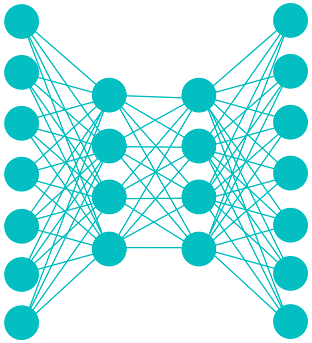
Lecture Notes for

Neural Networks and Machine Learning

CNN Circuits Continued

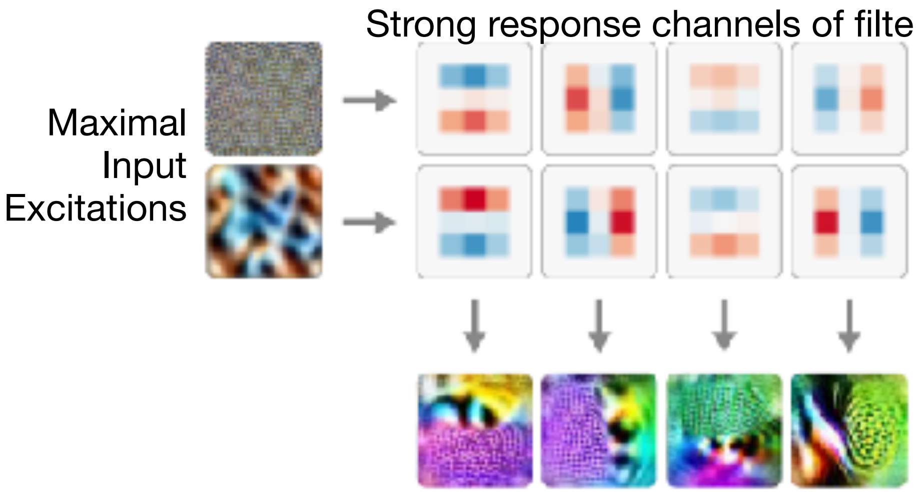


Logistics and Agenda

- Logistics
  - Grading Update
- Agenda
  - Finish Circuits
  - Student Paper Presentation
  - Next Time (or today, if time):
    - Fully Convolutional Networks

Review of Equivariant Circuits

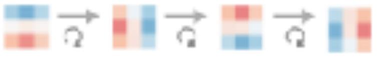
- Possible to reveal patterns of circuits via sets of weights



Strong response channels of filter

Maximal Input Excitations

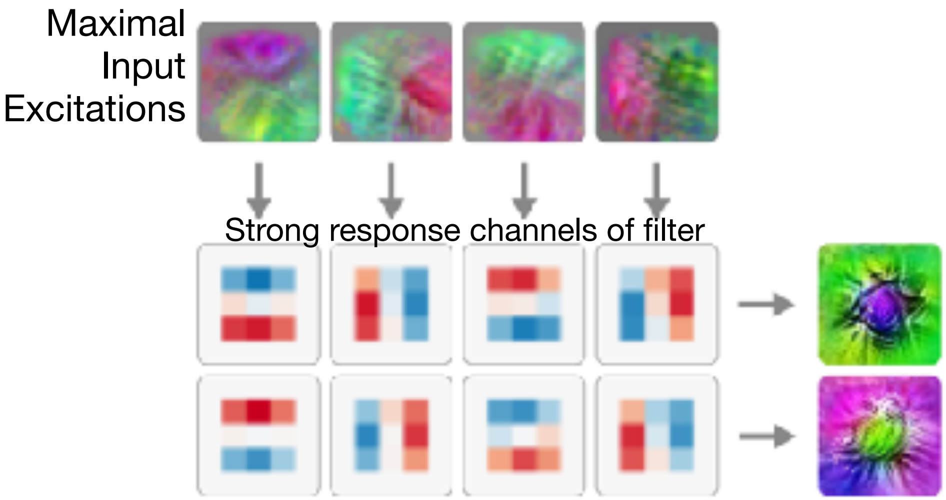
High-low frequency detectorsrespond to a high-frequency neuron factor on one side and low frequency on the other. Notice how the weights rotate:



This makes them rotationally equivariant.

🟥 positive (excitation)

🟦 negative (inhibition)



Maximal Input Excitations

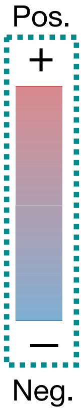
Strong response channels of filter

Rotational equivariance can be turned into invariance with the transpose of an invariant -> equivariant circuit.

Here, we see**color contrast units**(rotationally equivariant) combine to make **color center surround units**(rotationally invariant). Again, notice how the weights rotate, forming the same pattern we saw above with high-low frequency detectors, but with inputs and outputs swapped.

🟥 positive (excitation)

🟦 negative (inhibition)

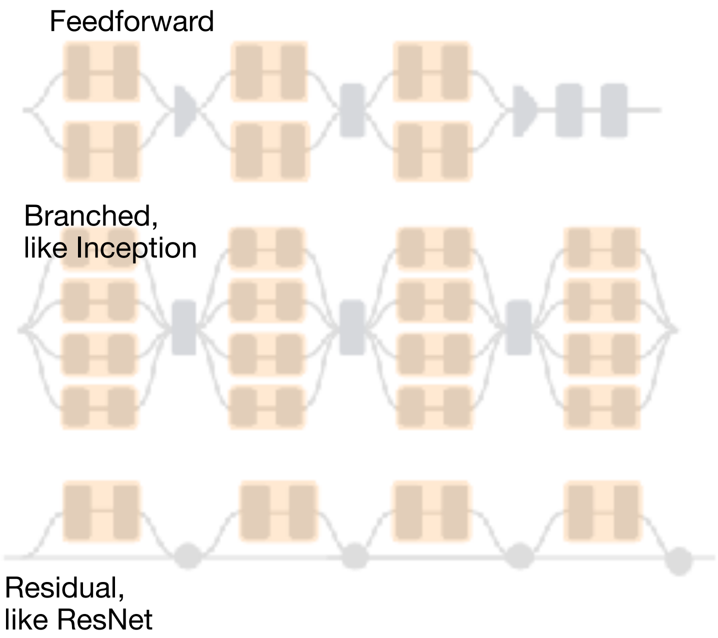


Pos.

