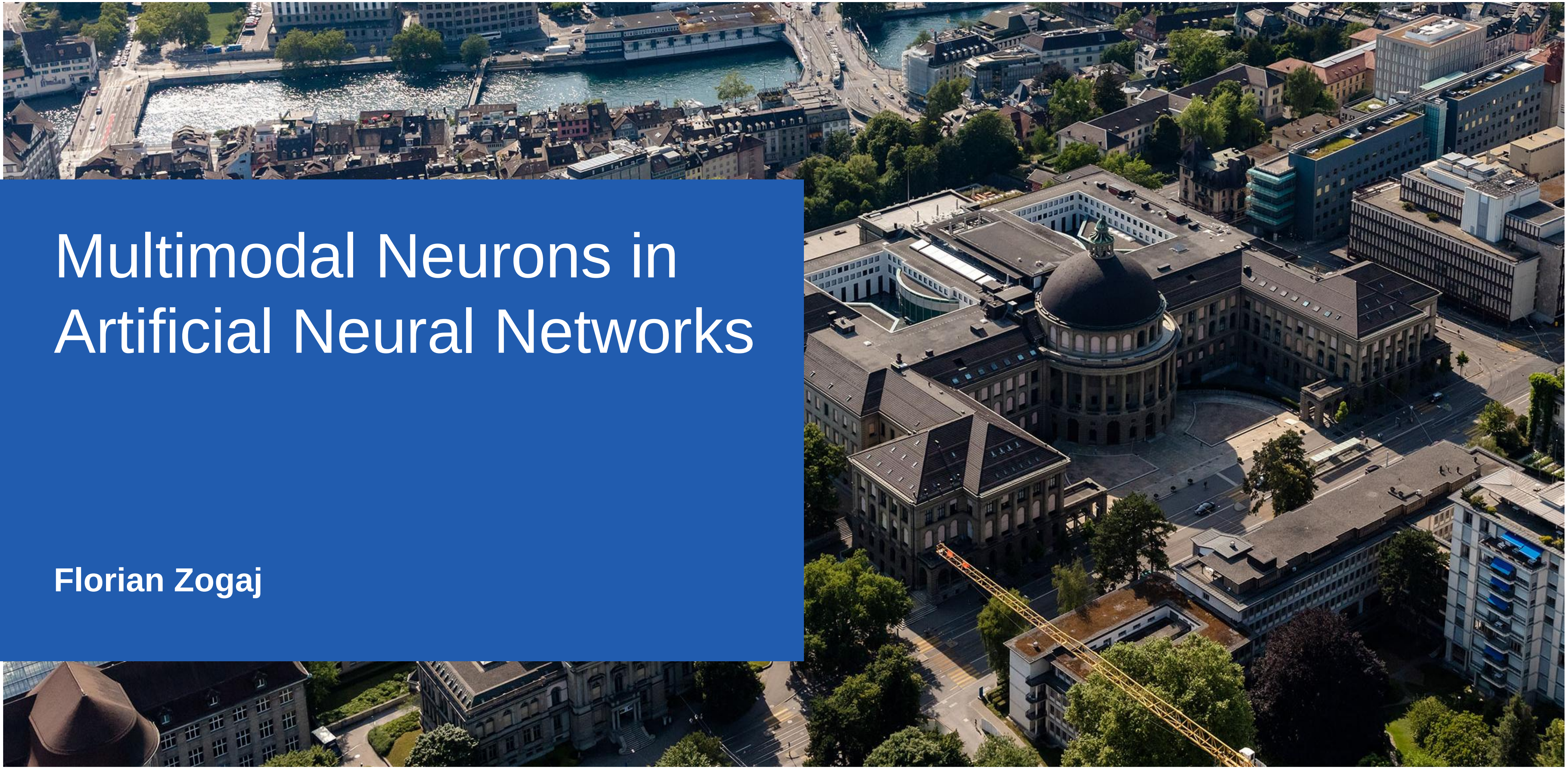
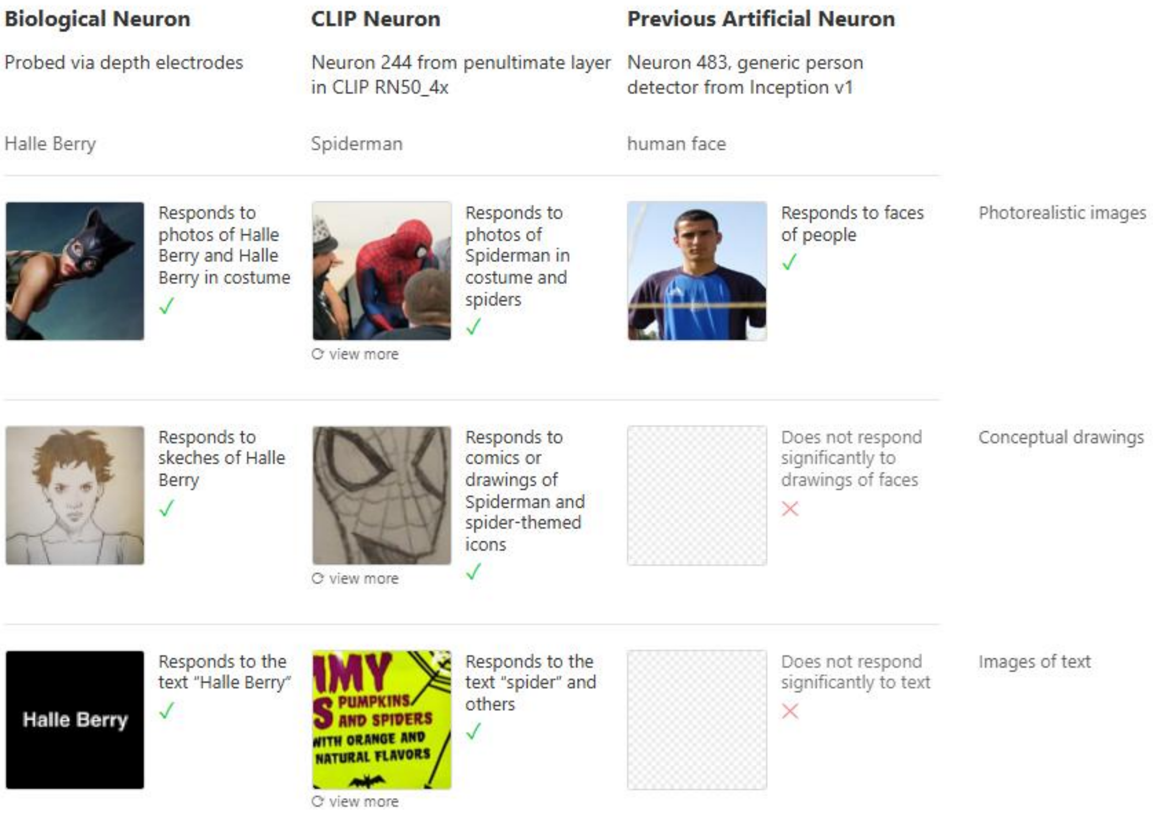
ETH zürich



Multimodal Neurons in Artificial Neural Networks

Florian Zogaj

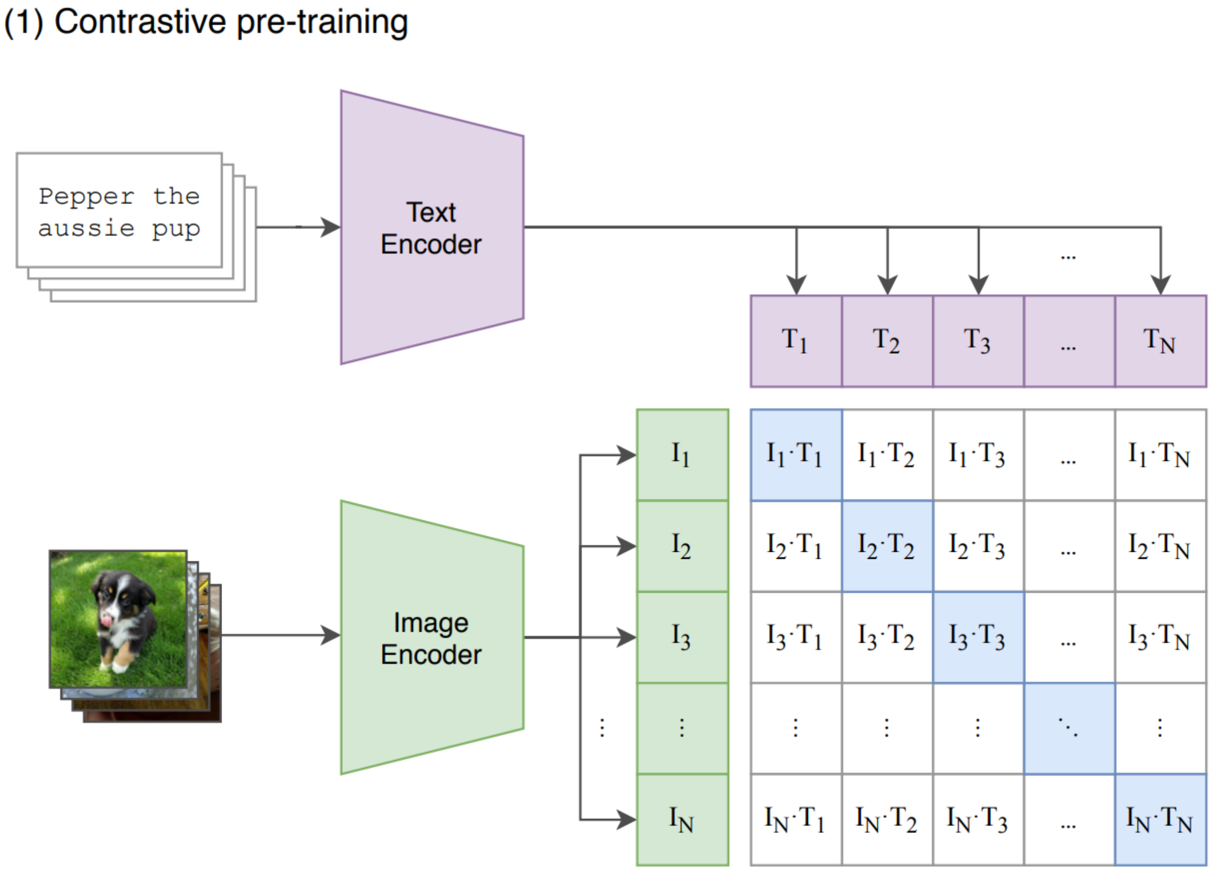
Multimodal neurons



Biological Neuron	CLIP Neuron	Previous Artificial Neuron	
Probed via depth electrodes	Neuron 244 from penultimate layer in CLIP RN50_4x	Neuron 483, generic person detector from Inception v1	
Halle Berry	Spiderman	human face	
Responds to photos of Halle Berry and Halle Berry in costume ✓	Responds to photos of Spiderman in costume and spiders ✓ ⟳ view more	Responds to faces of people ✓	Photorealistic images
Responds to skeches of Halle Berry ✓	Responds to comics or drawings of Spiderman and spider-themed icons ✓ ⟳ view more	Does not respond significantly to drawings of faces ✗	Conceptual drawings
Responds to the text "Halle Berry" ✓	Responds to the text "spider" and others ✓ ⟳ view more	Does not respond significantly to text ✗	Images of text

Goh, Gabriel, et al. "Multimodal neurons in artificial neural networks." *Distill* 6.3 (2021): e30.
Quiroga, R. Quian, et al. "Invariant visual representation by single neurons in the human brain." *Nature* 435.7045 (2005): 1102-1107

CLIP

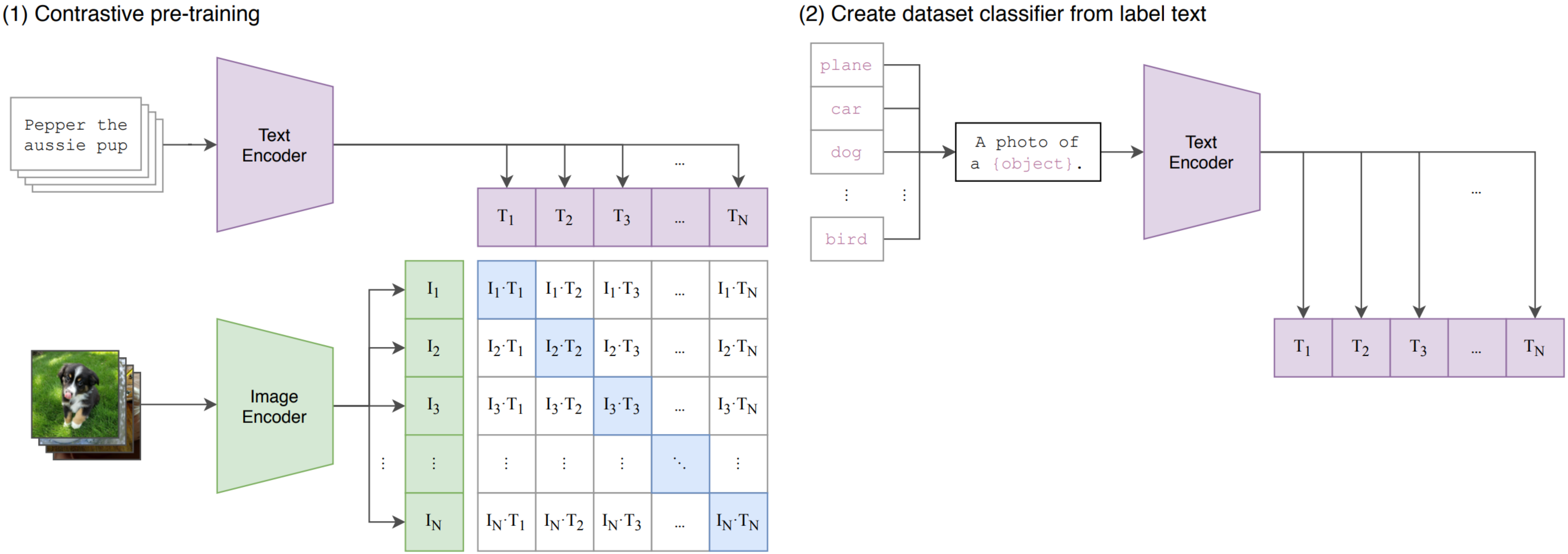


(1) Contrastive pre-training

Pepper the aussie pup
Text Encoder
Image Encoder
T_1
T_2
T_3
...
T_N
I_1
I_2
I_3
⋮
I_N
$I_1 \cdot T_1$ $I_1 \cdot T_2$ $I_1 \cdot T_3$ ... $I_1 \cdot T_N$
$I_2 \cdot T_1$ $I_2 \cdot T_2$ $I_2 \cdot T_3$ ... $I_2 \cdot T_N$
$I_3 \cdot T_1$ $I_3 \cdot T_2$ $I_3 \cdot T_3$ ... $I_3 \cdot T_N$
⋮ ⋮ ⋮ ⋱ ⋮
$I_N \cdot T_1$ $I_N \cdot T_2$ $I_N \cdot T_3$ ... $I_N \cdot T_N$

Radford, Alec, et al. "Learning transferable visual models from natural language supervision." *International conference on machine learning*. PmLR, 2021

