SUMMIT

Scaling Deep Learning Interpretability by Visualizing Activation & Attribution Summarizations

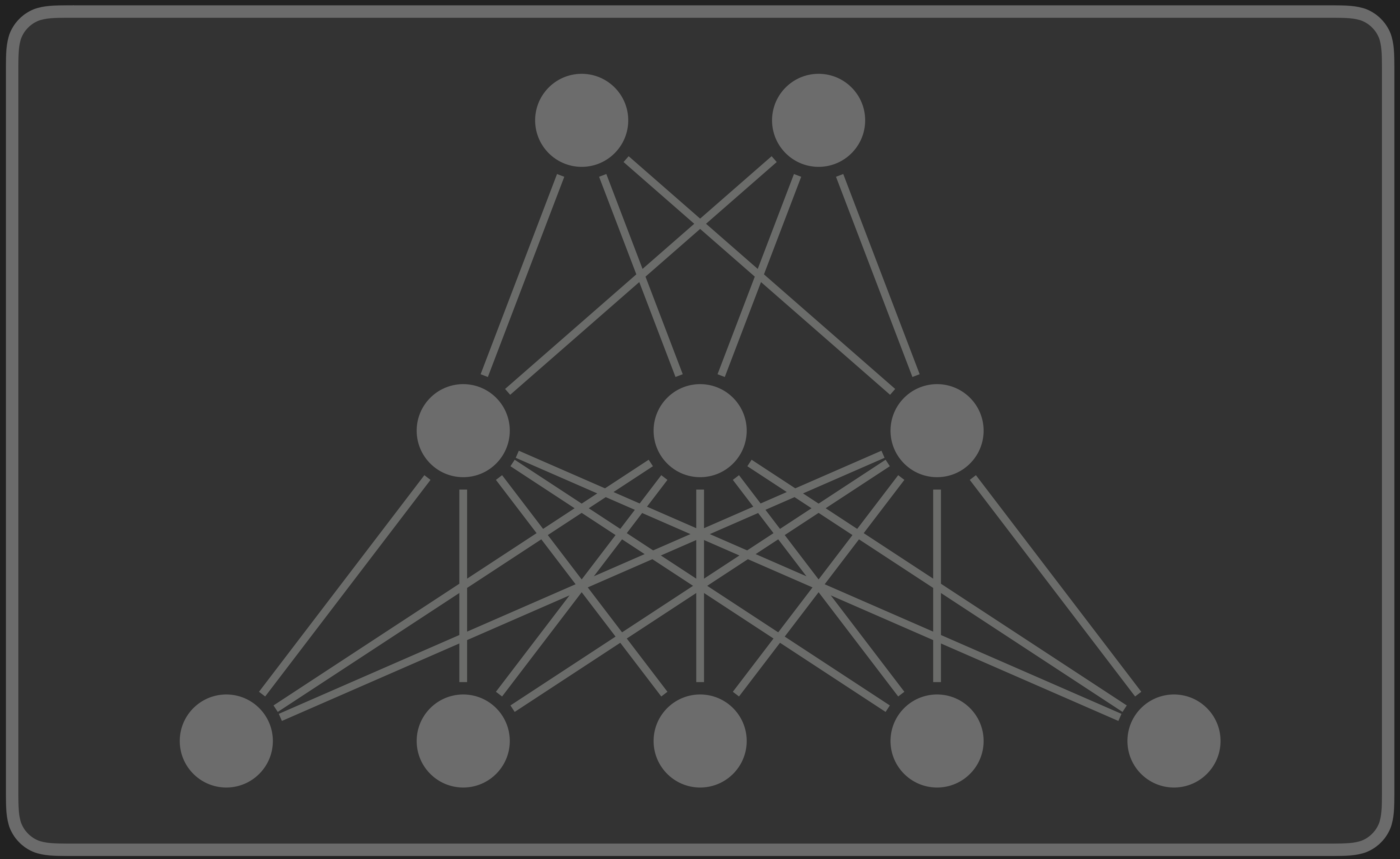
VAST 2019
Vancouver, Canada

Fred Hohman
@fredhohman
Georgia Tech

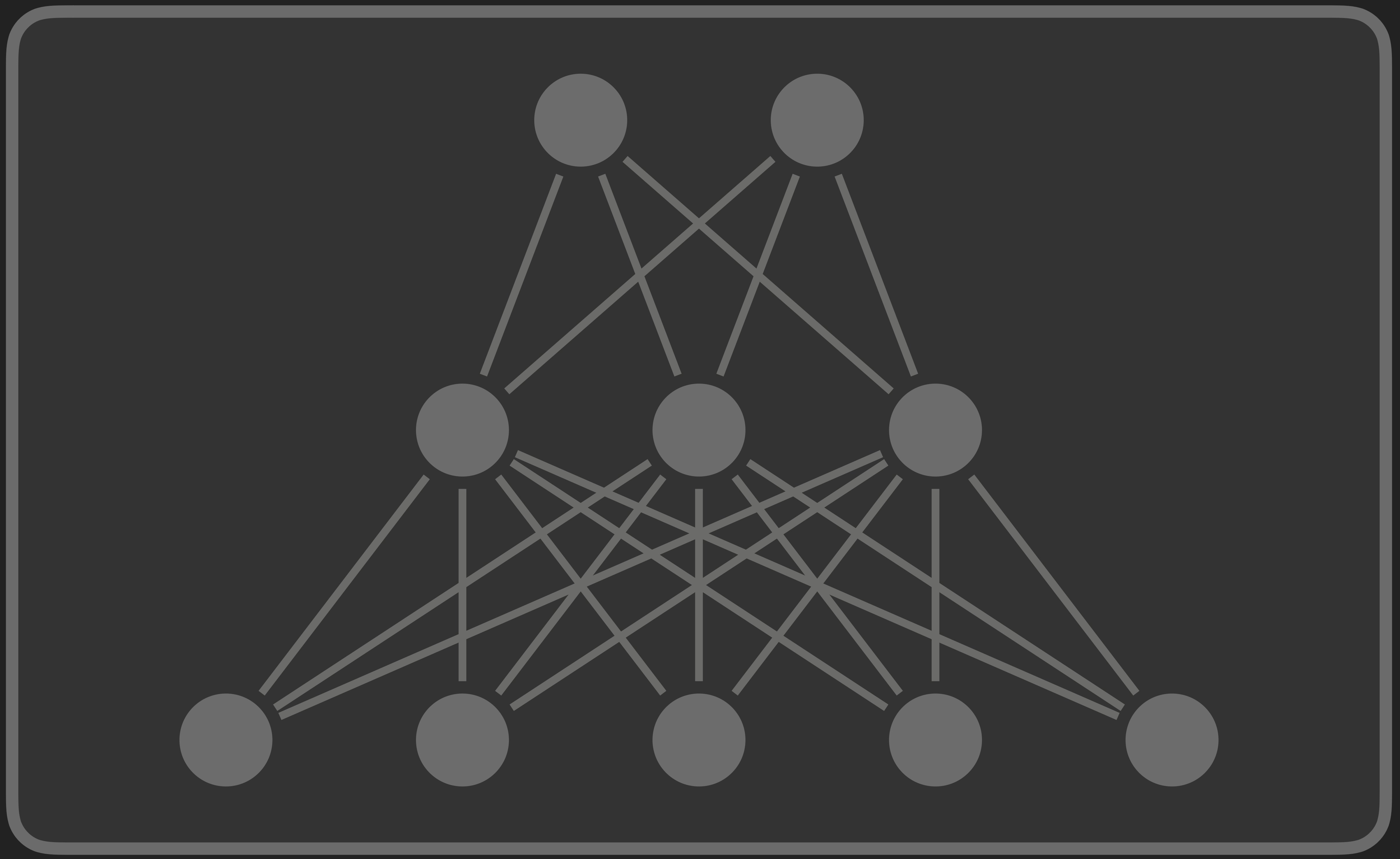
Haekyu Park
Georgia Tech

Polo Chau
Georgia Tech

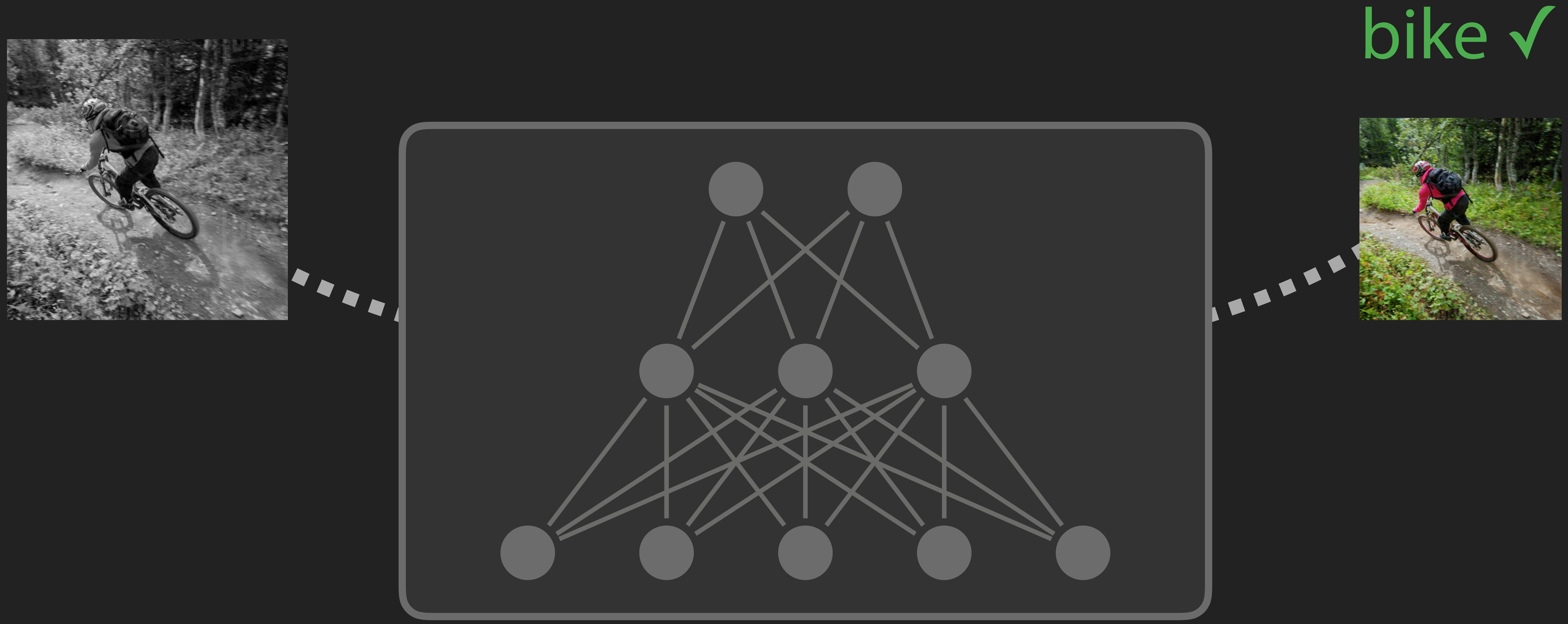
Caleb Robinson
Georgia Tech



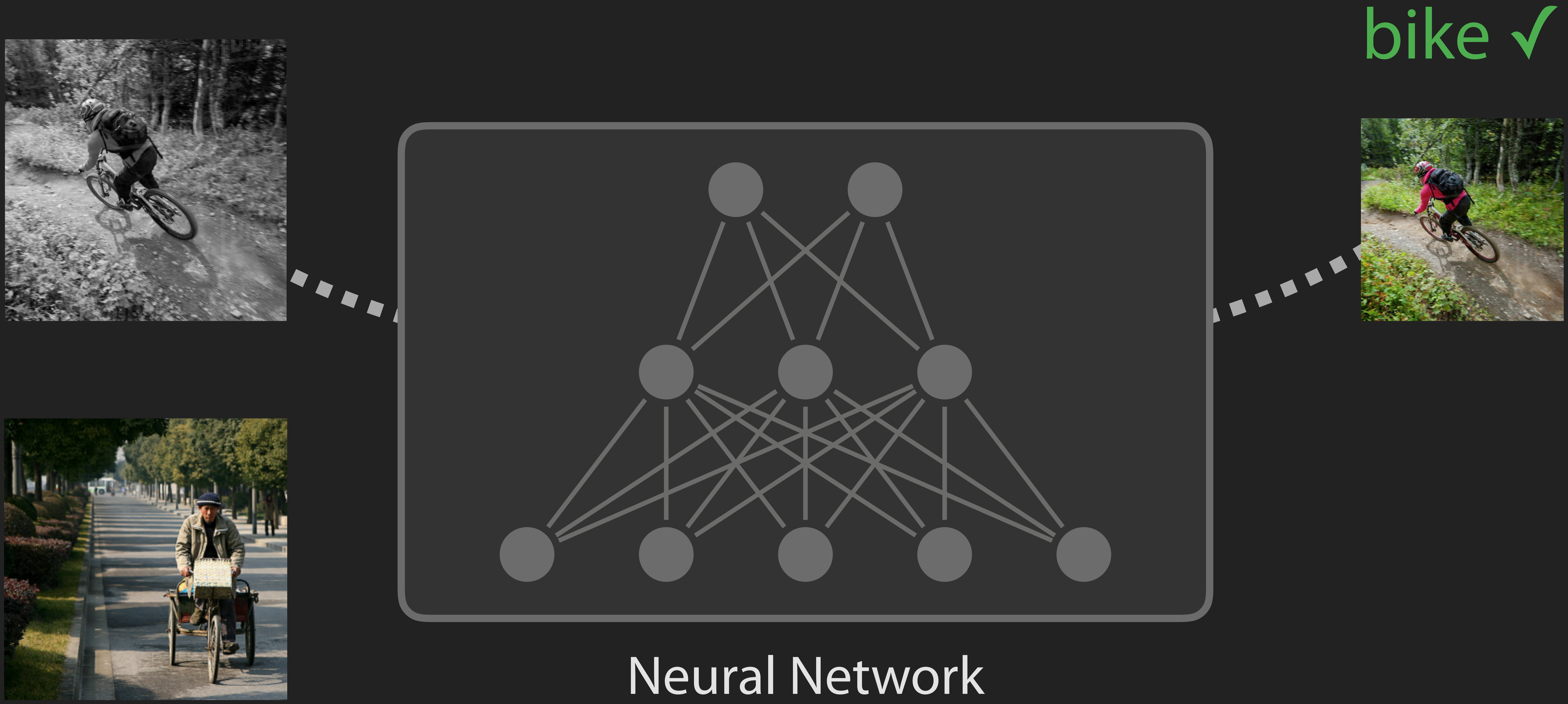
Neural Network

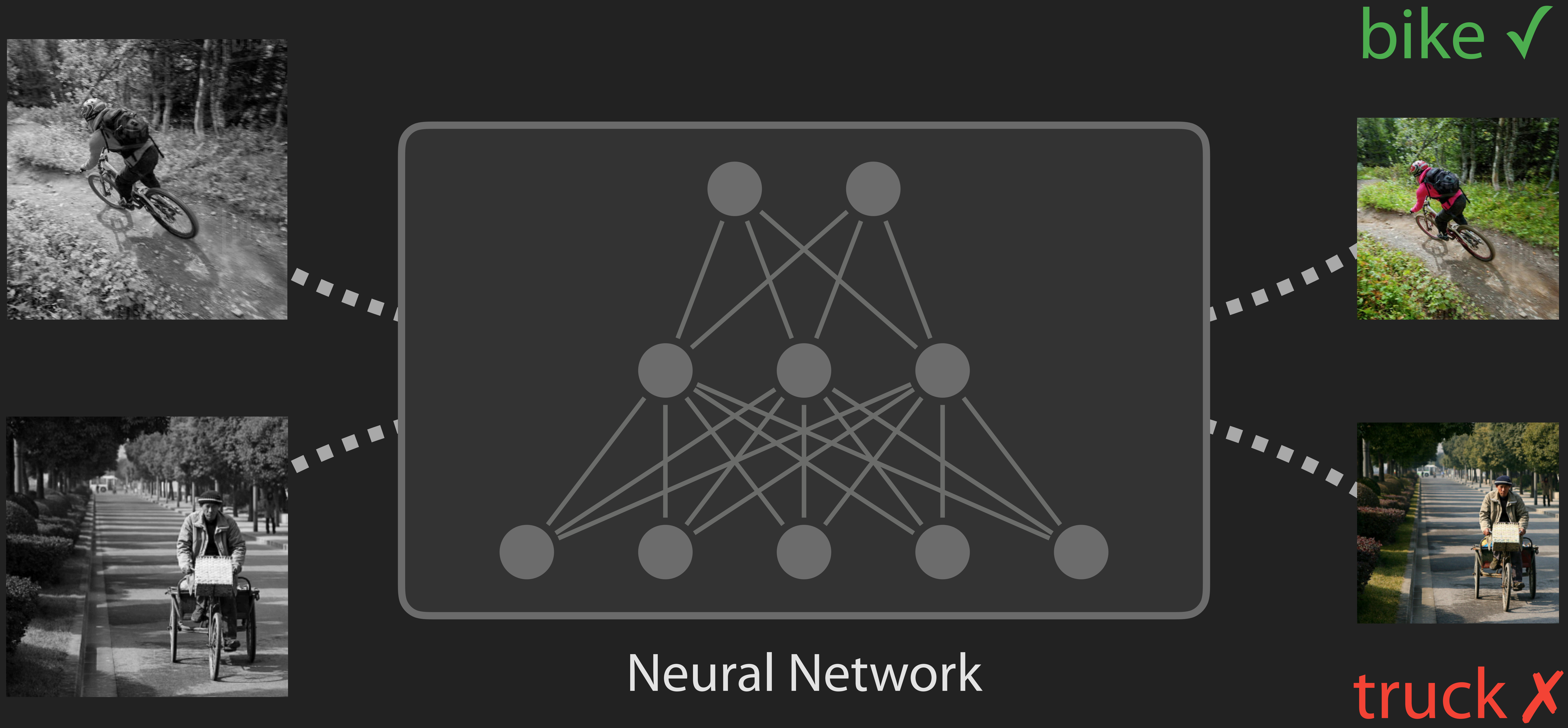


Neural Network



Neural Network


bike ✓
Neural Network


bike ✓
Neural Network
truck ✗

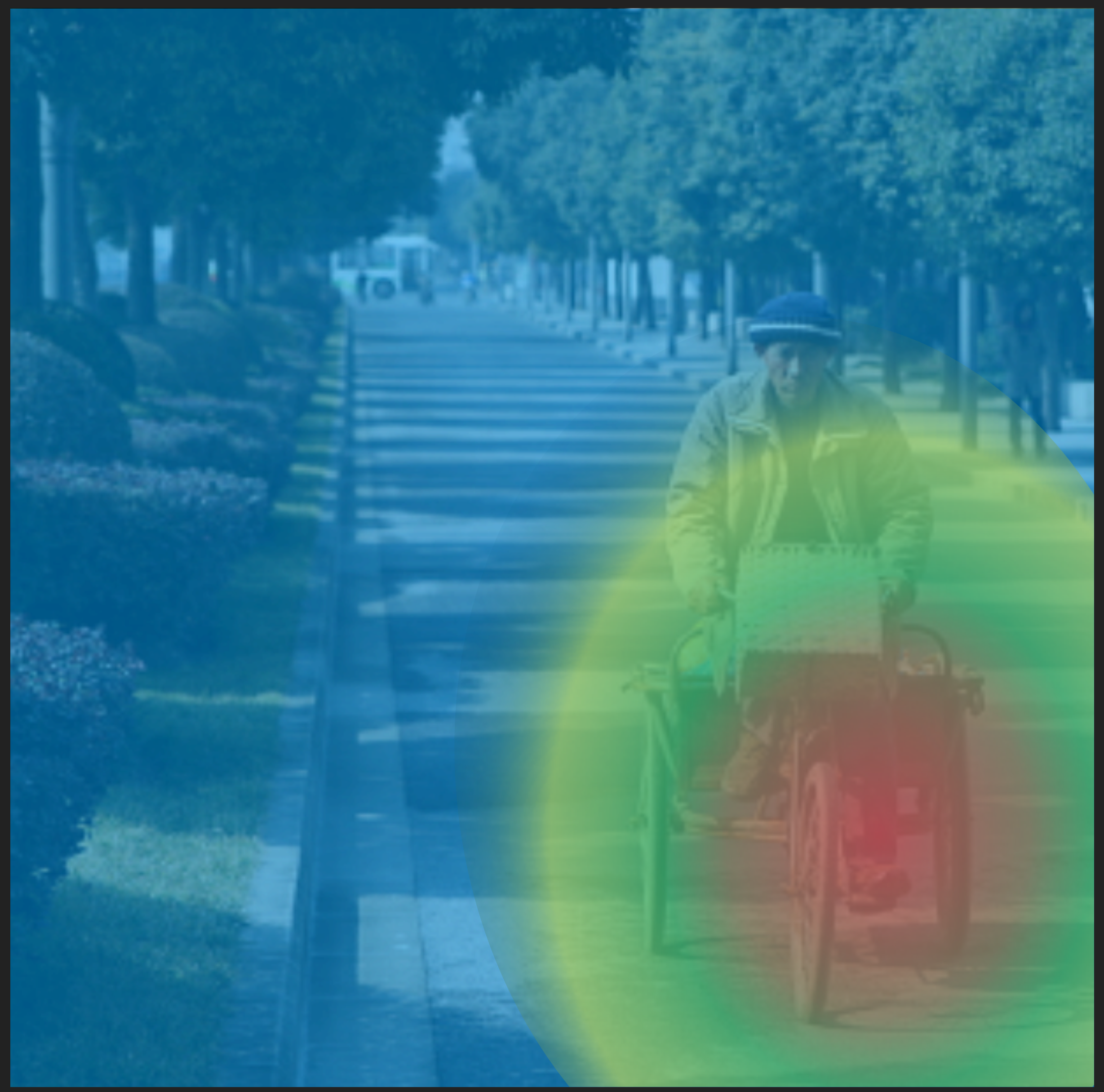
?

truck ✗

?

truck ✗

Attention

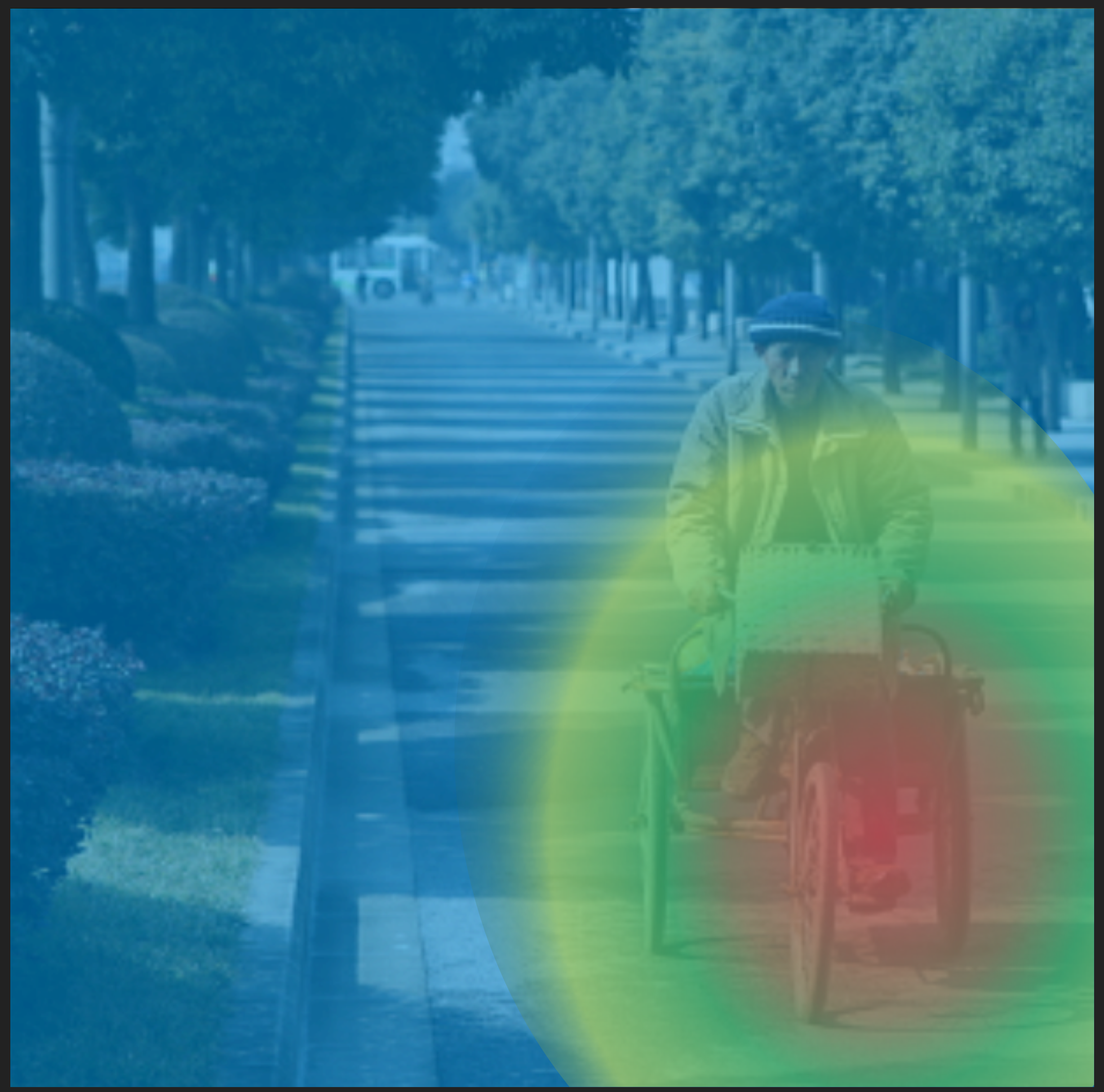


[Selvaraju, et al., ICCV, 2017]

?