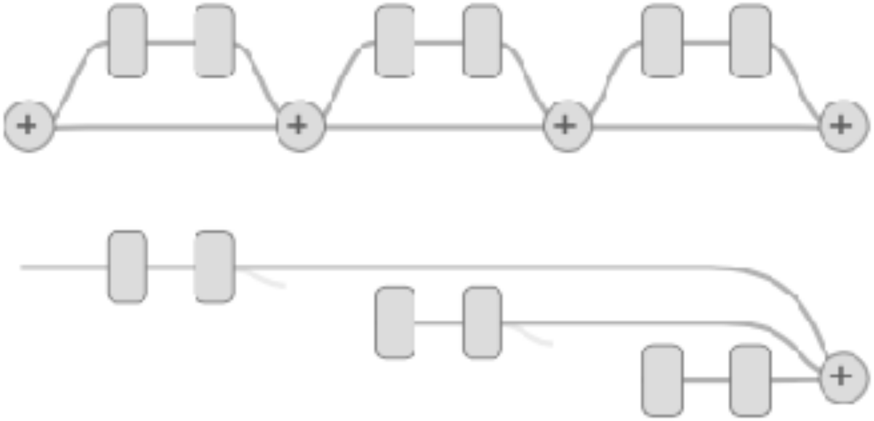
Neg.

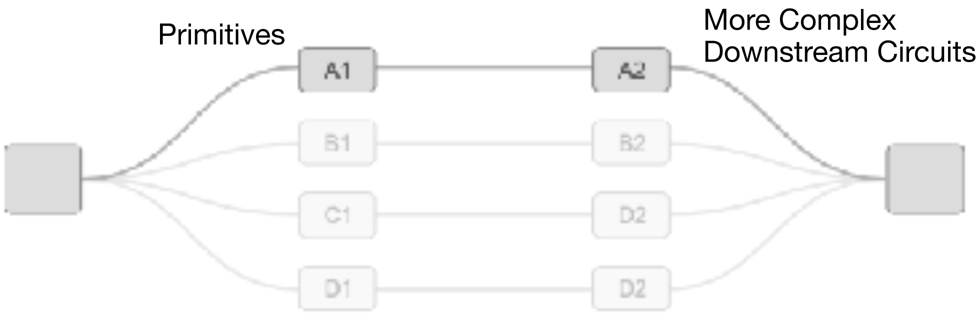
Olah, et al., “Naturally Occurring Equivariance in NN”, 2021.

Branch Specialization



Two ways of looking at residual networks



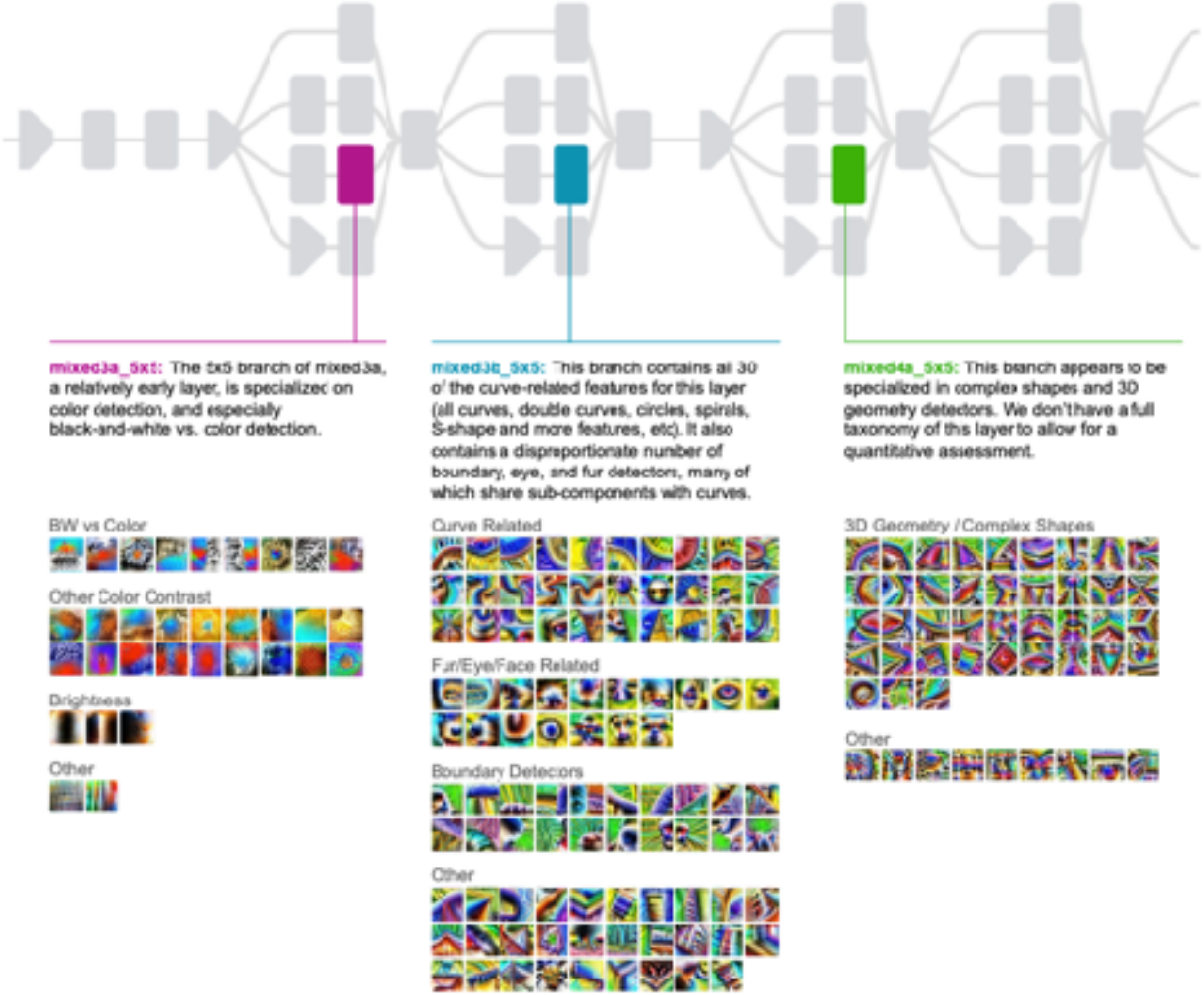


- Specialized branches are consistent across many architectures, support the idea of an interconnected graph of operations

https://distill.pub/2020/circuits/branch-specialization/

Voss, et al., “Branch Specialization", Distill, 2021.

Branch Specialization



mixed3a_5x5: The 5x5 branch of mixed3a, a relatively early layer, is specialized on color detection, and especially black-and-white vs. color detection.