CLIP



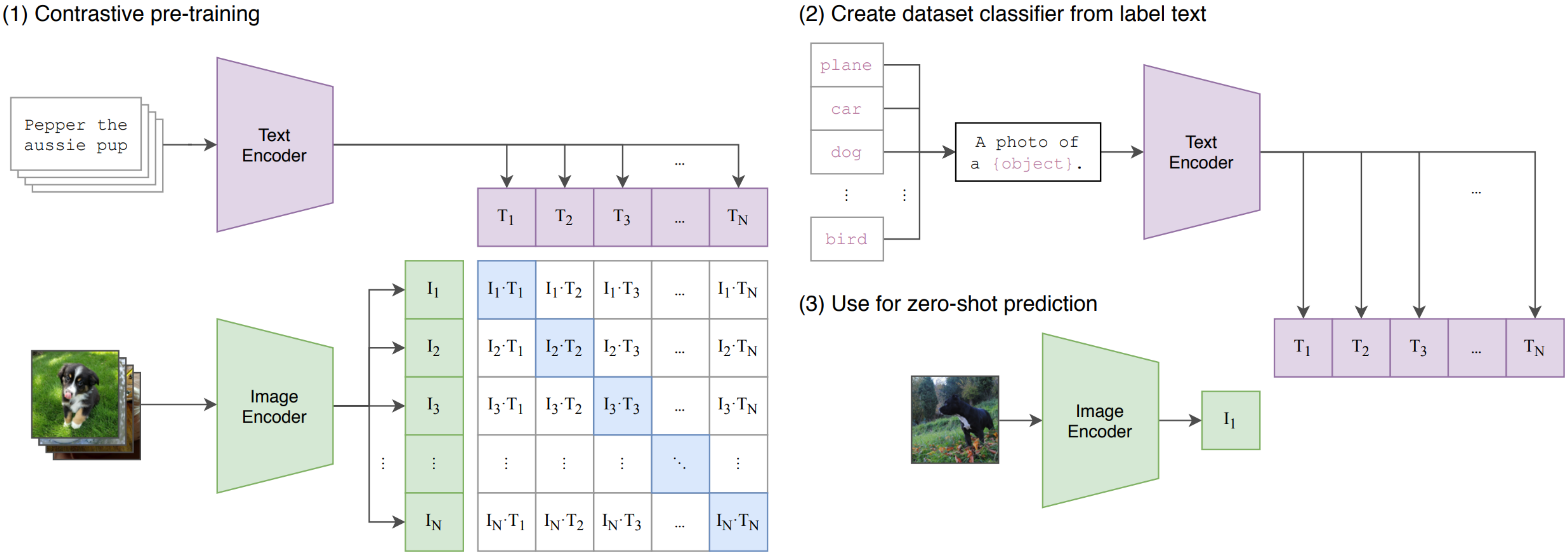
(1) Contrastive pre-training

	T_1	T_2	T_3	...	T_N
I_1	$I_1 \cdot T_1$	$I_1 \cdot T_2$	$I_1 \cdot T_3$	...	$I_1 \cdot T_N$
I_2	$I_2 \cdot T_1$	$I_2 \cdot T_2$	$I_2 \cdot T_3$	...	$I_2 \cdot T_N$
I_3	$I_3 \cdot T_1$	$I_3 \cdot T_2$	$I_3 \cdot T_3$	...	$I_3 \cdot T_N$
⋮	⋮	⋮	⋮	⋱	⋮
I_N	$I_N \cdot T_1$	$I_N \cdot T_2$	$I_N \cdot T_3$	...	$I_N \cdot T_N$

(2) Create dataset classifier from label text

Radford, Alec, et al. "Learning transferable visual models from natural language supervision." *International conference on machine learning*. PmLR, 2021

CLIP



(1) Contrastive pre-training

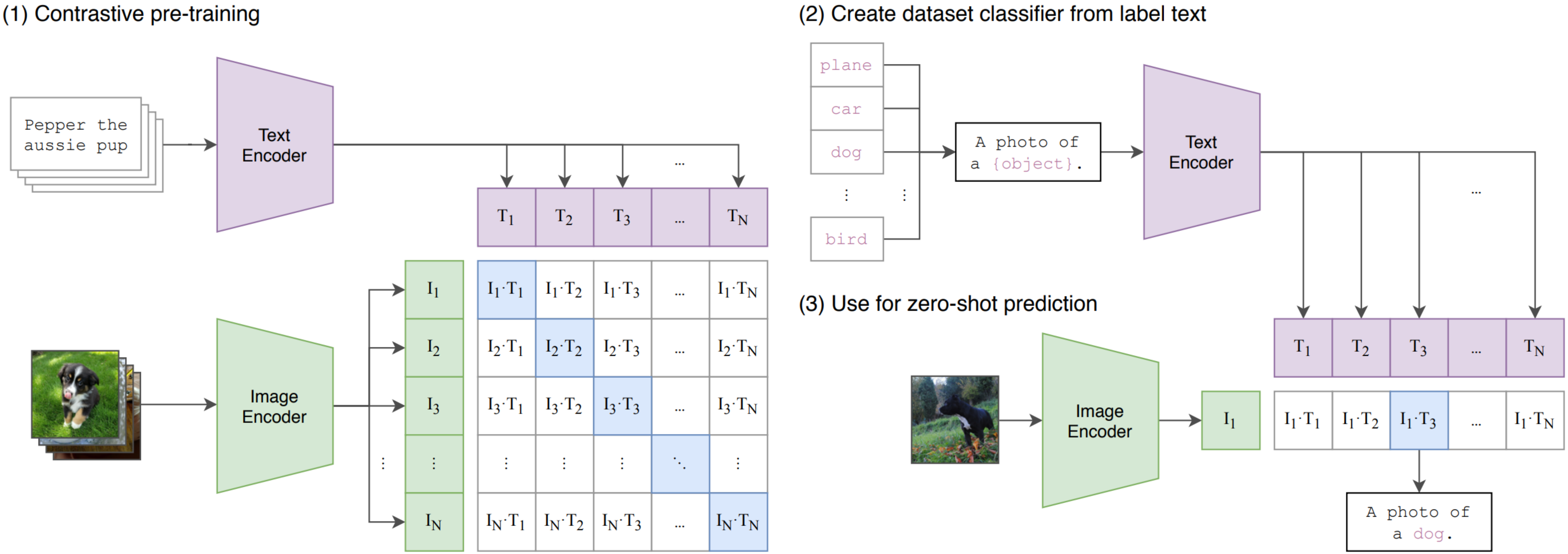
	T_1	T_2	T_3	...	T_N
I_1	$I_1 \cdot T_1$	$I_1 \cdot T_2$	$I_1 \cdot T_3$	...	$I_1 \cdot T_N$
I_2	$I_2 \cdot T_1$	$I_2 \cdot T_2$	$I_2 \cdot T_3$	...	$I_2 \cdot T_N$
I_3	$I_3 \cdot T_1$	$I_3 \cdot T_2$	$I_3 \cdot T_3$	...	$I_3 \cdot T_N$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
I_N	$I_N \cdot T_1$	$I_N \cdot T_2$	$I_N \cdot T_3$	...	$I_N \cdot T_N$

(2) Create dataset classifier from label text

(3) Use for zero-shot prediction

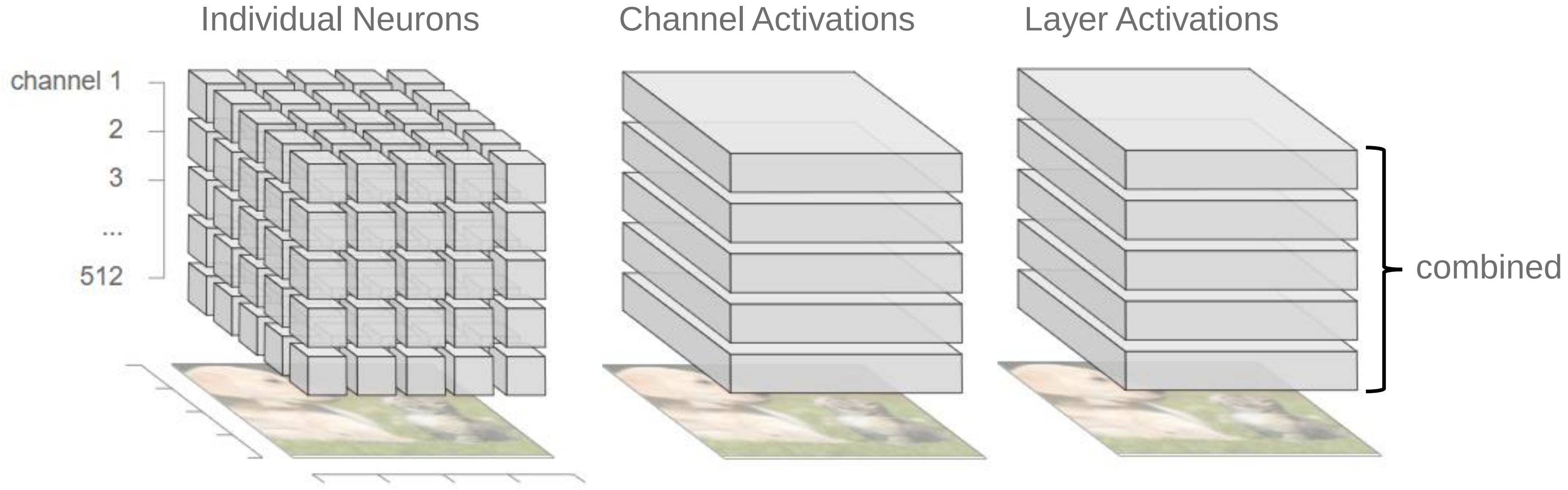
Radford, Alec, et al. "Learning transferable visual models from natural language supervision." *International conference on machine learning*. PmLR, 2021

CLIP


(1) Contrastive pre-training
Pepper the aussie pup
Text Encoder
T_1 T_2 T_3 ... T_N
Image Encoder
I_1 I_2 I_3 ⋮ I_N
$I_1 \cdot T_1$ $I_1 \cdot T_2$ $I_1 \cdot T_3$ ... $I_1 \cdot T_N$
$I_2 \cdot T_1$ $I_2 \cdot T_2$ $I_2 \cdot T_3$ ... $I_2 \cdot T_N$
$I_3 \cdot T_1$ $I_3 \cdot T_2$ $I_3 \cdot T_3$ ... $I_3 \cdot T_N$
⋮ ⋮ ⋮ ⋱ ⋮
$I_N \cdot T_1$ $I_N \cdot T_2$ $I_N \cdot T_3$ ... $I_N \cdot T_N$
(2) Create dataset classifier from label text
plane
car
dog
⋮
bird
A photo of a {object}.
Text Encoder
T_1 T_2 T_3 ... T_N
(3) Use for zero-shot prediction
Image Encoder
I_1
$I_1 \cdot T_1$ $I_1 \cdot T_2$ $I_1 \cdot T_3$ ... $I_1 \cdot T_N$
A photo of a dog.

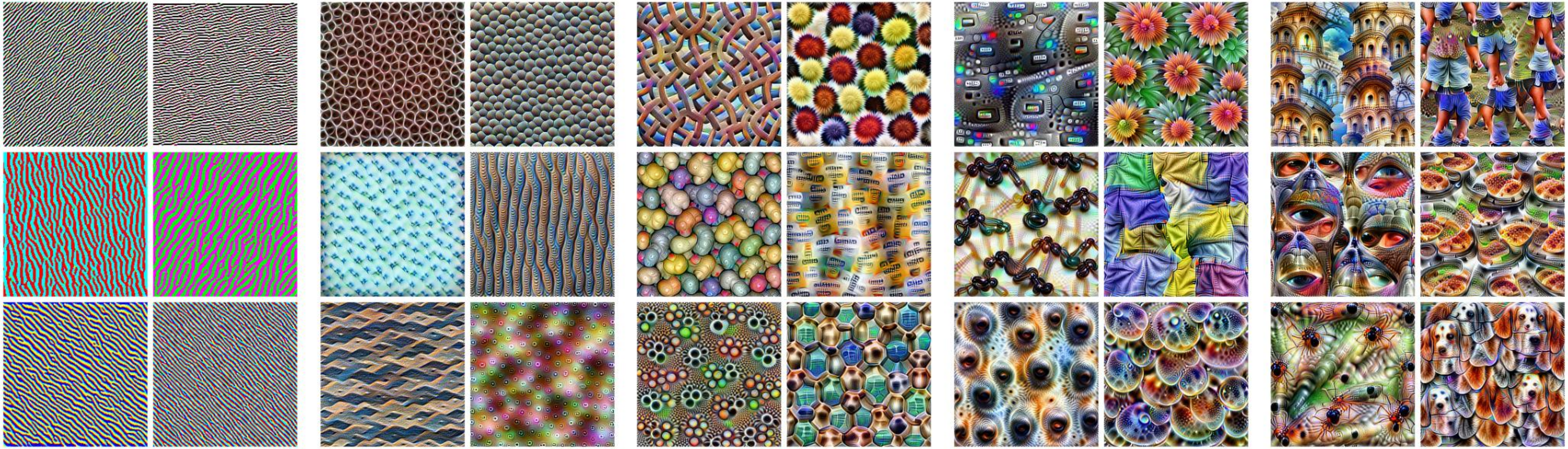
Radford, Alec, et al. "Learning transferable visual models from natural language supervision." *International conference on machine learning*. PmLR, 2021

What is a neuron, channel, layer?



Olah, Chris, et al. "The building blocks of interpretability." *Distill* 3.3 (2018): e10.

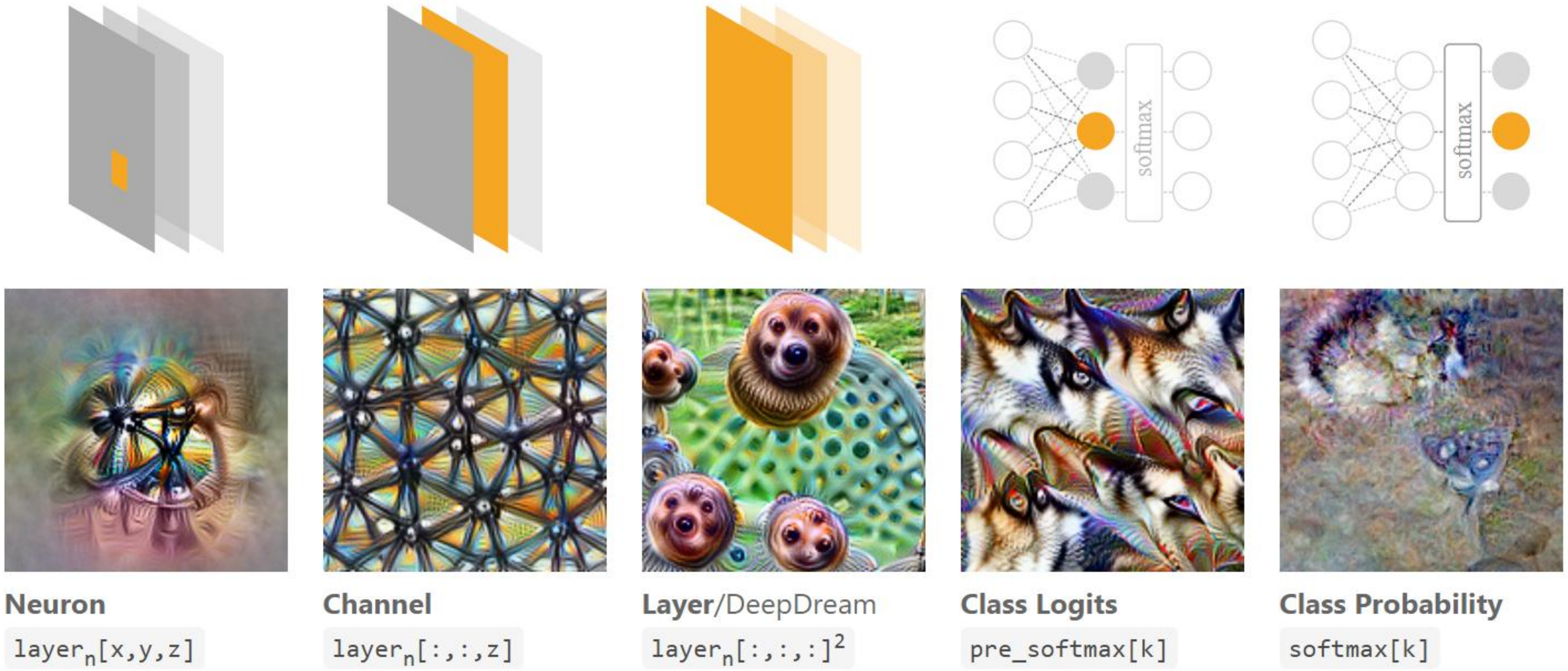
Feature Visualization



Edges (layer conv2d0) **Textures** (layer mixed3a) **Patterns** (layer mixed4a) **Parts** (layers mixed4b & mixed4c) **Objects** (layers mixed4d & mixed4e)

Olah, Chris, Alexander Mordvintsev, and Ludwig Schubert. "Feature visualization." *Distill* 2.11 (2017): e7.

