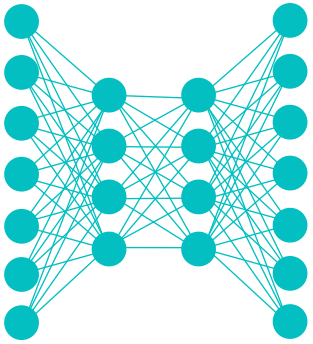
Lecture Notes for

# Neural Networks and Machine Learning

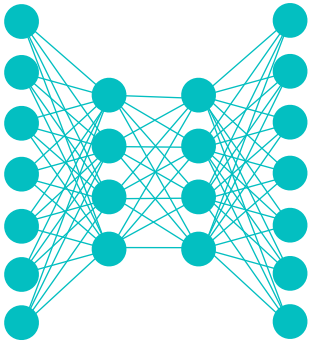
CNN Visualization



Lecture Notes for

# Neural Networks and Machine Learning

CNN Circuits



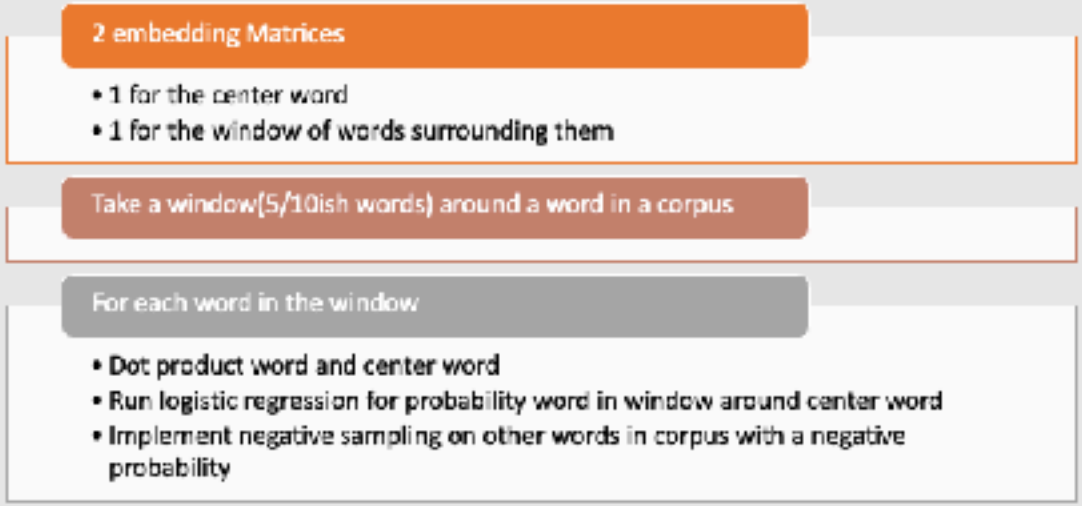
# Logistics and Agenda

- Logistics
  - Student Presentations (questions?)
    - If distance, can submit one page summary, rather than presentation
- Agenda
  - LastTime: Visualizing Convolutional Architectures
  - Today: Circuits in CNNs
  - Student Paper Presentation: Graph NN
  - Lab One Town Hall

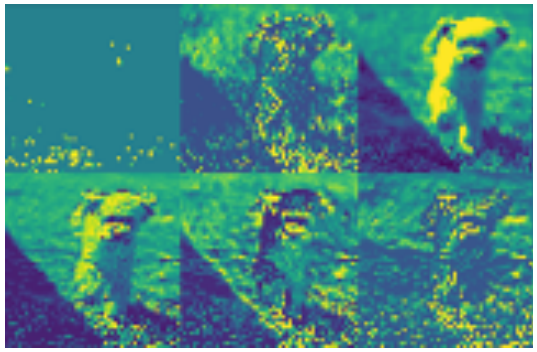


# Student Paper Presentation

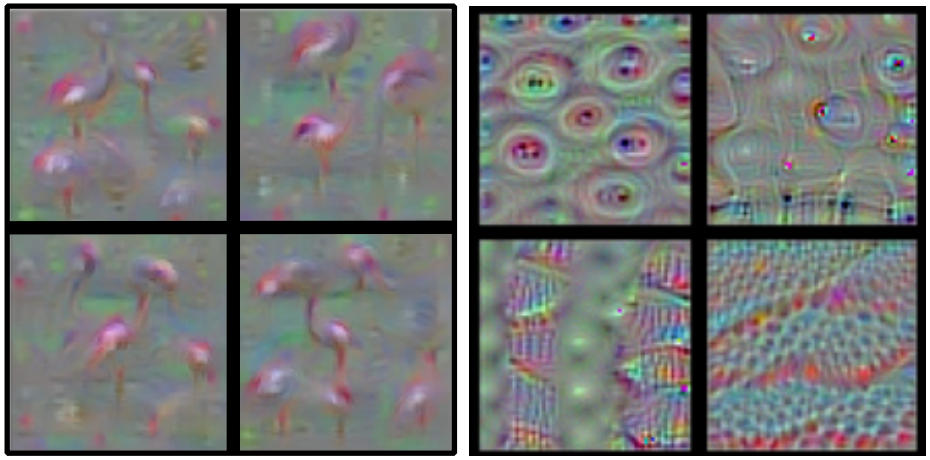
## Word2Vec Review (SkipGram)