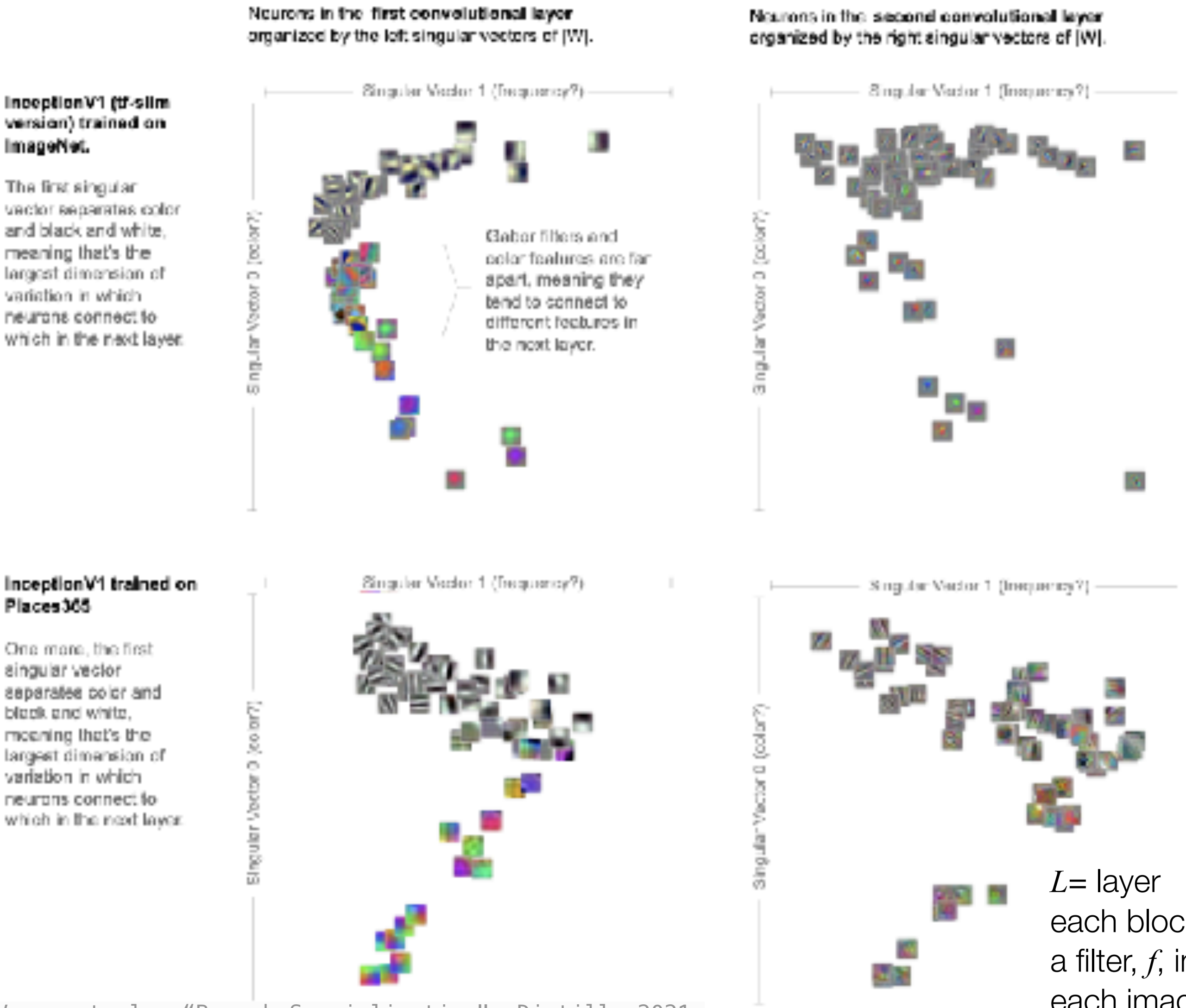
mixed3b_5x5: This branch contains all 30 of the curve-related features for this layer (all curves, double curves, circles, spirals, S-shape and more features, etc). It also contains a disproportionate number of boundary, eye, and fur detectors, many of which share sub-components with curves.

mixed4a_5x5: This branch appears to be specialized in complex shapes and 3D geometry detectors. We don't have a full taxonomy of this layer to allow for a quantitative assessment.

Motifs appear in branches. Similar clusters of operations can be found across different architectures

Voss, et al., "Branch Specialization", Distill, 2021.

Investigating Connection Clusters via SVD



Neurons in the first convolutional layer organized by the left singular vectors of |W|.

Neurons in the second convolutional layer organized by the right singular vectors of |W|.

InceptionV1 (tf-slim version) trained on ImageNet.

The first singular vector separates color and black and white, meaning that's the largest dimension of variation in which neurons connect to which in the next layer.

Gabor filters and color features are far apart, meaning they tend to connect to different features in the next layer.

InceptionV1 trained on Places365

One more, the first singular vector separates color and black and white, meaning that's the largest dimension of variation in which neurons connect to which in the next layer.

Singular Vector 1 (frequency?)
Singular Vector 0 (color?)
Singular Vector 1 (frequency?)
Singular Vector 0 (color?)
Singular Vector 1 (frequency?)
Singular Vector 0 (color?)
Singular Vector 1 (frequency?)
Singular Vector 0 (color?)

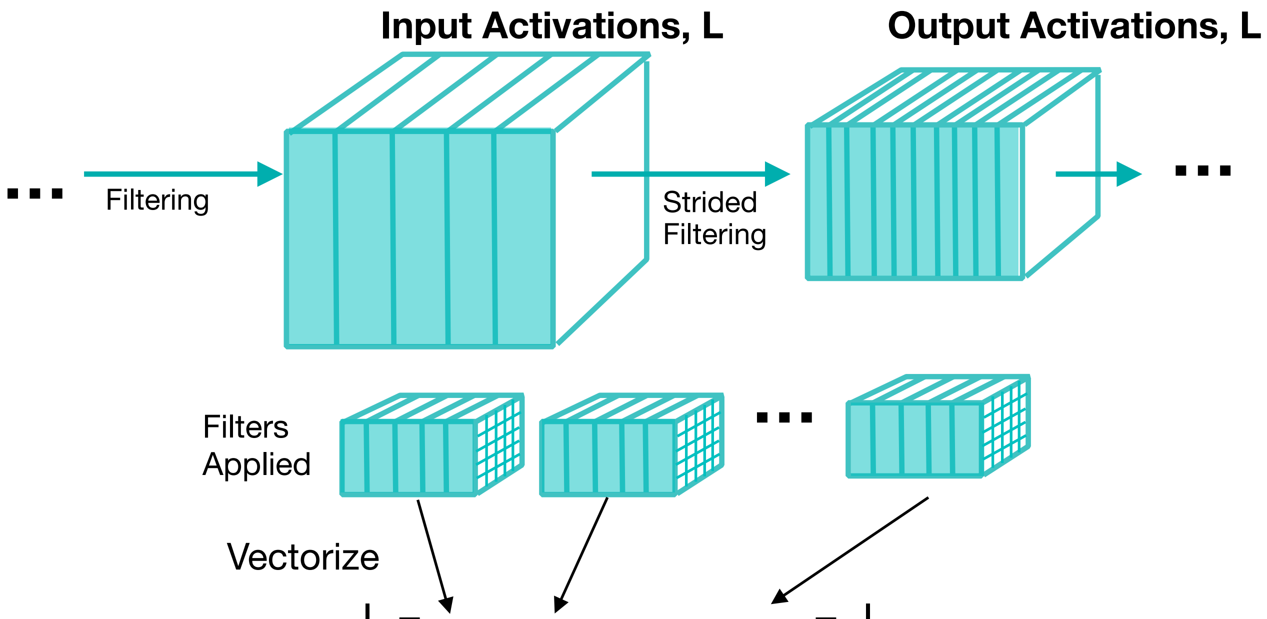
Voss, et al., “Branch Specialization", Distill, 2021.