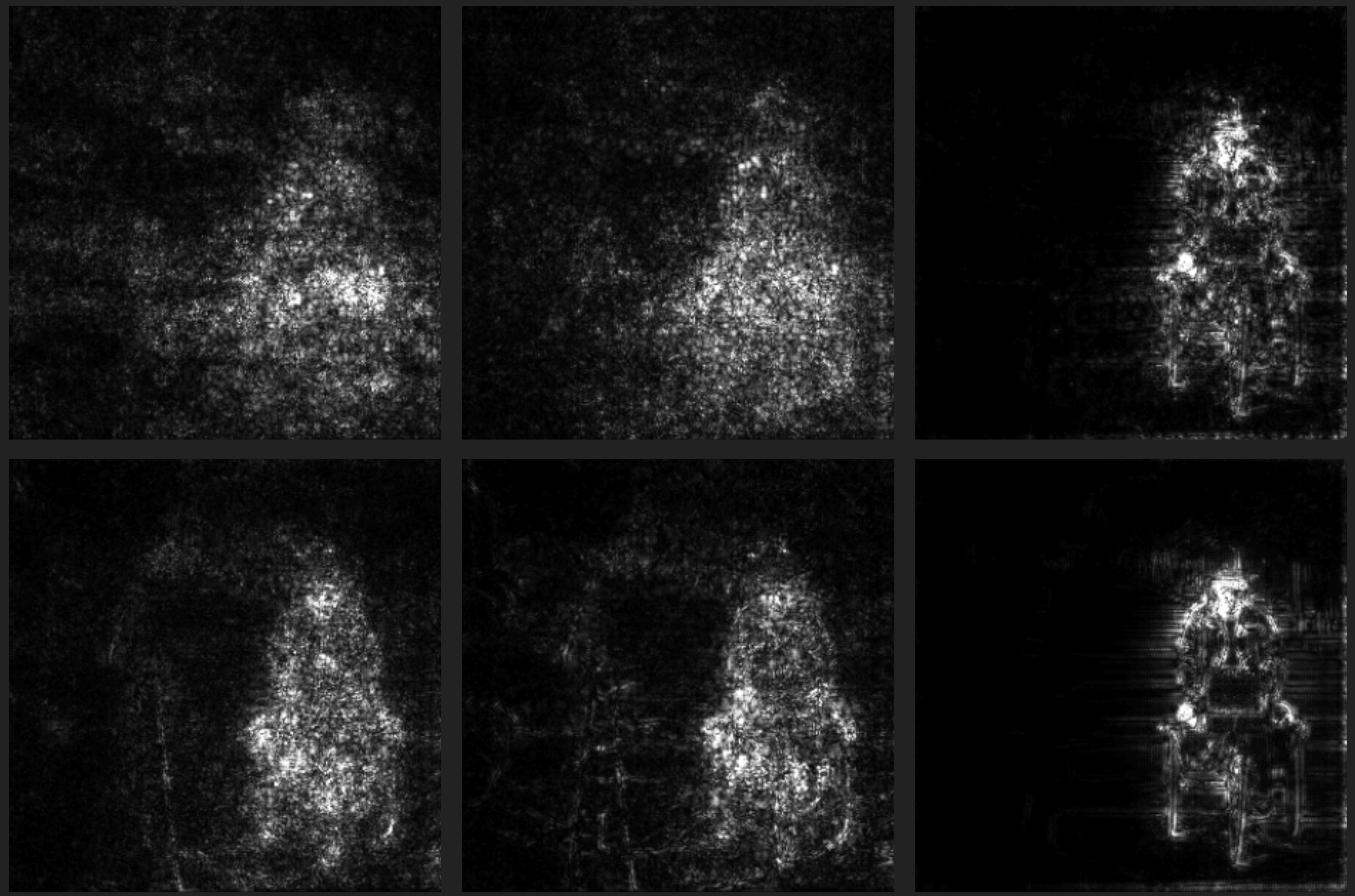
truck ✗

Attention



[Selvaraju, et al., ICCV, 2017]

Sensitivity



[Smilkov, et al., arXiv, 2017]


https://cs.stanford.edu/people/karpathy/cnnembed/

https://cs.stanford.edu/people/karpathy/cnnembed/

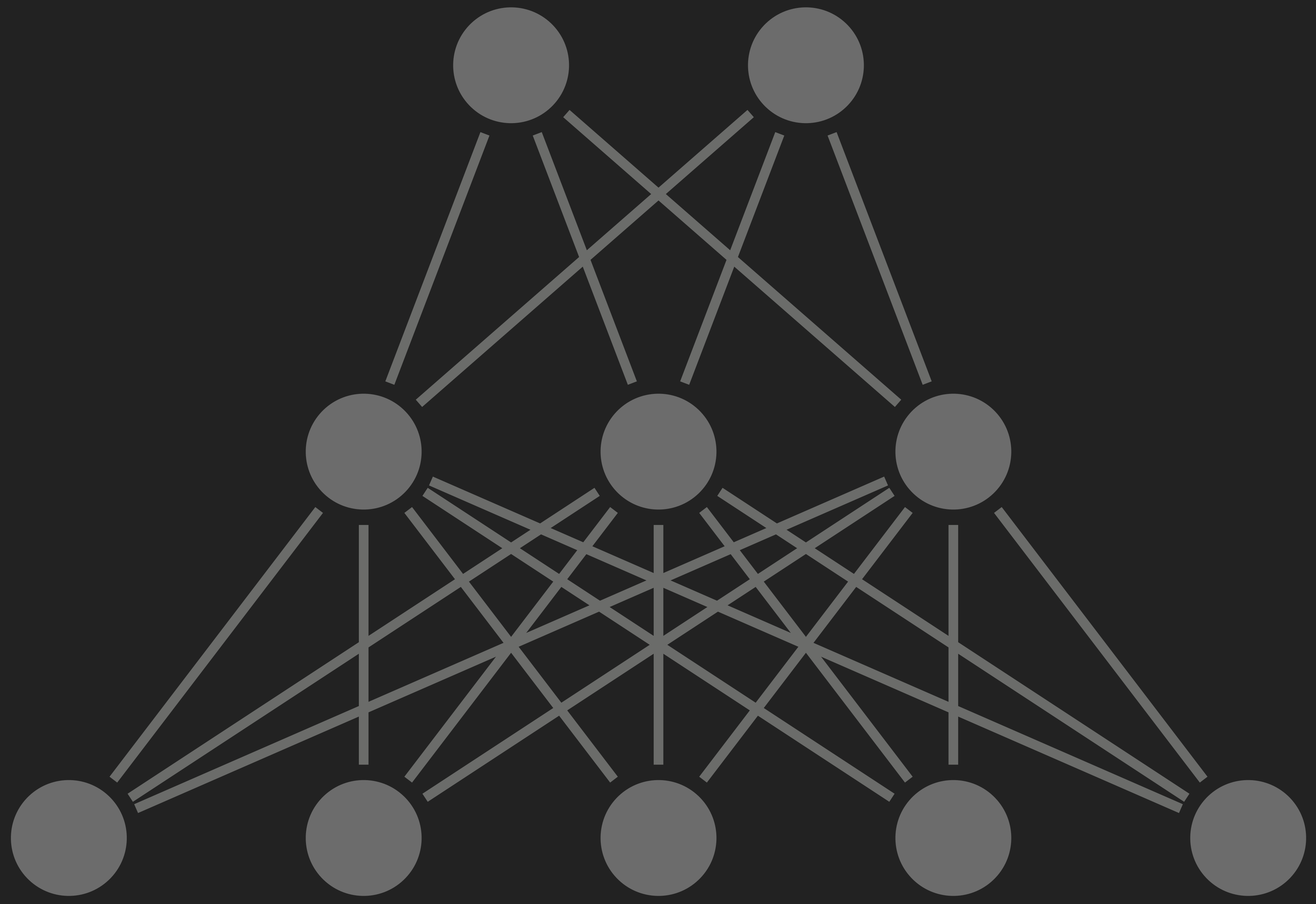
Summit

Scalably summarize and **interactively visualize** neural network feature representations for millions of images

Summit

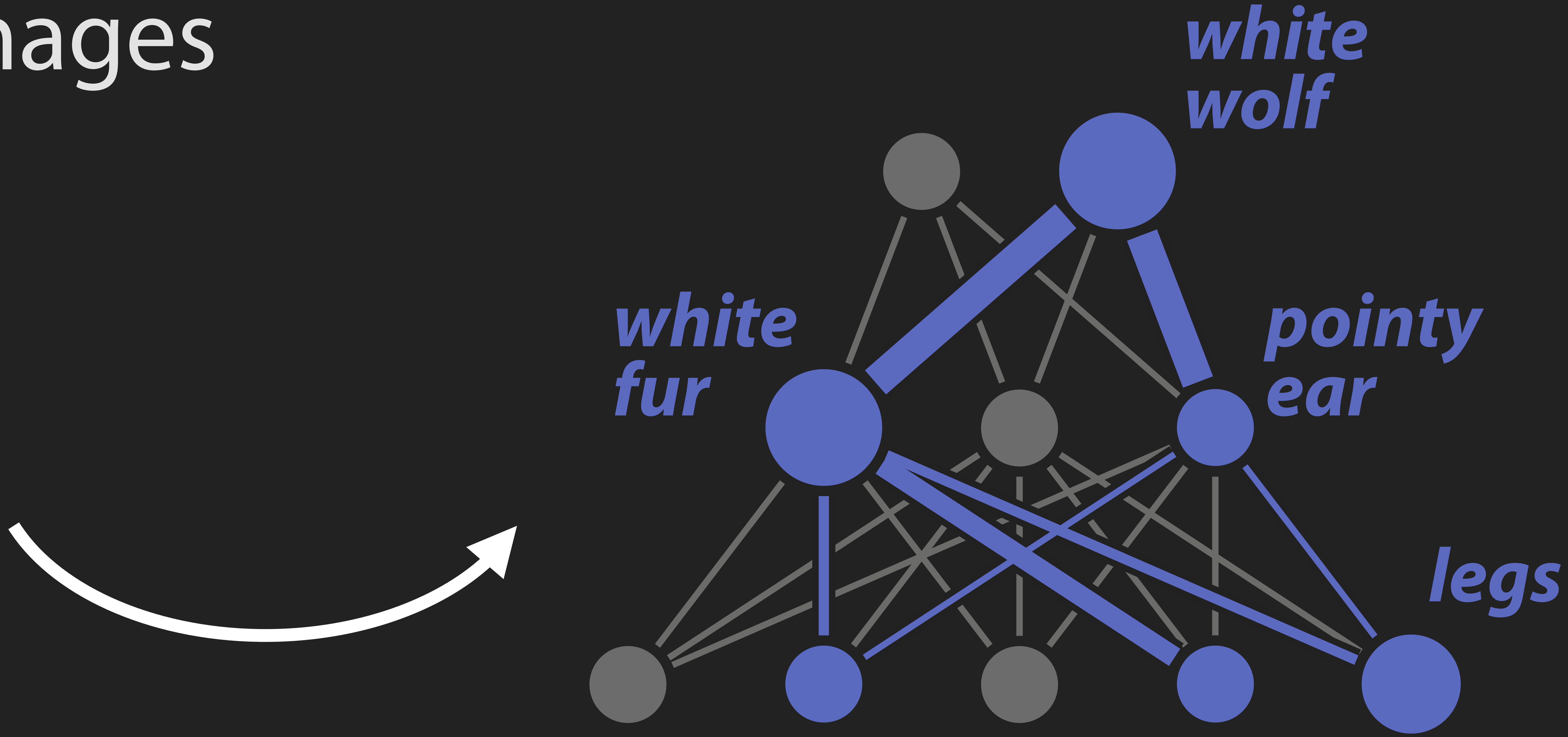
Scalably summarize and **interactively visualize** neural network feature representations for millions of images

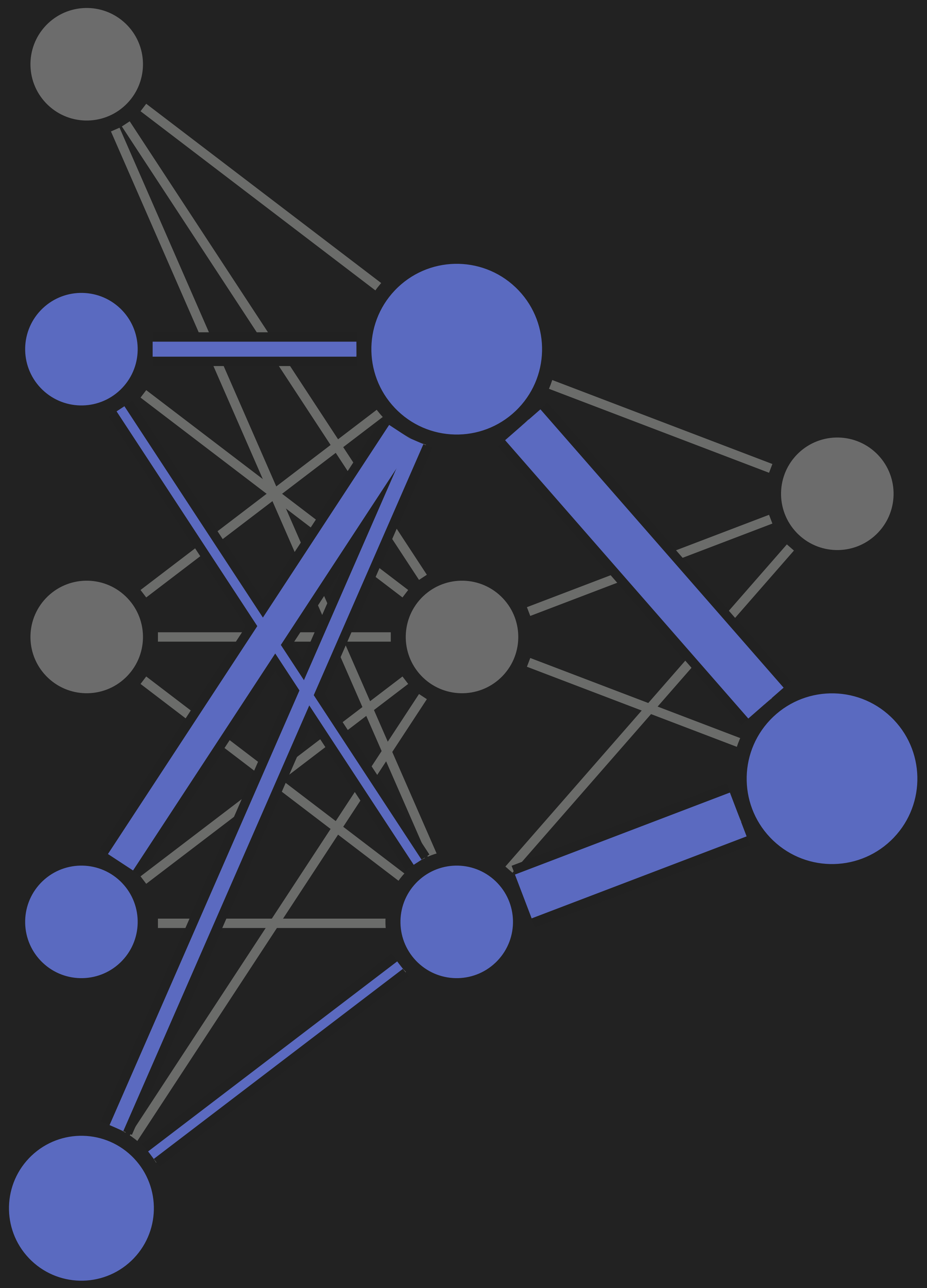
white wolf



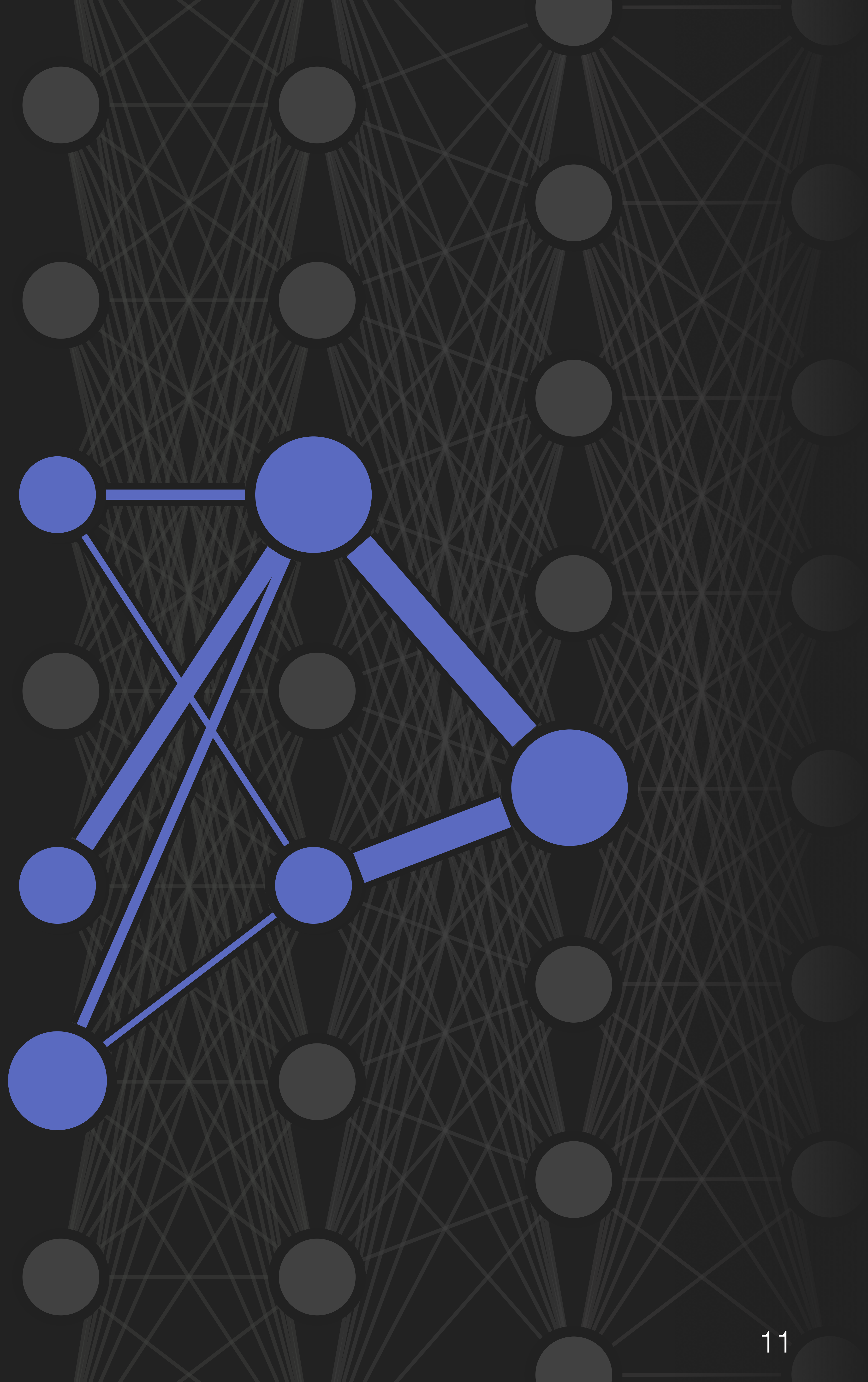
SUMMIT

Scalably summarize and **interactively visualize** neural network feature representations for millions of images

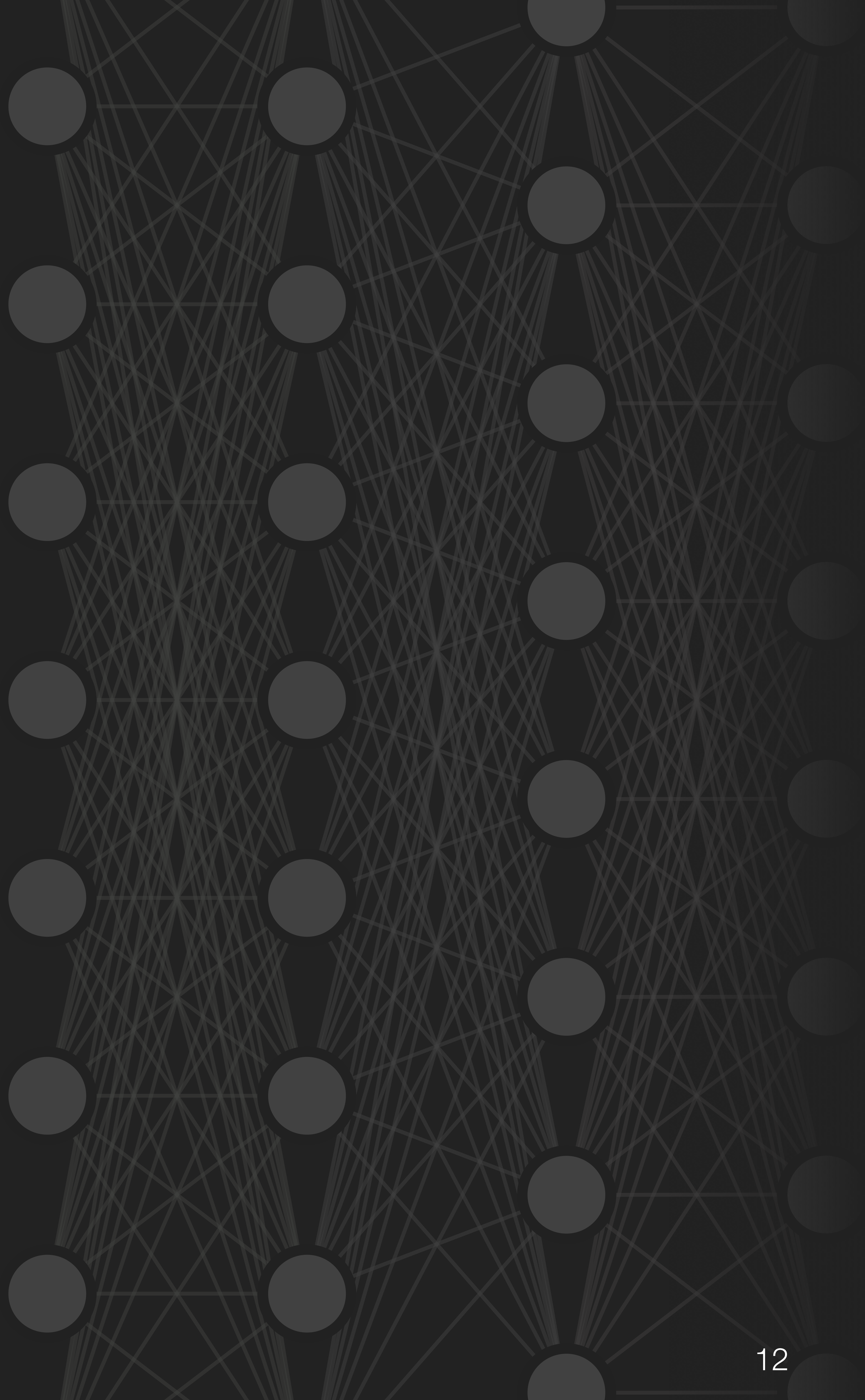




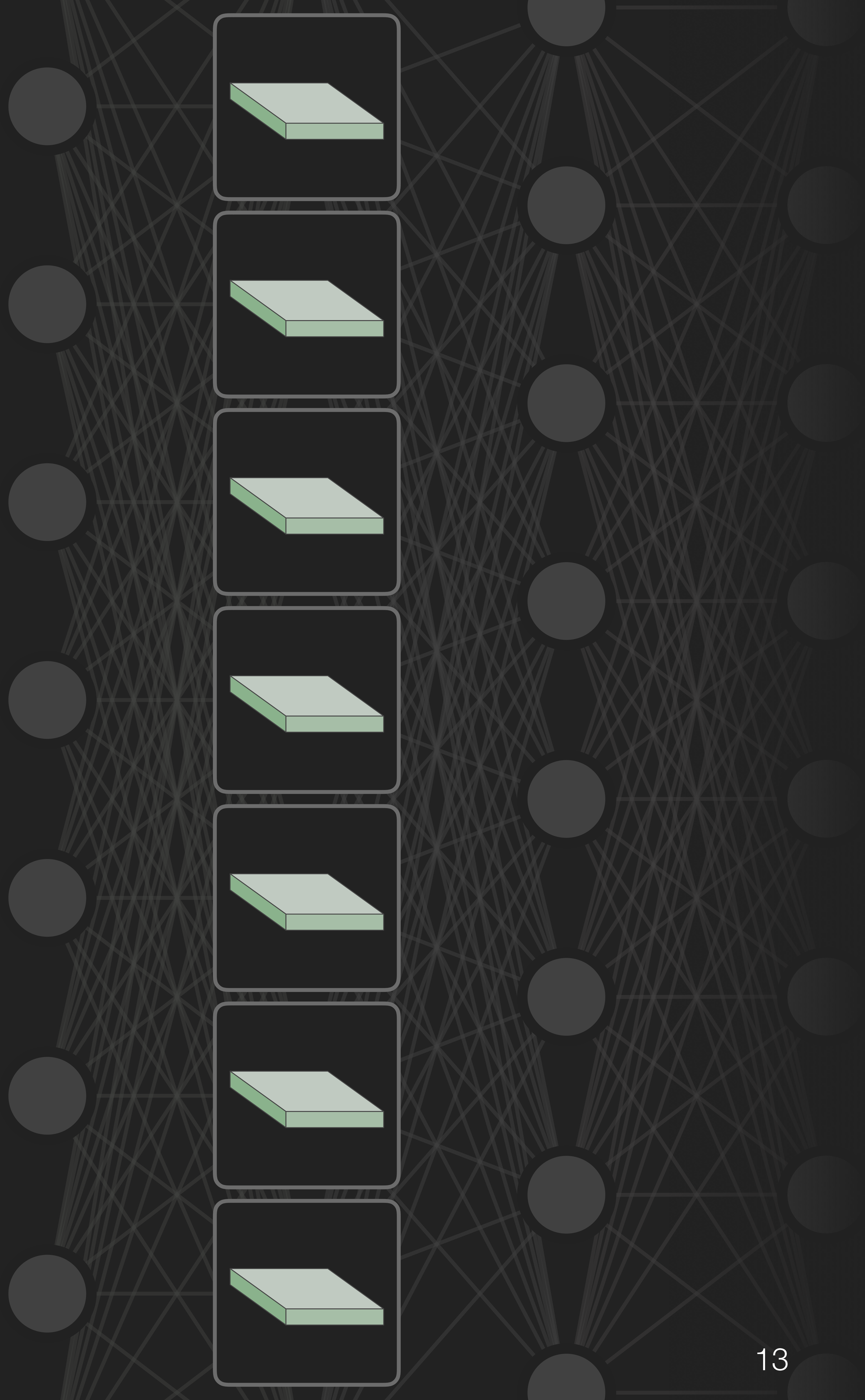
How do we make **attribution graphs**?



