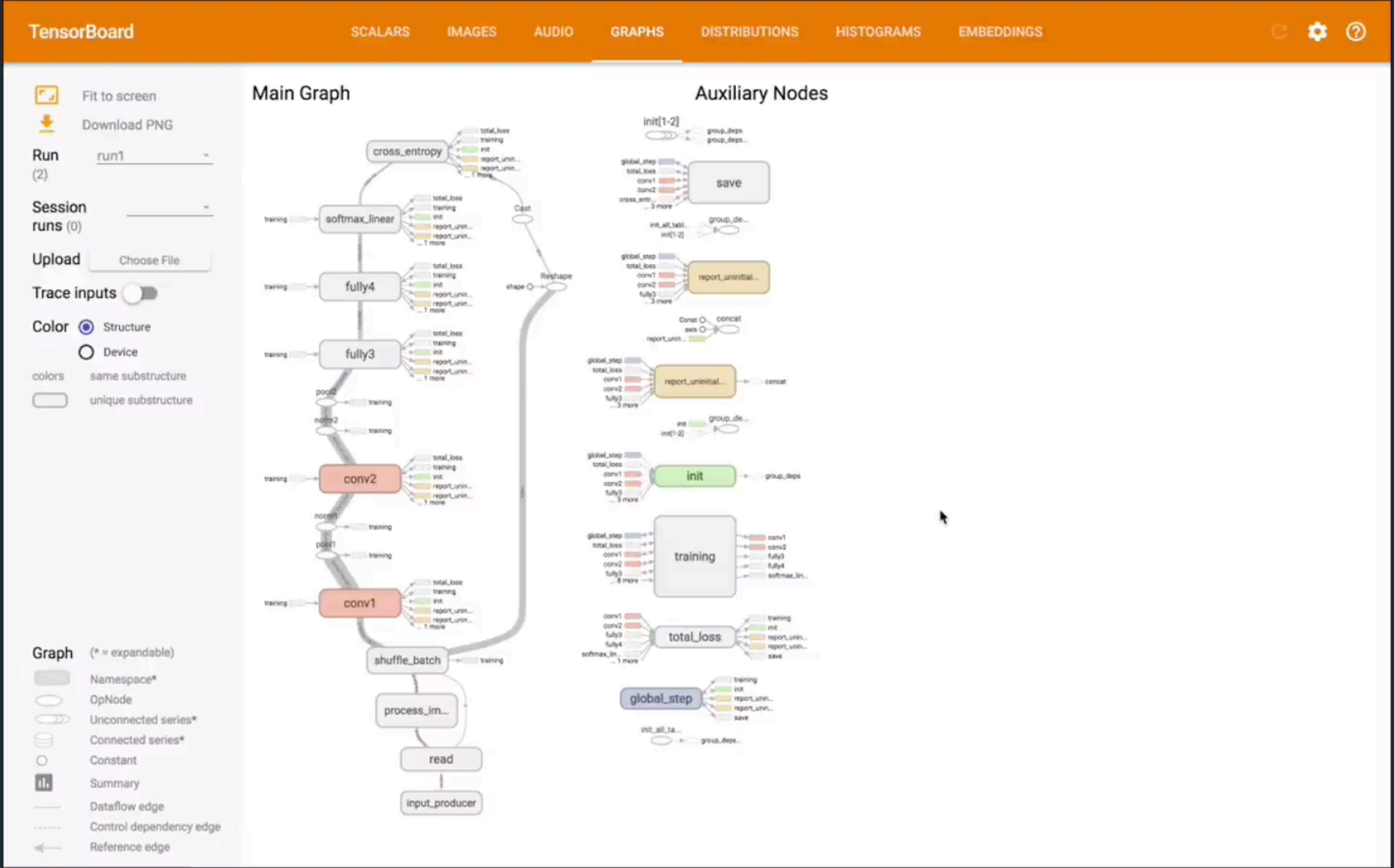
no one has managed to set it in writing, or (ii) the term interpretability is ill-defined, and thus claims regarding interpretability of various models may exhibit a quasi-scientific character. Our investigation of the literature suggests the latter to be the case. Both the motives for interpretability and the technical descriptions of interpretable models are diverse and occasionally discordant, suggesting that interpretability refers to more than one concept. In this paper,


ICE - CLEAN CYCLING COPY
Start
Session Features
Browse...
New...
Cycling Keywords
Cycling Clubs and Organizations
Cycling Lessons
Bike Safety and Maintenance
Cycling Tours and Events
motorcycle
Inactive Features
Shopping and Rentals
shop, store, rental, rentals, produ...
FEATURES: 6 ITEMS: 511
Find Items
Sample
0.75
Find Diverse
Find False Positives
Find False Negatives
Submit
Found items around score of 0.75.
http://www.tillamookrodeo.com
Tillamook County Rodeo
Welcome to the 26th Annual 2012 Tillamook County Rodeo
ROBERT WARREN TRUCKING
SATURDAY AND SUNDAY
0.25
http://www.70wildmiles.org/
70 Wild Miles
A charity triathlon; 70 miles, 1 day, 3 challenges
FERMANAGH
GLENCOE
BALA
0.75
Find Items
Featuring
Feature Suggestions
Review Labels
Standard Metrics
TEST TRAIN THRESHOLD: 0.0 1.0
41 223
38 209
Predicted
Positive Negative
Actual
Positive
Negative
34 7
11 27
TEST
185 38
54 155
TRAIN
TEST AUC
0.76 (+0.00)
TRAIN AUC
0.88 (+0.00)
Test Train
Precision-Recall Curves
Standard Metrics
ModelTracker [0.00, 1.00]
TEST AUC = 0.76 (+0.00) TRAIN AUC = 0.88 (+0.00)
Stanford
University

Test
Train
unicycling
triathlon
bike shop
motorcycles
blog
blog
2.84
3.97
4.52
4.12
92229
0.03795731
Similar Items Shown on Hover
Outliers
Changes from Previous Iteration
Interactive Prediction Threshold
Tags for Tracking

Dataset: mnist_randomforest.csv | 10 Classes, 5000 Instances, 5000 Shown | Acc.: 0.87, Prec.: 0.87, Recall: 0.87, TP: 436.0, FP: 64.0, TN: 4436.0, FN: 64.0 | 0 Hovered, 0 Selected

★	Image	TRUE	Assigned	Correct	Prediction Score	C0	C1	C2	C3	C4	C5	C6	C7
☆	7	C7	C7	1	0.5849	0.0103	0.0226	0.0646	0.0330	0.0621	0.0279	0.0383	0.5849
☆	1	C1	C1	1	0.3576	0.0111	0.3576	0.1180	0.0431	0.0450	0.0447	0.0589	0.0235
☆	7	C7	C7	1	0.6777	0.0185	0.0310	0.0485	0.0324	0.0328	0.0491	0.0152	0.6777
☆	9	C9	C9	1	0.2158	0.0146	0.0634	0.0233	0.1524	0.1435	0.1895	0.0541	0.0745
☆	5	C5	C5	1	0.2964	0.0346	0.2051	0.1039	0.1367	0.0286	0.2964	0.0606	0.0314
☆	3	C3	C3	1	0.2365	0.0121	0.0946	0.0367	0.2365	0.0617	0.1952	0.0599	0.1379
☆	2	C2	C2	1	0.3628	0.0069	0.1414	0.3628	0.1552	0.0153	0.0526	0.1246	0.0190
☆	1	C1	C1	1	0.8319	0.0017	0.8319	0.0342	0.0166	0.0087	0.0201	0.0123	0.0258
☆	2	C2	C2	1	0.7116	0.0192	0.0188	0.7116	0.0230	0.0390	0.0157	0.0597	0.0198
☆	1	C1	C1	1	0.7070	0.0014	0.7070	0.0384	0.0223	0.0234	0.0203	0.0194	0.0637
☆	7	C7	C7	1	0.7910	0.0081	0.0135	0.0240	0.0252	0.0417	0.0272	0.0079	0.7910
☆	1	C1	C1	1	0.8477	0.0014	0.8477	0.0253	0.0172	0.0068	0.0230	0.0165	0.0215
☆	6	C6	C6	1	0.3483	0.0364	0.0256	0.0358	0.0728	0.1373	0.1377	0.3483	0.0257
☆	8	C8	C8	1	0.3779	0.0115	0.1461	0.0314	0.0668	0.0879	0.0825	0.0701	0.0486
☆	2	C2	C2	1	0.3196	0.0638	0.0291	0.3196	0.1187	0.0546	0.0869	0.1316	0.0617
☆	0	C0	C0	1	0.9639	0.9639	0.0003	0.0020	0.0026	0.0008	0.0198	0.0053	0.0016
☆	4	C4	C4	1	0.5391	0.0044	0.0179	0.0161	0.0156	0.5391	0.0369	0.0449	0.1140