- Singular Value Decomposition (SVD) decomposes a matrix into three elements
  - $M = U\Sigma V^T$
  - U is eig-vec of MM^T
V is eig-vec of M^TM
  - U and V are orthogonal such that $UU^T=I \quad VV^T=I$
  - Σ is a diagonal matrix of the singular values
  - These values characterize the variability in a matrix

L= layer
each block is
a filter, f, in layer
each image is the optimized excitation

$$SVD(|\mathbf{W}^{(L)}|)$$

Structure of Each Tensor: Channels x Rows x Columns



Input Activations, L

Output Activations, L

… Filtering

Strided Filtering

…

Filters Applied

…

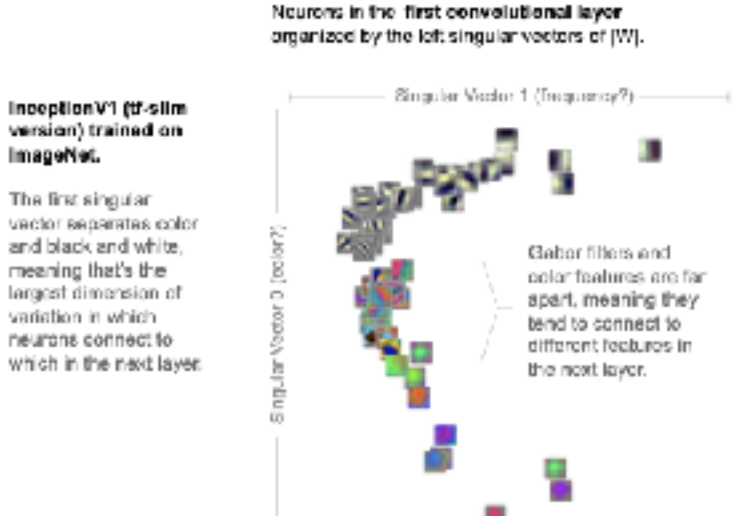
Vectorize

$$\mathbf{M} = |\mathbf{W}^{(L)}| = \left| \begin{bmatrix} \uparrow & \uparrow & \uparrow & \uparrow \\ \mathbf{f}_0 & \mathbf{f}_1 & \dots & \mathbf{f}_K \\ \downarrow & \downarrow & \downarrow & \downarrow \end{bmatrix} \right|$$

$$\downarrow$$

$$SVD(|\mathbf{W}^{(L)}|) \longrightarrow$$

- Singular Value Decomposition (SVD) decomposes a matrix into three elements
  - $M = U\Sigma V^T$
  - U is eig-vec of MM^T V is eig-vec of M^TM
  - U and V are orthogonal such that $UU^T=I \quad VV^T=I$
  - Σ is a diagonal matrix of the singular values
  - These values characterize the variability in a matrix



InceptionV1 (tf-slim version) trained on ImageNet.

The first singular vector separates color and black and white, meaning that's the largest dimension of variation in which neurons connect to which in the next layer.

Neurons in the first convolutional layer organized by the left singular vectors of |W|.

Gabor filters and color features are far apart, meaning they tend to connect to different features in the next layer.

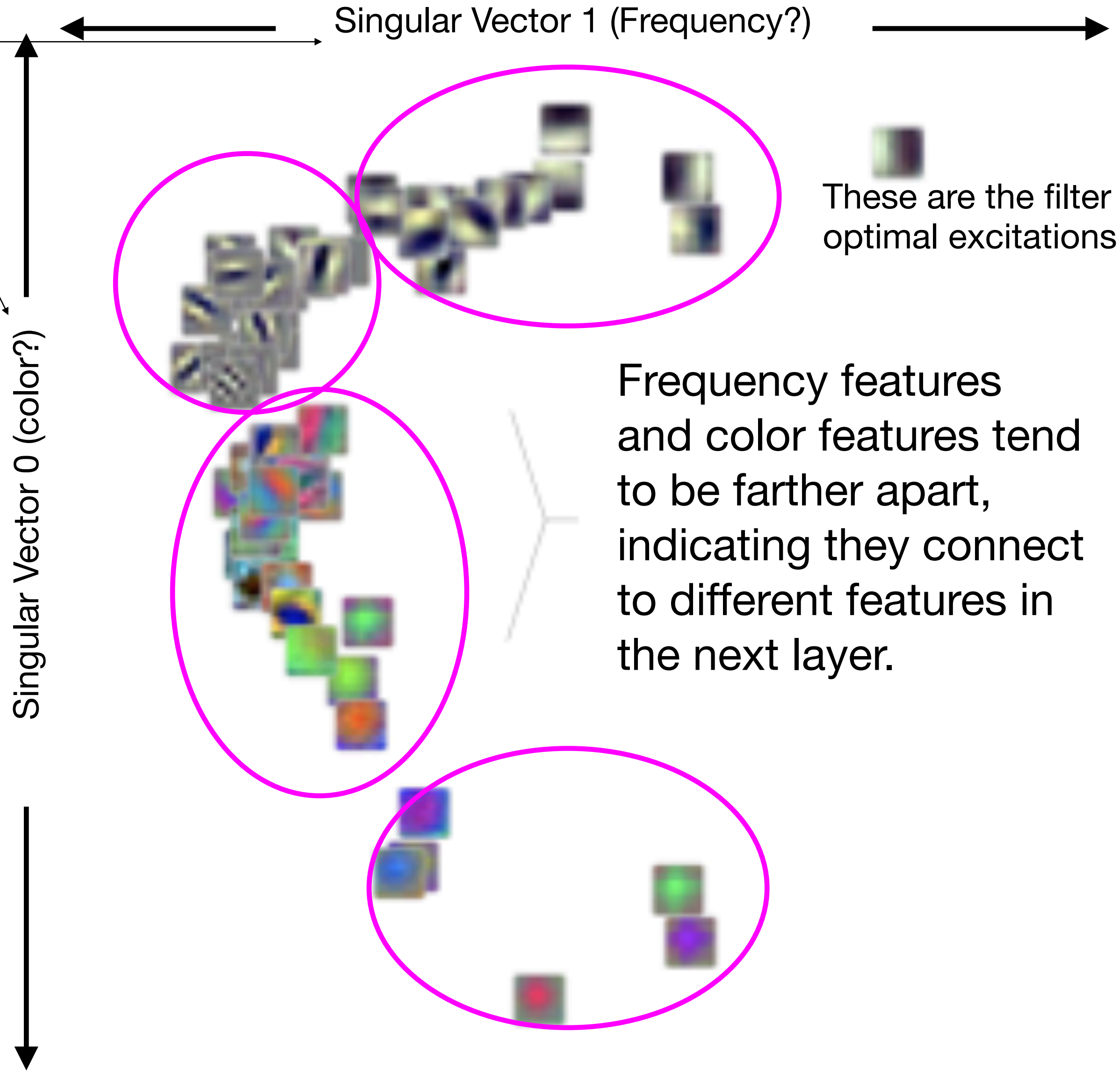
$$\mathbf{U}^{(L)} = \begin{bmatrix} \uparrow & \uparrow & & \uparrow \\ \mathbf{e}_0 & \mathbf{e}_1 & \dots & \mathbf{e}_K \\ \downarrow & \downarrow & & \downarrow \end{bmatrix}$$