Neuron, Channel and Layer activations

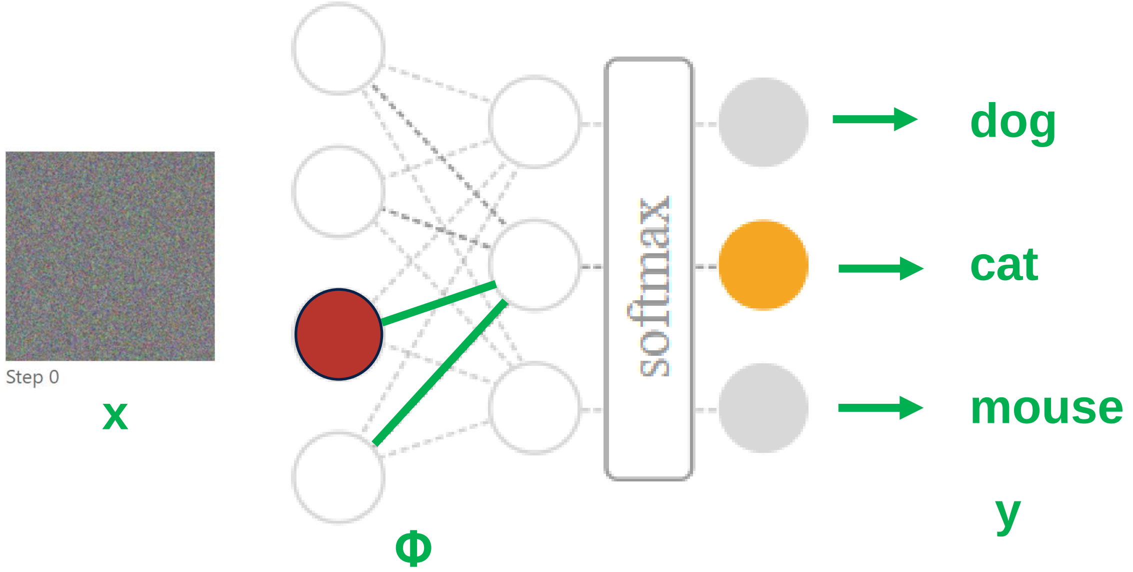


> Feature visualization is a 'powerful tool for understanding what our models are doing' - Geoffrey Hinton, Google Researcher and Pioneer in Neural Networks.

Olah, Chris, Alexander Mordvintsev, and Ludwig Schubert. "Feature visualization." *Distill* 2.11 (2017): e7.

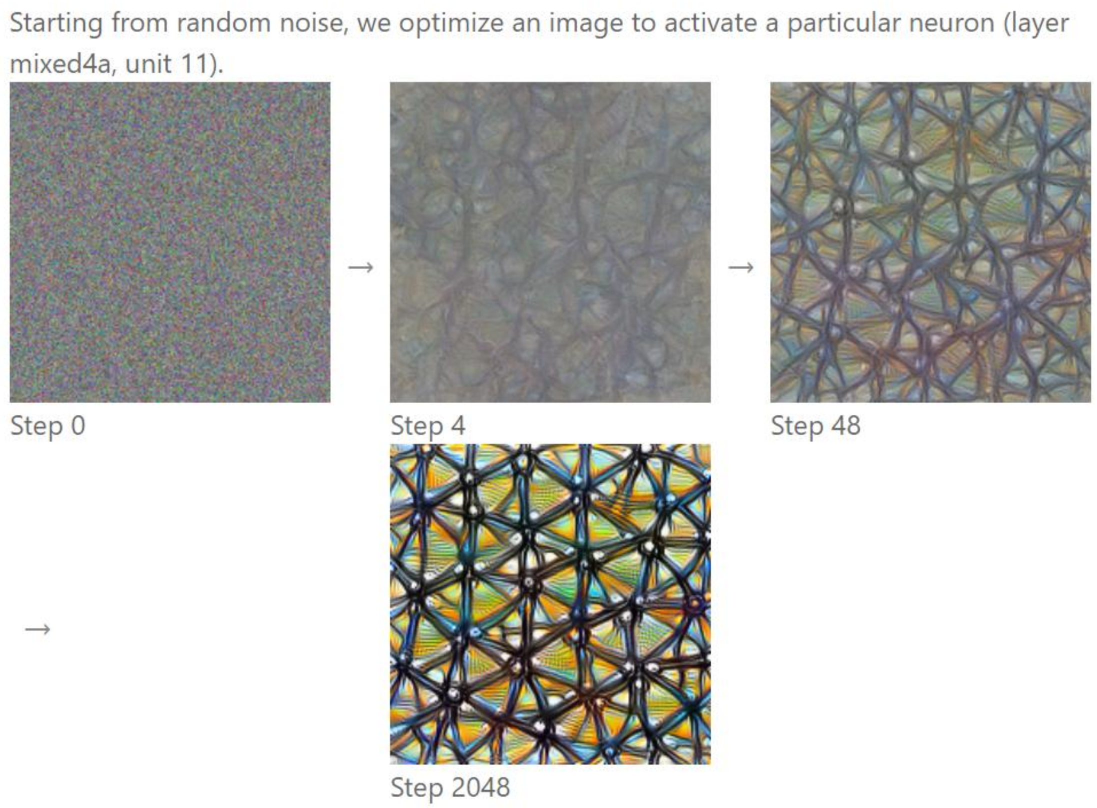
Feature Visualization

- Optimize for neuron



Olah, Chris, Alexander Mordvintsev, and Ludwig Schubert. "Feature visualization." *Distill* 2.11 (2017): e7.

Feature Visualization



Starting from random noise, we optimize an image to activate a particular neuron (layer mixed4a, unit 11).

Step 0

Step 4

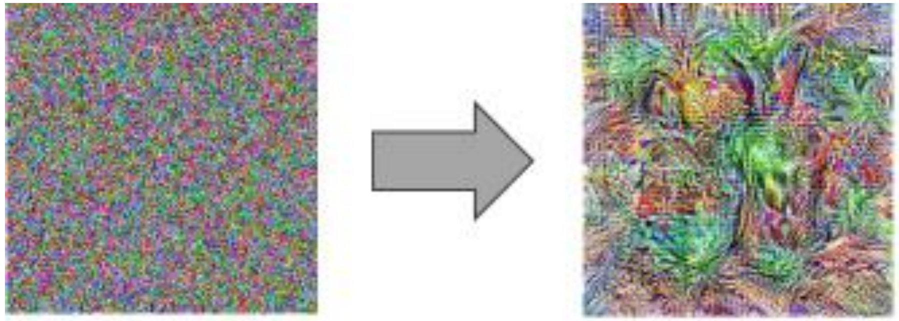
Step 48

Step 2048

Olah, Chris, Alexander Mordvintsev, and Ludwig Schubert. "Feature visualization." *Distill* 2.11 (2017): e7.

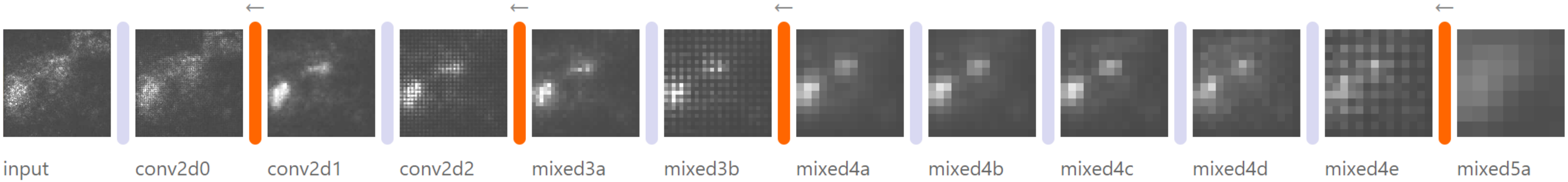
Activation Maximization

$$x_{t+1} = x_t + \varepsilon_1 \cdot \frac{\partial_{\mathrm{a}}(\theta, x_t)}{\partial_{x_t}}$$



a) Random initialization

b) Adversarial example

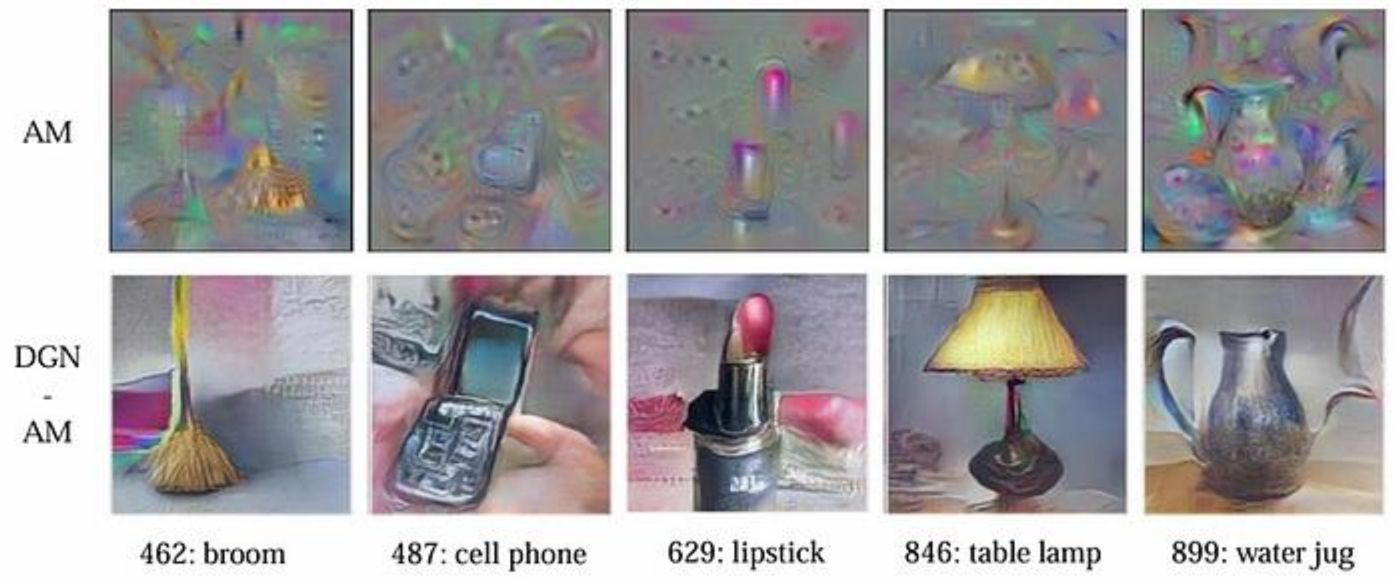


Olah, Chris, Alexander Mordvintsev, and Ludwig Schubert. "Feature visualization." *Distill* 2.11 (2017): e7.

Activation Maximization - Regularization

- Frequency penalization
- Transformation robustness
- Learned priors

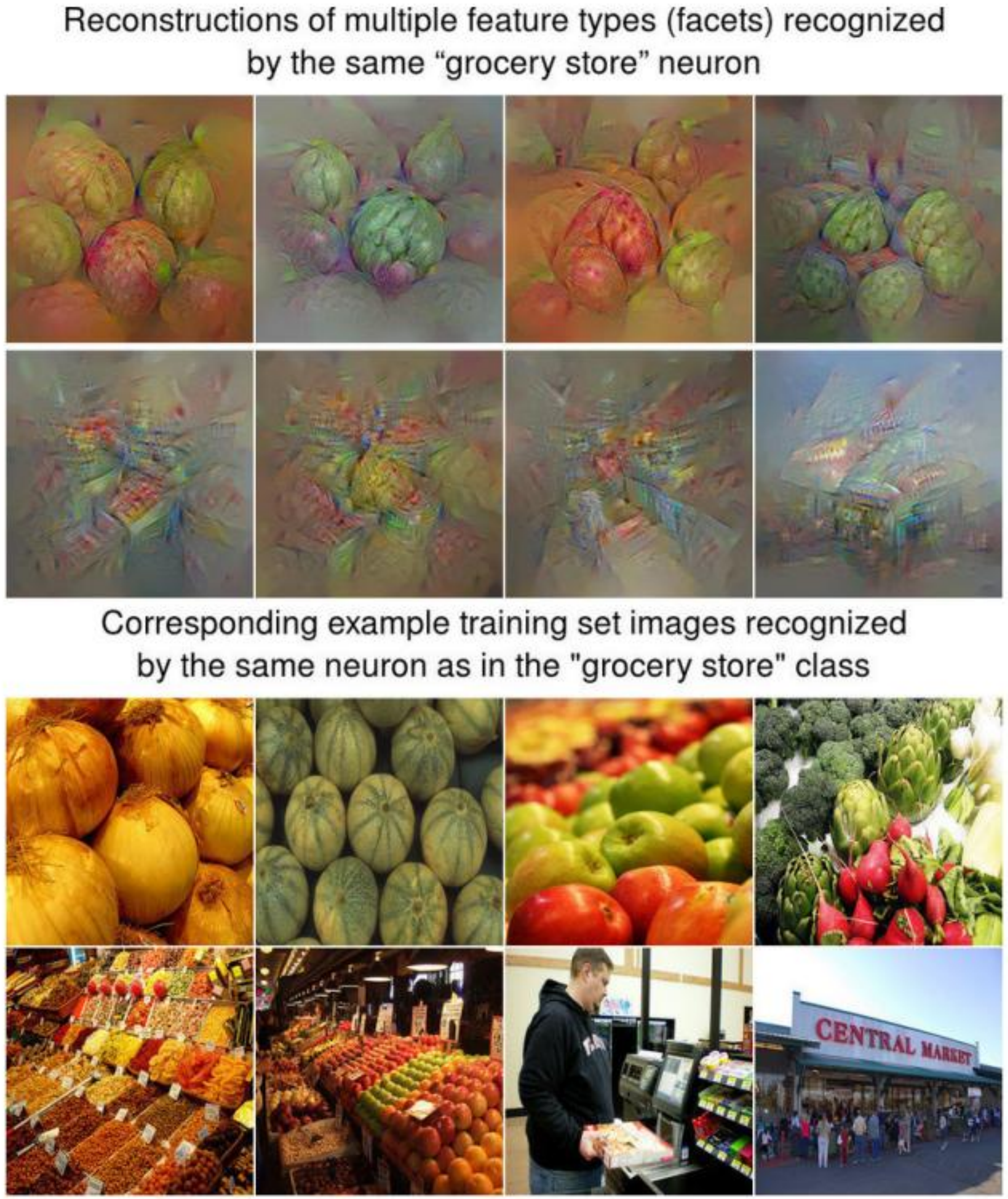
$$x_{t+1} = r(x_t) + \varepsilon_1 \cdot \frac{\partial_{\mathrm{a}}(x_t)}{\partial_{x_t}}$$



Qin, Zhuwei, et al. "How convolutional neural network see the world-A survey of convolutional neural network visualization methods." arXiv preprint arXiv:1804.11191 (2018).