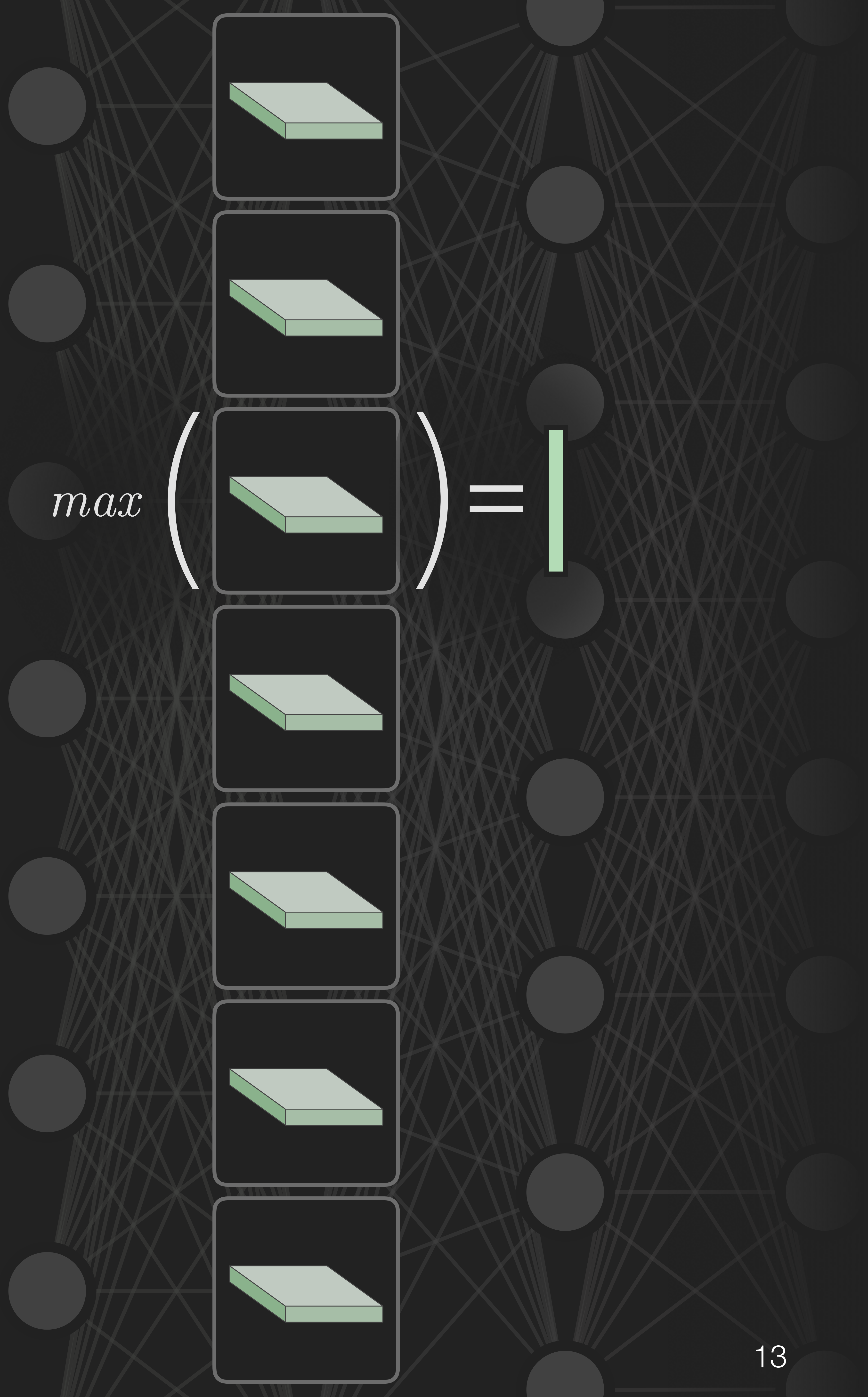
How do we make **attribution graphs**?



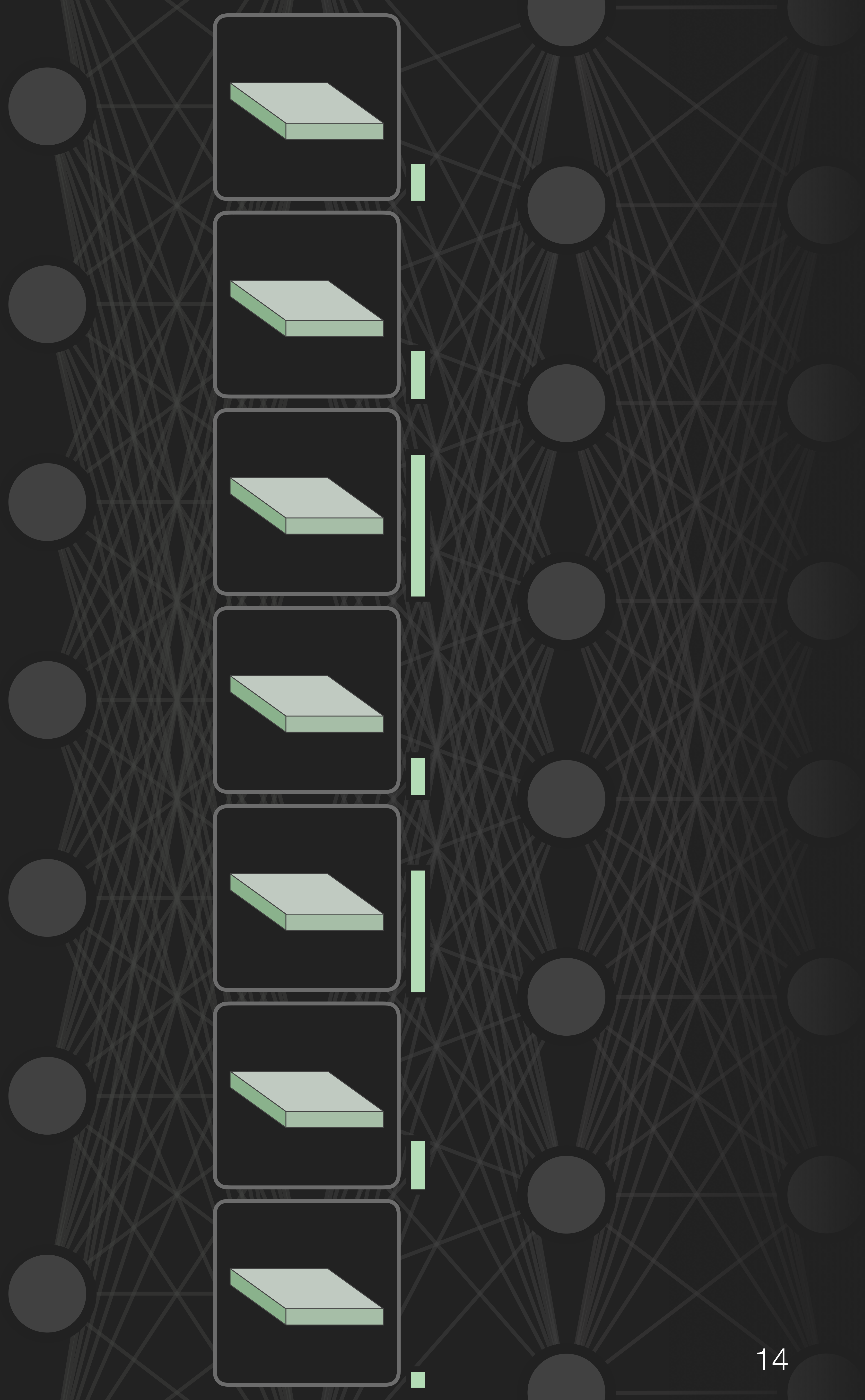
How do we make **attribution graphs**?



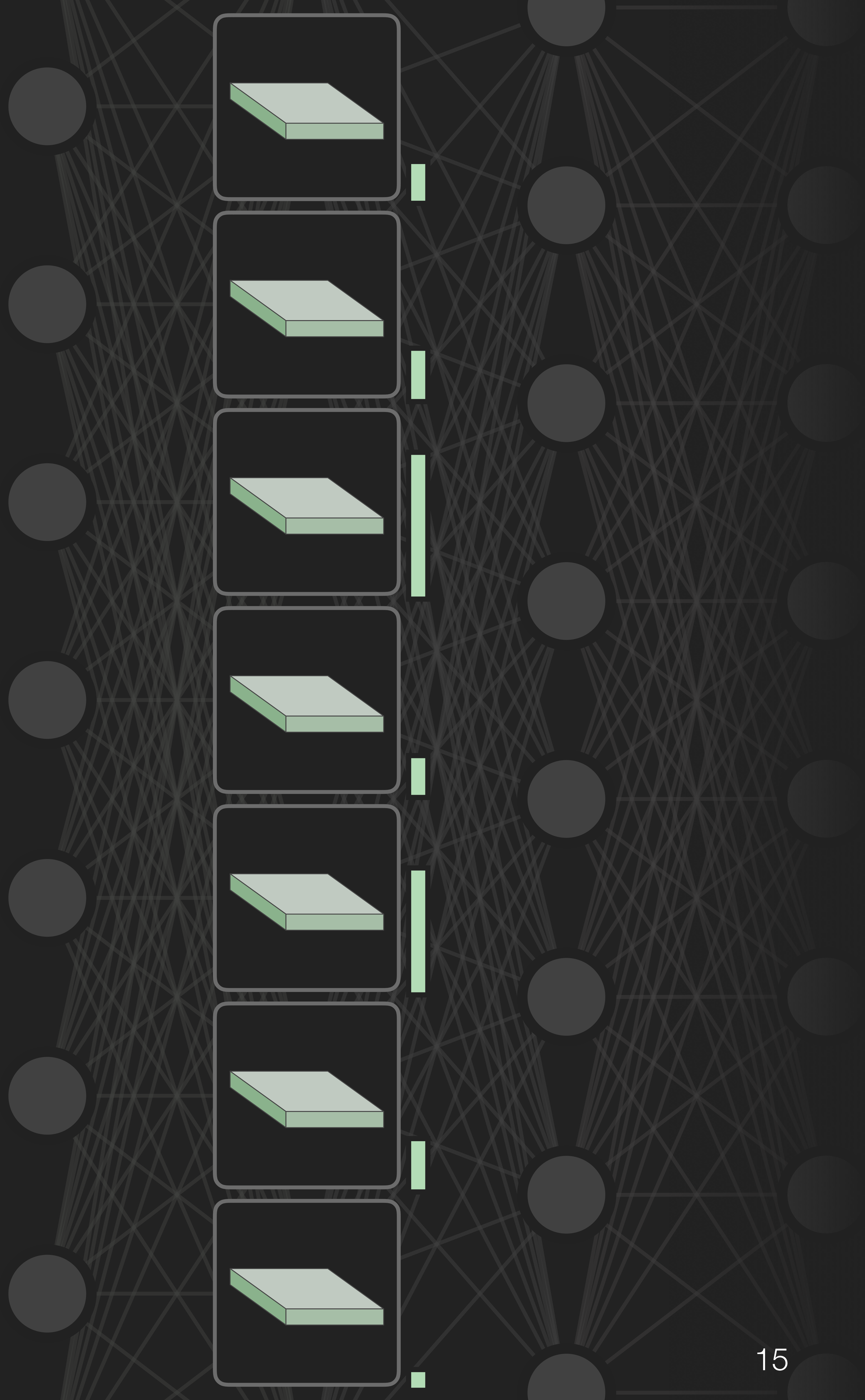
Aggregate network
activations (nodes)

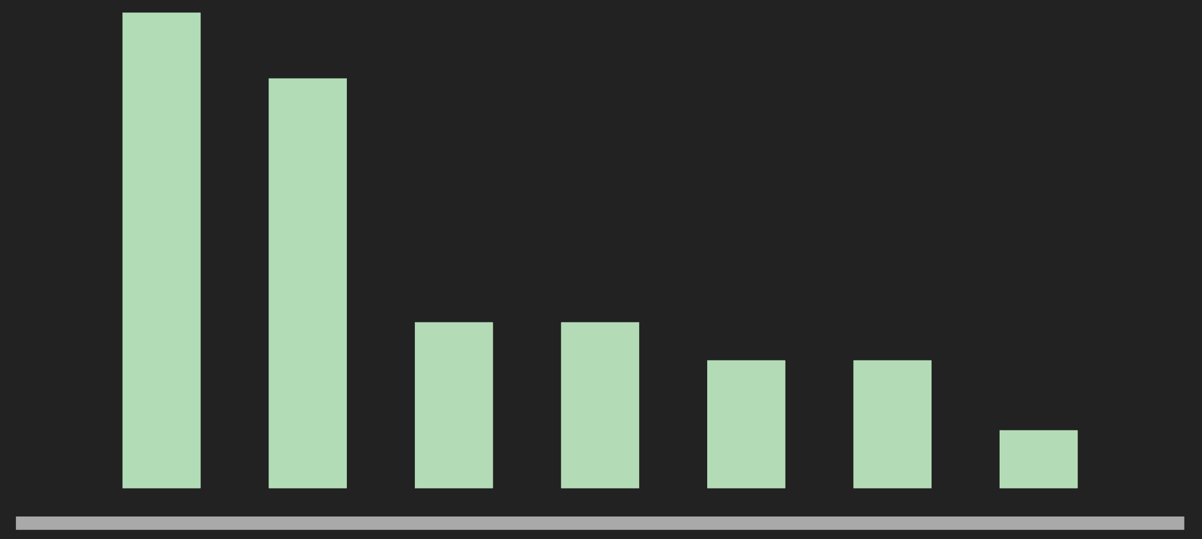


Aggregate network
activations (nodes)

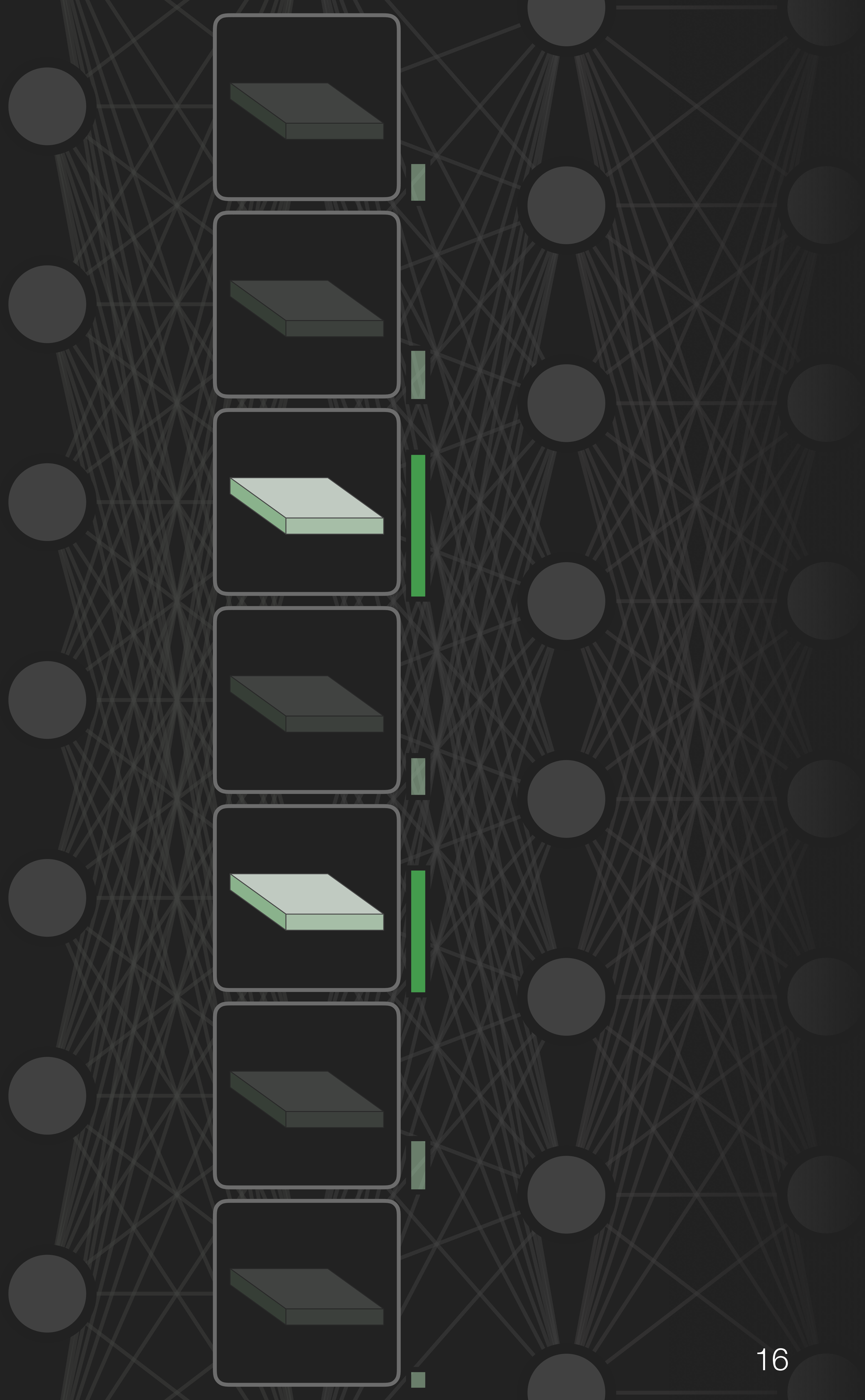


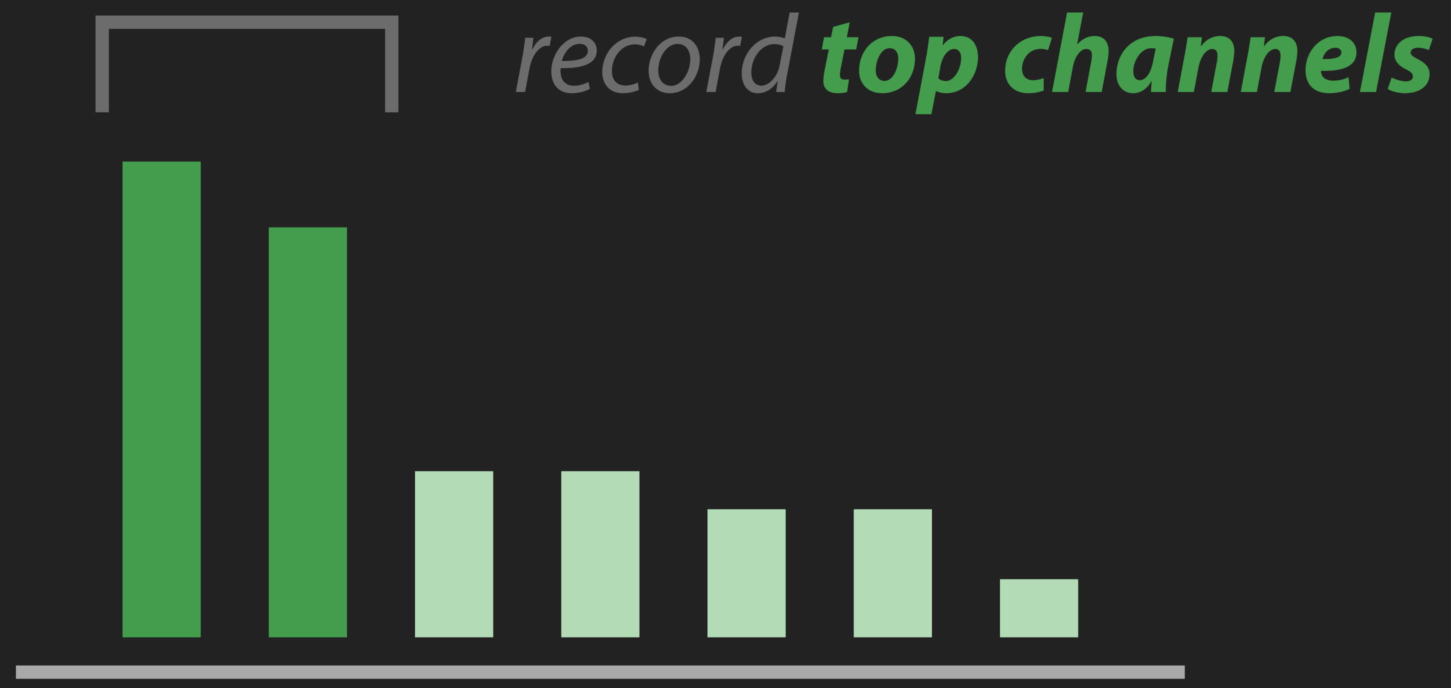
Aggregate network
activations (nodes)