- $M = U\Sigma V^T$
- U is eig-vec of MM^T

$$SVD(|\mathbf{W}^{(L)}|)$$

$$SVD(\mathbf{M})$$

$$\left| \begin{bmatrix} \uparrow & \uparrow & \uparrow & \uparrow \\ \mathbf{f}_0 & \mathbf{f}_1 & \dots & \mathbf{f}_K \\ \downarrow & \downarrow & \downarrow & \downarrow \end{bmatrix} \right|$$

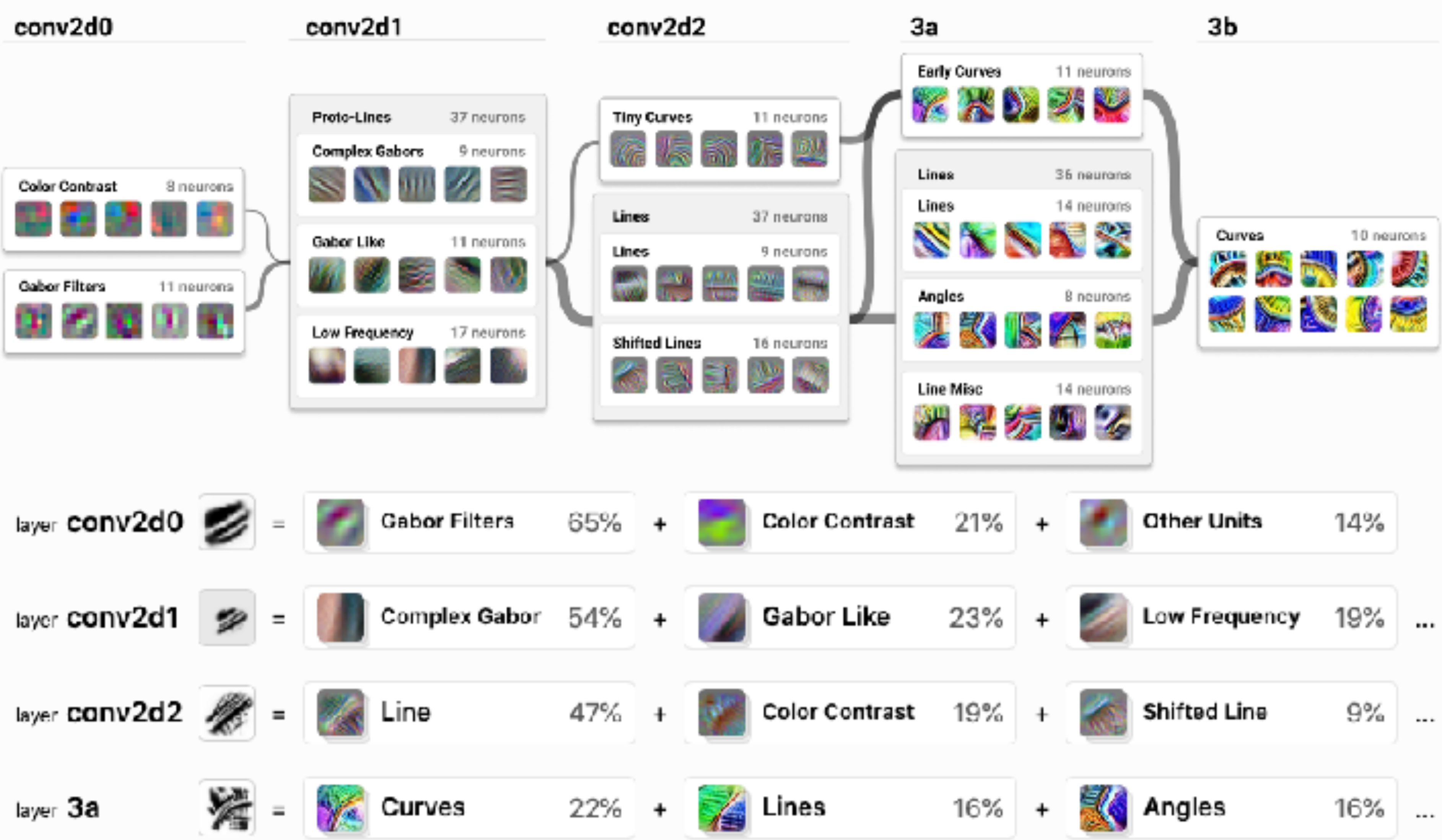


These are the filter optimal excitations

Frequency features and color features tend to be farther apart, indicating they connect to different features in the next layer.

Clusters tend to be motifs, and
separation of clusters reveals connections between layers (like edges in a graph)