Multi-faceted neurons

- Diverse images as seed
- Add diversity term to optimization
- Faceted feature visualization

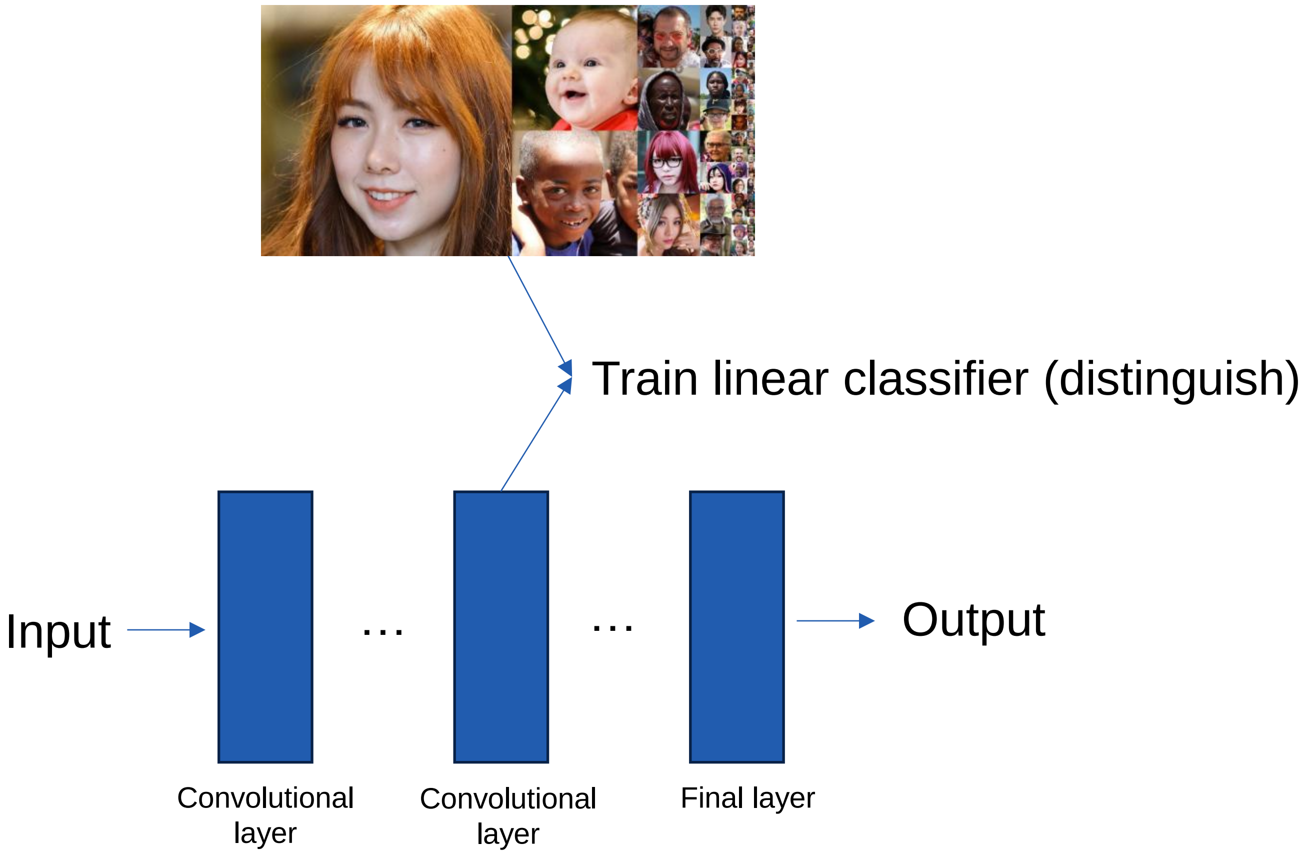


Reconstructions of multiple feature types (facets) recognized by the same “grocery store” neuron

Corresponding example training set images recognized by the same neuron as in the "grocery store" class

Qin, Zhuwei, et al. "How convolutional neural network see the world-A survey of convolutional neural network visualization methods." arXiv preprint arXiv:1804.11191 (2018).

Faceted Feature Visualization



Train linear classifier (distinguish)

Input

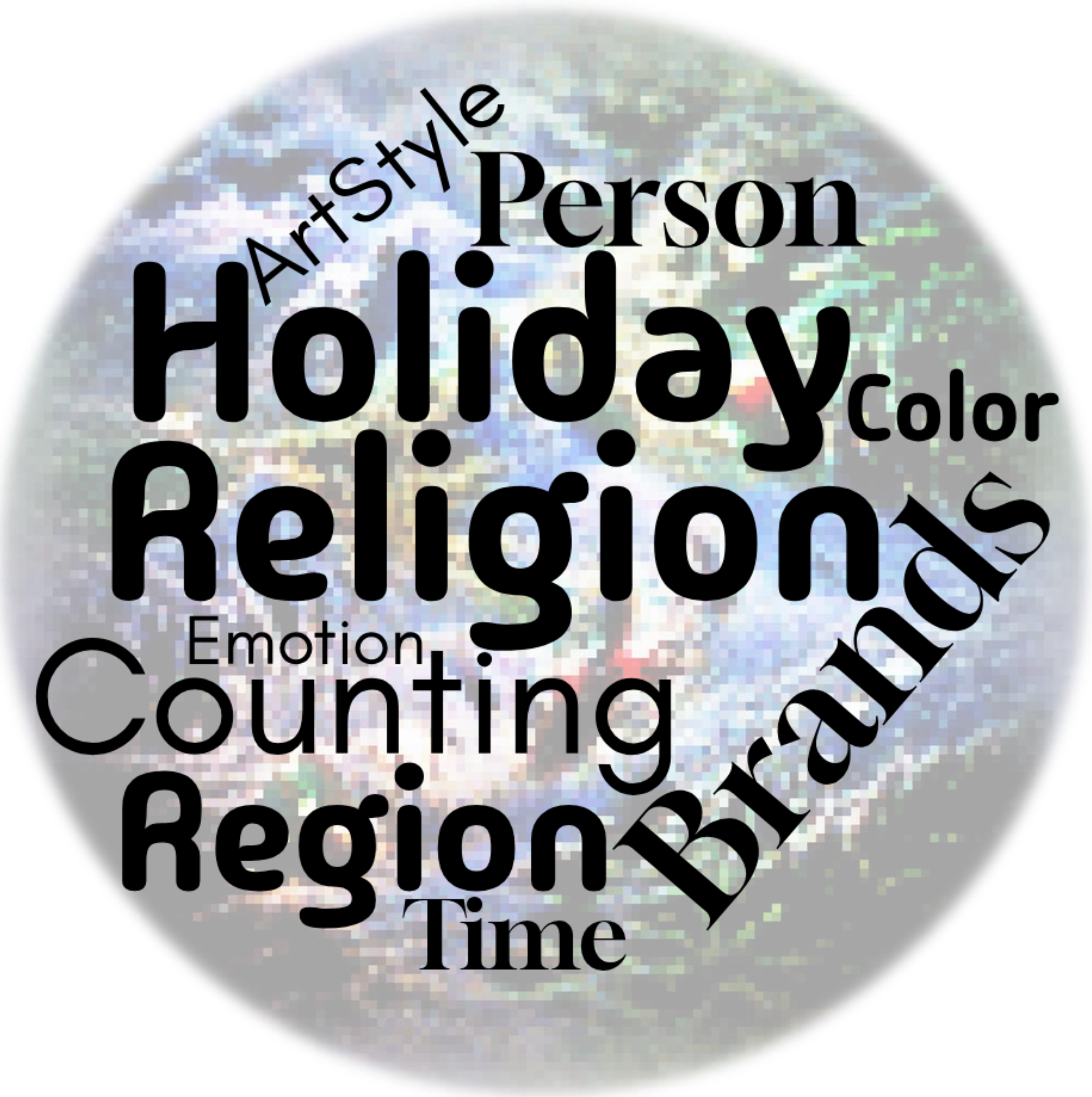
Output

$$\mathrm{f}(\mathrm{g}(\mathrm{x})) + w^T \left(g(x)\ \odot\ \nabla f(g(x))\right)$$

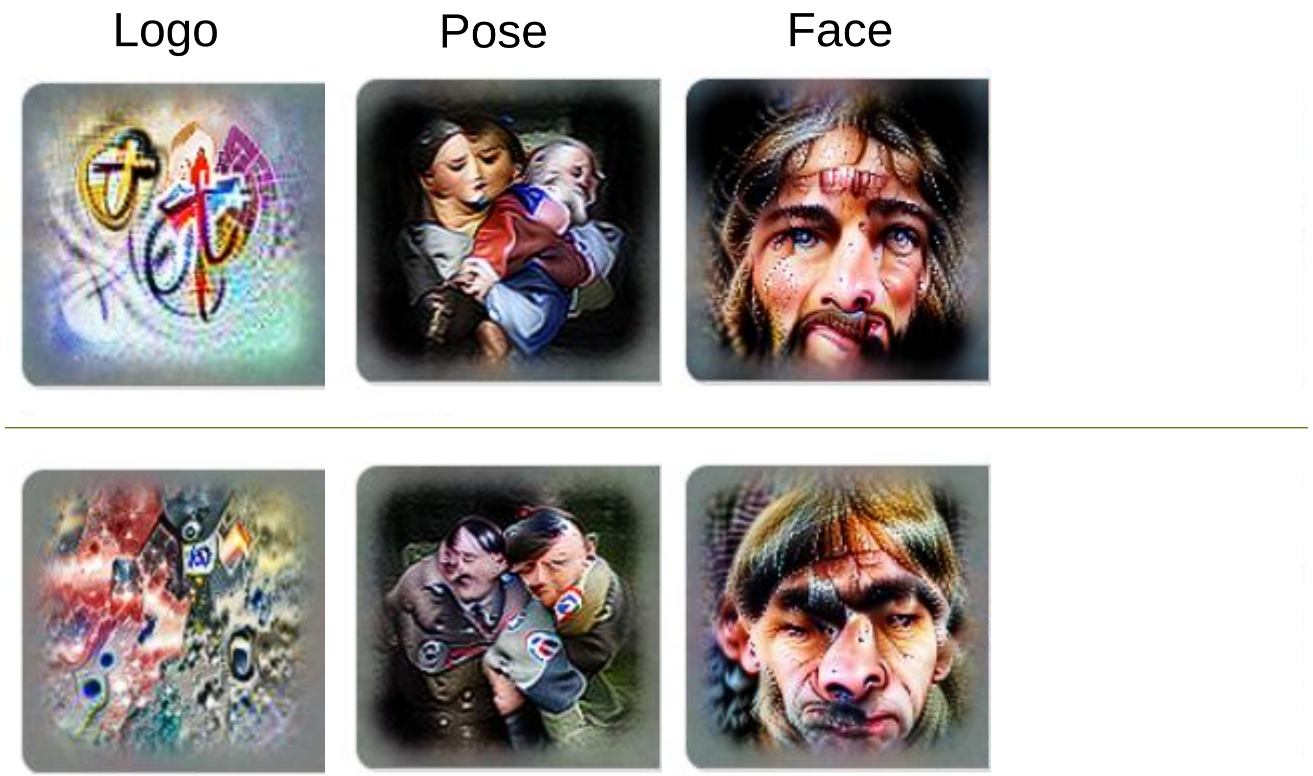
g(x)	activations up to intermediate layer
f(g(x))	activations of all layers
w	trained weights from linear classifier

Goh, Gabriel, et al. "Multimodal neurons in artificial neural networks." *Distill* 6.3 (2021): e30.

CLIP RN50-x4 – What kind of neurons were found?

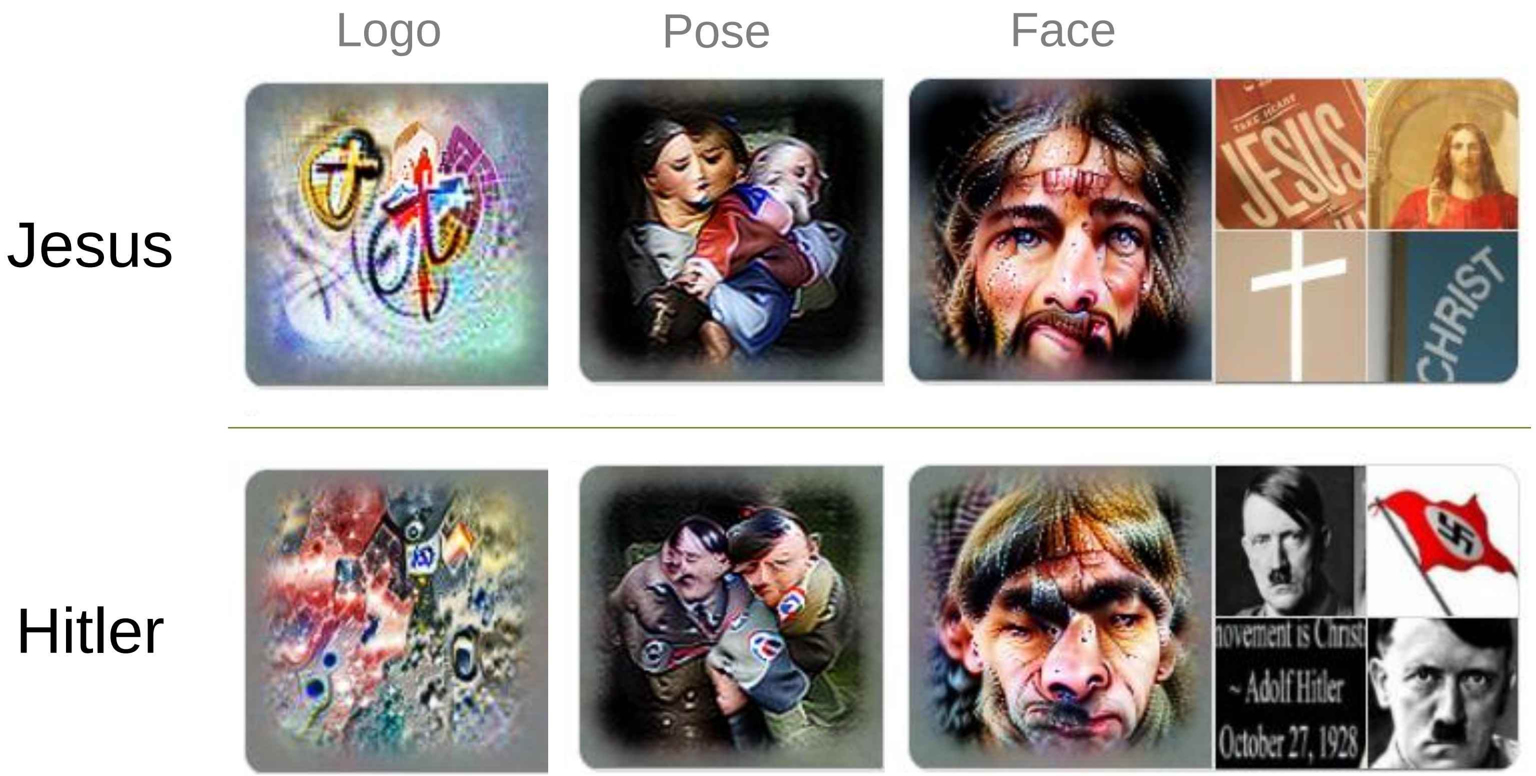


Person Neurons



Goh, Gabriel, et al. "Multimodal neurons in artificial neural networks." *Distill* 6.3 (2021): e30.