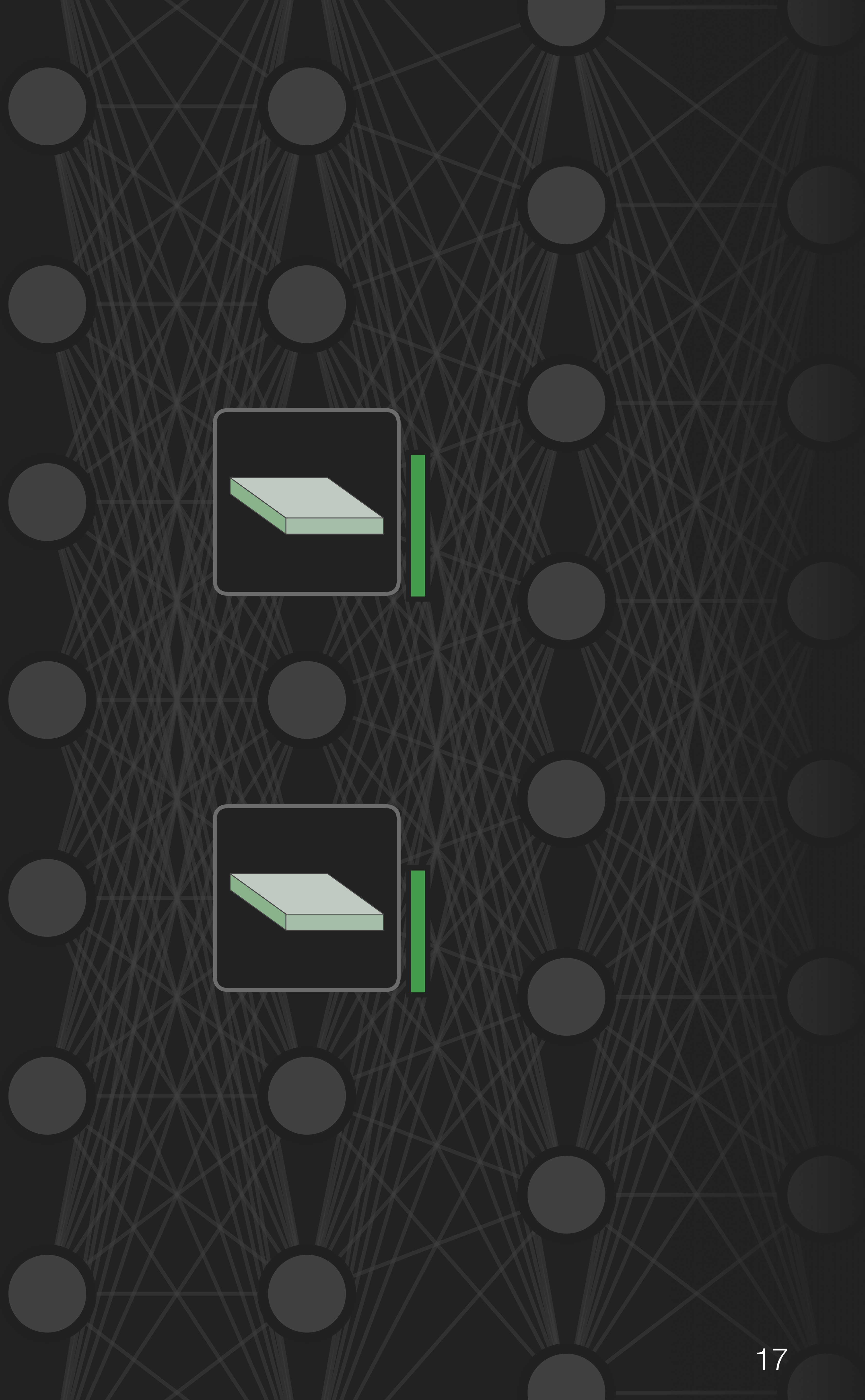


Aggregate network
activations (nodes)

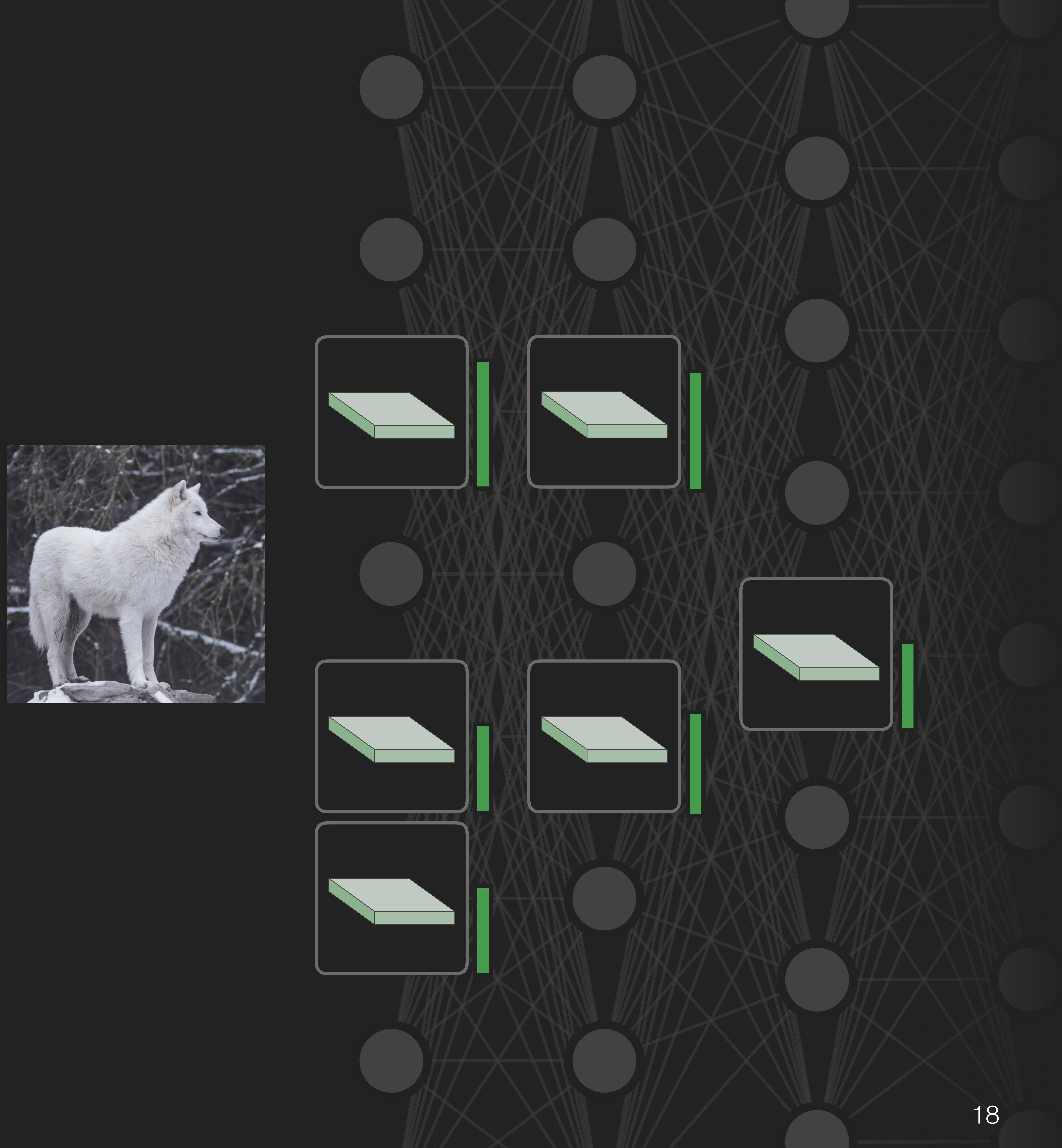




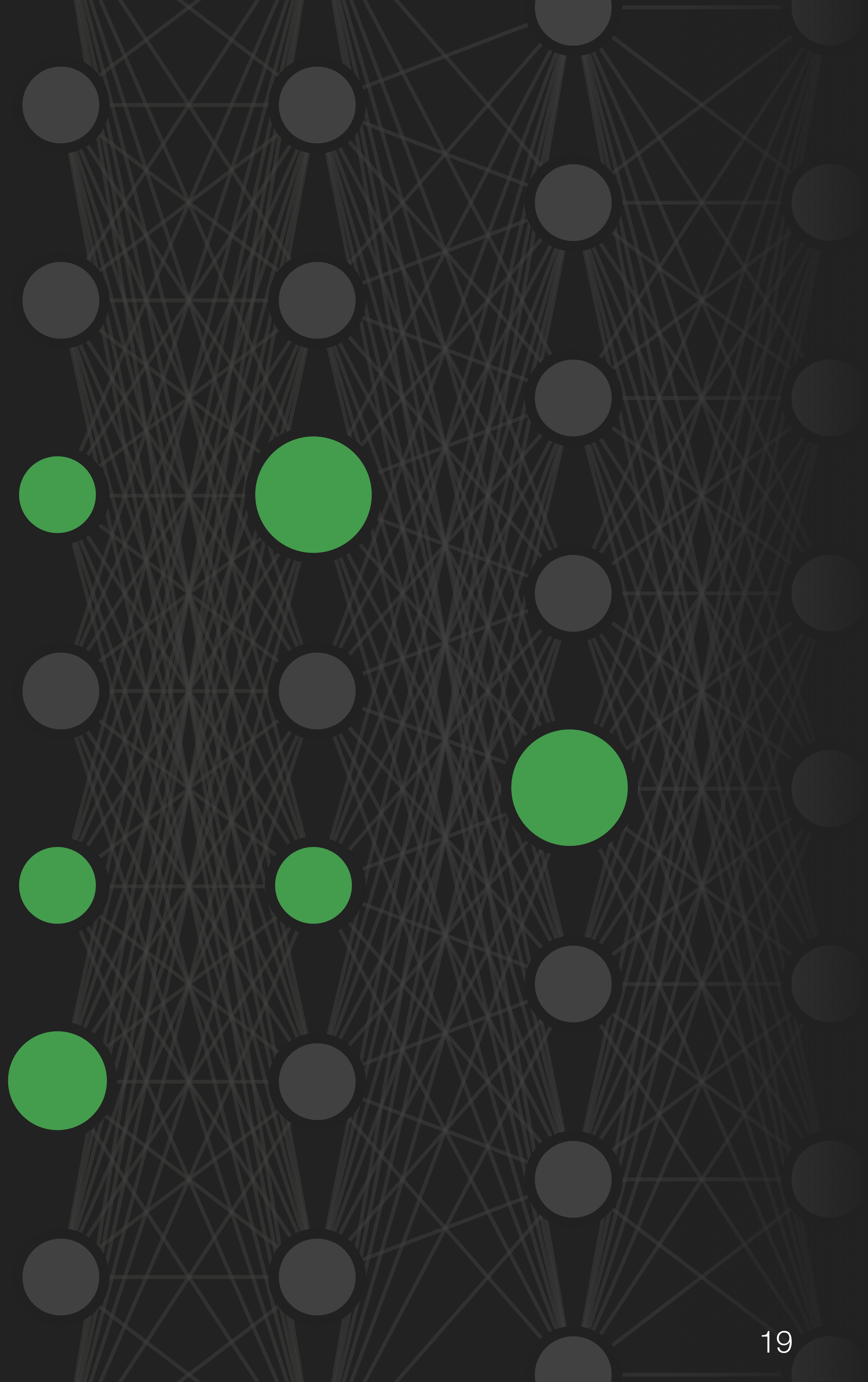
Aggregate network
activations (nodes)



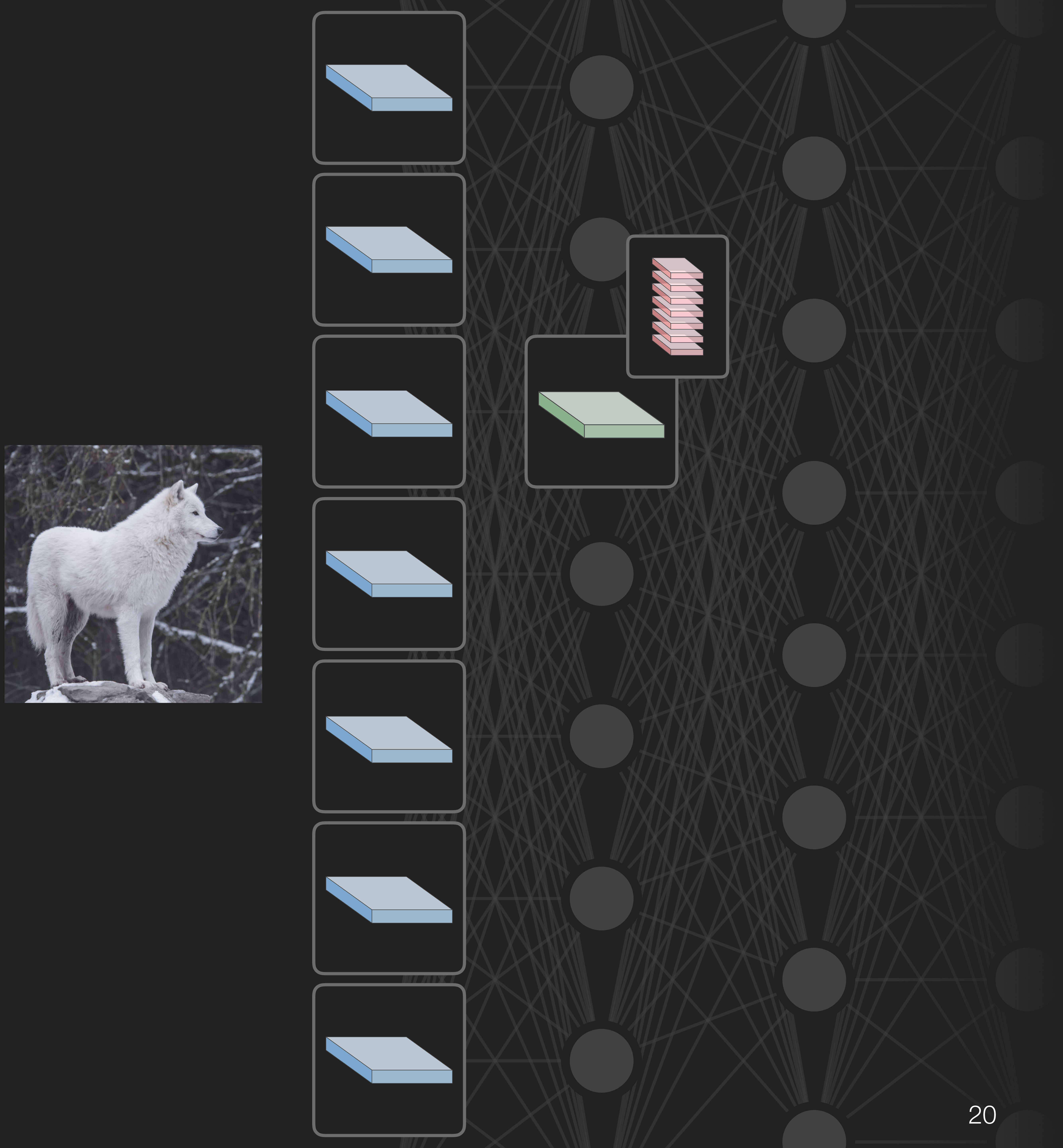
Aggregate network
activations (nodes)



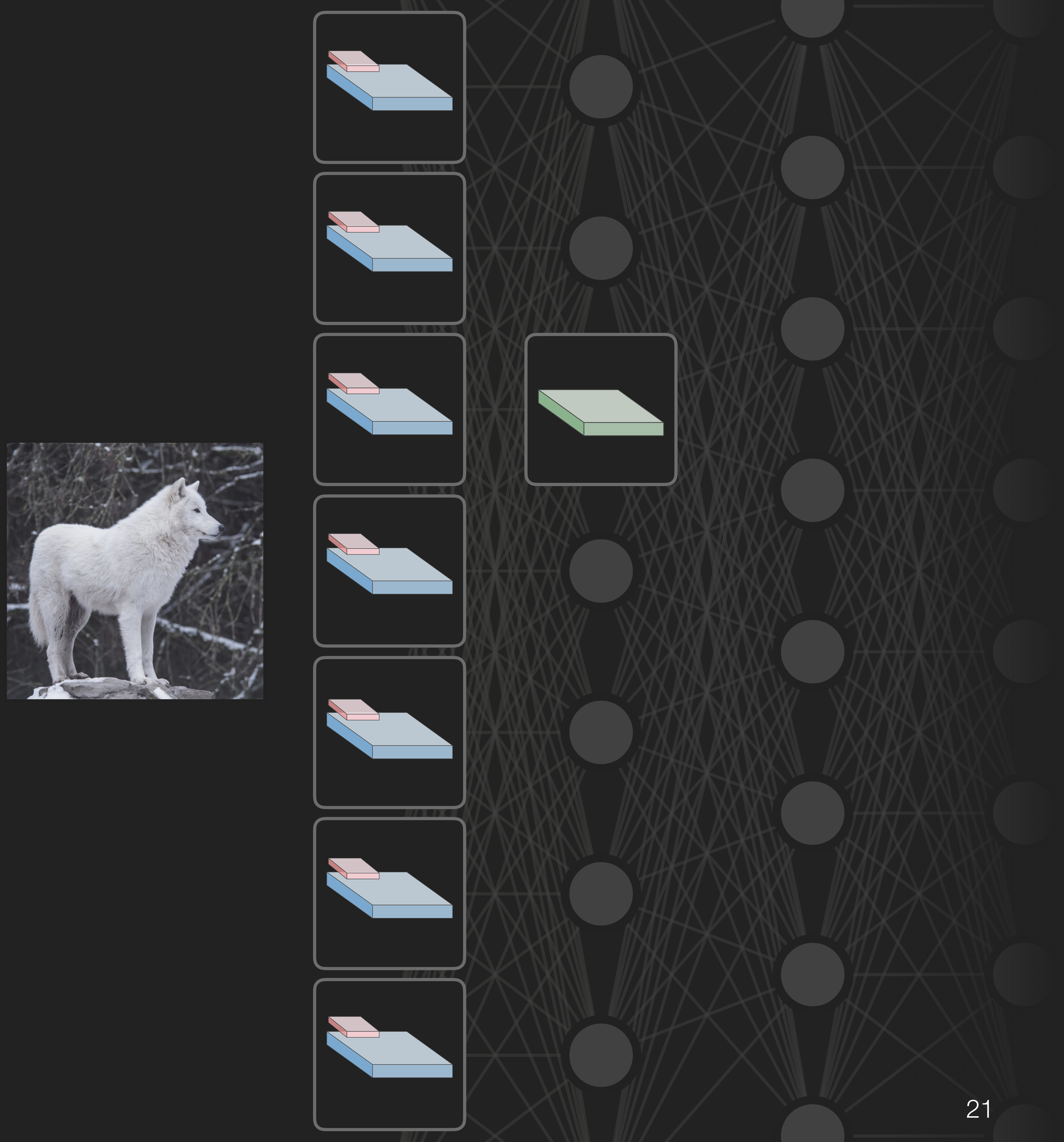
Aggregate network **activations** (nodes)



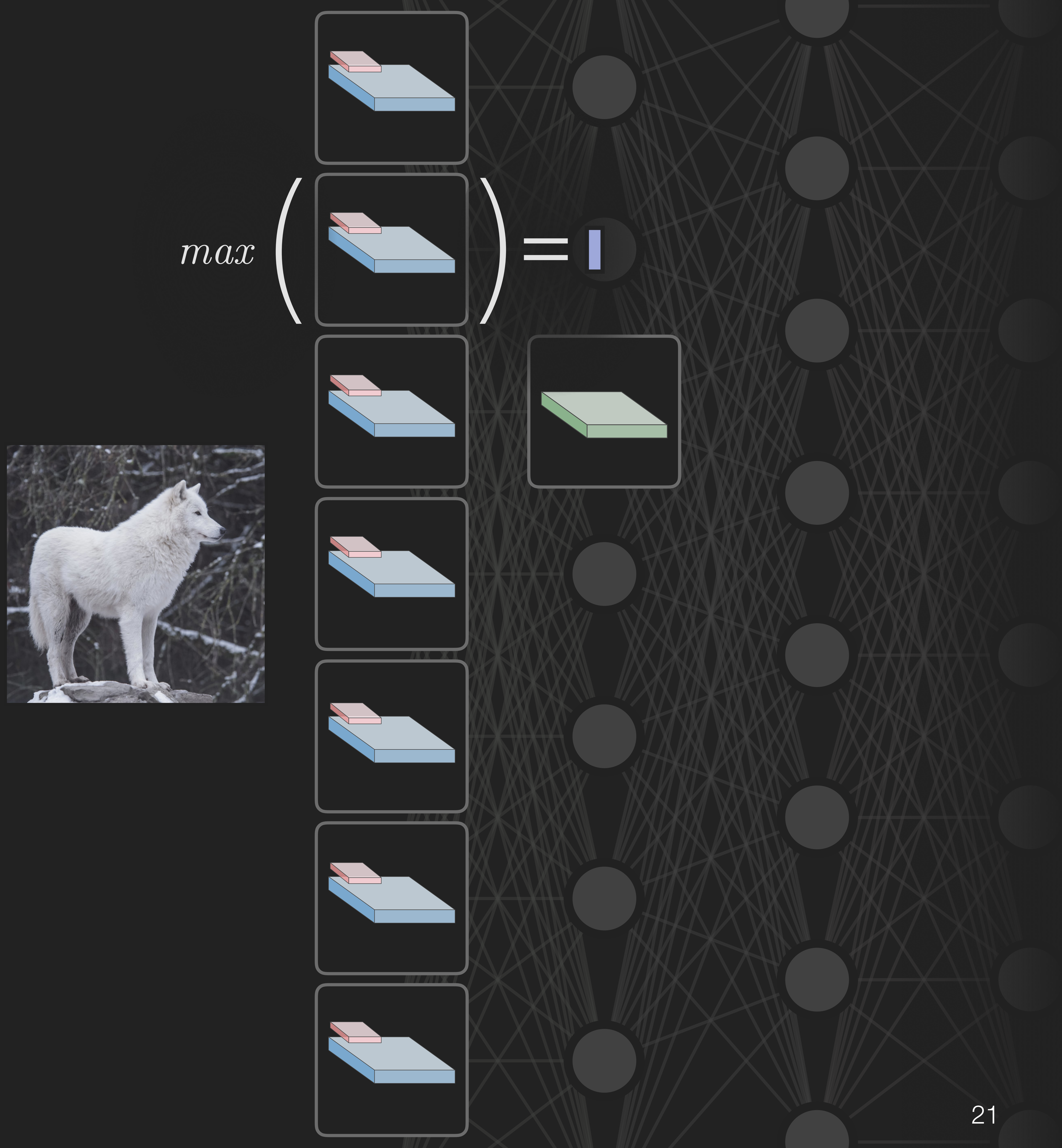
Aggregate network
activations (nodes)



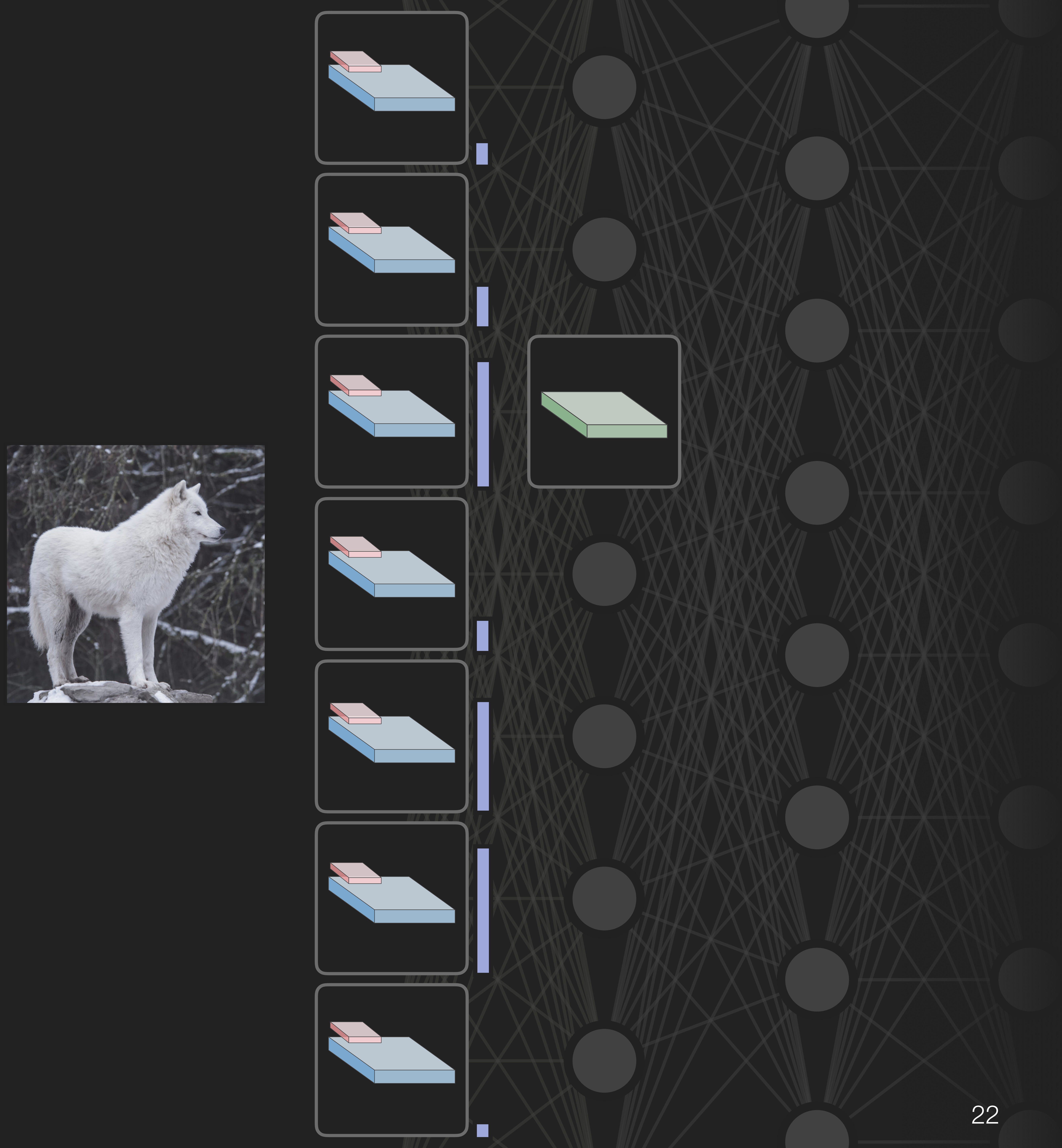
Aggregate network **influences** (edges)