TensorBoard
SCALARS
IMAGES
AUDIO
GRAPHS
DISTRIBUTIONS
HISTOGRAMS
EMBEDDINGS
Fit to screen
Download PNG
Run
(2)
run1
Session runs (0)
Upload
Choose File
Trace inputs
Color
Structure
Device
colors
same substructure
unique substructure
Graph
(* = expandable)
Namespace*
OpNode
Unconnected series*
Connected series*
Constant
Summary
Dataflow edge
Control dependency edge
Reference edge
Main Graph
cross_entropy
softmax_linear
fully4
fully3
conv2
conv1
shuffle_batch
process_im...
read
input_producer
Cast
Reshape
Auxiliary Nodes
init[1-2]
save
report_uninitial...
concat
init
training
total_loss
global_step


FBLearner
ActiVis: Visualization of Deep Neural Networks #15782570
COMPUTATION GRAPH VISUALIZATION
Concat
Tanh
Dropout
CNN_1/Dropout:0
Operator node
Blob node
Blob node w/ activation
Concat
FC
fc_0
Relu
INSTANCE SELECTION
Left column shows correctly classified instances.
Right column shows misclassified instances, with border colors indicating predicted classes.
DESC
ENTY
ABBR
HUM
NUM
LOC
From
FC
Selected
fc_0
To
Relu
NEURON ACTIVATION
Each row represents a group of instances. Each column is a neuron. Columns sorted by activation strength for Neuron idx
By class
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22
DESC
ENTY
ABBR
HUM
NUM
LOC
By user-defined filters
Contain 'Where'
Contain 'located'
Contain 'How many'
Contain 'How'
By instance ID
PROJECTED

Seq2Seq-Vis: A visual debugging tool for sequence-to-sequence models

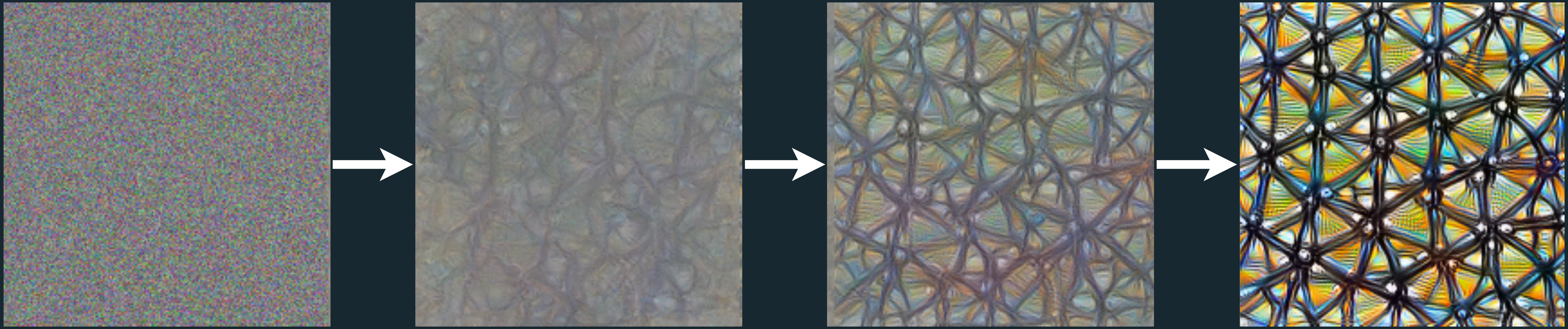
H. Strobelt, S. Gehrmann, M. Behrisch, A. Perer, H. Pfister, A. Rush

IBM Research

HARVARD
School of Engineering
and Applied Sciences

Feature Visualization

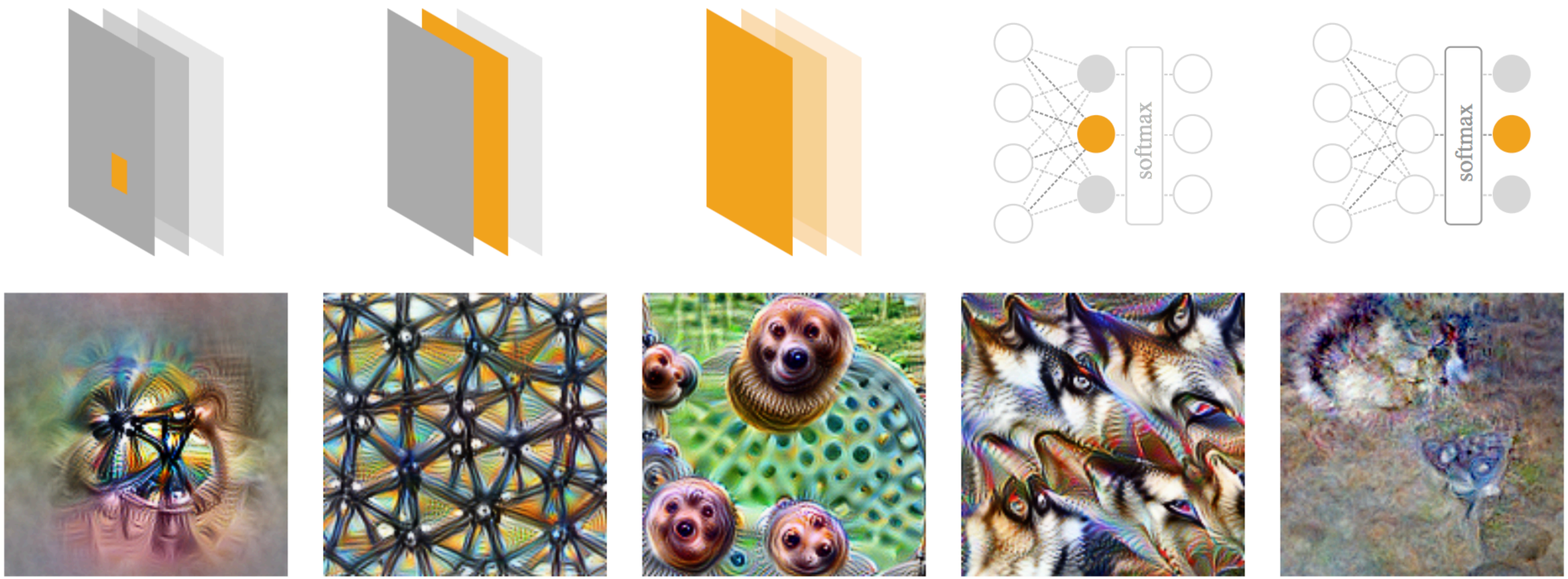
Olah, Mordvintsev, and Schubert. Distill, 2017.
https://distill.pub/2017/feature-visualization/



Step 0 Step 4 Step 48 Step 2,048

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]`

Simple Optimization

Optimization with diversity reveals four different, curvy facets. *Layer mixed4a, Unit 97*