Person Neurons



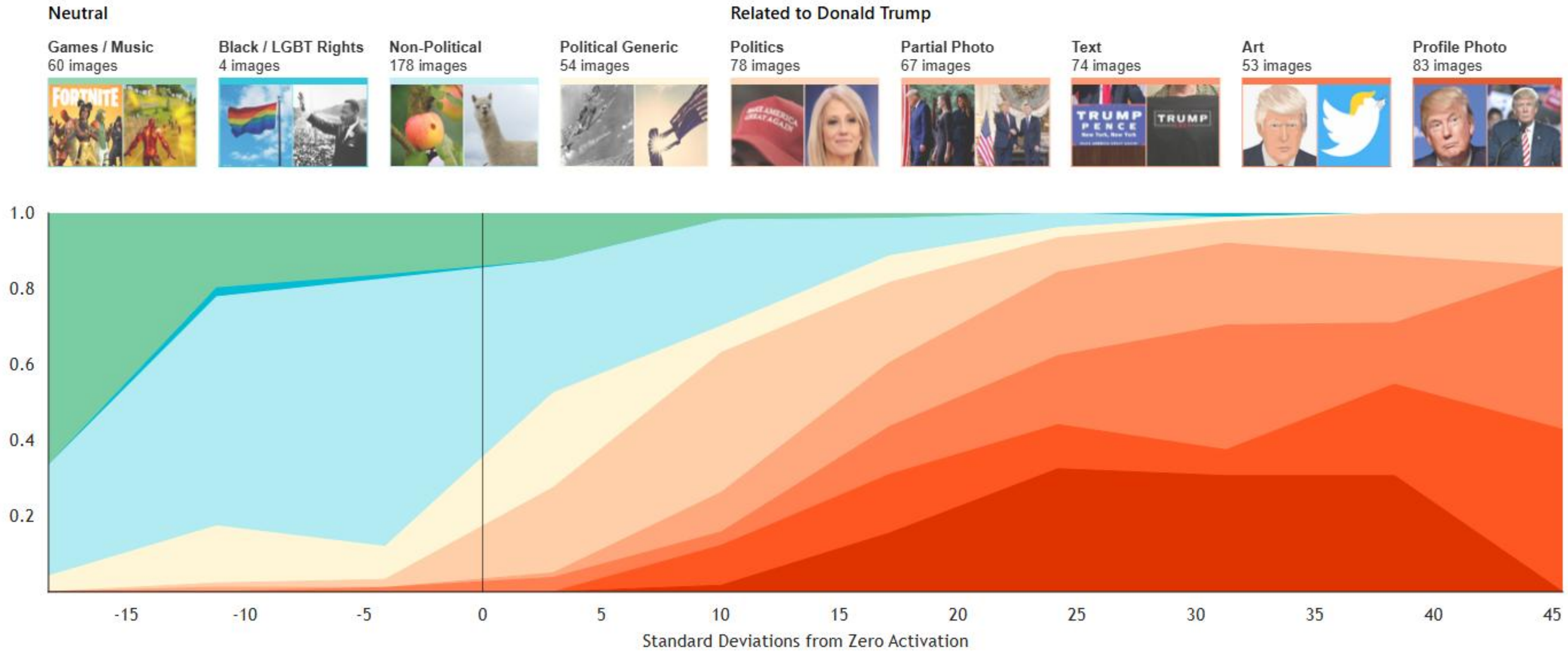
Logo Pose Face

Jesus

Hitler

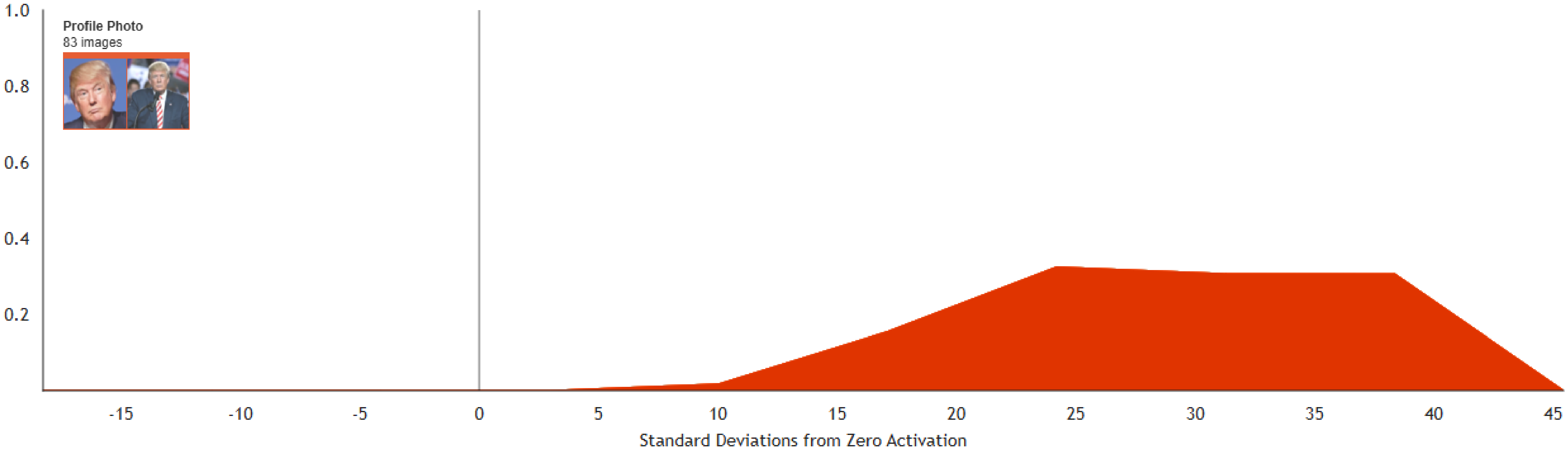
Goh, Gabriel, et al. "Multimodal neurons in artificial neural networks." *Distill* 6.3 (2021): e30.

Case study – Donald Trump neuron activation



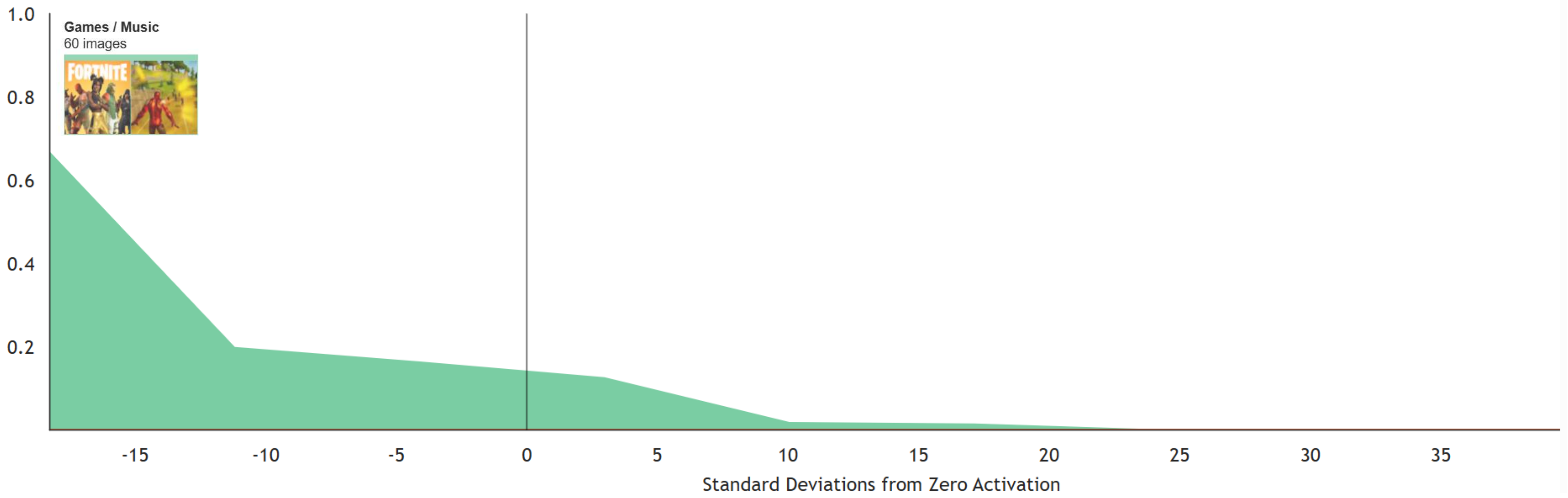
Goh, Gabriel, et al. "Multimodal neurons in artificial neural networks." *Distill* 6.3 (2021): e30.

Case study – Donald Trump neuron activation



Goh, Gabriel, et al. "Multimodal neurons in artificial neural networks." *Distill* 6.3 (2021): e30.

Case study – Donald Trump neuron activation



Goh, Gabriel, et al. "Multimodal neurons in artificial neural networks." *Distill* 6.3 (2021): e30.

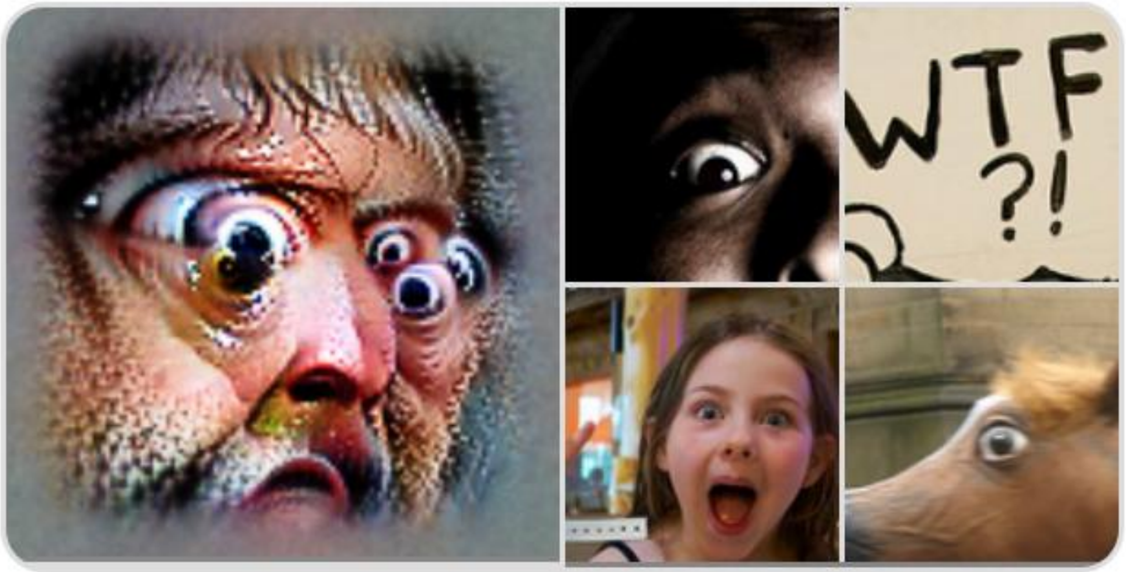
Emotion Neurons

- Facial expression
- Body language
- Text

Goh, Gabriel, et al. "Multimodal neurons in artificial neural networks." *Distill* 6.3 (2021): e30.

Emotion Neurons

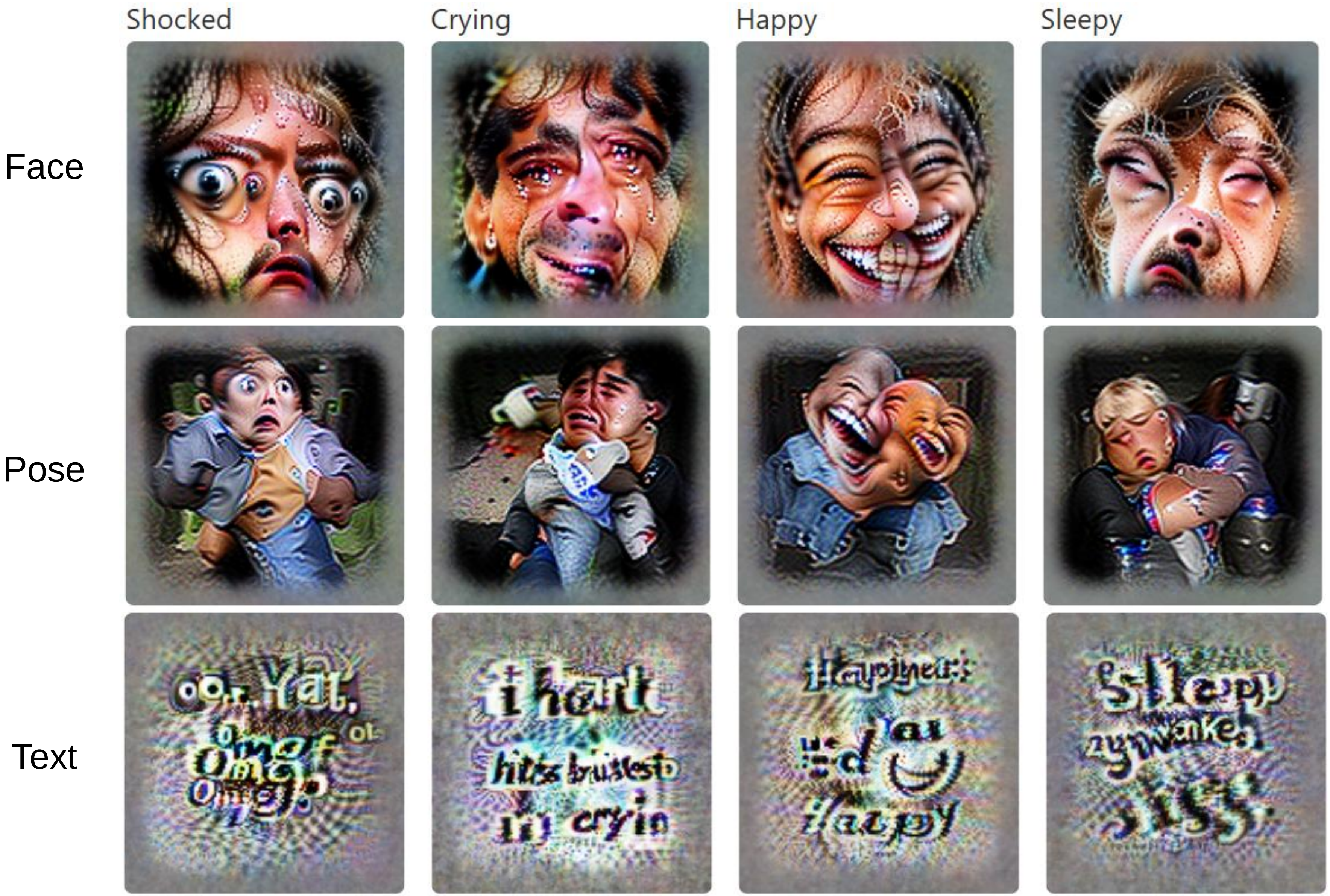
- Facial expression
- Body language
- Text



Surprise / Shock

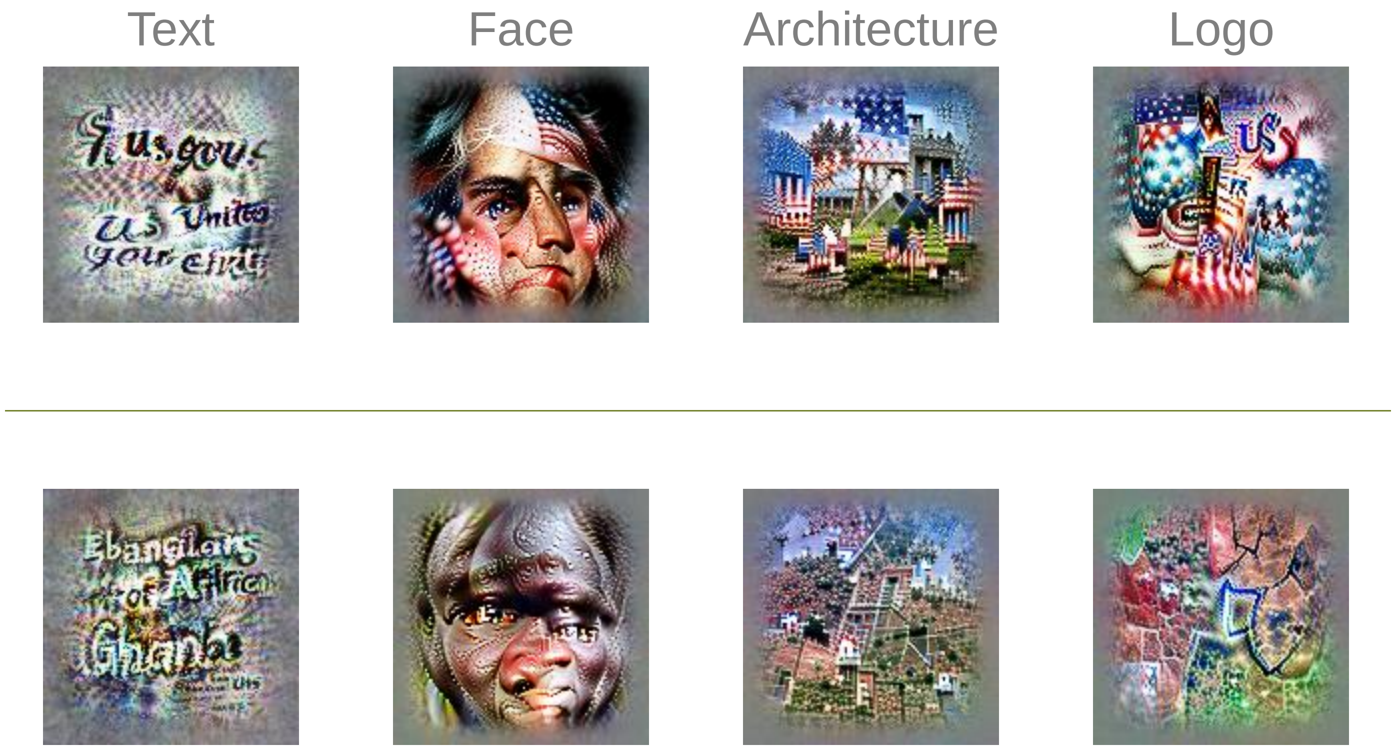
Goh, Gabriel, et al. "Multimodal neurons in artificial neural networks." *Distill* 6.3 (2021): e30.