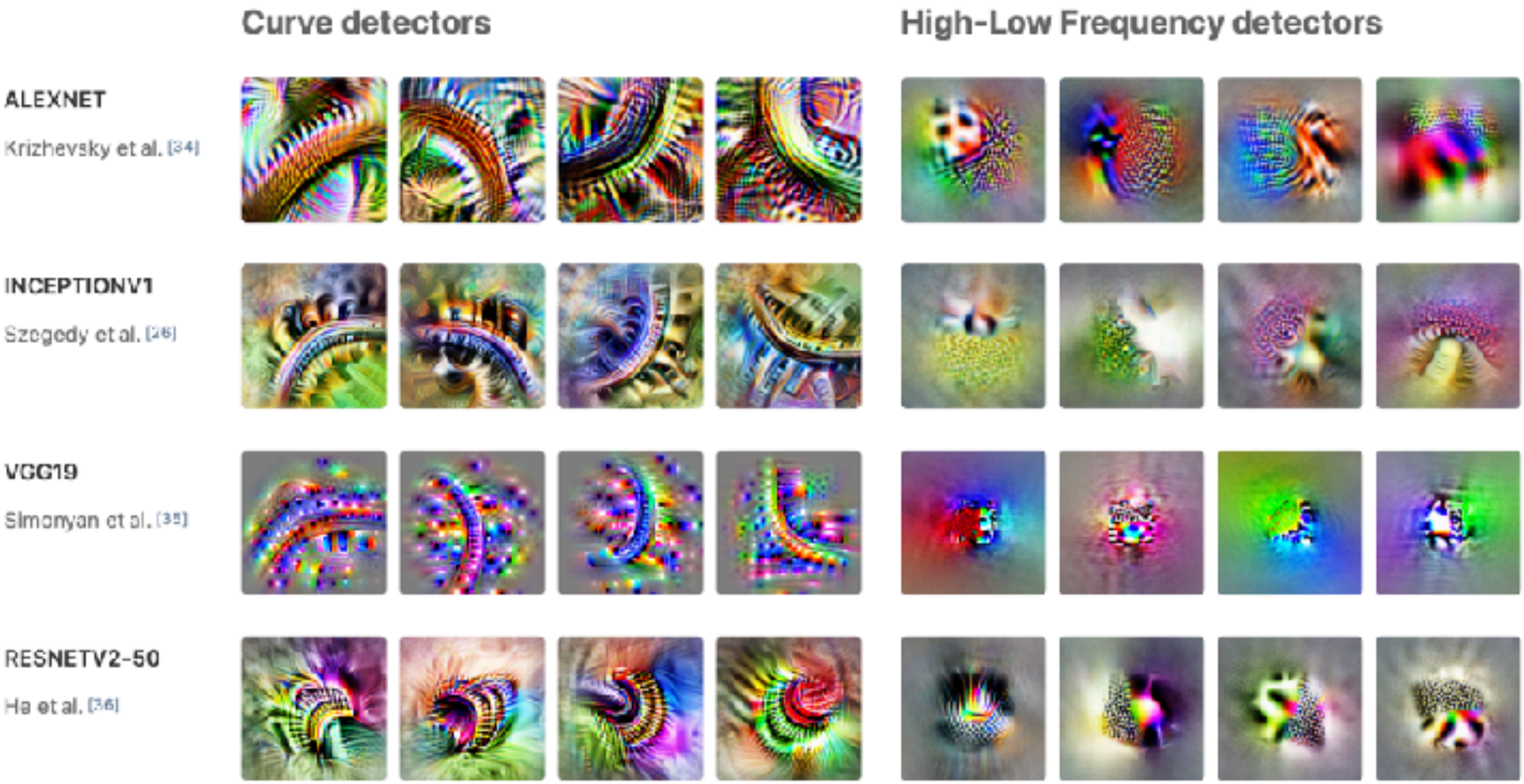
Neural Nets: Directed Graph of Circuits



Voss, et al., “Branch Specialization", Distill, 2021.

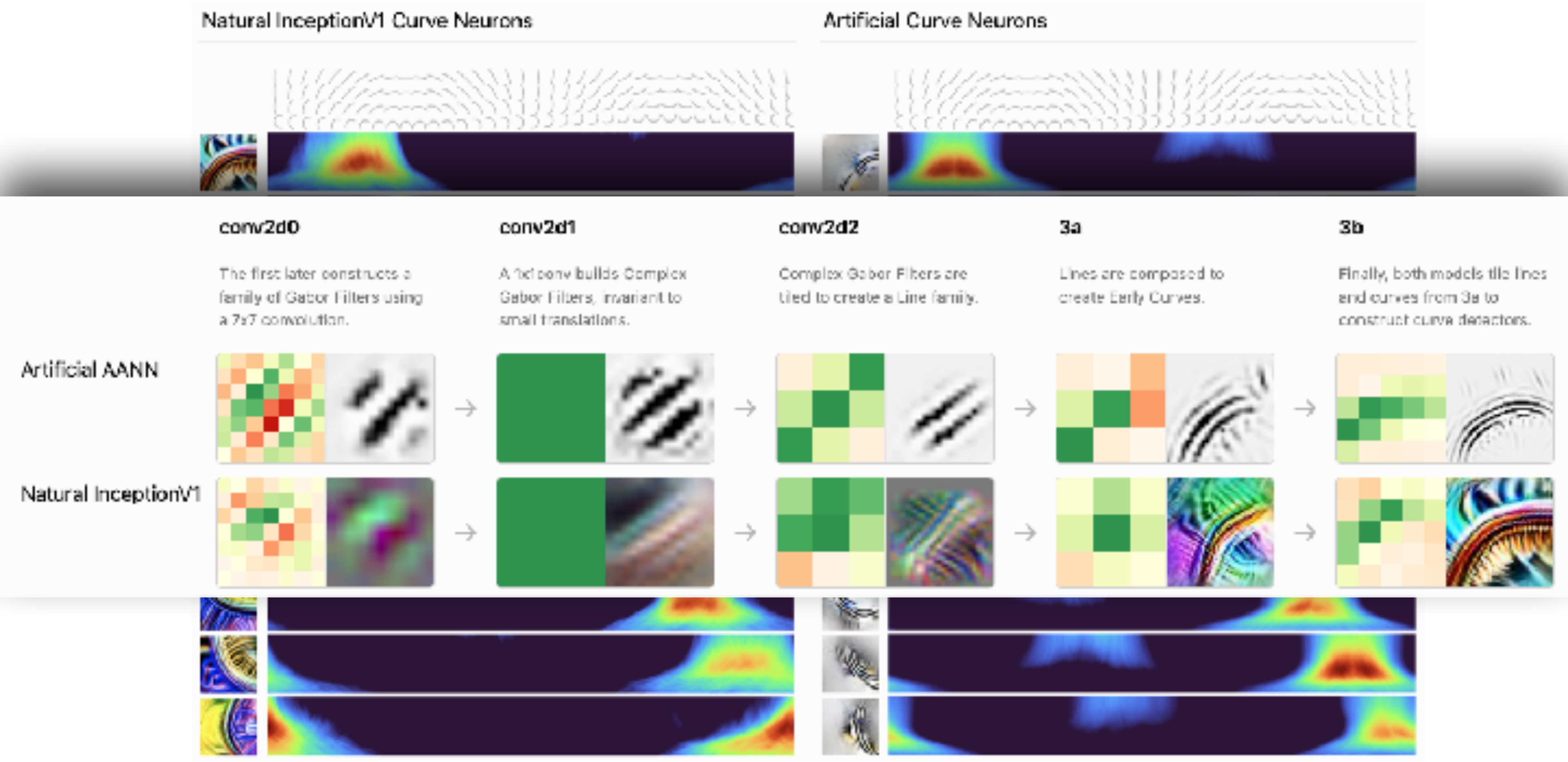
Universality of Circuits

- Analogous features and circuits form across models and tasks



Reverse Engineering a Circuit

- With assumption of what feature is, a circuit can be implemented by hand that nearly identically follows the assumed functionality



conv2d0

The first later constructs a family of Gabor Filters using a 7x7 convolution.

conv2d1

A 1x1 conv builds Complex Gabor Filters, invariant to small translations.

conv2d2

Complex Gabor Filters are tiled to create a Line family.