Dataset examples

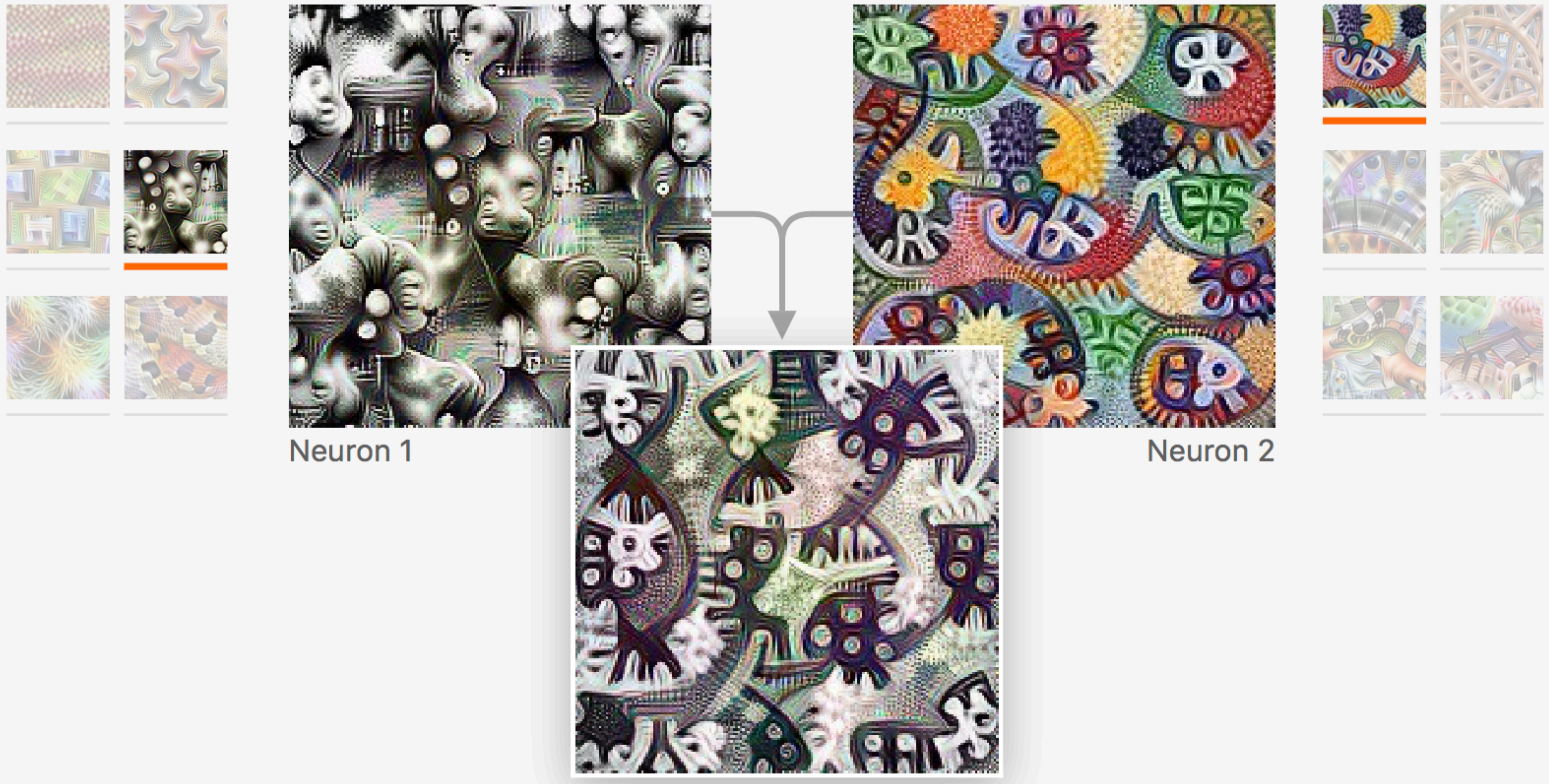
Simple Optimization

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

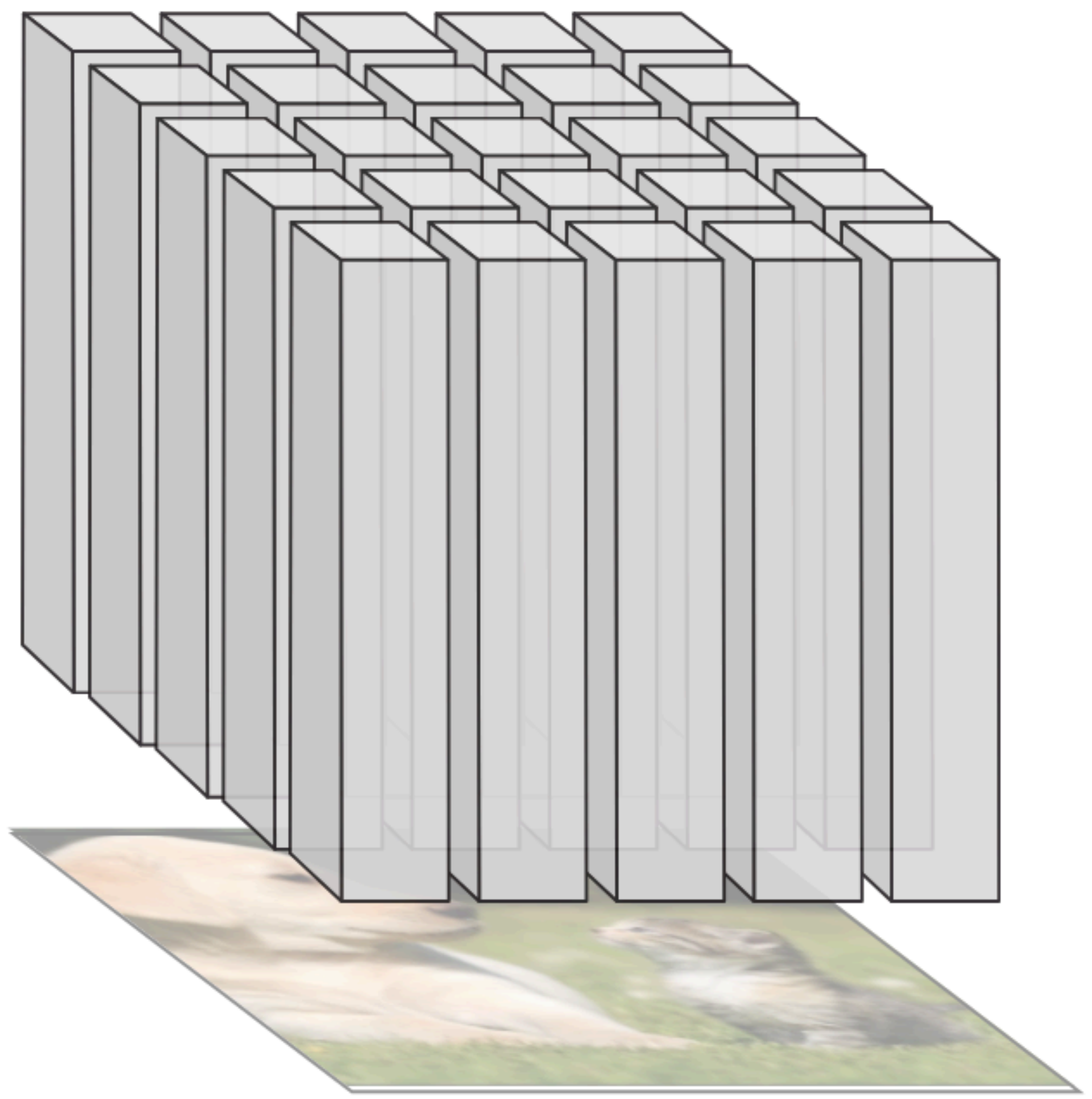
Dataset examples

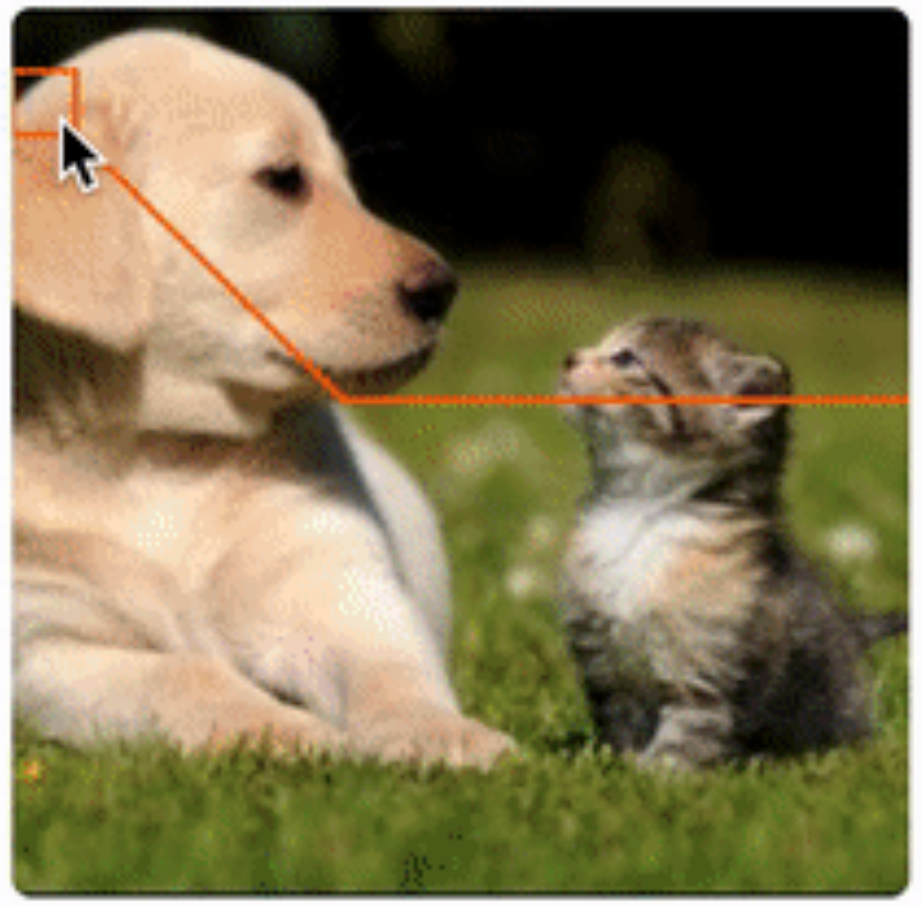
By jointly optimizing two neurons we can get a sense of how they interact.