Emotion Neurons



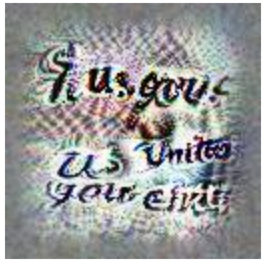
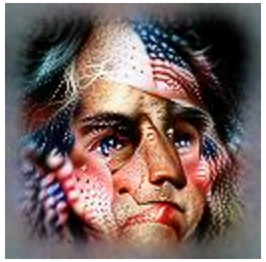
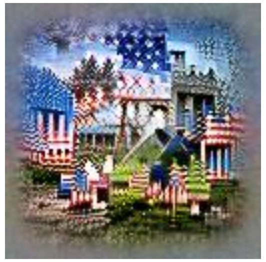
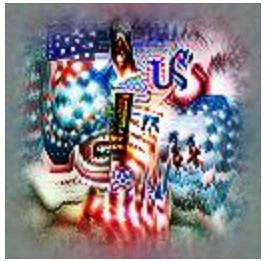
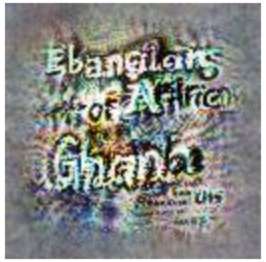
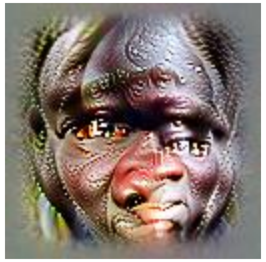
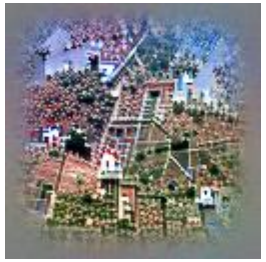
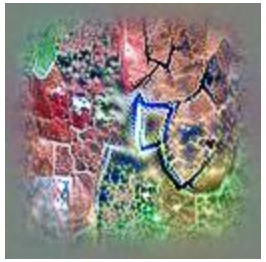
Goh, Gabriel, et al. "Multimodal neurons in artificial neural networks." *Distill* 6.3 (2021): e30.

Region Neurons



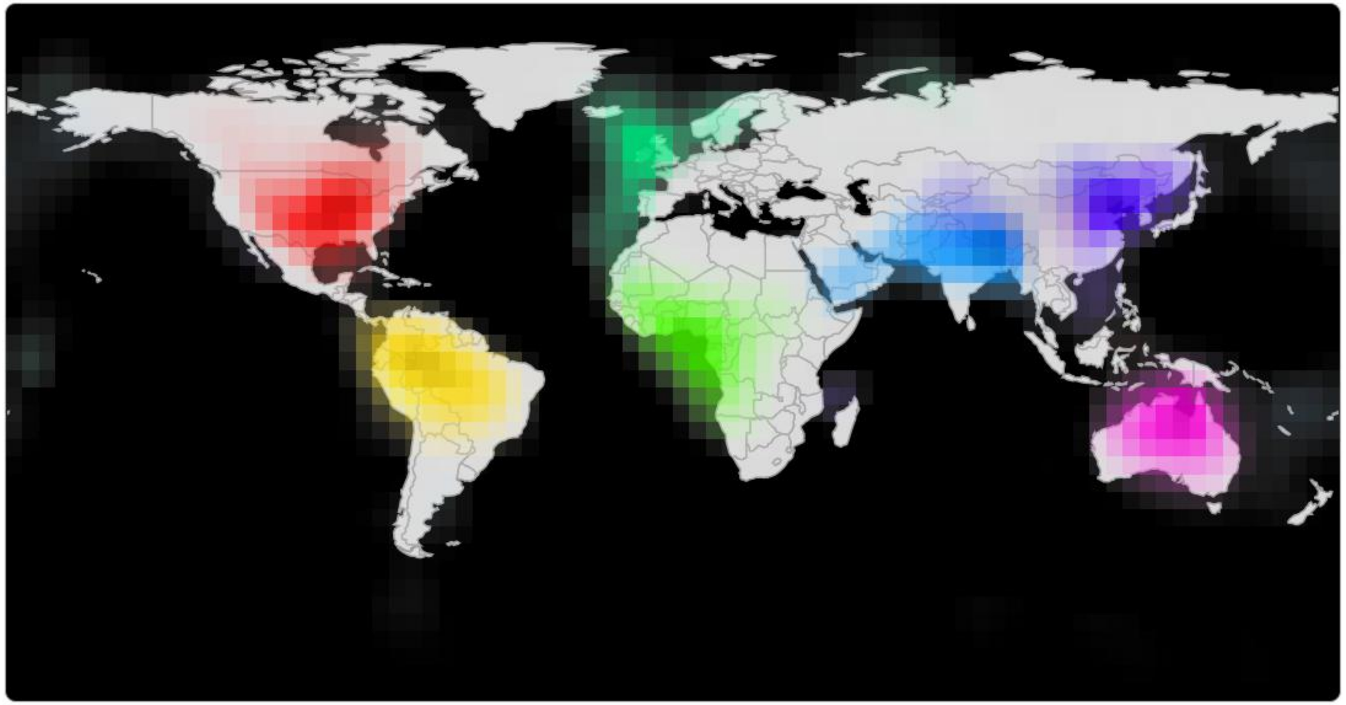
Goh, Gabriel, et al. "Multimodal neurons in artificial neural networks." *Distill* 6.3 (2021): e30.

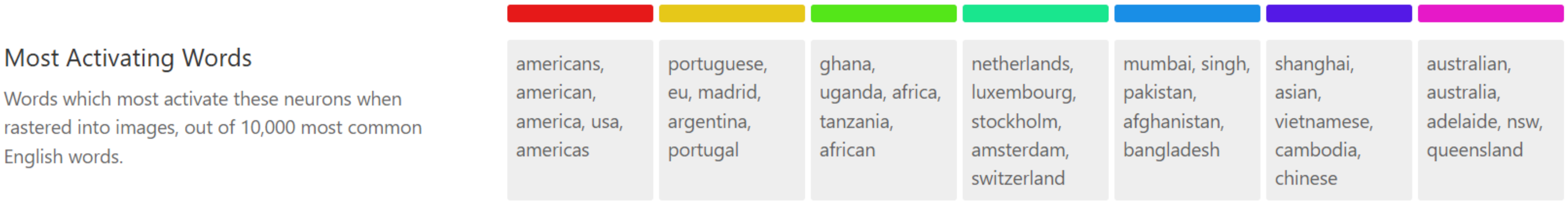
Region Neurons

	Text	Face	Architecture	Logo
USA				
West Africa				

Goh, Gabriel, et al. "Multimodal neurons in artificial neural networks." *Distill* 6.3 (2021): e30.

Region Neurons





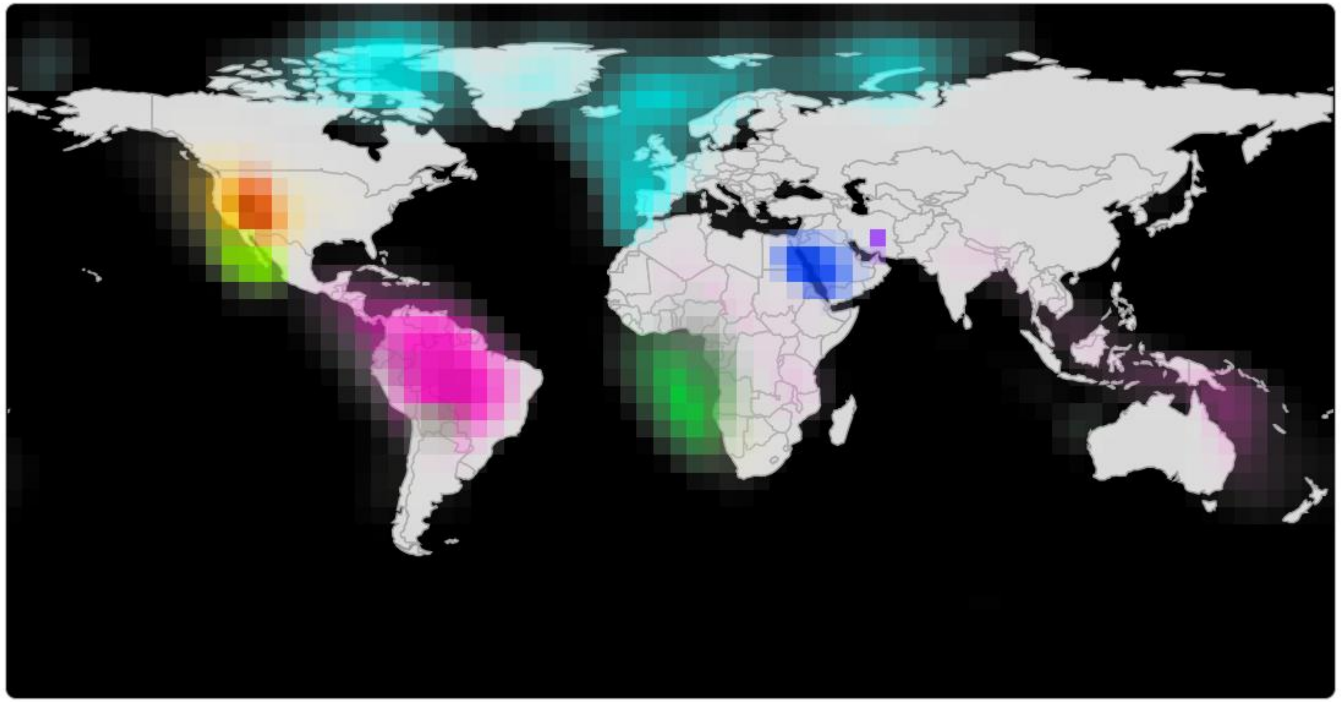
Most Activating Words

Words which most activate these neurons when rastered into images, out of 10,000 most common English words.

🟥	🟨	🟩	🟩	🟦	🟪	🟪
americans, american, america, usa, americas	portuguese, eu, madrid, argentina, portugal	ghana, uganda, africa, tanzania, african	netherlands, luxembourg, stockholm, amsterdam, switzerland	mumbai, singh, pakistan, afghanistan, bangladesh	shanghai, asian, vietnamese, cambodia, chinese	australian, australia, adelaide, nsw, queensland

Goh, Gabriel, et al. "Multimodal neurons in artificial neural networks." *Distill* 6.3 (2021): e30.

Region Neurons



Most Activating Words

Words which most activate these neurons when rastered into images, out of 10,000 most common English words.

🟥	🟧	🟩	🟩	🟦	🟦	🟪	🟪
angel, angels, wings, heaven, angeles	entrepreneurs, entrepreneur, founder, startup, starter	immigrants, immigration, borders, border, refugees	lions, jaguar, tigers, eagles, tiger	blanket, jackets, jacket, wrap, arctic	saudi, terrorists, terrorism, terrorist, allah	muslim, muslims, allah, islamic, islam	muslims, somalia, ethiopia, aboriginal, muslim

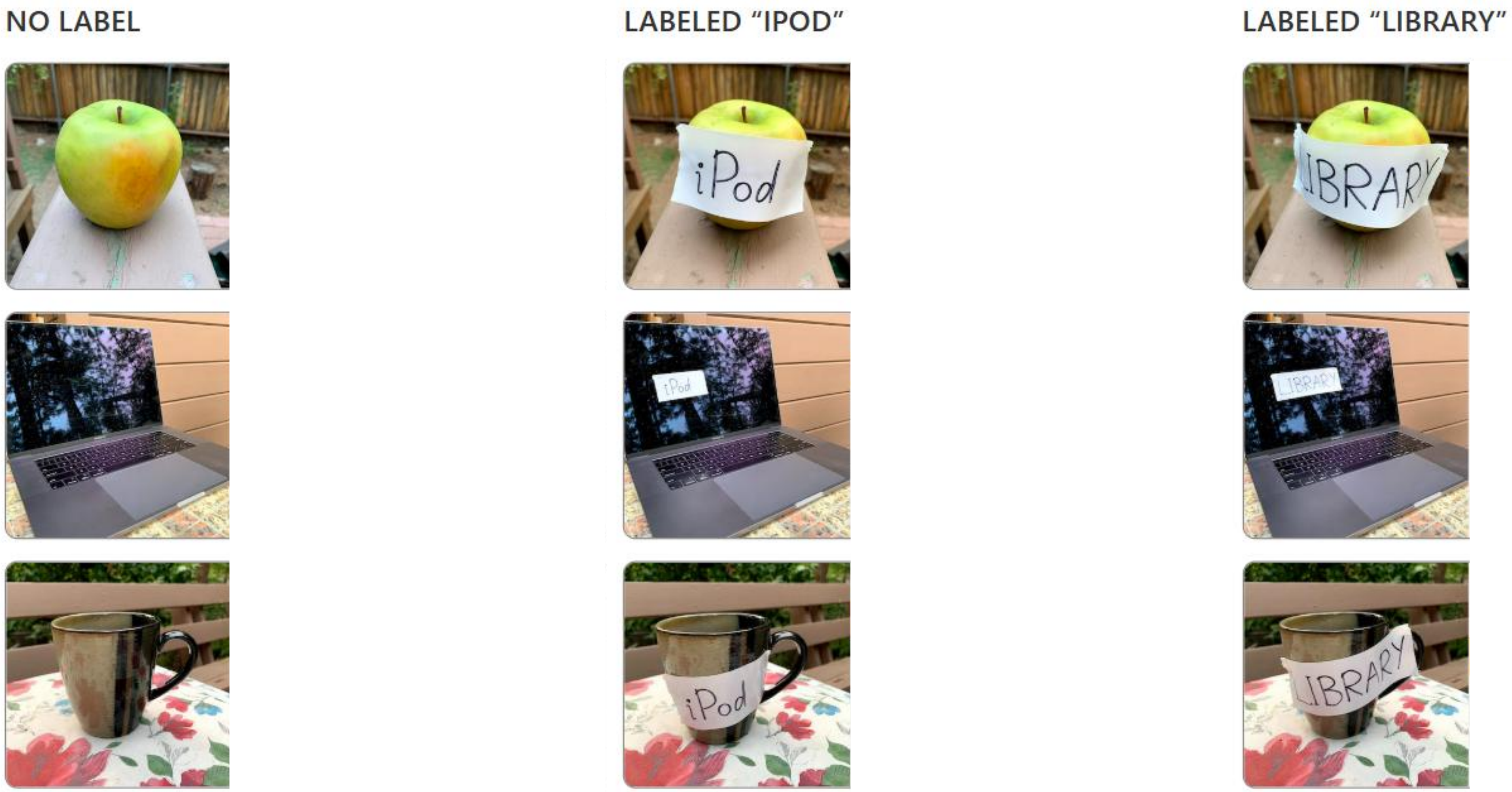
Goh, Gabriel, et al. "Multimodal neurons in artificial neural networks." *Distill* 6.3 (2021): e30.