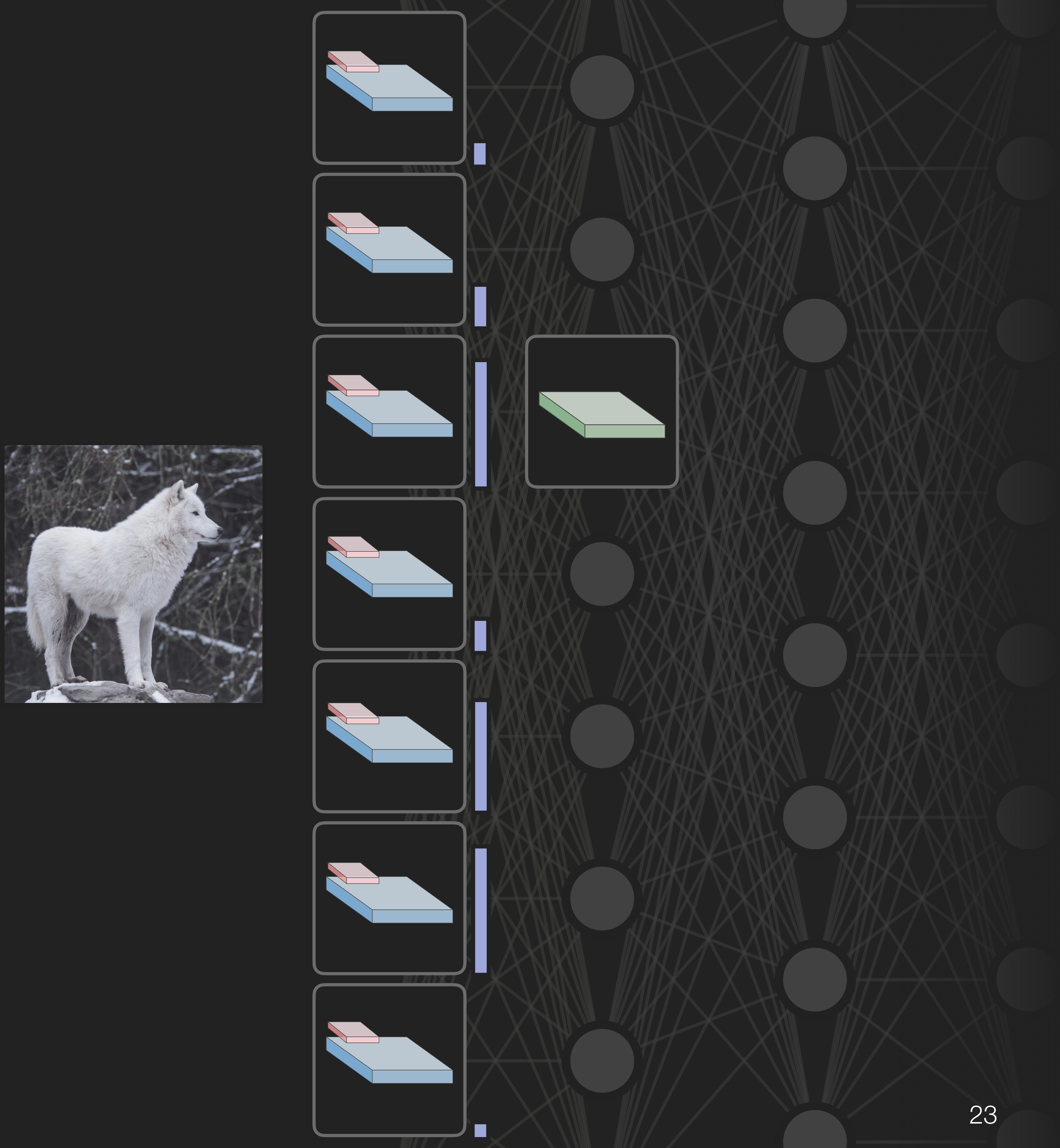
Aggregate network
influences (edges)



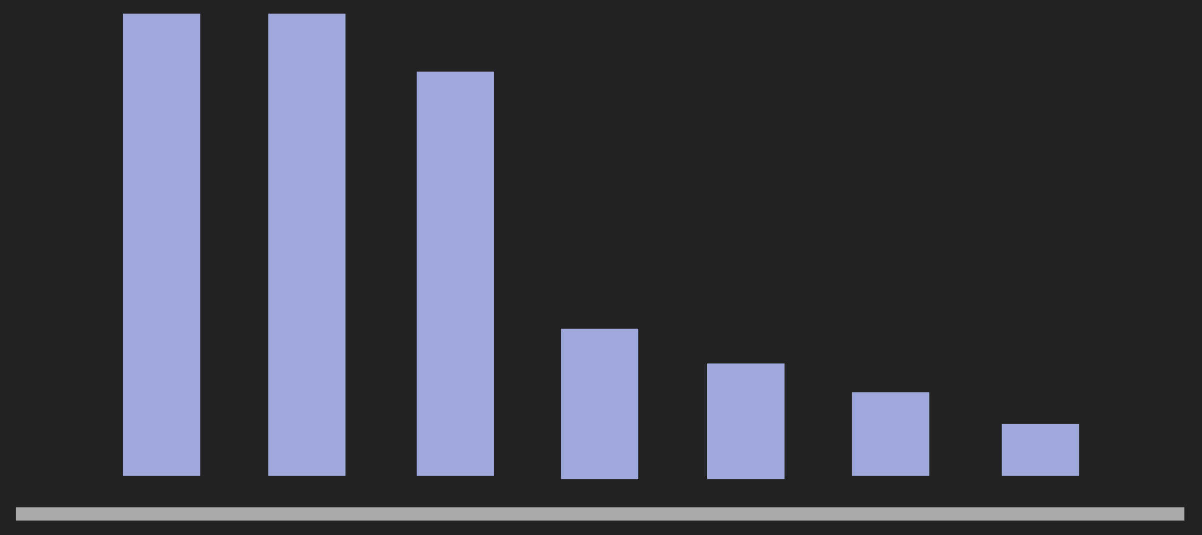
Aggregate network
influences (edges)

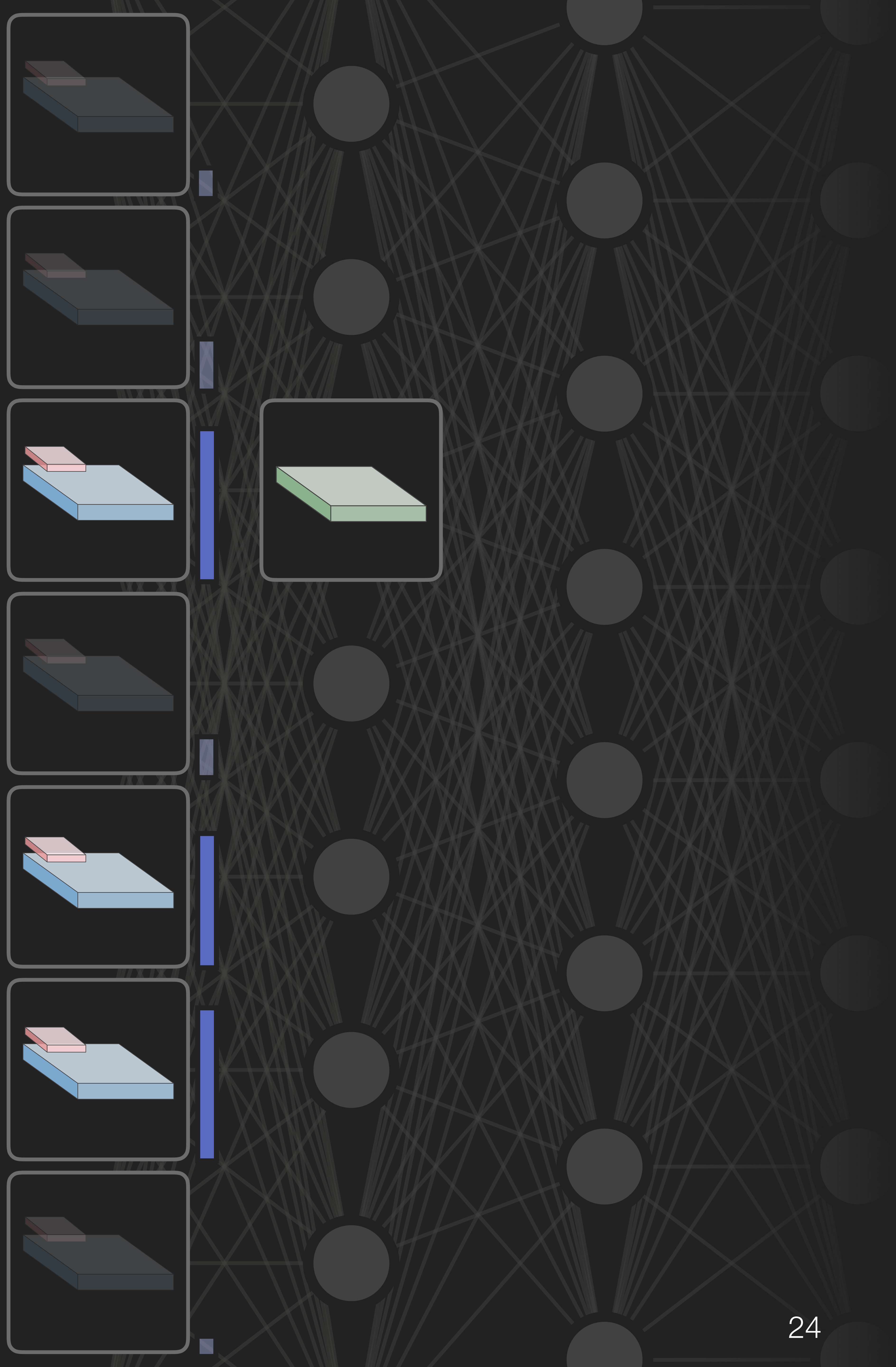


Aggregate network **influences** (edges)

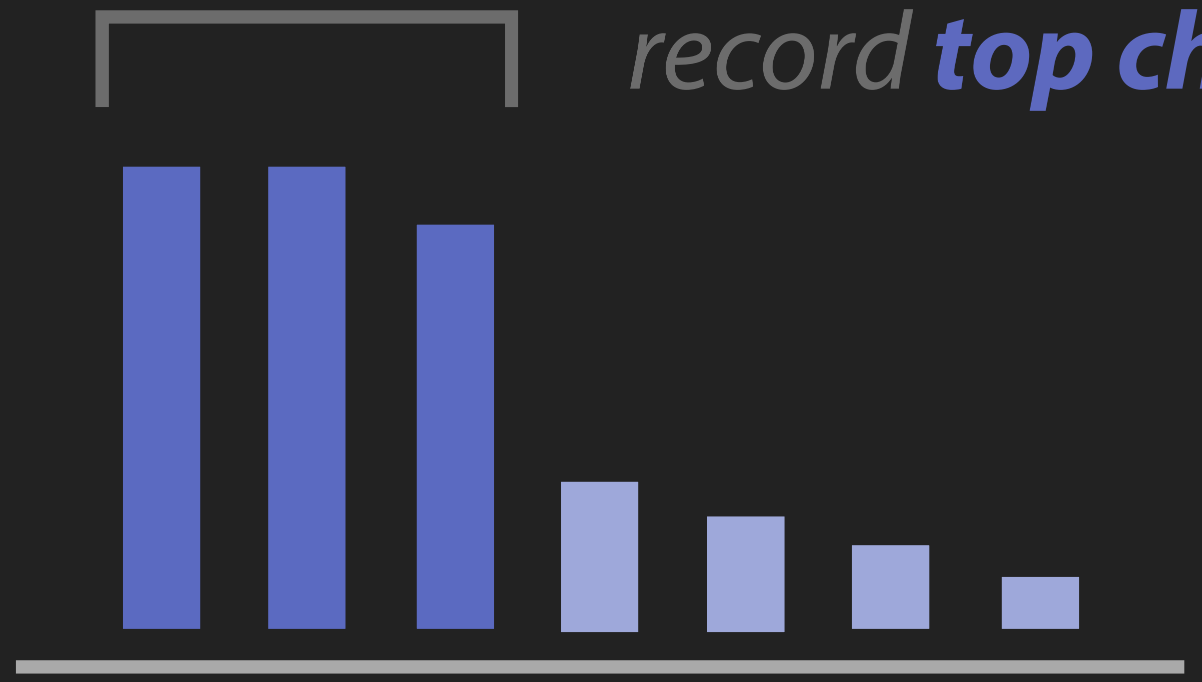


Aggregate network
influences (edges)

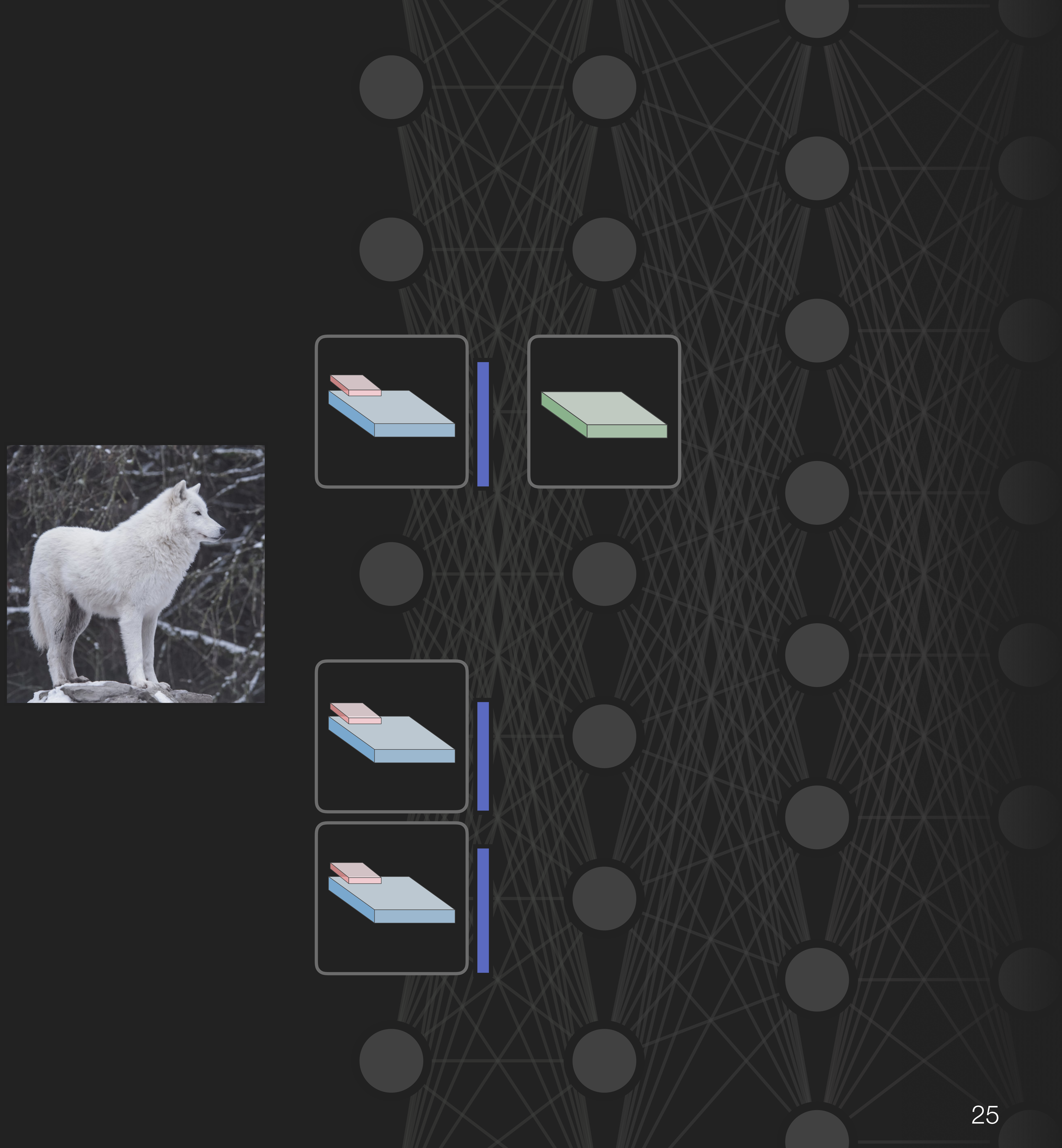




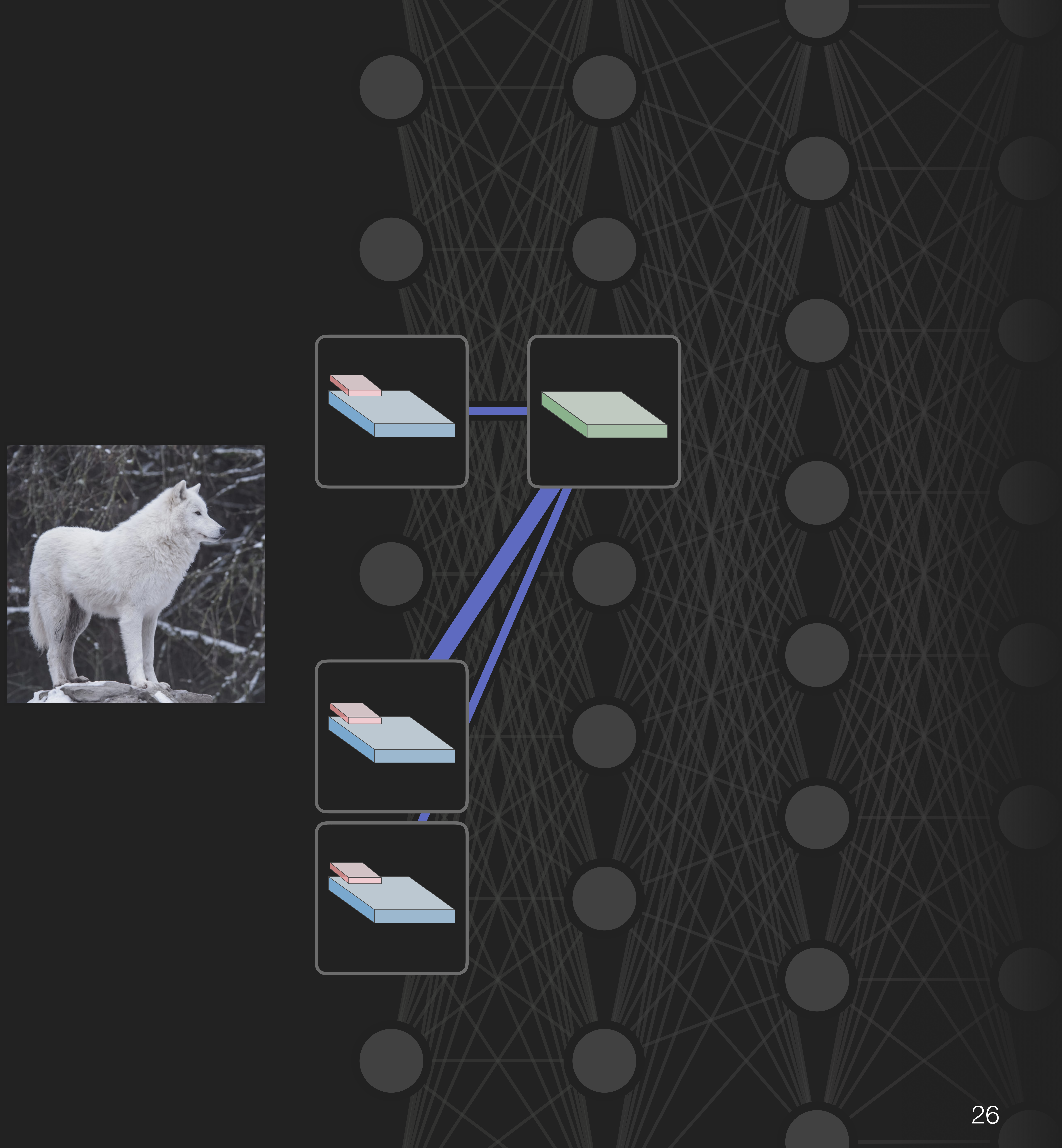
Aggregate network **influences** (edges)



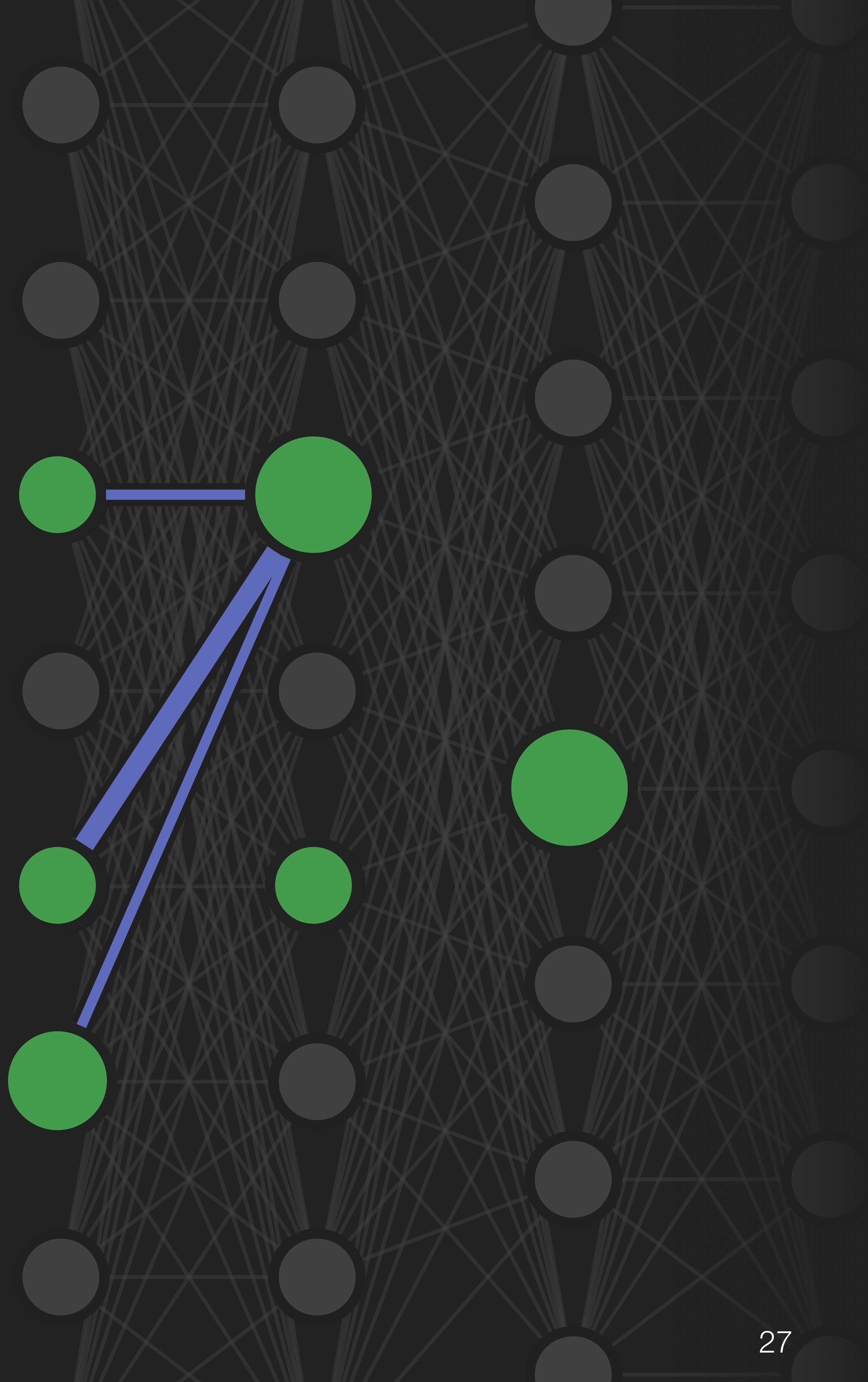
record ***top channels***



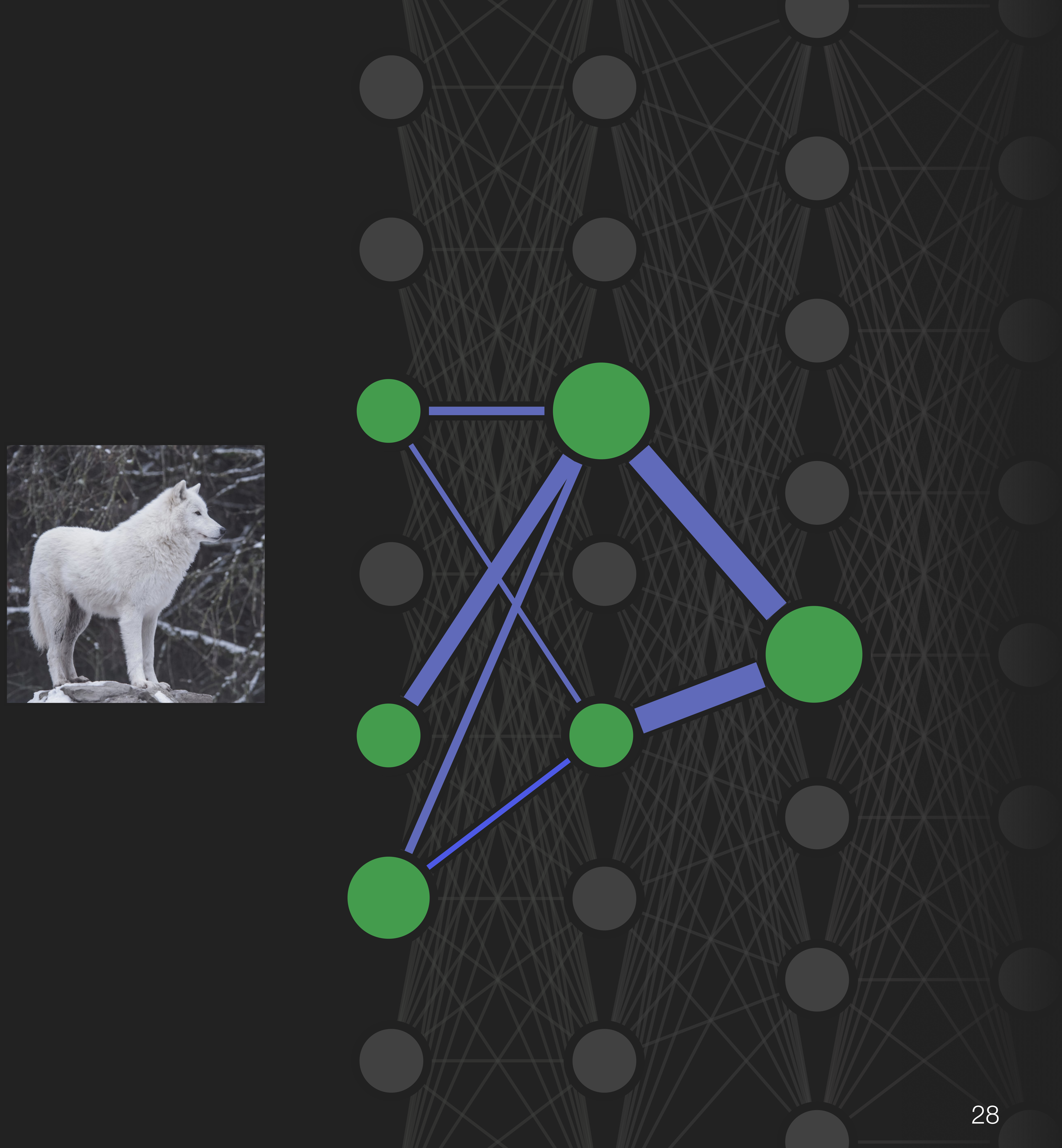
Aggregate network **influences** (edges)