REPRODUCE IN A NOTEBOOK



Spatial Activations

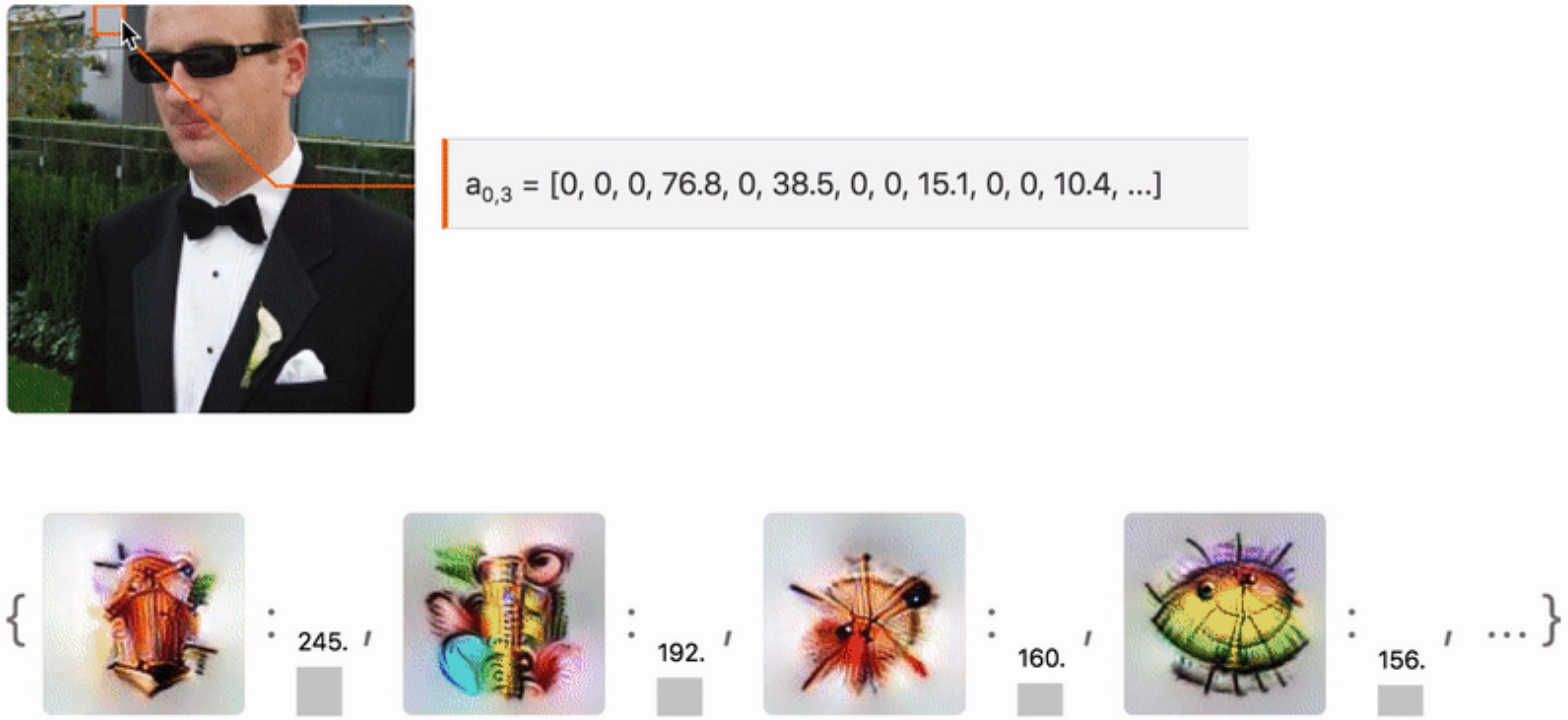




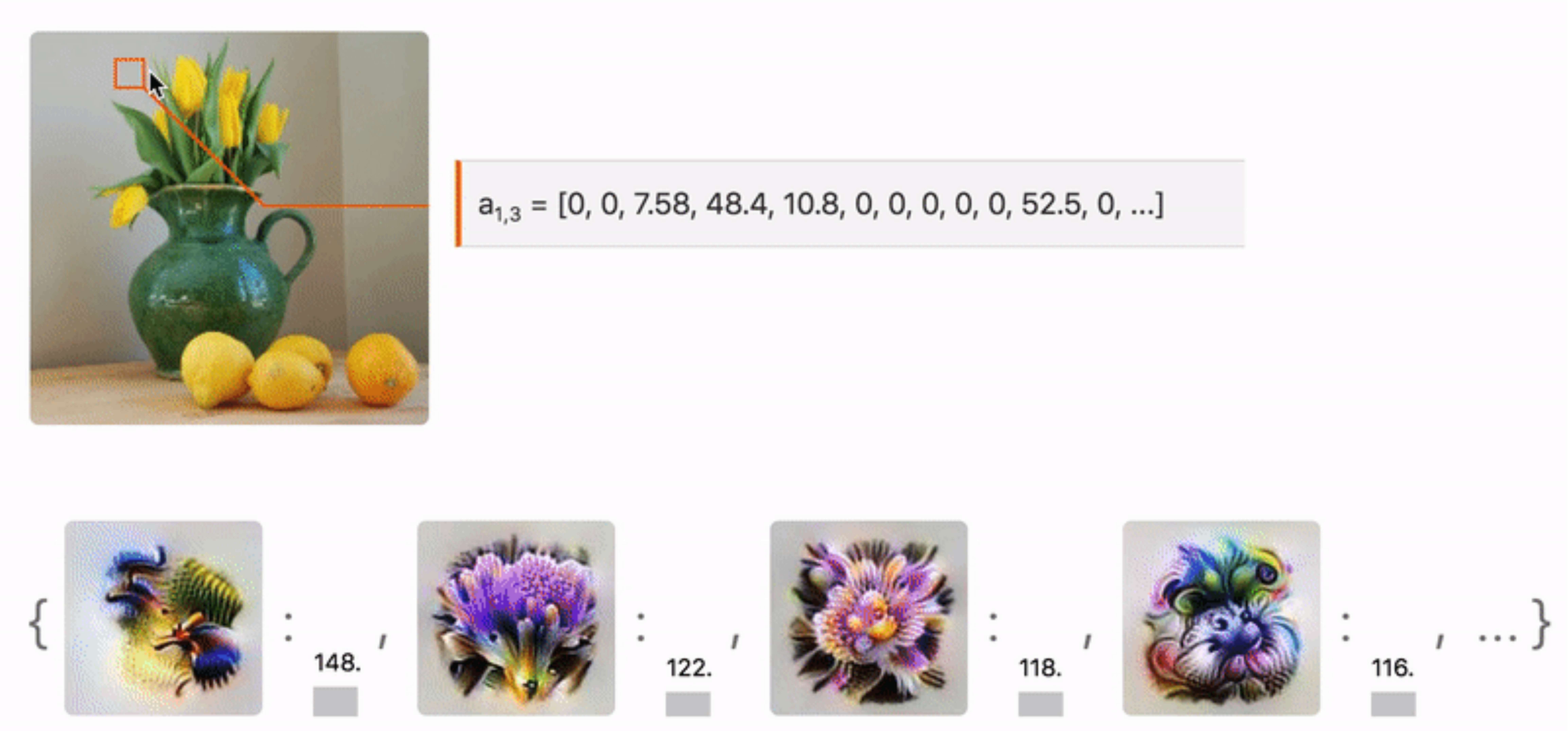
$a_{1,0} = [0, 0, 0, 0, 49.6, 0, 43.6, 30.2, 119.8, 62.7, 0, 51...$