Typographic attacks – Zero shot

NO LABEL

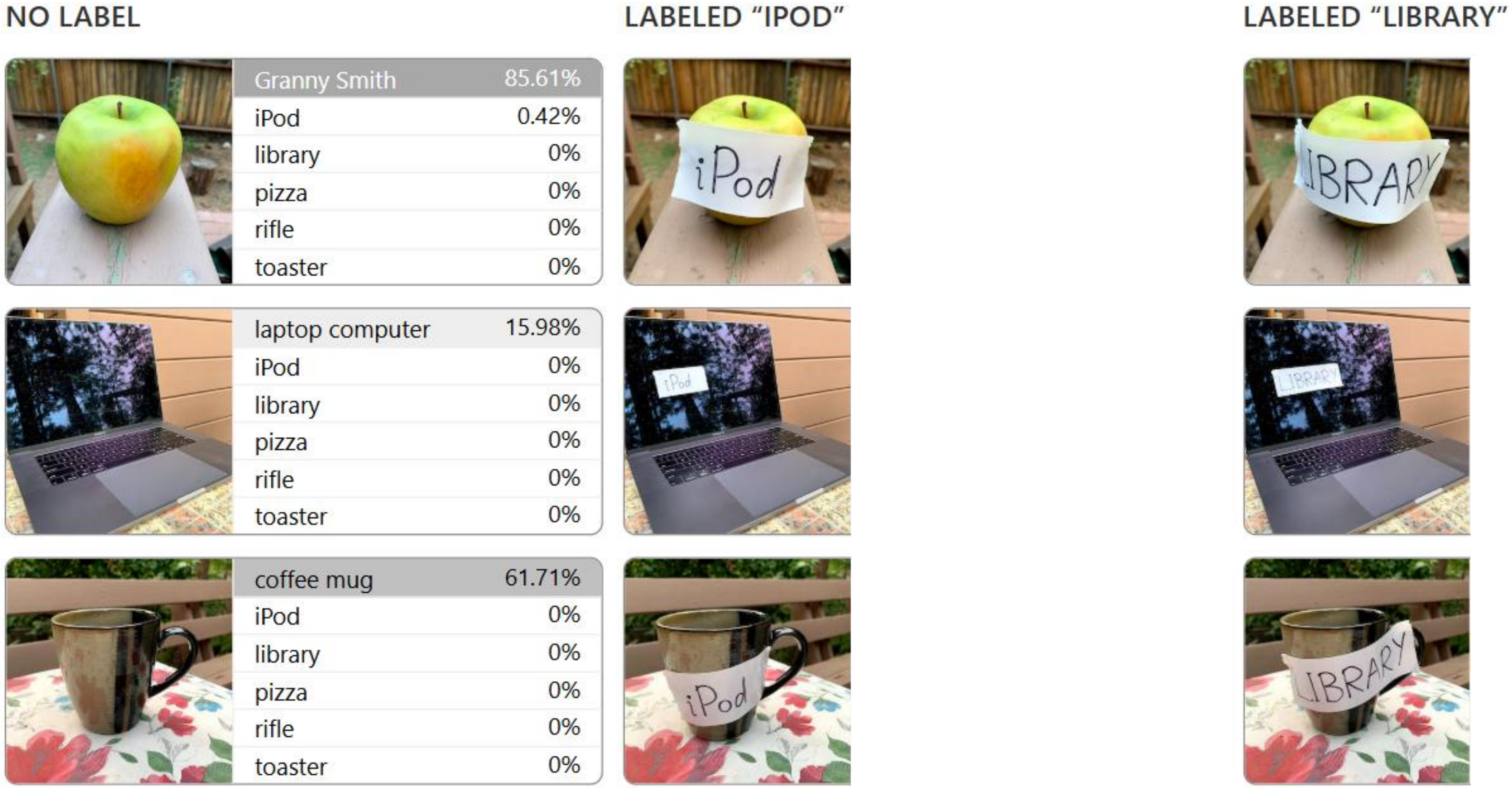
Goh, Gabriel, et al. "Multimodal neurons in artificial neural networks." *Distill* 6.3 (2021): e30.

Typographic attacks – Zero shot



Goh, Gabriel, et al. "Multimodal neurons in artificial neural networks." *Distill* 6.3 (2021): e30.

Typographic attacks – Zero shot



NO LABEL

LABELED “IPOD”

LABELED “LIBRARY”

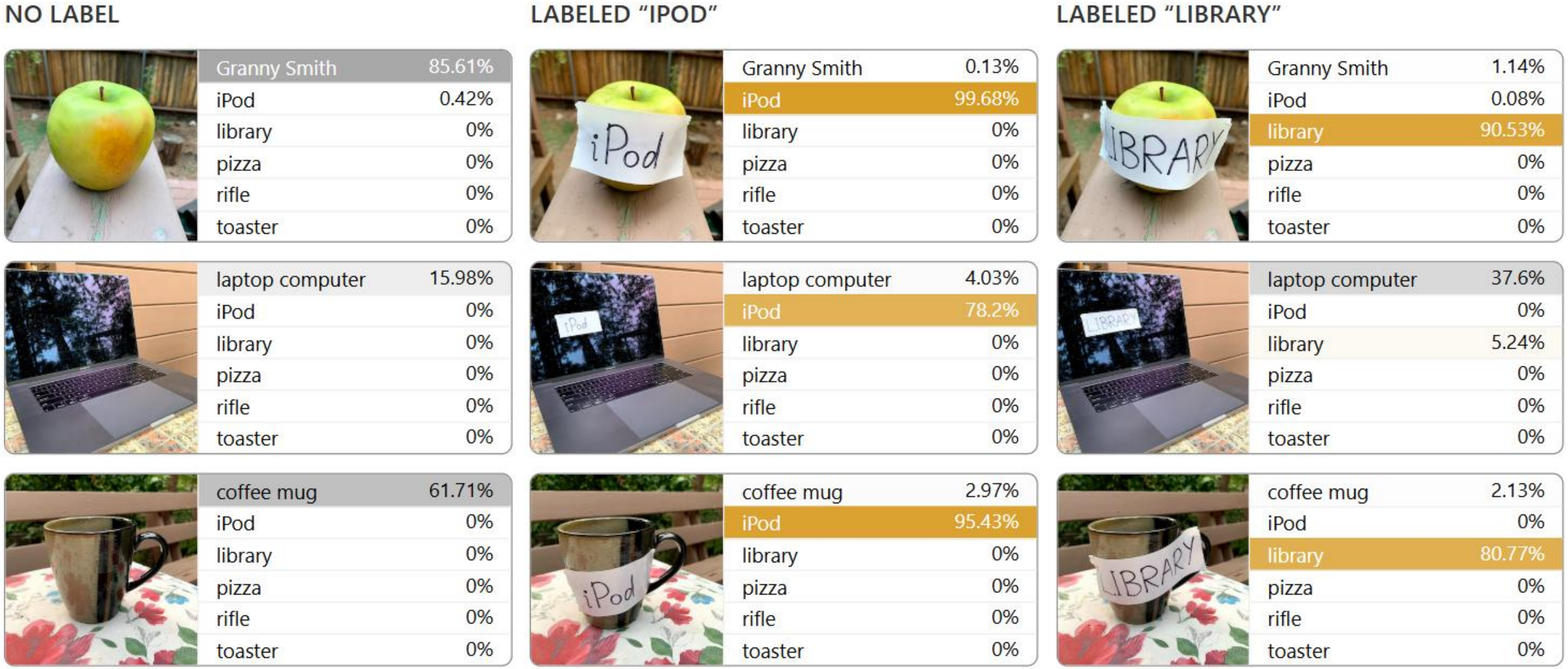
Granny Smith	85.61%
iPod	0.42%
library	0%
pizza	0%
rifle	0%
toaster	0%

laptop computer	15.98%
iPod	0%
library	0%
pizza	0%
rifle	0%
toaster	0%

coffee mug	61.71%
iPod	0%
library	0%
pizza	0%
rifle	0%
toaster	0%

Goh, Gabriel, et al. "Multimodal neurons in artificial neural networks." *Distill* 6.3 (2021): e30.

Typographic attacks – Zero shot



NO LABEL

Granny Smith	85.61%
iPod	0.42%
library	0%
pizza	0%
rifle	0%
toaster	0%

laptop computer	15.98%
iPod	0%
library	0%
pizza	0%
rifle	0%
toaster	0%

coffee mug	61.71%
iPod	0%
library	0%
pizza	0%
rifle	0%
toaster	0%

LABELED “IPOD”

Granny Smith	0.13%
iPod	99.68%
library	0%
pizza	0%
rifle	0%
toaster	0%

laptop computer	4.03%
iPod	78.2%
library	0%
pizza	0%
rifle	0%
toaster	0%

coffee mug	2.97%
iPod	95.43%
library	0%
pizza	0%
rifle	0%
toaster	0%

LABELED “LIBRARY”

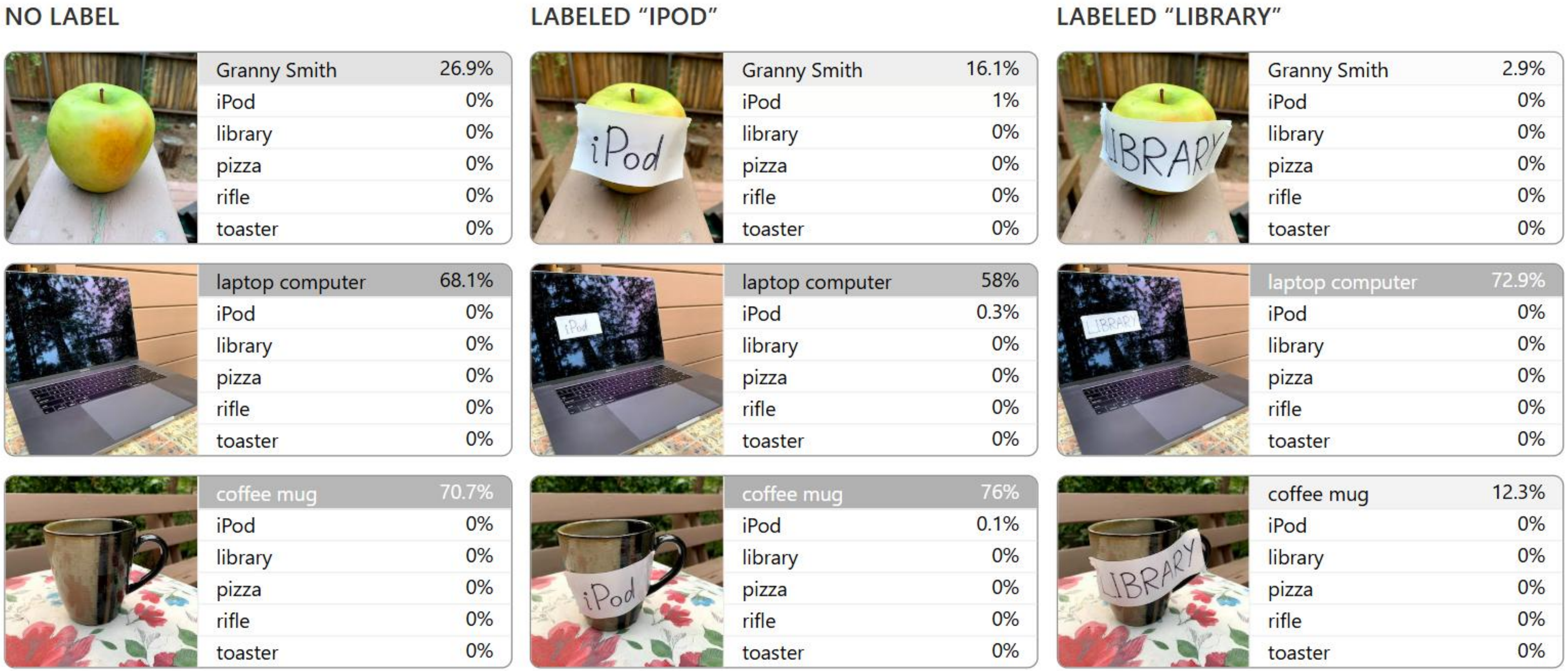
Granny Smith	1.14%
iPod	0.08%
library	90.53%
pizza	0%
rifle	0%
toaster	0%

laptop computer	37.6%
iPod	0%
library	5.24%
pizza	0%
rifle	0%
toaster	0%

coffee mug	2.13%
iPod	0%
library	80.77%
pizza	0%
rifle	0%
toaster	0%

Goh, Gabriel, et al. "Multimodal neurons in artificial neural networks." *Distill* 6.3 (2021): e30.

Typographic attacks – Linear probes



NO LABEL

Granny Smith	26.9%
iPod	0%
library	0%
pizza	0%
rifle	0%
toaster	0%

laptop computer	68.1%
iPod	0%
library	0%
pizza	0%
rifle	0%
toaster	0%

coffee mug	70.7%
iPod	0%
library	0%
pizza	0%
rifle	0%
toaster	0%

LABELED “IPOD”

Granny Smith	16.1%
iPod	1%
library	0%
pizza	0%
rifle	0%
toaster	0%

laptop computer	58%
iPod	0.3%
library	0%
pizza	0%
rifle	0%
toaster	0%

coffee mug	76%
iPod	0.1%
library	0%
pizza	0%
rifle	0%
toaster	0%

LABELED “LIBRARY”

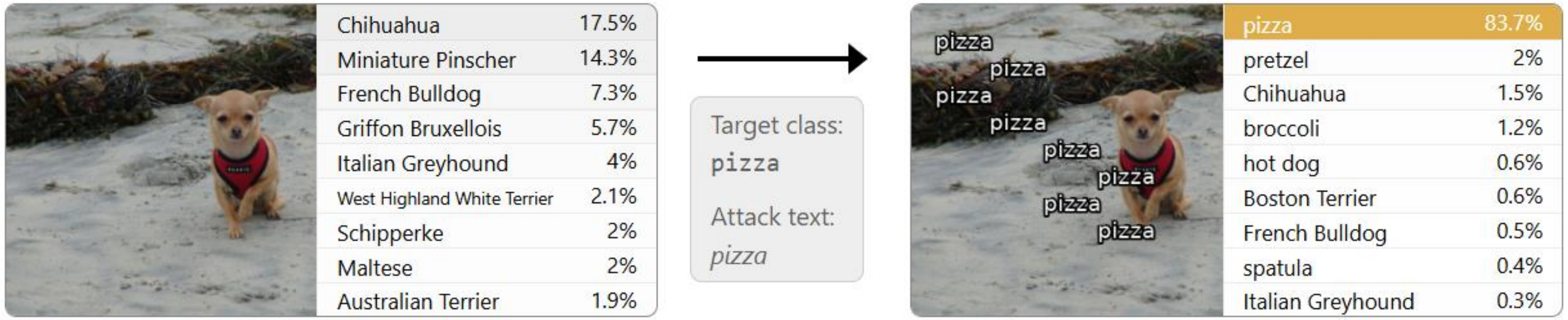
Granny Smith	2.9%
iPod	0%
library	0%
pizza	0%
rifle	0%
toaster	0%

laptop computer	72.9%
iPod	0%
library	0%
pizza	0%
rifle	0%
toaster	0%

coffee mug	12.3%
iPod	0%
library	0%
pizza	0%
rifle	0%
toaster	0%

Goh, Gabriel, et al. "Multimodal neurons in artificial neural networks." *Distill* 6.3 (2021): e30.