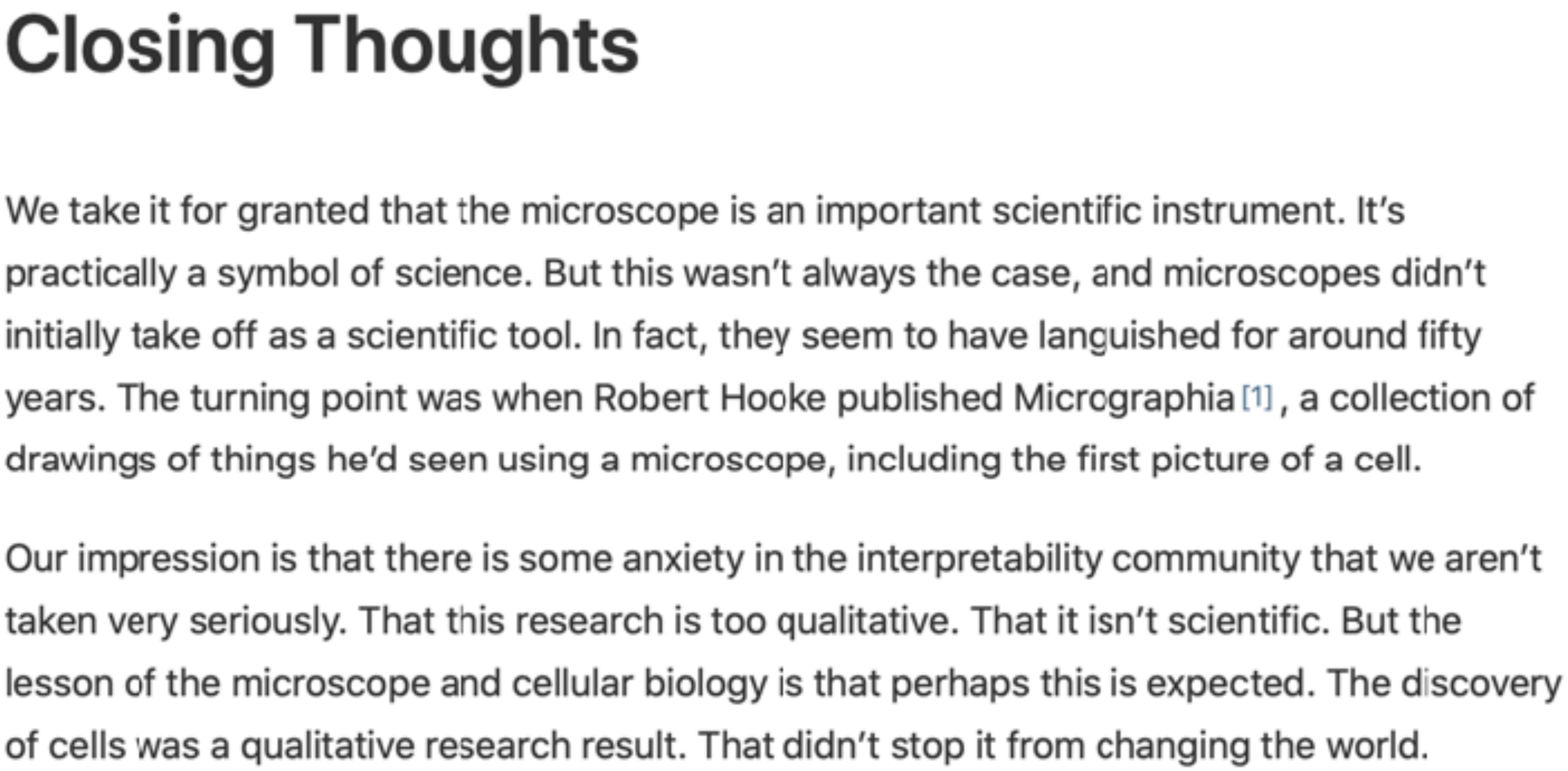
3a

Lines are composed to create Early Curves.

3b

Finally, both models tile lines and curves from 3a to construct curve detectors.