Semantic Dictionaries

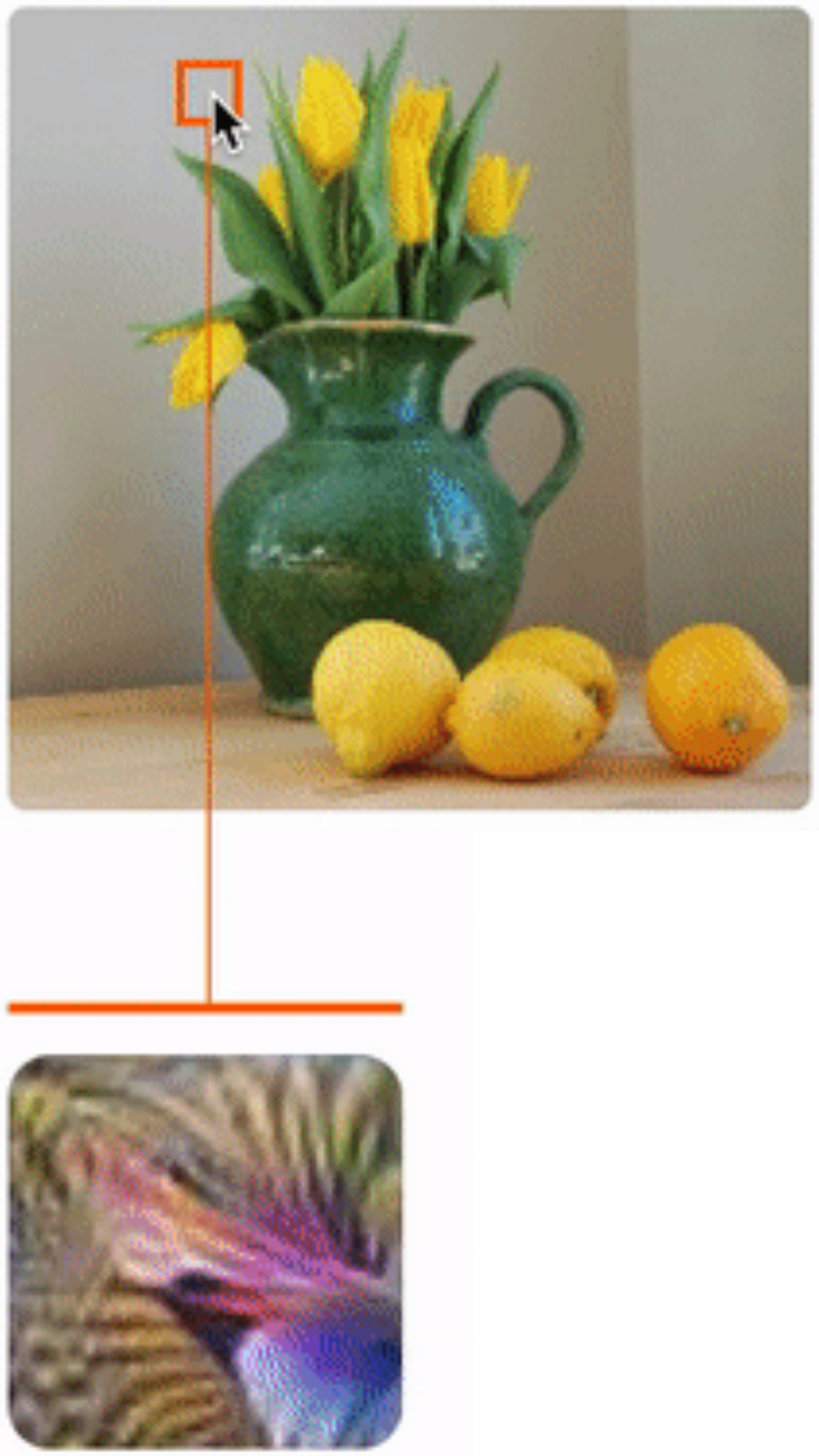
Semantic Dictionaries



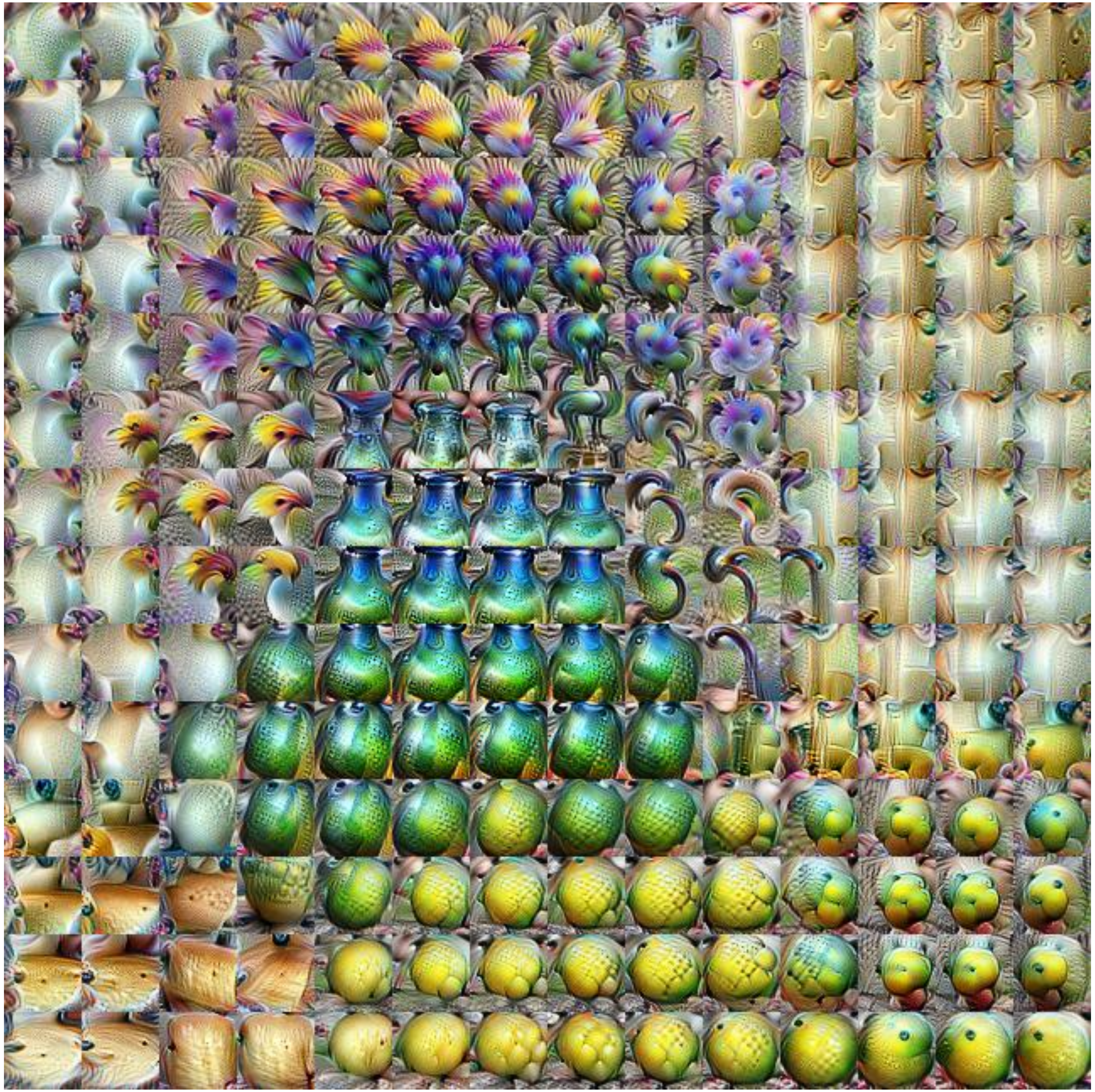
Semantic Dictionaries

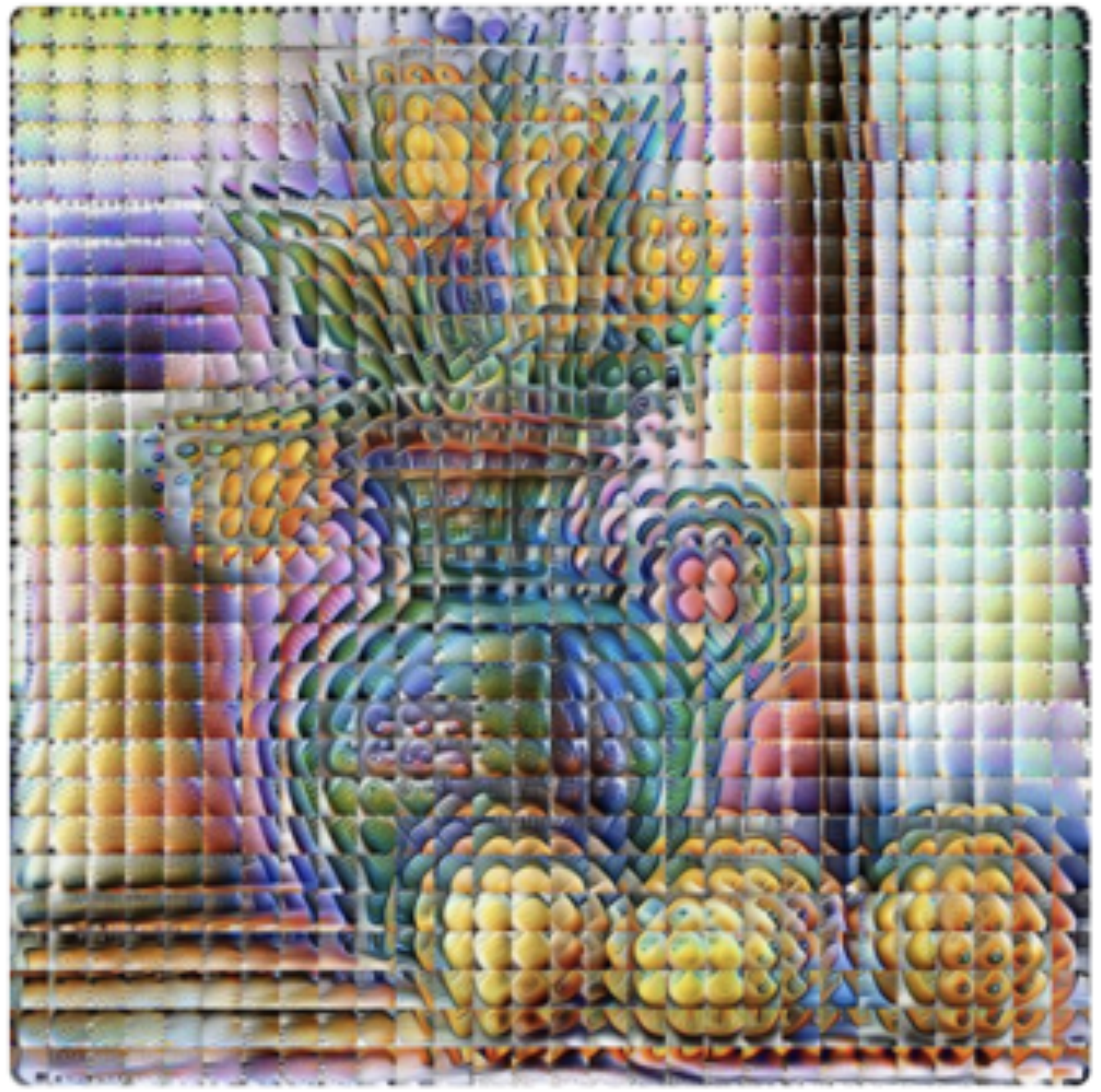



148.