Aggregate network
influences (edges)



Combine **activations** and **influences**



Combine **activations** and **influences**

Further summarize graph
personalized PageRank

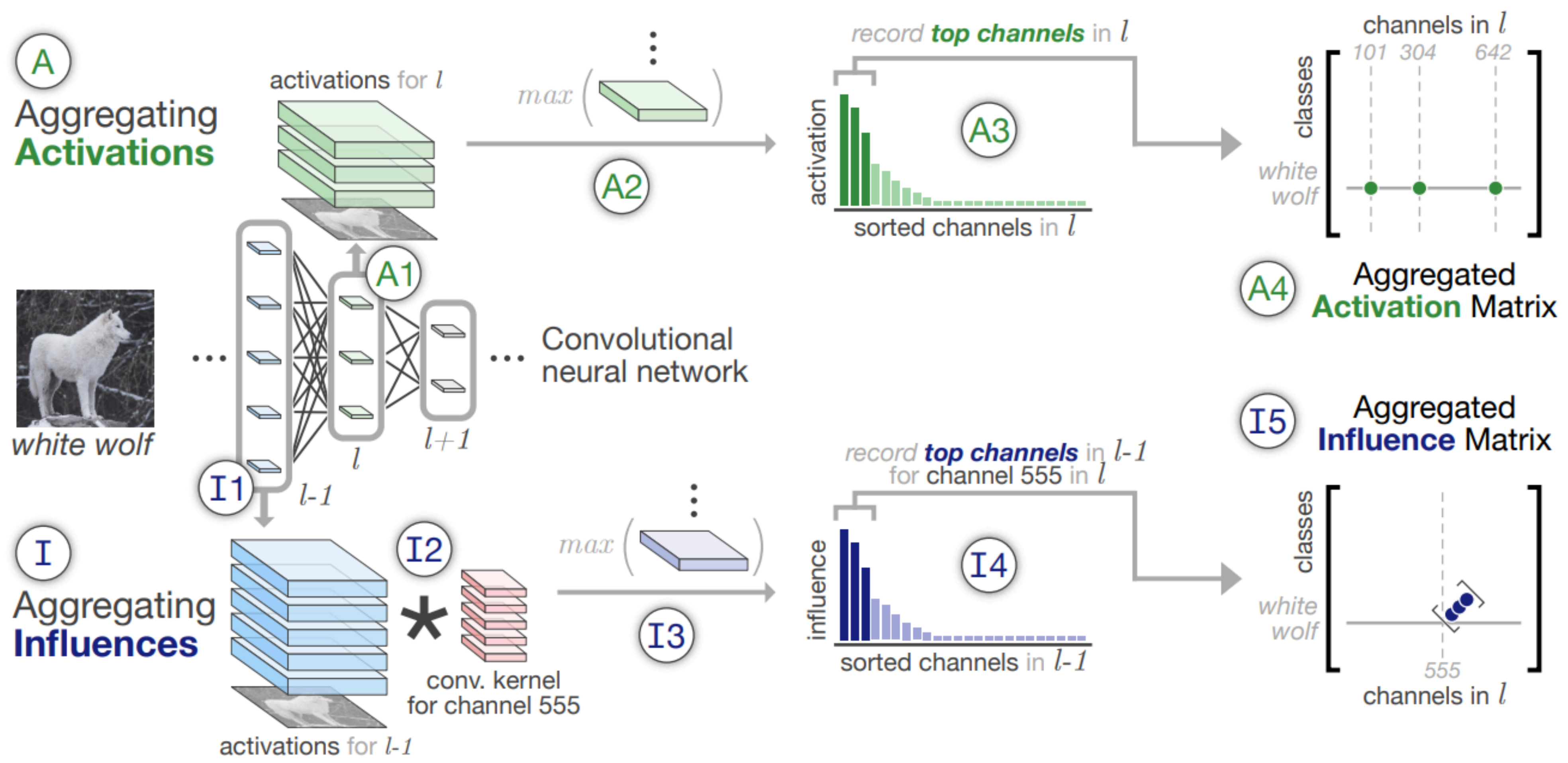


Fig. 4. Our approach for aggregating activations and influences for a layer l. **Aggregating Activations**: (**A1**) given activations at layer l, (**A2**) compute the max of each 2D channel, and (**A3**) record the top activated channels into an (**A4**) aggregated activation matrix, which tells us which channels in a layer most activate and represent every class in the model. **Aggregating Influences**: (**I1**) given activations at layer $l-1$, (**I2**) convolve them with a convolutional kernel from layer l, (**I3**) compute the max of each resulting 2D activation map, and (**I4**) record the top most influential channels from layer $l-1$ that impact channels in layer l into an (**I5**) aggregated influence matrix, which tells us which channels in the previous layer most influence a particular channel in the next layer.

6.2 Aggregating Inter-layer Influences

Aggregating activations at each convolutional layer in a network will only give a local description of which channels are important for each class, i.e., from examining A^l we will not know *how* certain channels come to be the most representative for a given class. Thus, we need a way to calculate how the activations from the channels of a previous layer, $l-1$, **influence** the activations at the current layer, l. In dense layers, this influence is trivial to compute: the activation at a neuron in l is computed as the weighted sum of activations from neurons in $l-1$. The influence of a single neuron from $l-1$ is then proportional to the activation of that neuron multiplied by the associated weight to the neuron being examined from l. In convolutional layers, calculating this influence is more complicated: the activations at a channel in l are computed as the 3D convolution of all of the channels from $l-1$ with a learned kernel tensor. This operation can be broken down (shown formally later in this section) as a summation of the 2D convolutions of each channel in $l-1$ with a corresponding slice of the appropriate kernel. The summations of 2D convolutions are similar in structure to the weighted-summations performed by dense layers, however the corresponding "influence" of a single channel from $l-1$ on the output of a particular channel in l is a 2D feature map. We can summarize this feature map into a scalar influence value by using any type of reduce operation, which we discuss further below.

We propose a method for (1) quantifying the *influence* a channel from a previous layer has on the activations of a channel in a following layer,

the j^{th} kernel, and the resulting maps are summed to produce a single channel in Y. We care about the 2D quantity $X_{:,:,i} * K^{(j)}_{:,:,i}$ as it contains exactly the contributions of a *single* channel from the previous layer to a channel in the current layer.

Second, we must summarize the quantity $X_{:,:,i} * K^{(j)}_{:,:,i}$ into a scalar influence value. Similarly discussed in Sect. 6.1, this can be done in many ways, e.g., by summing all values, applying the Frobenius norm, or taking the maximum value. Each of these summarization methods (i.e., 2D to 1D reduce operations) may lend itself well to exposing interesting connections between channels later in our pipeline. We chose to (**I3**) take the maximum value of $X_{:,:,i} * K^{(j)}_{:,:,i}$ as our measure of influence for the image classification task, since this task intuitively considers the largest magnitude of a feature, e.g., how strongly a "dog ear" or "car wheel" feature is expressed, instead of summing values for example, which might indicate how many places in the image a "dog ear" or "car wheel" is being expressed. Also, this mirrors our approach for aggregating activations above.

Lastly, we must aggregate these influence values between channel pairs in consecutive layers, for all images in a given class, i.e., create the proposed I^l matrix from the pairwise channel influence values. This process mirrors the aggregation described previously (Sect. 6.1), and we follow the same framework. Let L^l_{ij} be the scalar influence value computed by the previous step *for a single image in class c*, between channel i in layer $l-1$ and channel j in layer l. We increment an entry

ivations

es

marize graph

d PageRank

Feature Visualization

Feature Visualization

What kind of input would cause a neuron to maximally activate?

Feature Visualization

What kind of input would cause a neuron to maximally activate?

mixed4b, 409

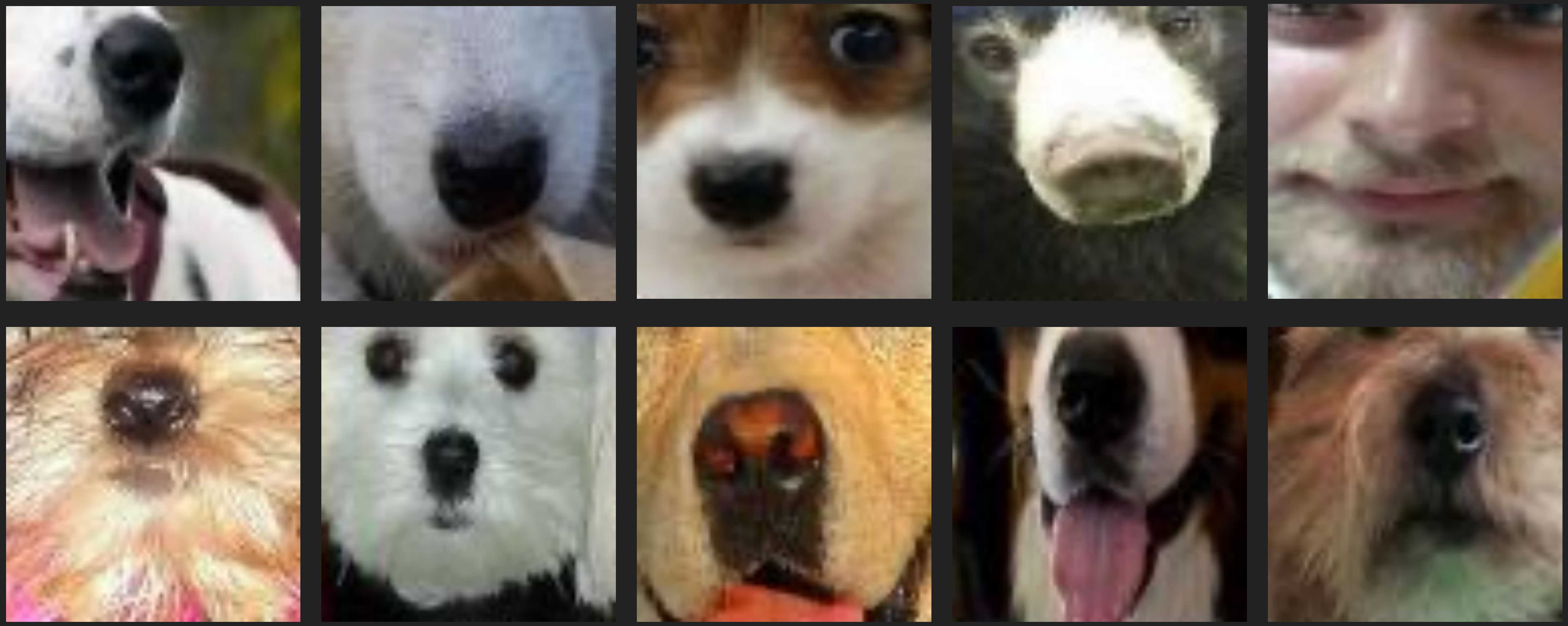
Generate examples: *starting from random noise, optimize an image to activate a particular neuron*

Feature Visualization

What kind of input would cause a neuron to maximally activate?

mixed4b, 409

Generate examples: *starting from random noise, optimize an image to activate a particular neuron*



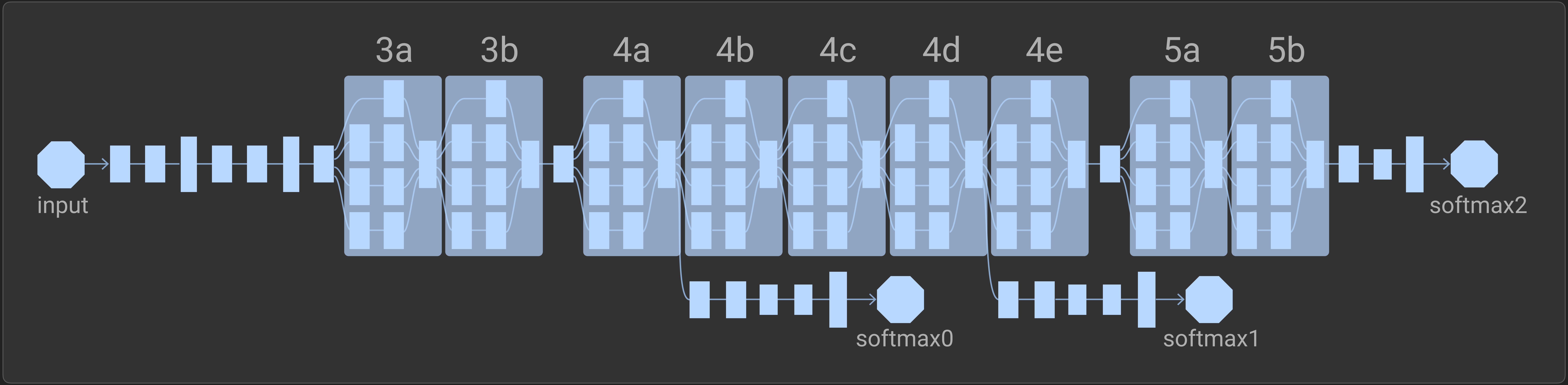
[Olah, et al., Distill, 2017]

Demo

Demo

InceptionV1
Large-scale,
prevalent CNN

ImageNet (ILSVRC)
~1.3M images
1,000 classes



[Olah, et al., Distill, 2017]


SUMMIT
MODEL
InceptionV1
DATASET
ImageNet
CLASSES
1,000
INSTANCES
1,281,024
LAYER
mixed
3a
3b
4a
4b
4c
4d
4e
5a
5b
CLASS
white_wolf
INSTANCES
1299
ACCURACY
81.8%
PROBABILITIES
FILTER GRAPH
ADJUST WIDTH
ADJUST HEIGHT
white wolf
white wolf
white wolf
81.8%
red wolf
69.9%
timber wolf
64.2%
arctic fox
87.1%
lion
87.1%
chow
87.1%
rottweiler
76.6%
silky terrier
63.3%


SUMMIT
MODEL
InceptionV1
DATASET
ImageNet
CLASSES
1,000
INSTANCES
1,281,024
LAYER
mixed
3a
3b
4a
4b
4c
4d
4e
5a
5b
CLASS
white_wolf
INSTANCES
1299
ACCURACY
81.8%
PROBABILITIES
FILTER GRAPH
ADJUST WIDTH
ADJUST HEIGHT
white wolf
white wolf
white wolf 81.8%
red wolf 69.9%
timber wolf 64.2%
arctic fox 87.1%
lion 87.1%
chow 87.1%
rottweiler 76.6%
silky terrier 63.3%

Unexpected Features

Unexpected Features

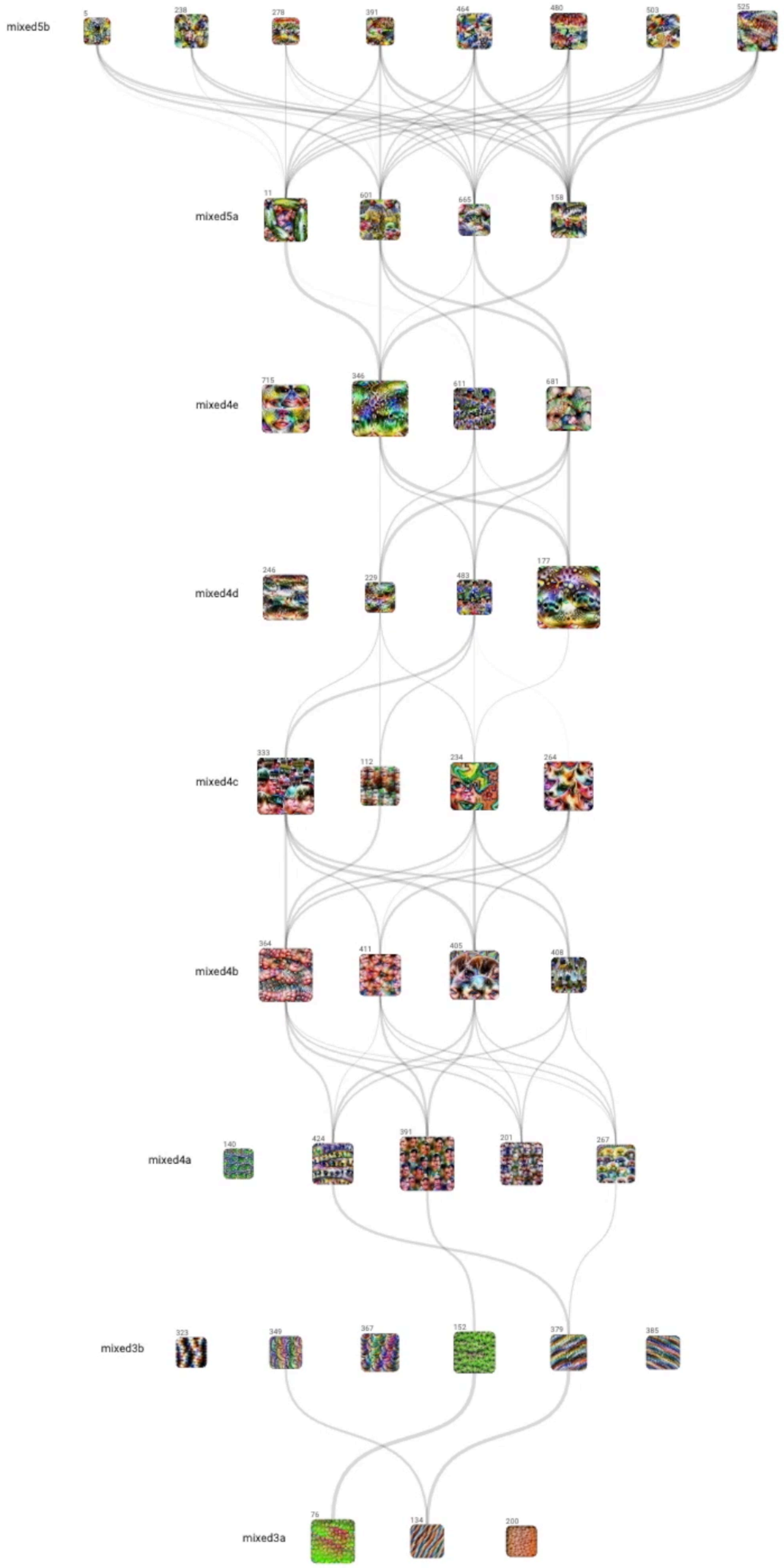
tench

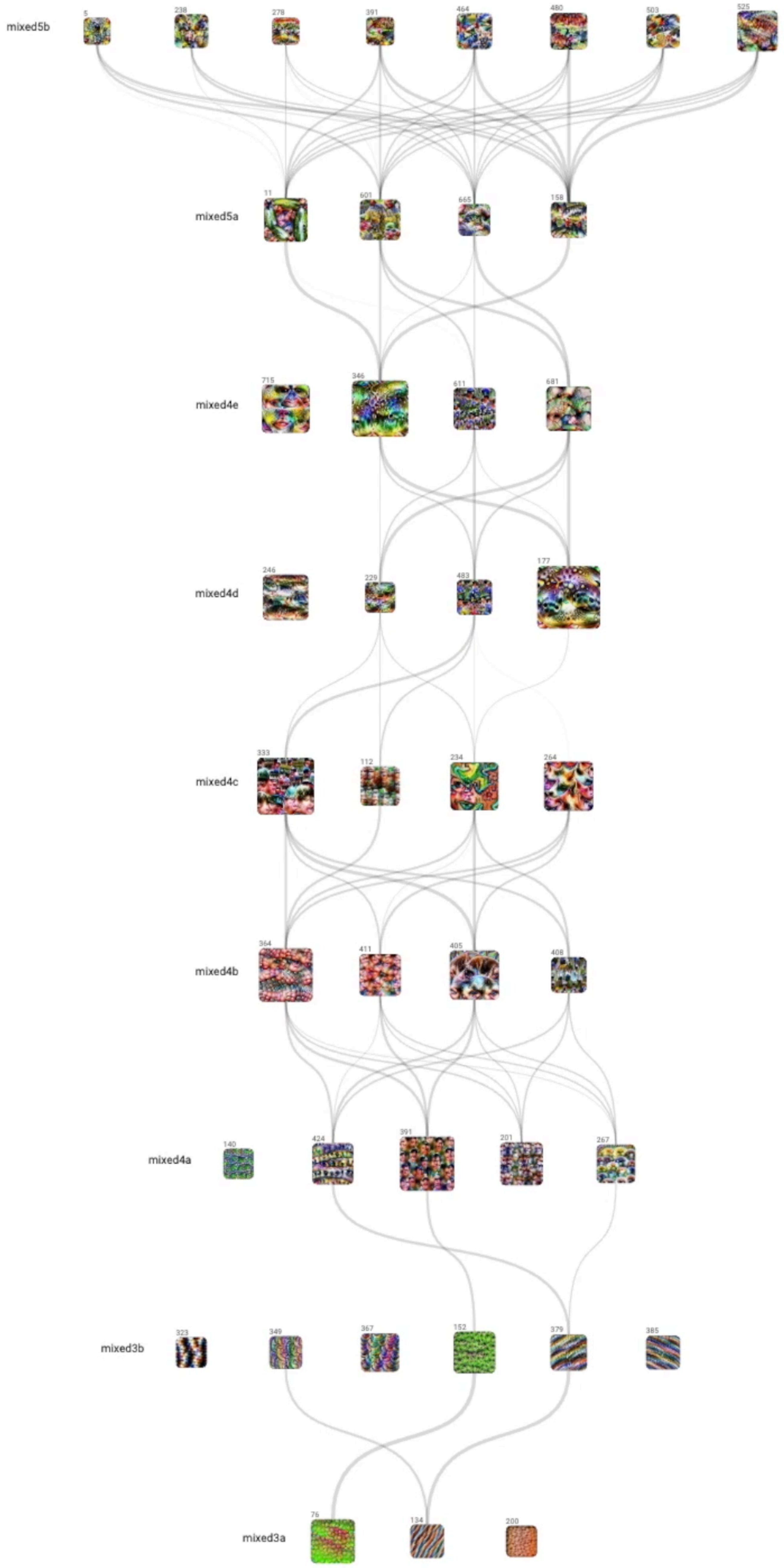
Unexpected Features

tench

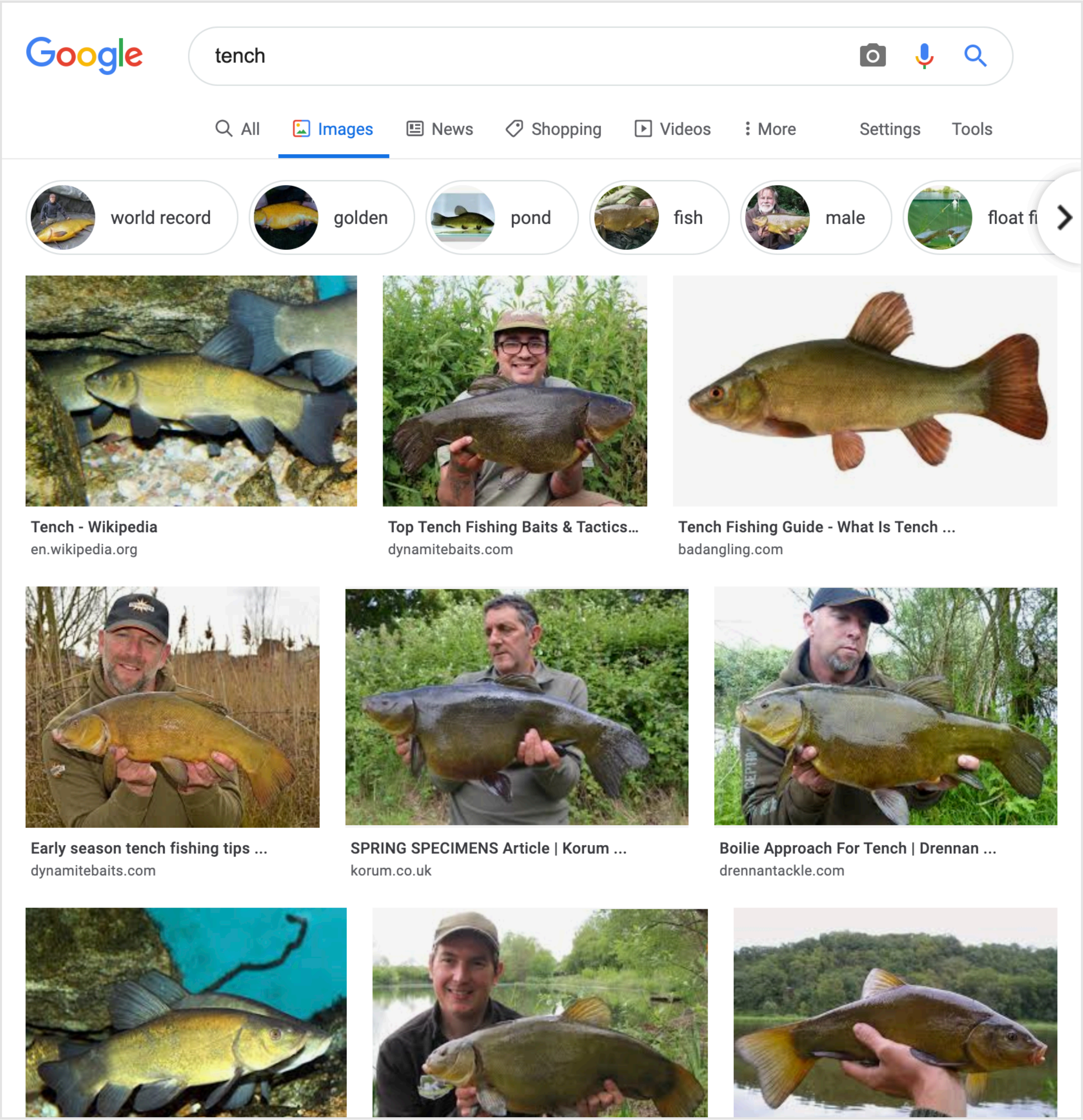
What features has a neural network learned for ***tench***?