Closing Thoughts from OpenAI Researchers



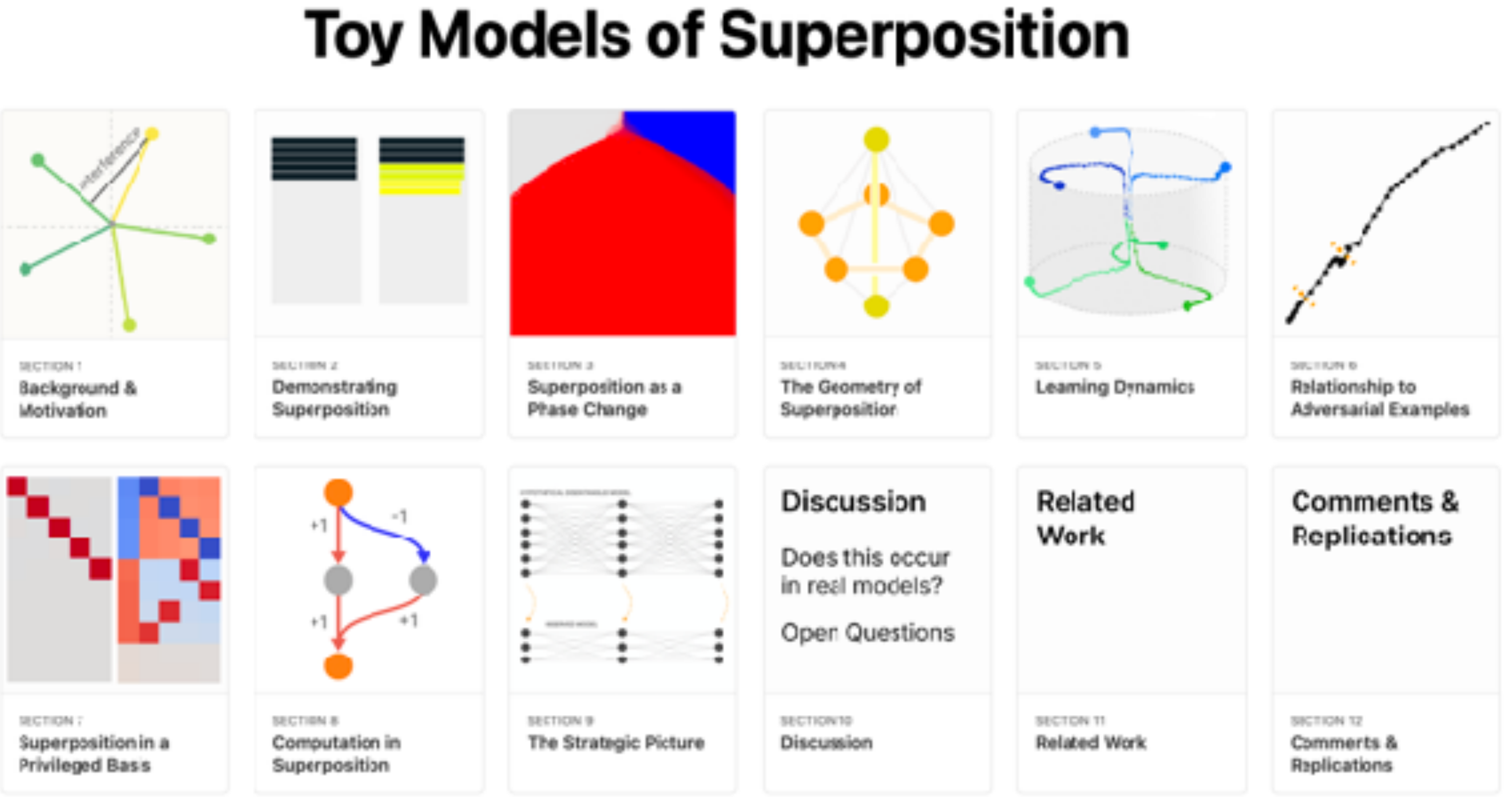
Closing Thoughts

We take it for granted that the microscope is an important scientific instrument. It's practically a symbol of science. But this wasn't always the case, and microscopes didn't initially take off as a scientific tool. In fact, they seem to have languished for around fifty years. The turning point was when Robert Hooke published Micrographia [1], a collection of drawings of things he'd seen using a microscope, including the first picture of a cell.

Our impression is that there is some anxiety in the interpretability community that we aren't taken very seriously. That this research is too qualitative. That it isn't scientific. But the lesson of the microscope and cellular biology is that perhaps this is expected. The discovery of cells was a qualitative research result. That didn't stop it from changing the world.

https://distill.pub/2020/circuits/zoom-in/

Paper Presentation



Toy Models of Superposition

SECTION 1
Background & Motivation

SECTION 2
Demonstrating Superposition

SECTION 3
Superposition as a Phase Change

SECTION 4
The Geometry of Superposition

SECTION 5
Learning Dynamics

SECTION 6
Relationship to Adversarial Examples

SECTION 7
Superposition in a Privileged Basis

SECTION 8
Computation in Superposition

SECTION 9
The Strategic Picture

Discussion

Does this occur in real models?

Open Questions

SECTION 10
Discussion

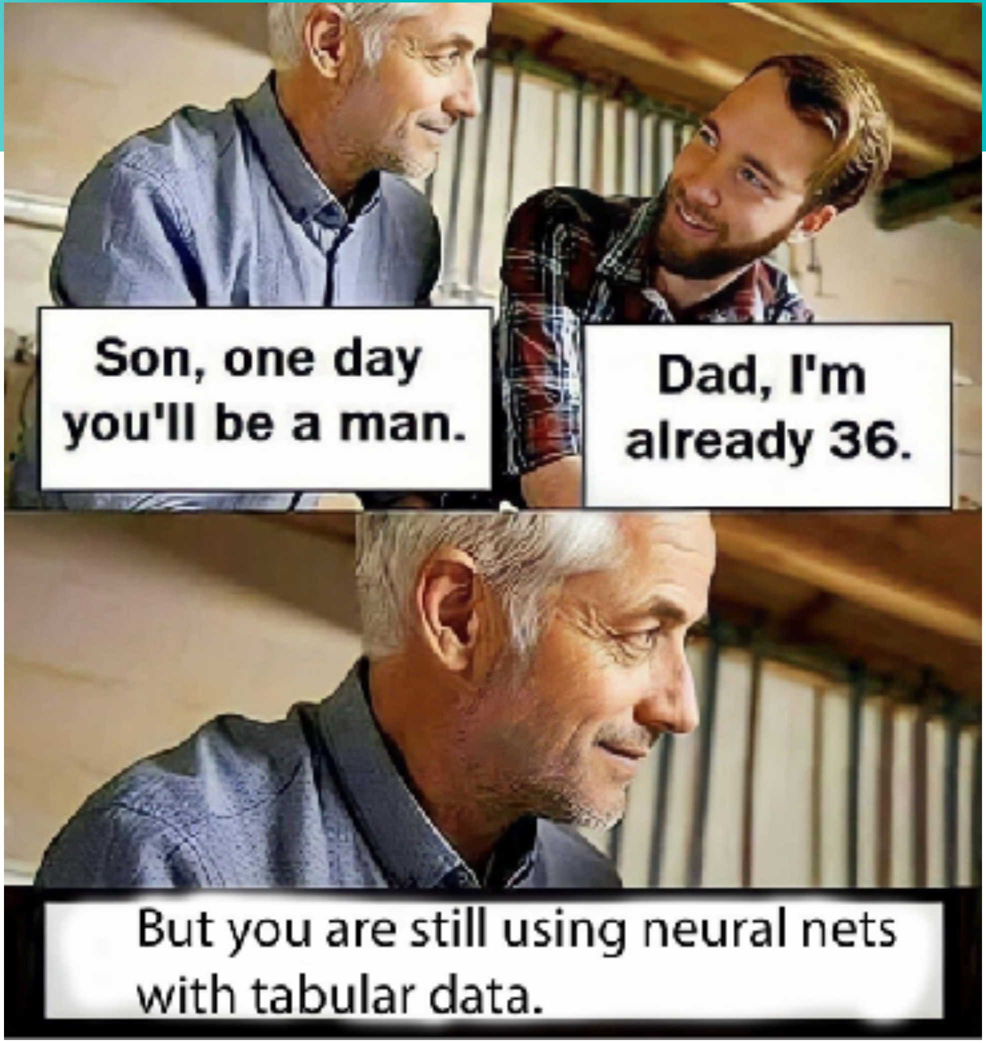
Related Work

SECTION 11
Related Work

Comments & Replications

SECTION 12
Comments & Replications

Lab Three Town Hall Review


Son, one day you'll be a man.
Dad, I'm already 36.
But you are still using neural nets with tabular data.

Lecture Notes for

Neural Networks and Machine Learning

CNN Circuits

Next Time:
Fully Convolutional Learning

Reading: Chollet 5.4