122.
118.
116.
Activation Vector
Channels

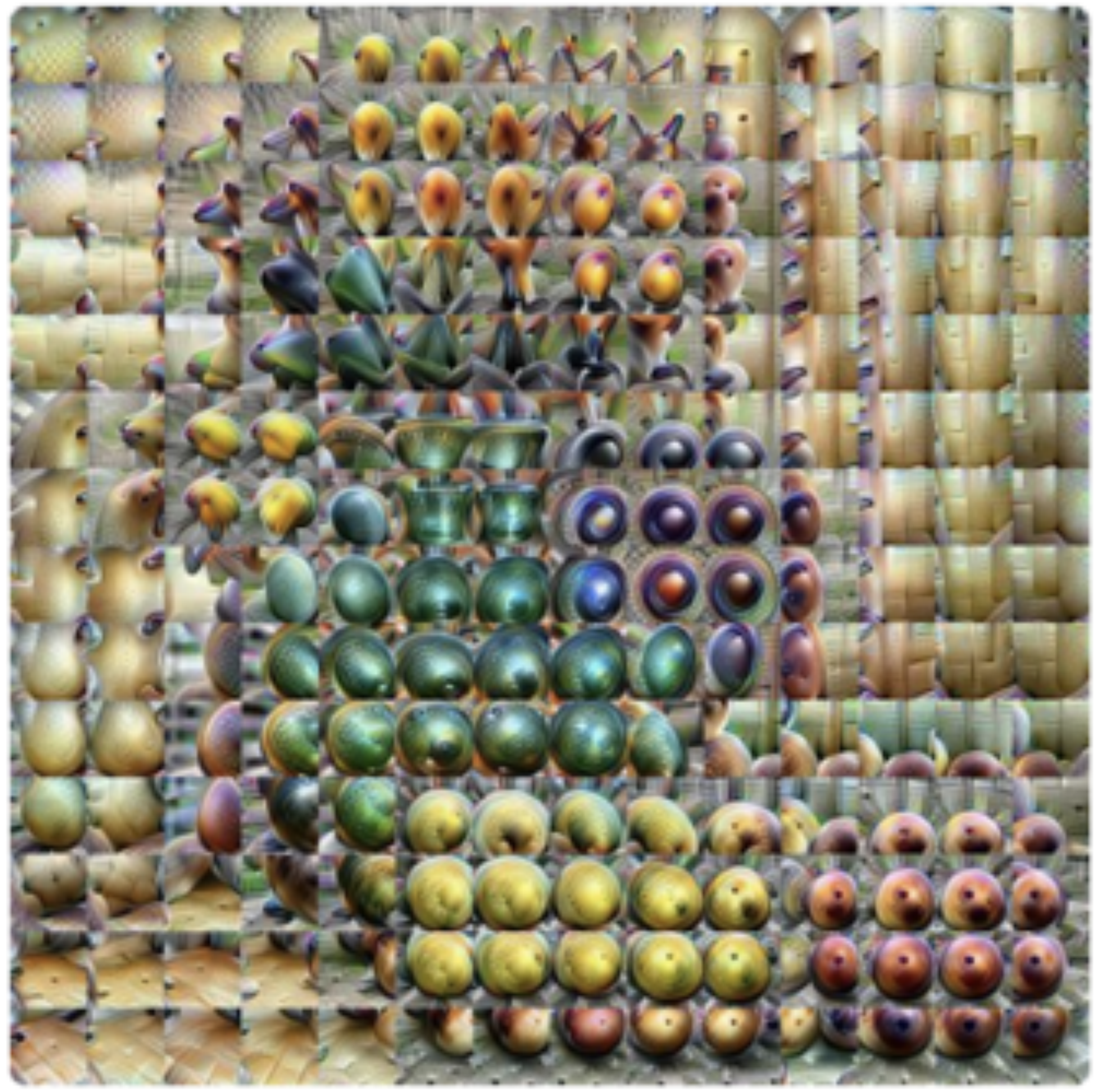


Activation Vector

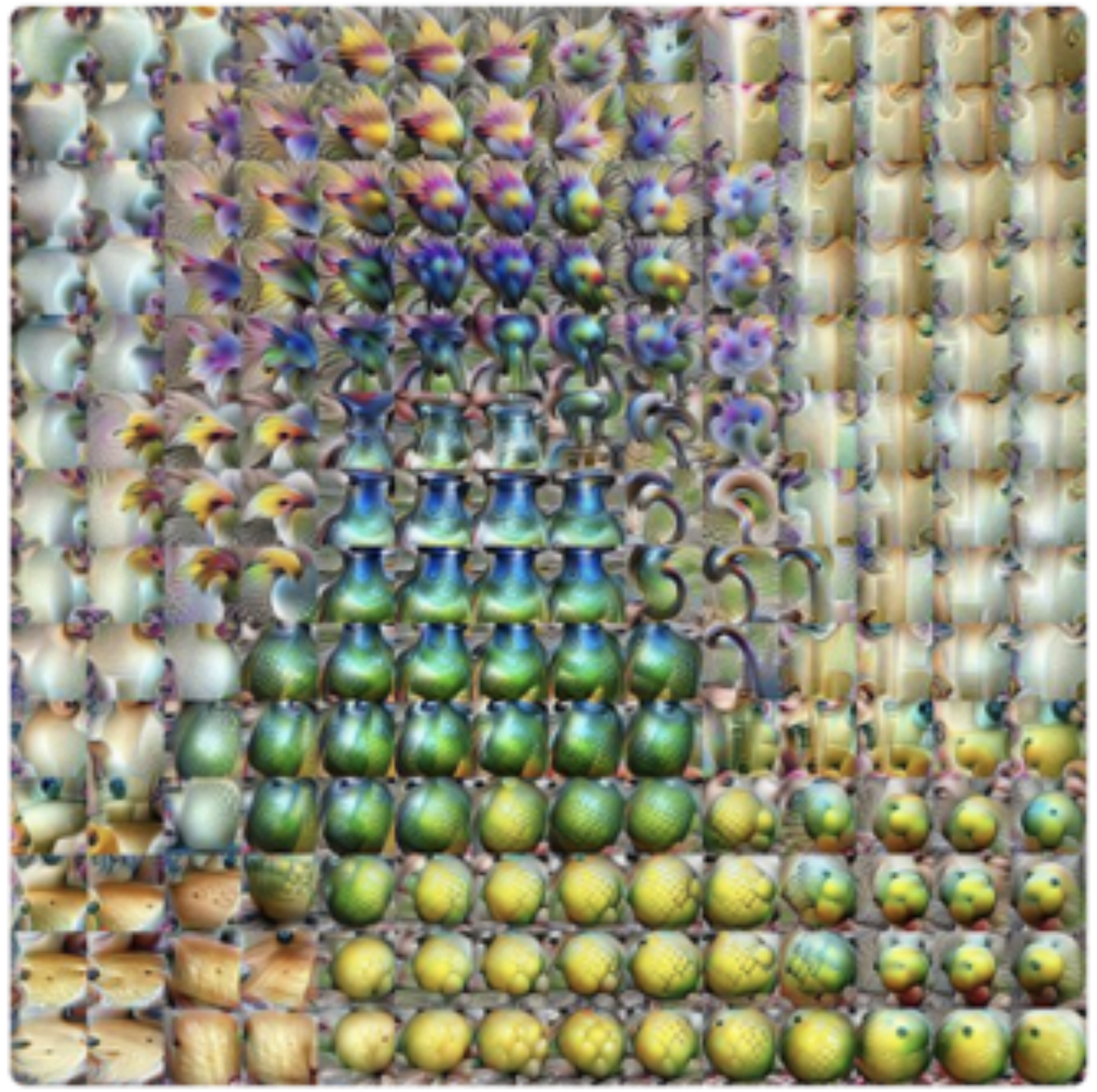




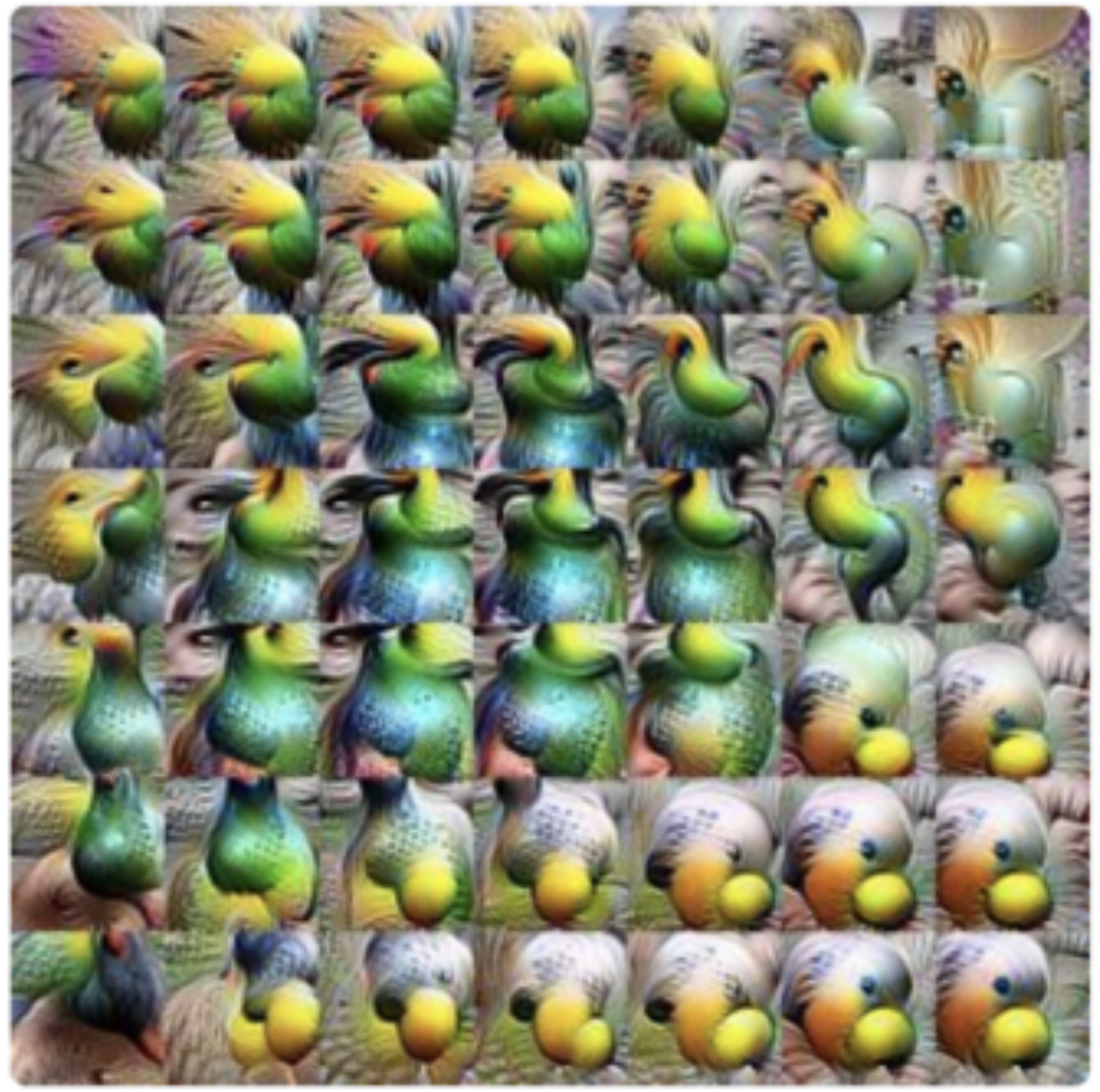
MIXED3A