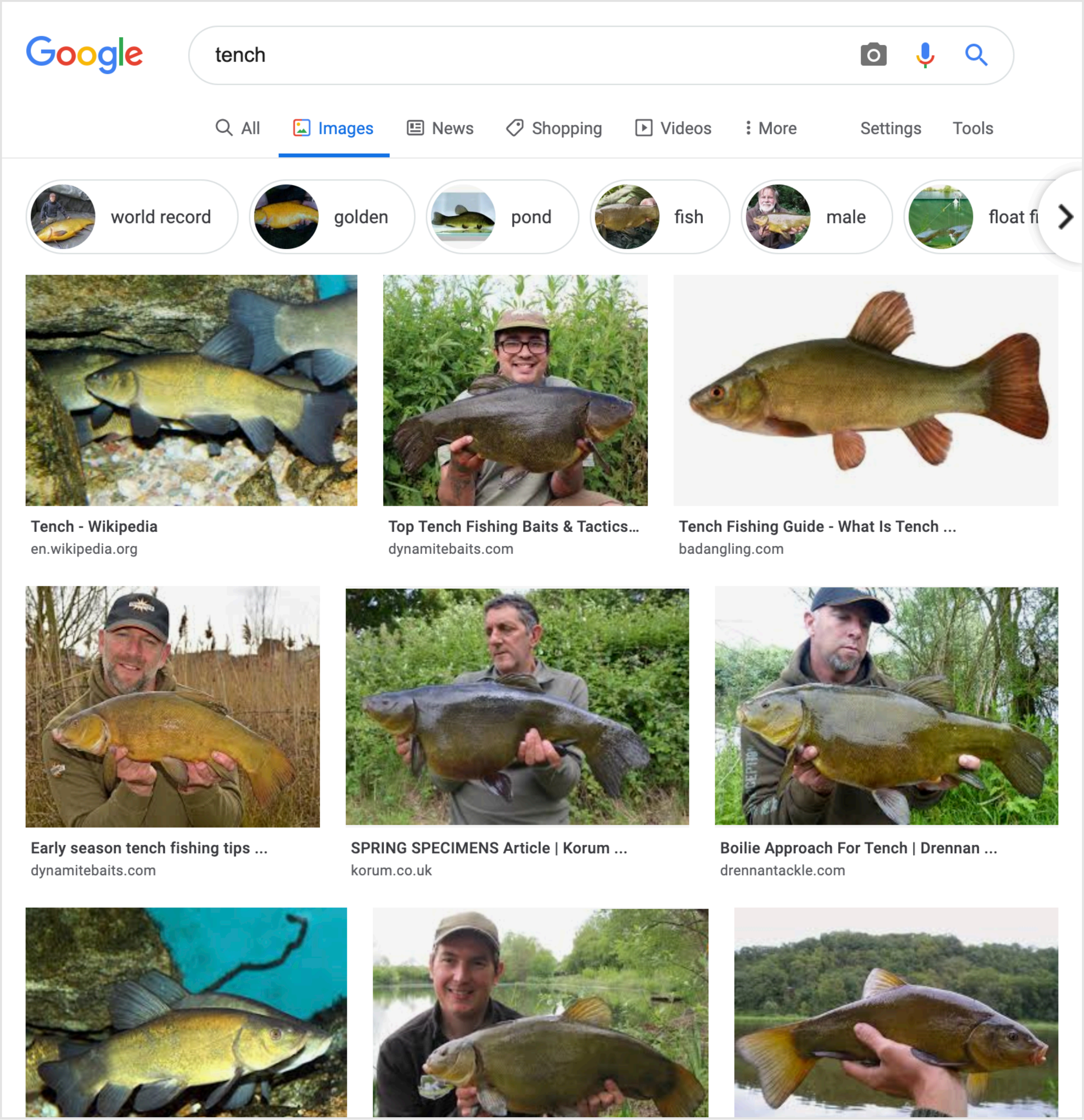
How are those features related?

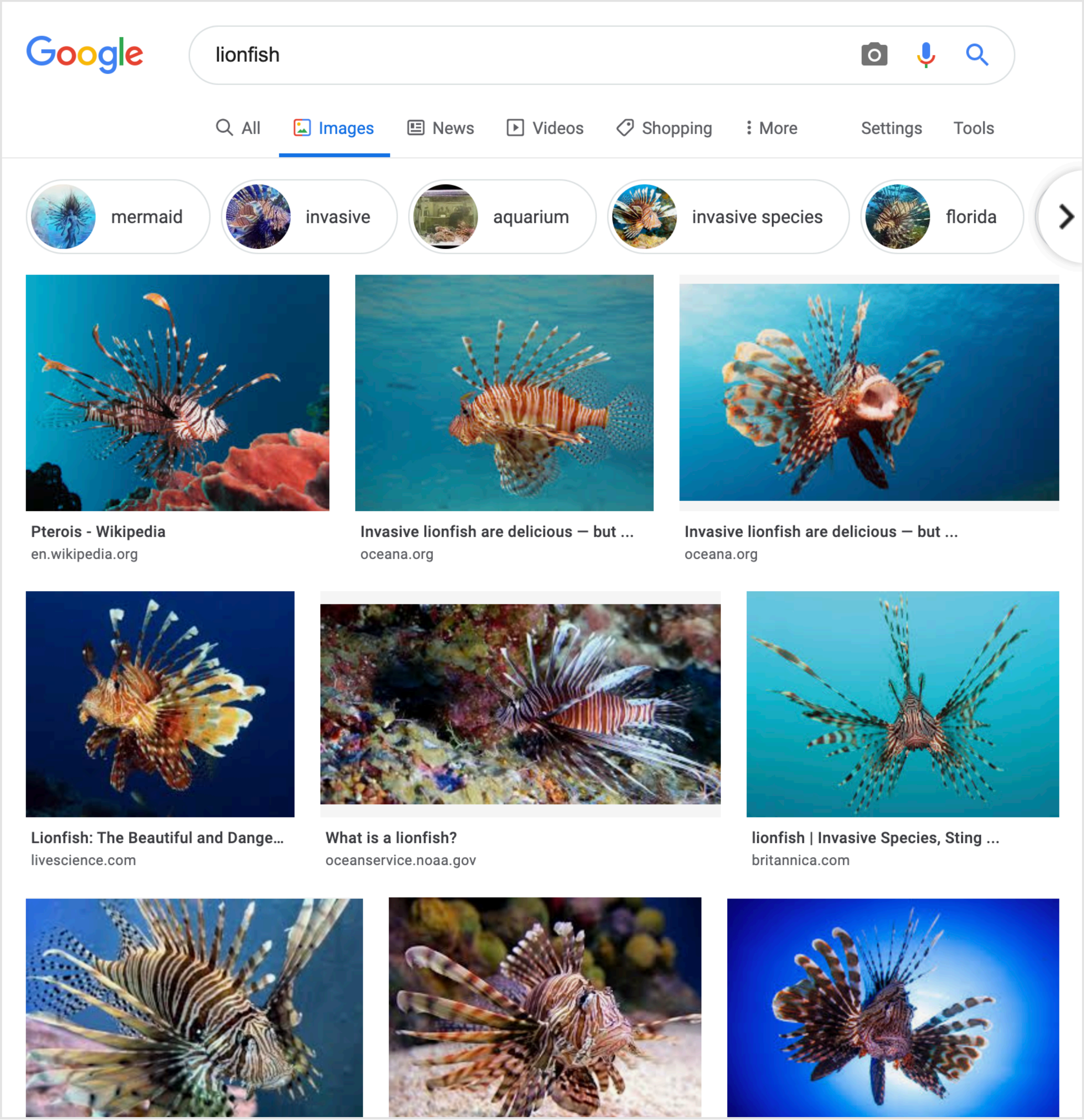

mixed5b
mixed5a
mixed4e
mixed4d
mixed4c
mixed4b
mixed4a
mixed3b
mixed3a


mixed5b
mixed5a
mixed4e
mixed4d
mixed4c
mixed4b
mixed4a
mixed3b
mixed3a


Data is important too!


Google
tench
All
Images
News
Shopping
Videos
More
Settings
Tools
world record
golden
pond
fish
male
float fi
Tench - Wikipedia
en.wikipedia.org
Top Tench Fishing Baits & Tactics...
dynamitebaits.com
Tench Fishing Guide - What Is Tench ...
badangling.com
Early season tench fishing tips ...
dynamitebaits.com
SPRING SPECIMENS Article | Korum ...
korum.co.uk
Boilie Approach For Tench | Drennan ...
drennantackle.com

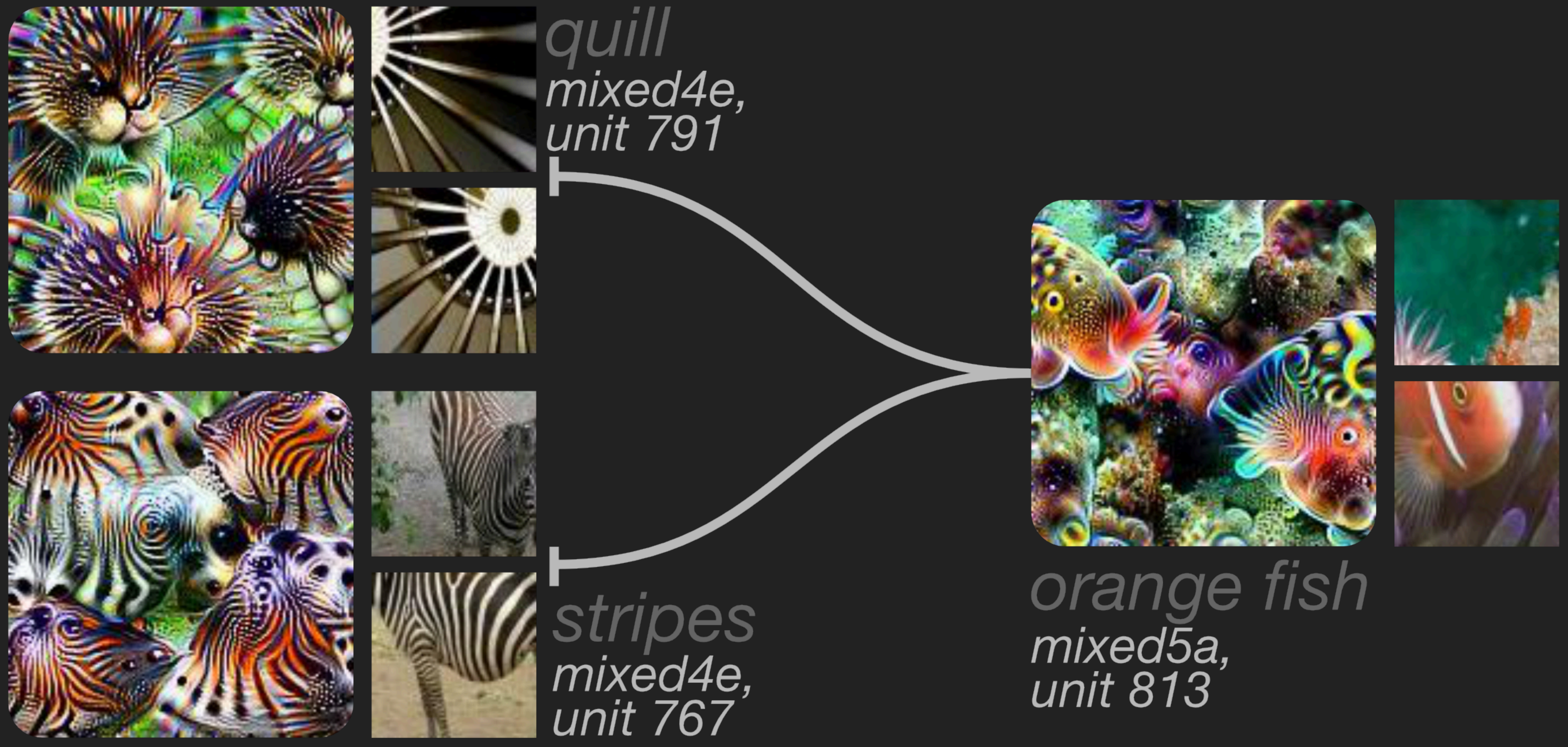

Google
tench
All
Images
News
Shopping
Videos
More
Settings
Tools
world record
golden
pond
fish
male
float fi
Tench - Wikipedia
en.wikipedia.org
Top Tench Fishing Baits & Tactics...
dynamitebaits.com
Tench Fishing Guide - What Is Tench ...
badangling.com
Early season tench fishing tips ...
dynamitebaits.com
SPRING SPECIMENS Article | Korum ...
korum.co.uk
Boilie Approach For Tench | Drennan ...
drennantackle.com


Google
lionfish
All
Images
News
Videos
Shopping
More
Settings
Tools
mermaid
invasive
aquarium
invasive species
florida
Pterois - Wikipedia
en.wikipedia.org
Invasive lionfish are delicious — but ...
oceana.org
Invasive lionfish are delicious — but ...
oceana.org
Lionfish: The Beautiful and Dange...
livescience.com
What is a lionfish?
oceanservice.noaa.gov
lionfish | Invasive Species, Sting ...
britannica.com

~~Un~~expected Features

lionfish

No more people features.
But few "fish" features! Mostly textures.



Attribution graph substructure from ***lionfish*** class.

Discriminable Features

Discriminable Features

Do neural network feature representations align with people's expectations?

Discriminable Features

Do neural network feature representations align with people's expectations?

brown bear

Discriminable Features

Do neural network feature representations align with people's expectations?

brown bear

black bear

Discriminable Features

Do neural network feature representations align with people's expectations?

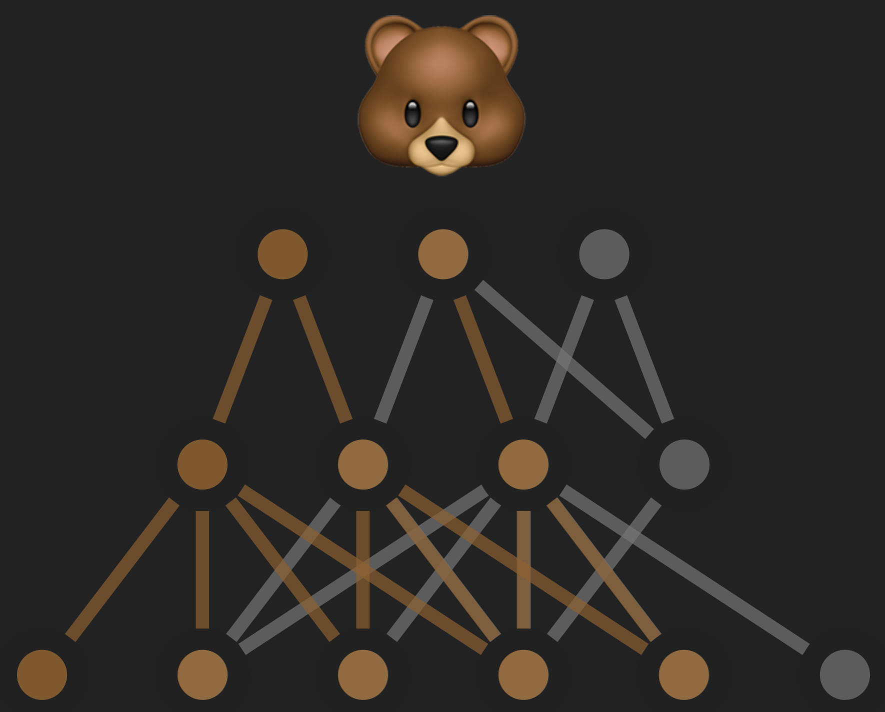
brown bear

black bear

Discriminable Features

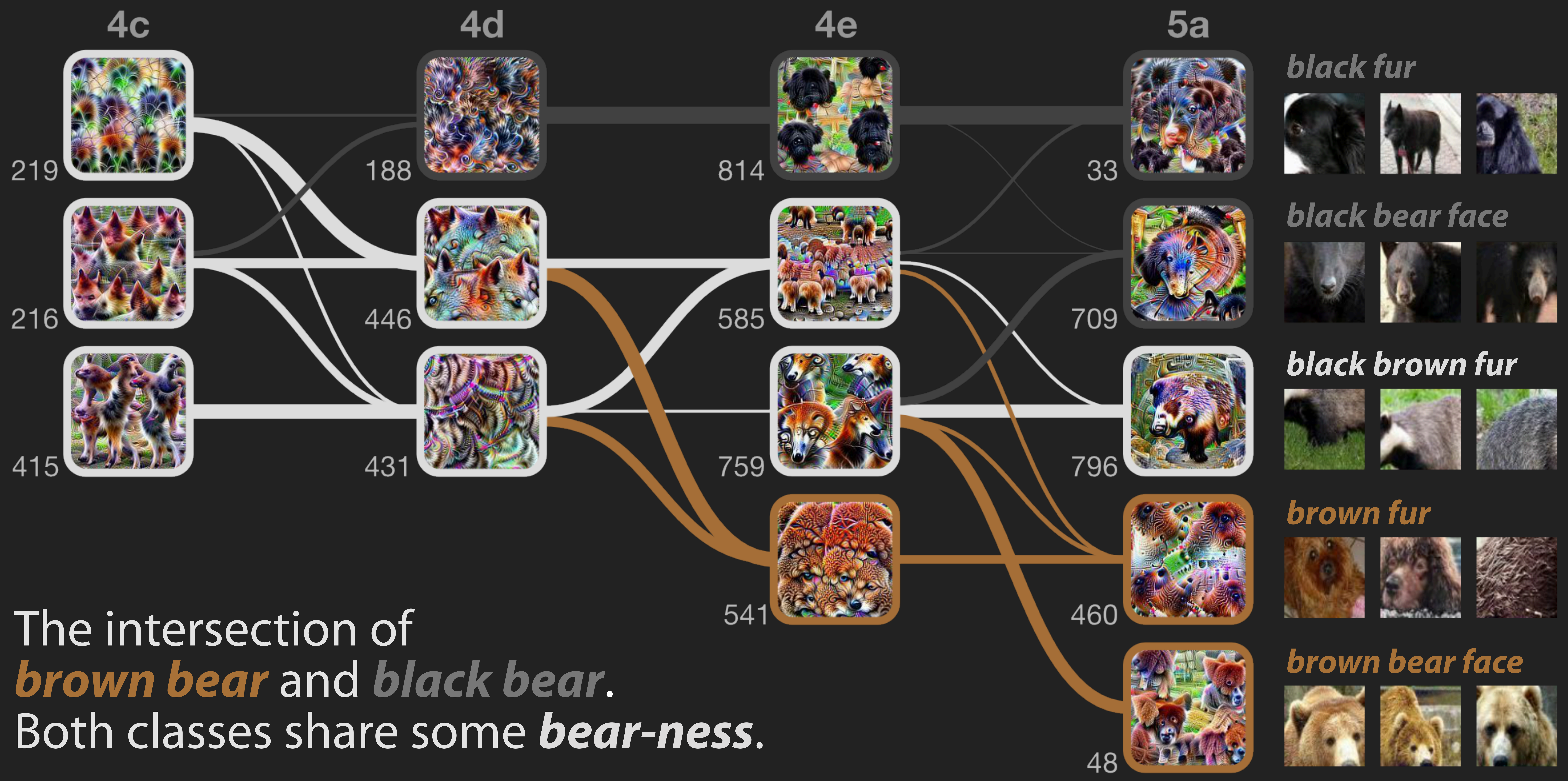
Do neural network feature representations align with people's expectations?

brown bear



intersection

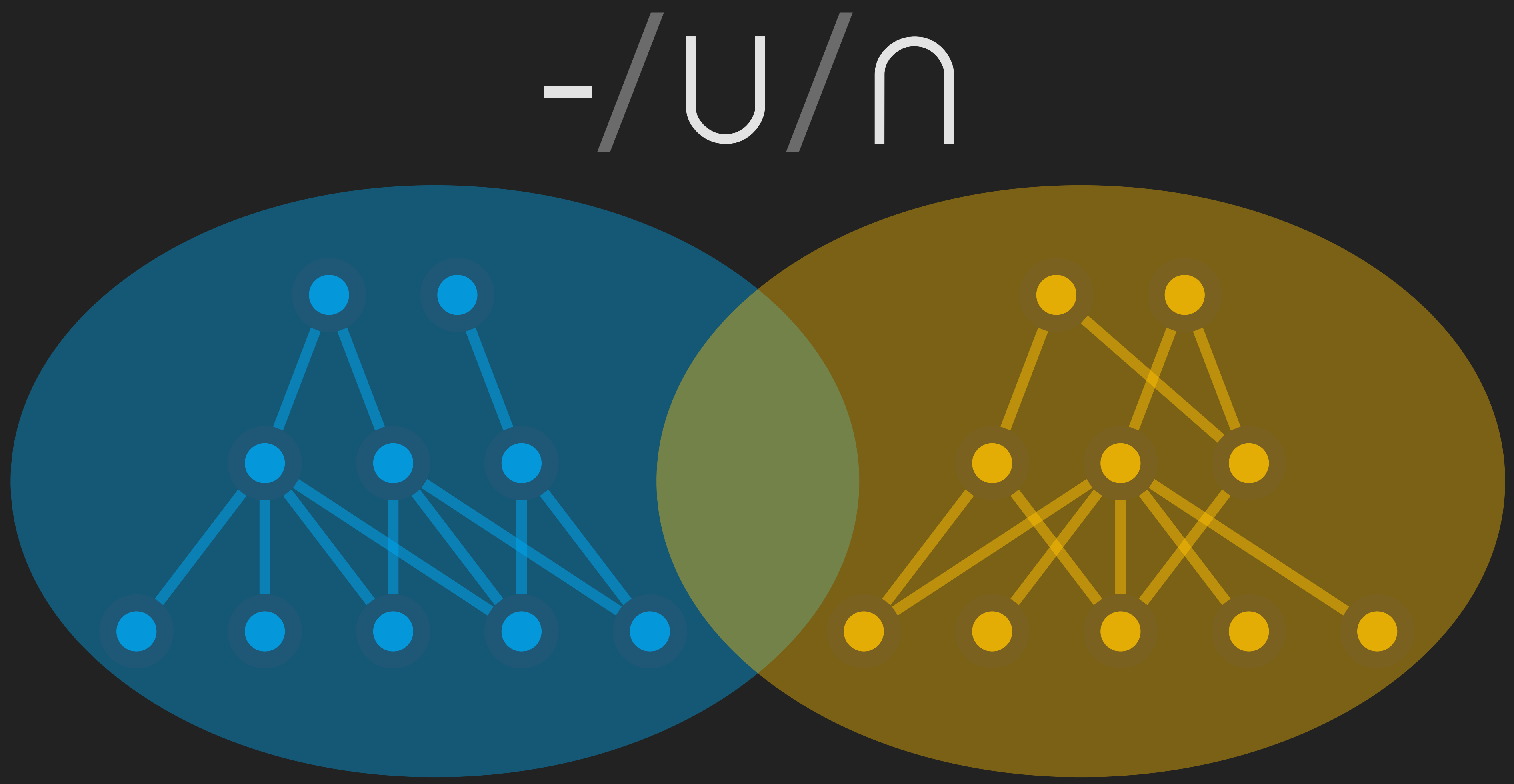
black bear



The intersection of
brown bear and ***black bear***.
Both classes share some ***bear-ness***.