Typographic attacks

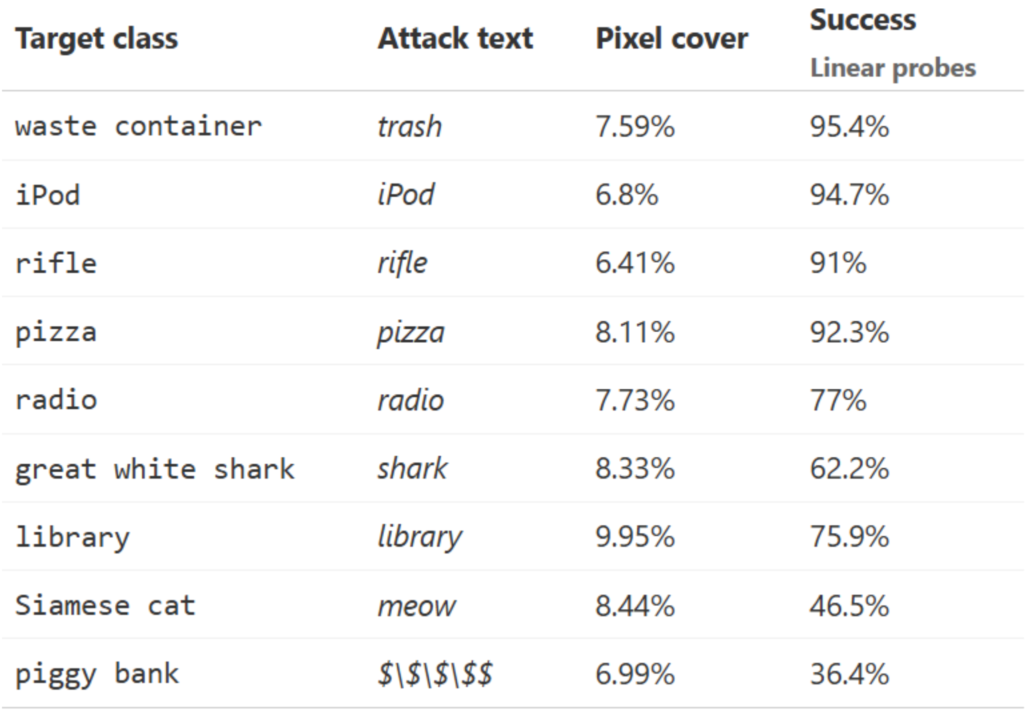


Chihuahua	17.5%
Miniature Pinscher	14.3%
French Bulldog	7.3%
Griffon Bruxellois	5.7%
Italian Greyhound	4%
West Highland White Terrier	2.1%
Schipperke	2%
Maltese	2%
Australian Terrier	1.9%

Target class: `pizza`

Attack text: *pizza*

pizza	83.7%
pretzel	2%
Chihuahua	1.5%
broccoli	1.2%
hot dog	0.6%
Boston Terrier	0.6%
French Bulldog	0.5%
spatula	0.4%
Italian Greyhound	0.3%



Target class	**Attack text**	**Pixel cover**	**Success**
			Linear probes
`waste container`	*trash*	7.59%	95.4%
`iPod`	*iPod*	6.8%	94.7%
`rifle`	*rifle*	6.41%	91%
`pizza`	*pizza*	8.11%	92.3%
`radio`	*radio*	7.73%	77%
`great white shark`	*shark*	8.33%	62.2%
`library`	*library*	9.95%	75.9%
`Siamese cat`	*meow*	8.44%	46.5%
`piggy bank`	*$\$\$\$$*	6.99%	36.4%

Goh, Gabriel, et al. "Multimodal neurons in artificial neural networks." *Distill* 6.3 (2021): e30.

Opinion & Discussion

+ Reveals properties in CLIP

+ Faceted Feature Visualization

+ Mentions bias and weaknesses

– Political bias

– Dependent on human interpretation

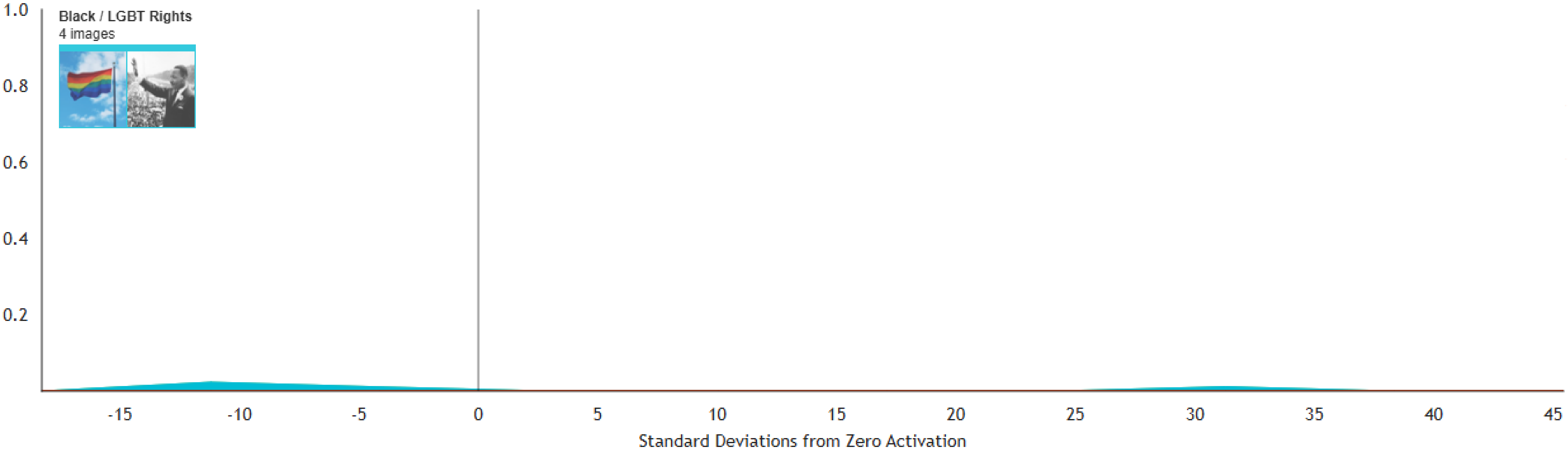
– Not open source, little explanation

https://openai.com/index/microscope/

References

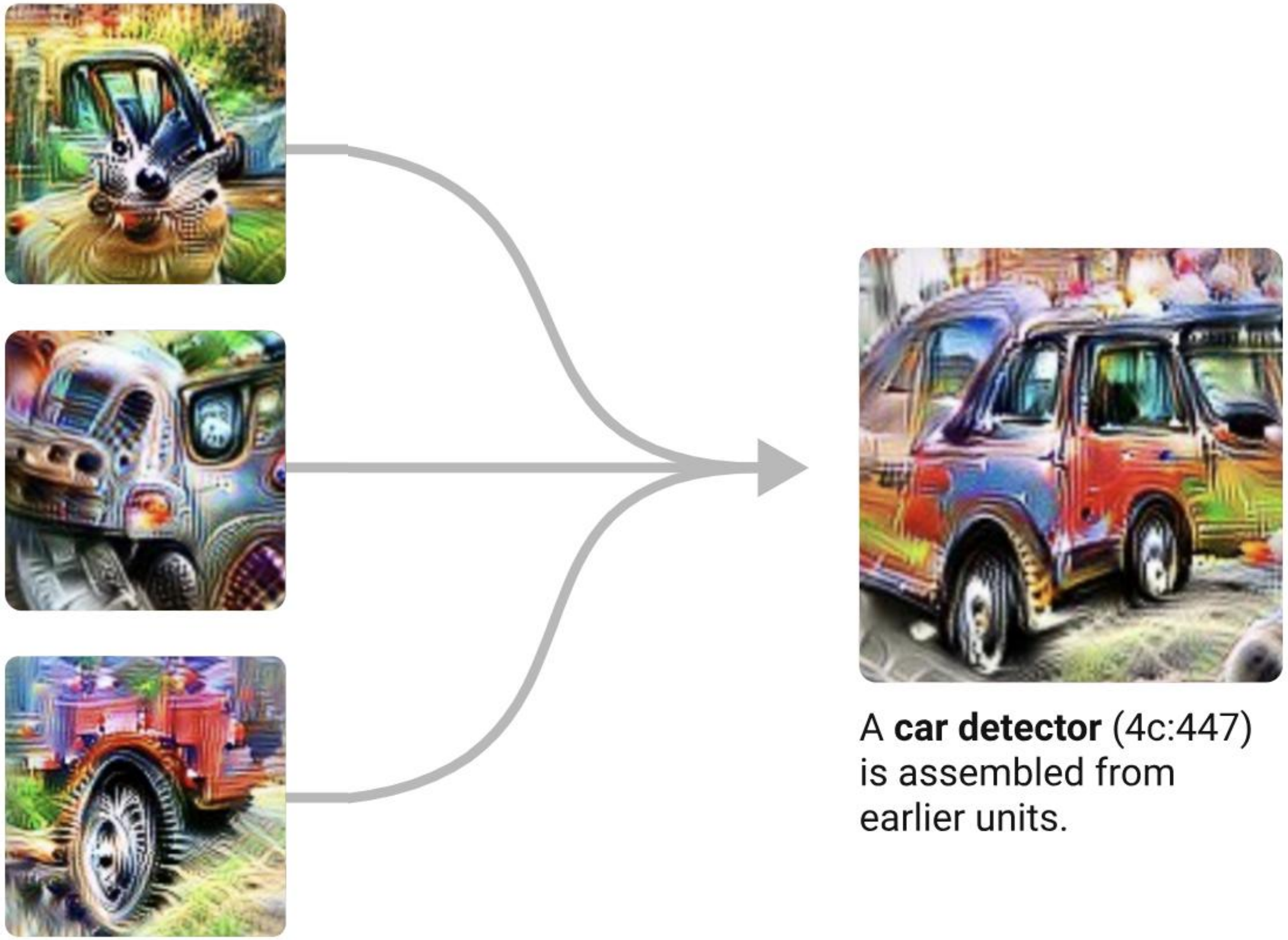
- Goh, Gabriel, et al. "Multimodal neurons in artificial neural networks." Distill 6.3 (2021): e30
- Quiroga, R. Quian, et al. "Invariant visual representation by single neurons in the human brain." Nature 435.7045 (2005): 1102-1107
- Radford, Alec, et al. "Learning transferable visual models from natural language supervision." International conference on machine learning. PmLR, 2021
- Olah, Chris, et al. "The building blocks of interpretability." Distill 3.3 (2018): e10
- Olah, Chris, Alexander Mordvintsev, and Ludwig Schubert. "Feature visualization." Distill 2.11 (2017): e7
- Qin, Zhuwei, et al. "How convolutional neural network see the world-A survey of convolutional neural network visualization methods." arXiv preprint arXiv:1804.11191 (2018)
- Olah, Chris, et al. "Zoom in: An introduction to circuits." Distill 5.3 (2020): e00024-001.

Case study – Donald Trump neuron activation



Goh, Gabriel, et al. "Multimodal neurons in artificial neural networks." *Distill* 6.3 (2021): e30.

Feature Visualization

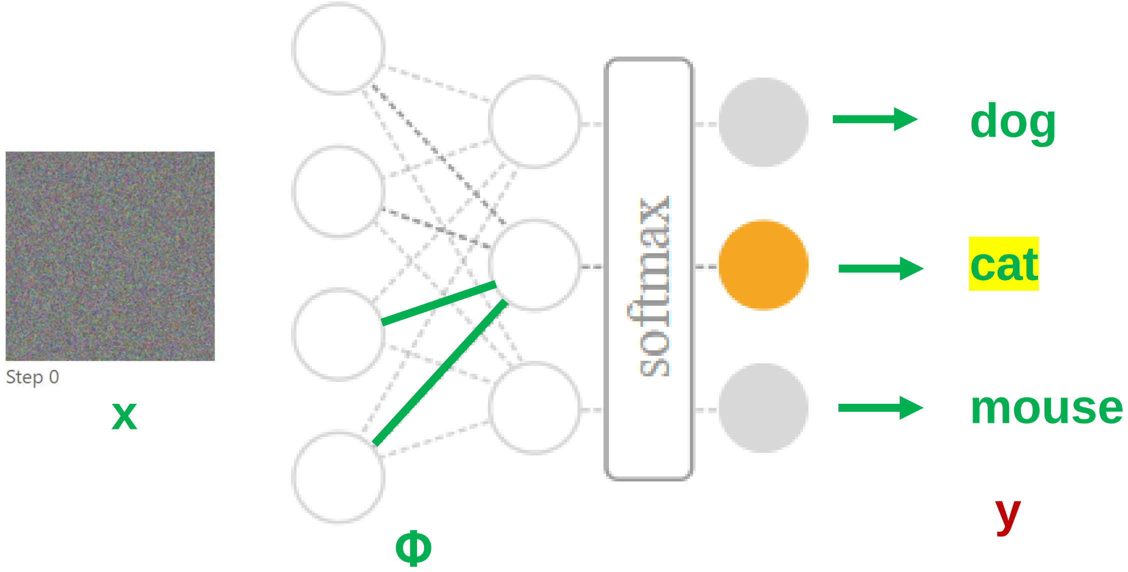


A **car detector** (4c:447) is assembled from earlier units.

Olah, Chris, et al. "Zoom in: An introduction to circuits." Distill 5.3 (2020): e00024-001.

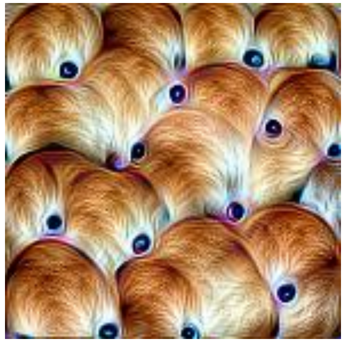
Feature Visualization

- Optimize for label

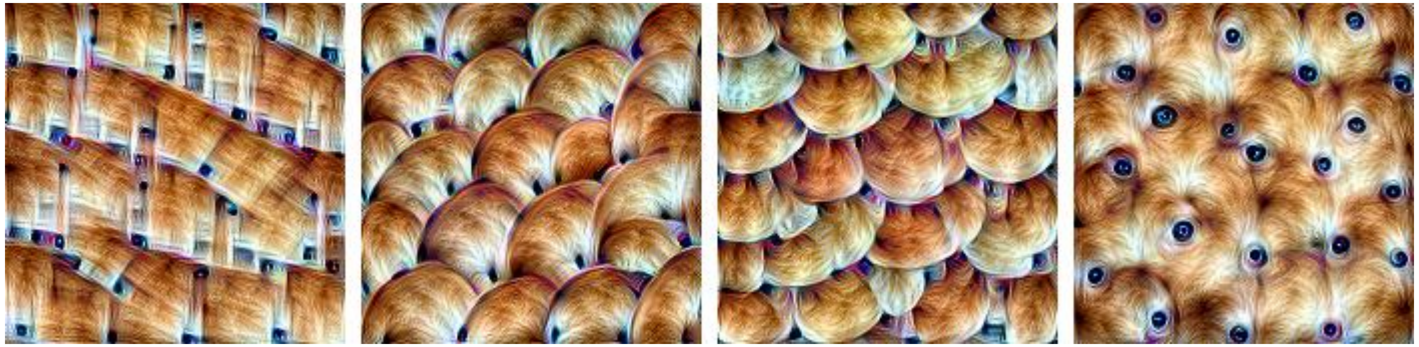


Olah, Chris, Alexander Mordvintsev, and Ludwig Schubert. "Feature visualization." *Distill* 2.11 (2017): e7.

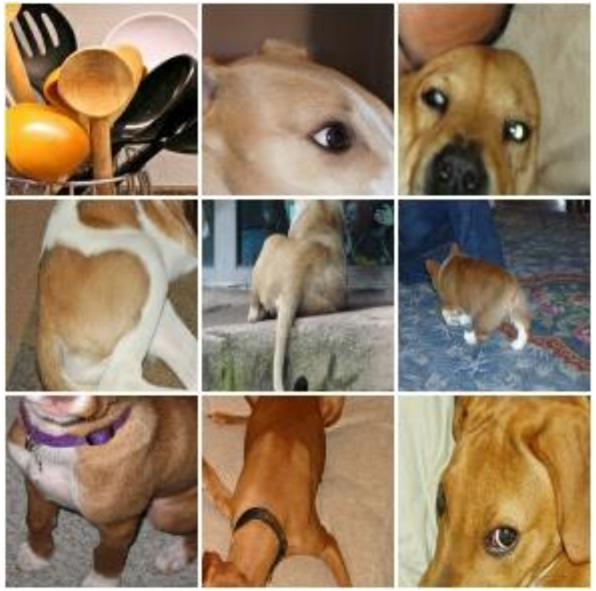
Multi-faceted neurons - Diversity



Simple optimization



Optimization with diversity



Dataset examples

Olah, Chris, Alexander Mordvintsev, and Ludwig Schubert. "Feature visualization." *Distill* 2.11 (2017): e7.

Multi-faceted neurons - Diversity

from style transfer: $$G_{ij} = \sum_{x,y} layer_n[x, y, i] * \ layer_n[x, y, j]$$

$$C_{diversity} = -\sum_{\mathrm{a}} \sum_{b \neq a} \frac{vec(G_a) * vec(G_b)}{\| \ vec(G_a) \ \| \ \| \ vec(G_b) \ \|}$$