MIXED4A



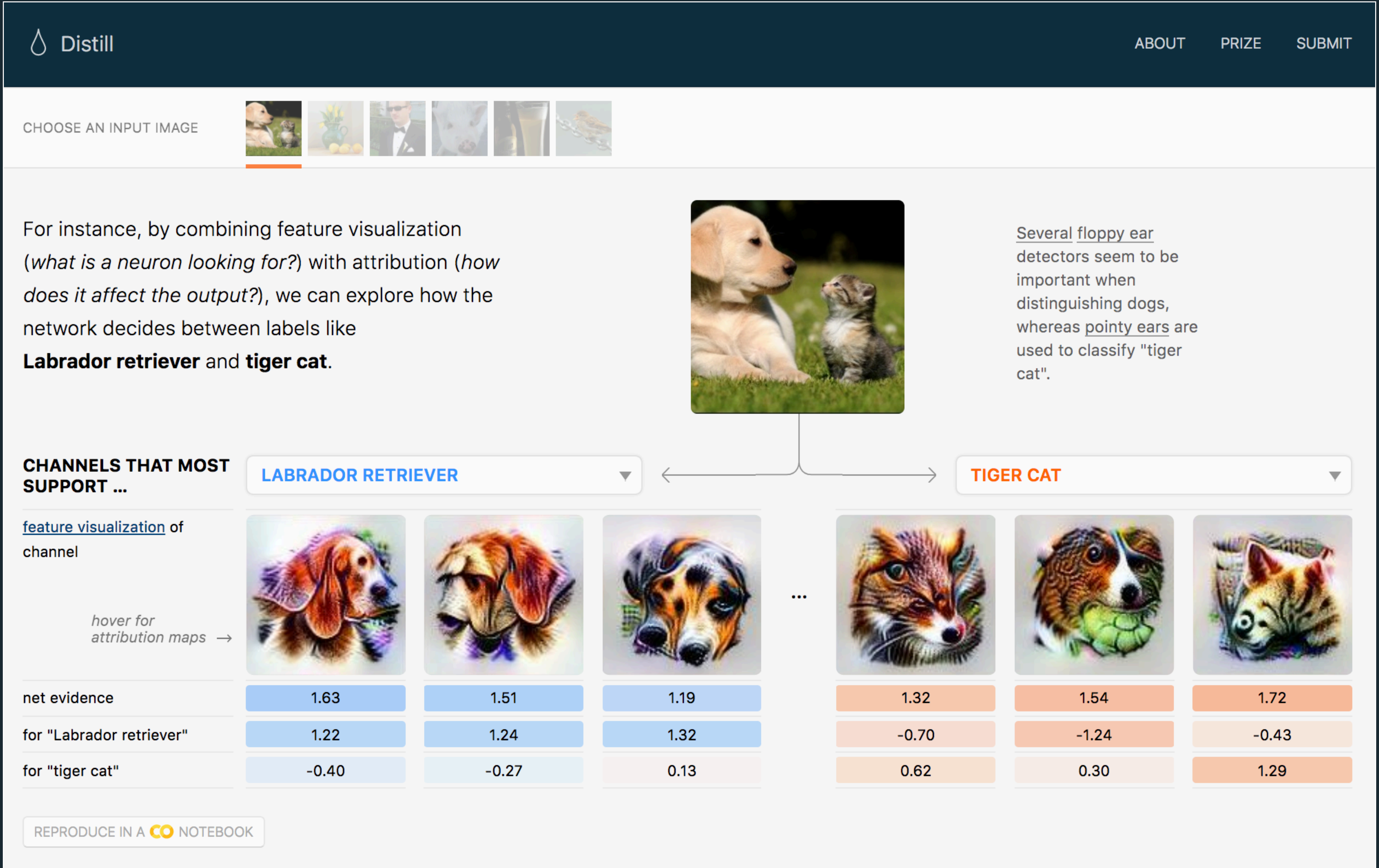
MIXED4D



MIXED5A

The Building Blocks of Interpretability

Olah, Satyanarayan, et al. Distill, 2018.
https://distill.pub/2018/building-blocks/