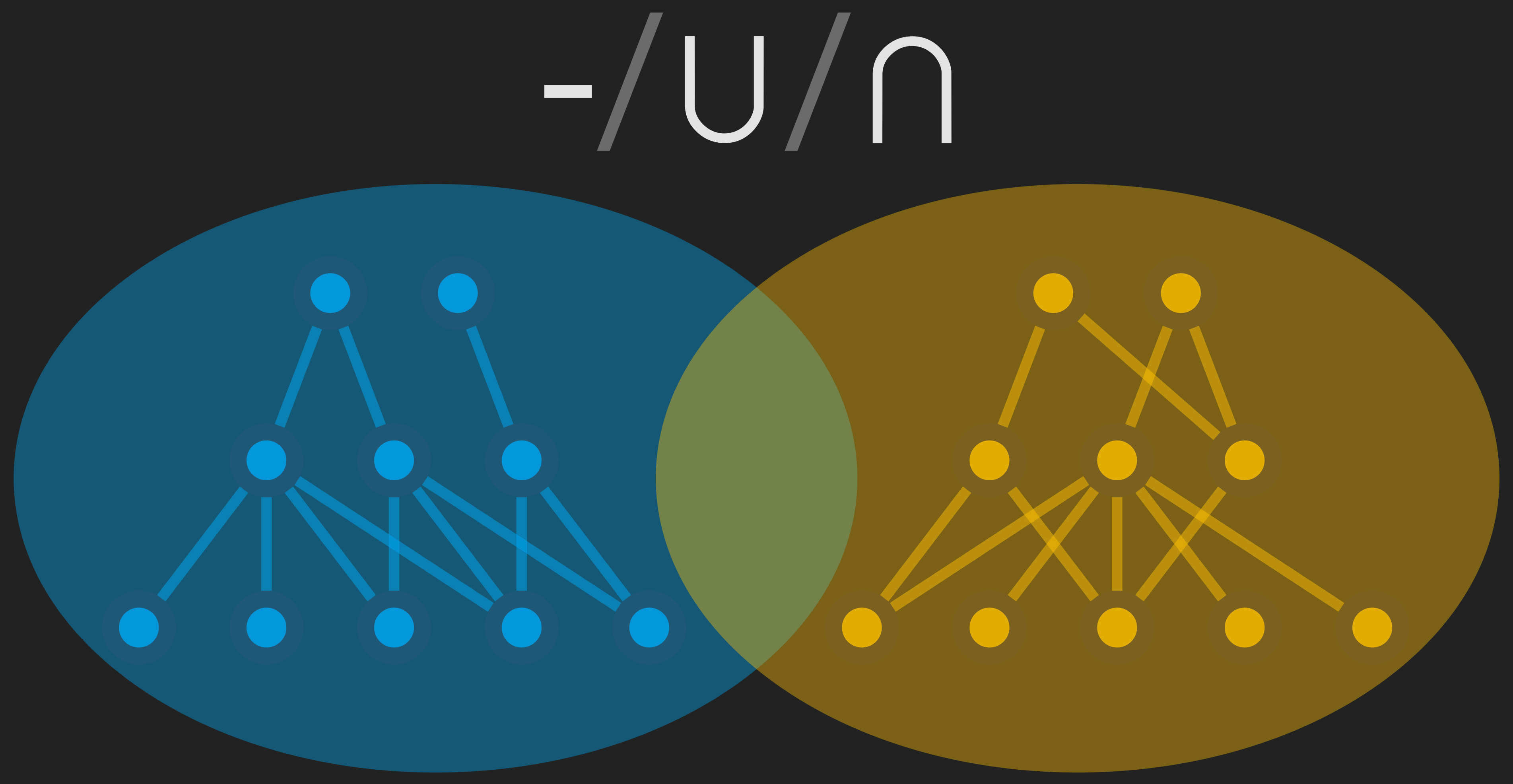
Future Work

Future Work

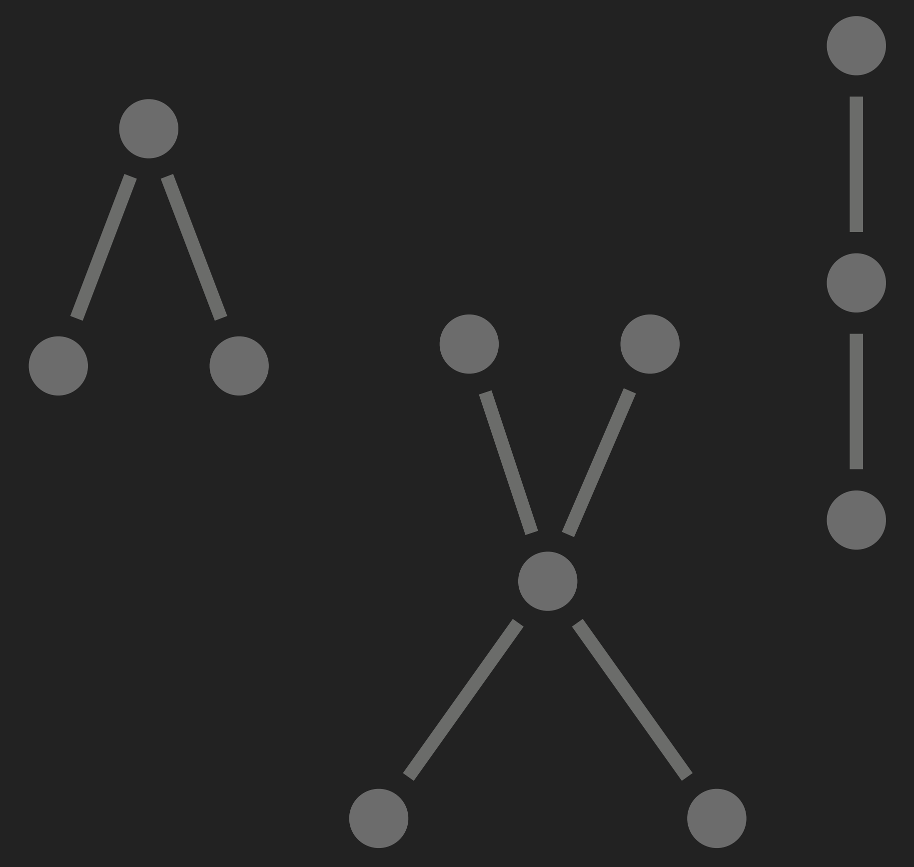


Interactive attribution graph comparison

Future Work

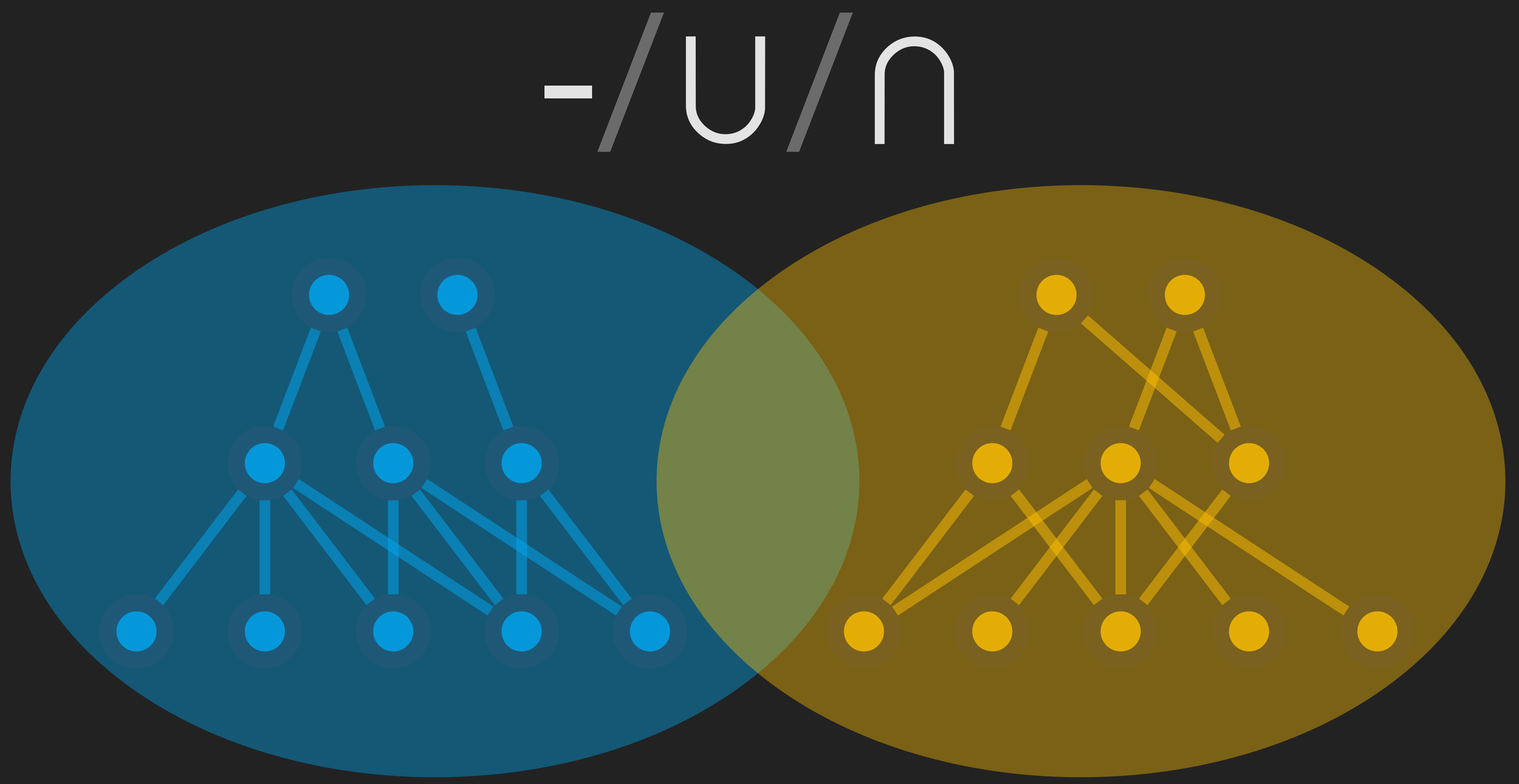


Interactive attribution
graph comparison

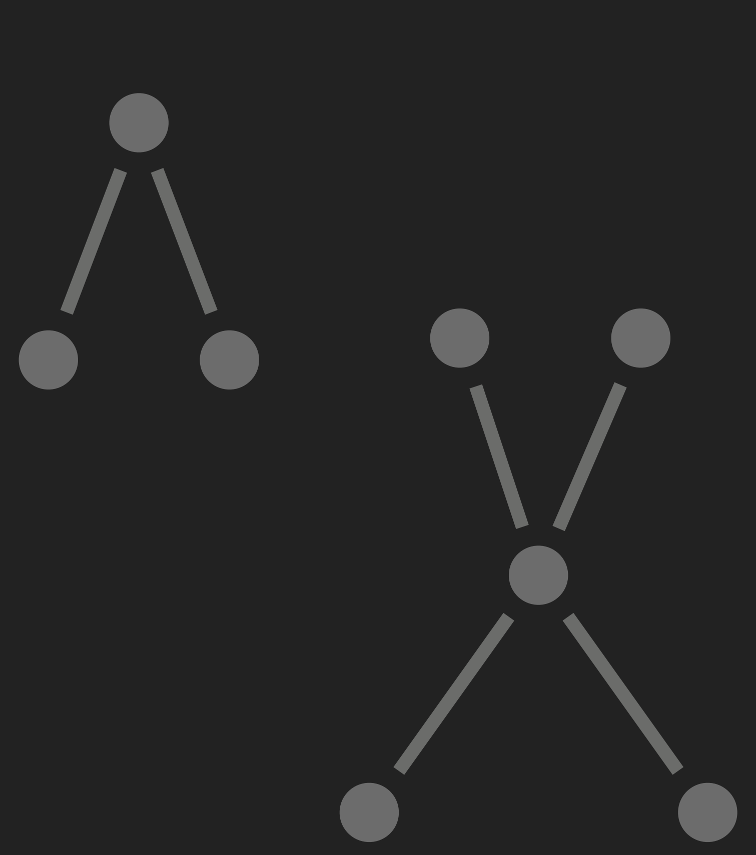


Mining for
subgraphs motifs

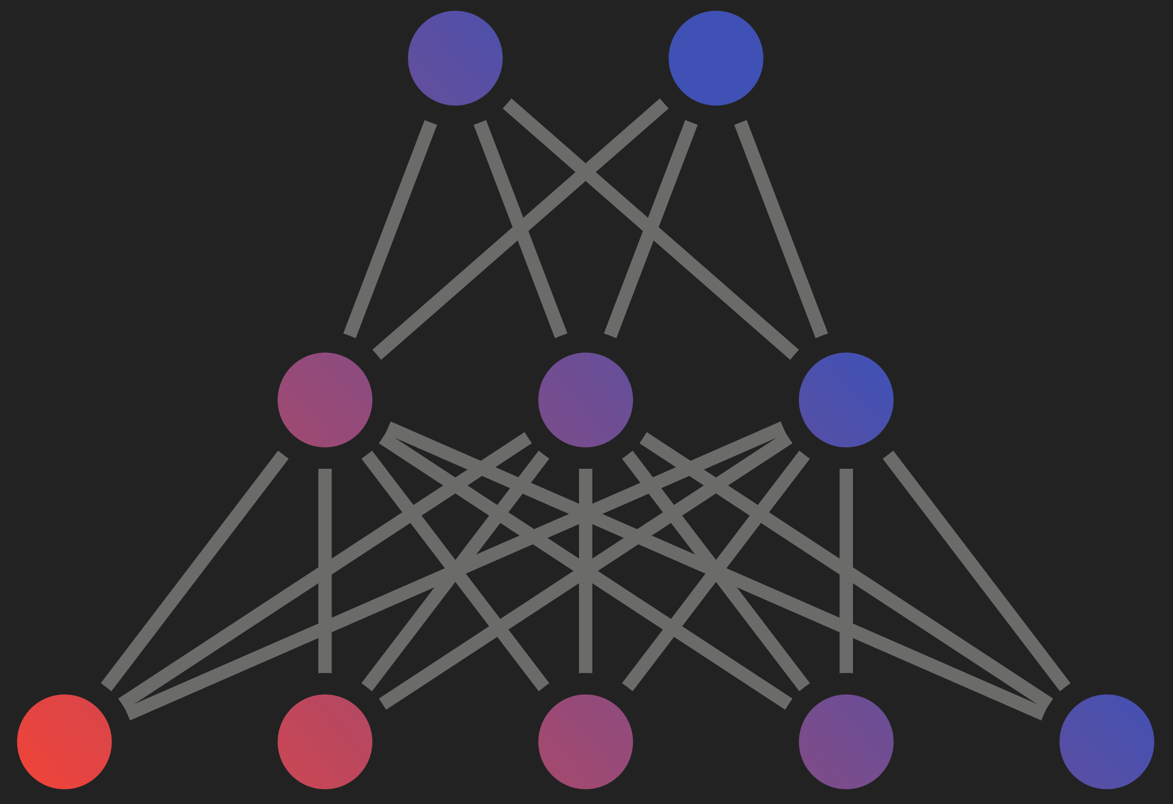
Future Work



Interactive attribution graph comparison



Mining for subgraphs motifs



Adversarial attacks


SUMMIT
MODEL
InceptionV1
DATASET
ImageNet
CLASSES
1,000
INSTANCES
1,281,024
What is SUMMIT?
LAYER
mixed
3a
3b
4a
4b
4c
4d
4e
5a
5b
CLASS
white_wolf
INSTANCES
1299
ACCURACY
81.8%
PROBABILITIES
FILTER GRAPH
ADJUST WIDTH
ADJUST HEIGHT
white wolf
white wolf
white wolf 81.8%
red wolf 69.9%
timber wolf 64.2%
arctic fox 87.1%
lion 87.1%


SUMMIT
MODEL
InceptionV1
DATASET
ImageNet
CLASSES
1,000
INSTANCES
1,281,024
What is SUMMIT?
LAYER
mixed
3a
3b
4a
4b
4c
4d
4e
5a
5b
CLASS
white_wolf
INSTANCES
1299
ACCURACY
81.8%
PROBABILITIES
FILTER GRAPH
ADJUST WIDTH
ADJUST HEIGHT
white wolf
white wolf
white wolf
81.8%
red wolf
69.9%
timber wolf
64.2%
arctic fox
87.1%
lion
87.1%


SUMMIT
MODEL
InceptionV1
DATASET
ImageNet
CLASSES
1,000
INSTANCES
1,281,024
What is SUMMIT?
LAYER
mixed
3a
3b
4a
4b
4c
4d
4e
5a
5b
CLASS
white_wolf
INSTANCES
1299
ACCURACY
81.8%
PROBABILITIES
FILTER GRAPH
ADJUST WIDTH
ADJUST HEIGHT
white wolf
white wolf
white wolf
81.8%
red wolf
69.9%
timber wolf
64.2%
arctic fox
87.1%
lion
87.1%

What is SUMMIT?

Understanding how neural networks make predictions remains a fundamental challenge. Existing work on interpreting neural network predictions for images often focuses on explaining predictions for single images or neurons, yet predictions are computed from millions of weights optimized over millions of images—such explanations can easily miss a bigger picture.

We present **SUMMIT**, an interactive visualization that scalably summarizes what features a deep learning model has learned and how those features interact to make predictions.

How does it work?

SUMMIT introduces two new scalable summarization techniques that aggregate activations and neuron-influences to create *attribution graphs:* a class-specific visualization that simultaneously highlights *what* features a neural network detects and *how* they are related.

Our work joins a growing body of open-access research that aims to use interactive visualization to explain complex inner workings of modern machine learning techniques. We believe our summarization approach that builds entire class representations is an important step for developing higher-level explanations for neural networks. We hope our work will inspire deeper engagement from both the information visualization and machine learning communities to further develop human-centered tools for artificial intelligence.

Credits

Summit was created by Fred Hohman, Haekyu Park, Caleb Robinson, and Polo Chau at Georgia Tech. We also thank Nilaksh Das and the Georgia Tech Visualization Lab for their support and constructive feedback. This work is supported by a NASA Space Technology Research Fellowship and NSF grants IIS-1563816, CNS-1704701, and TWC-1526254.

Summit: Scaling Deep Learning Interpretability by Visualizing Activation and Attribution Summarizations
Fred Hohman, Haekyu Park, Caleb Robinson, and Duen Horng (Polo) Chau.
IEEE Transactions on Visualization and Computer Graphics (TVCG, Proc. VAST'19). 2020.

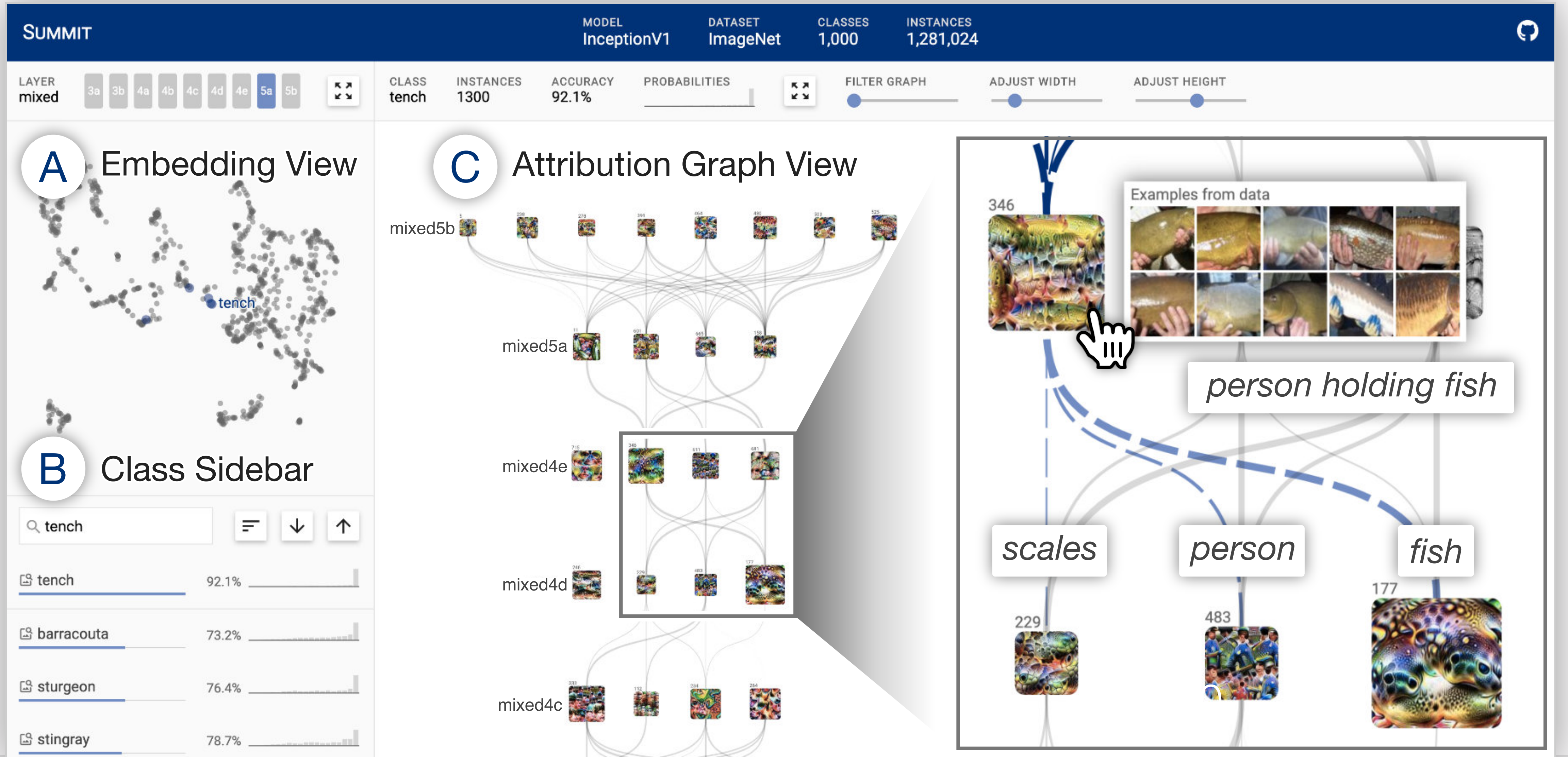
- ⛰️ **Live demo:** fredhohman.com/summit
- 📘 **Paper:** https://fredhohman.com/papers/19-summit-vast.pdf
- 🎥 **Video:** https://youtu.be/J4GMLvoH1ZU
- 💻 **Code:** https://github.com/fredhohman/summit
- 📺 **Slides:** coming October 2019!

SUMMIT

Visualizing Activation and Attribution Summarizations

fredhohman.com/summit

Demo | Paper | Video | Code | Slides



Thanks!


Fred Hohman
@fredhohman
Georgia Tech


Haekyu Park
Georgia Tech


Caleb Robinson
Georgia Tech


Polo Chau
Georgia Tech

Georgia Tech | Polo Club of Data Science | NASA | NSF

We thank Nilaksh Das, the GT Vis Lab, and the reviewers for their constructive feedback. Funded by the NSF and a NASA Fellowship.