Summit

Scaling Deep Learning Interpretability by Visualizing Activation and Attribution Summarizations

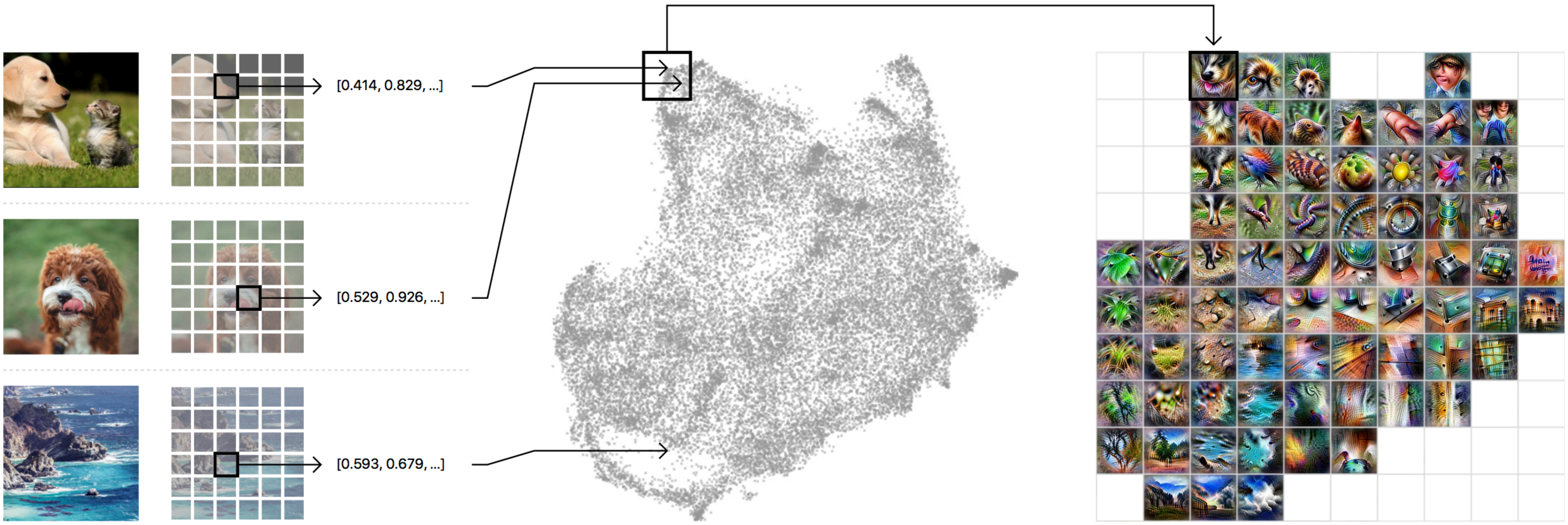
Fred Hohman, Haekyu Park, Caleb Robinson, Duen Horng (Polo) Chau

IEEE VIS 2019
Vancouver, Canada

Georgia Tech

Exploring Neural Networks with Activation Atlases

By using feature inversion to visualize millions of activations from an image classification network, we create an explorable *activation atlas* of features the network has learned which can reveal how the network typically represents some concepts.

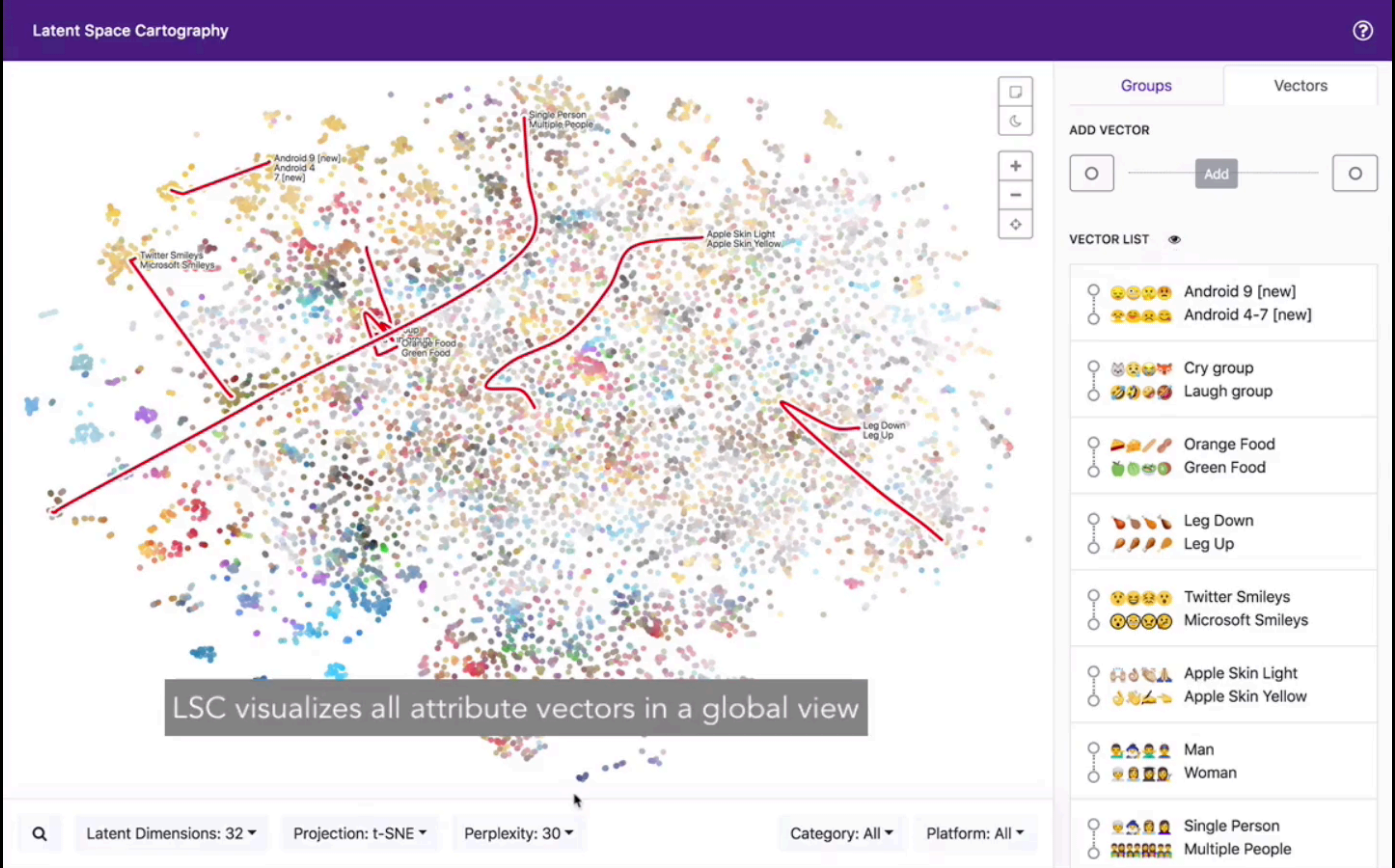


A randomized set of one million images is fed through the network, collecting one random spatial activation per image.

The activations are fed through UMAP to reduce them to two dimensions. They are then plotted, with similar activations placed near each other.

We then draw a grid and average the activations that fall within a cell and run feature inversion on the averaged activation. We also optionally size the grid cells according to the density of the number of activations that are averaged within.

TRY IN A NOTEBOOK


Latent Space Cartography
Single Person
Multiple People
Android 9 [new]
Android 4
7 [new]
Apple Skin Light
Apple Skin Yellow
Twitter Smileys
Microsoft Smileys
Orange Food
Green Food
Leg Down
Leg Up
LSC visualizes all attribute vectors in a global view
Groups
Vectors
ADD VECTOR
Add
VECTOR LIST
Android 9 [new]
Android 4-7 [new]
Cry group
Laugh group
Orange Food
Green Food
Leg Down
Leg Up
Twitter Smileys
Microsoft Smileys
Apple Skin Light
Apple Skin Yellow
Man
Woman
Single Person
Multiple People
Latent Dimensions: 32
Projection: t-SNE
Perplexity: 30
Category: All
Platform: All