# Review: our visualization toolset

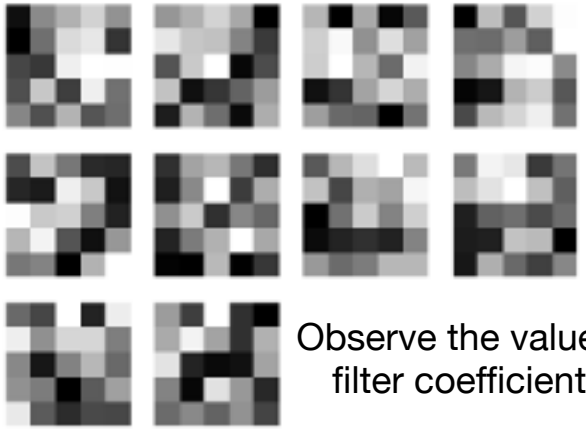


Visualize Activation
in response to input image



Visualize input maximized
to activate a certain class of filter

Use final convolutional layer to
see most influential part of input



Observe the value of
filter coefficients

# Circuits and Features

We believe that neural networks consist of meaningful, understandable features. Early layers contain features like edge or curve detectors, while later layers have features like floppy ear detectors or wheel detectors. The community is divided on whether this is true. While many researchers treat the existence of meaningful neurons as an almost trivial fact—there's even a small literature studying them [15, 2, 16, 17, 4, 18, 19]—many others are deeply skeptical and believe that past cases of neurons that seemed to track meaningful latent variables were mistaken [20, 21, 22, 23, 24]. [3] Nevertheless, thousands of hours of studying individual neurons have led us to believe the typical case is that neurons (or in some cases, other directions in the vector space of neuron activations) are understandable.

Cammarata, et al., "Thread: Circuits", Distill, 2020.


# Why Visualize Trained CNN Architectures?

**From OpenAI**: Chris Olah, Nick Cammarata, Ludwig Schubert, Gabriel Goh, Michael Petrov, Shan Carter

Many important transition points in the history of science have been moments when science "zoomed in." At these points, we develop a visualization or tool that allows us to see the world in a new level of detail, and a new field of science develops to study the world through this lens.

For example, microscopes let us see cells, leading to cellular biology. Science zoomed in. Several techniques including x-ray crystallography let us see DNA, leading to the molecular revolution. Science zoomed in. Atomic theory. Subatomic particles. Neuroscience. Science zoomed in.

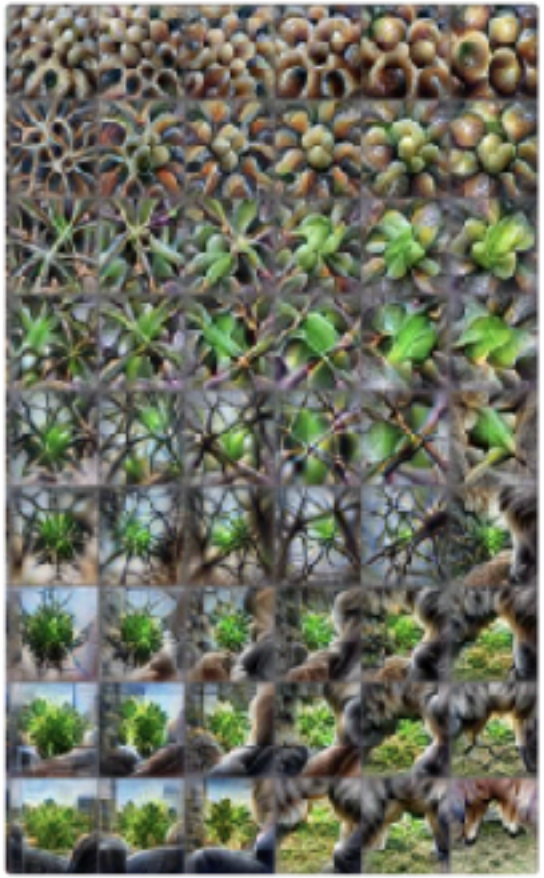
These transitions weren't just a change in precision: they were qualitative changes in what the objects of scientific inquiry are. For example, cellular biology isn't just more careful zoology. It's a new kind of inquiry that dramatically shifts what we can understand.

The famous examples of this phenomenon happened at a very large scale, but it can also be the more modest shift of a small research community realizing they can now study their topic in a finer grained level of detail.

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



# Speculative Claims for Circuits



THREE SPECULATIVE CLAIMS ABOUT NEURAL NETWORKS

**Claim 1: Features**

Features are the fundamental unit of neural networks.

They correspond to directions. [1] These features can be rigorously studied and understood.

**Claim 2: Circuits**

Features are connected by weights, forming circuits. [2]

These circuits can also be rigorously studied and understood.

**Claim 3: Universality**

Analogous features and circuits form across models and tasks.

Left: An activation atlas [13] visualizing part of the space neural network features can represent.

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



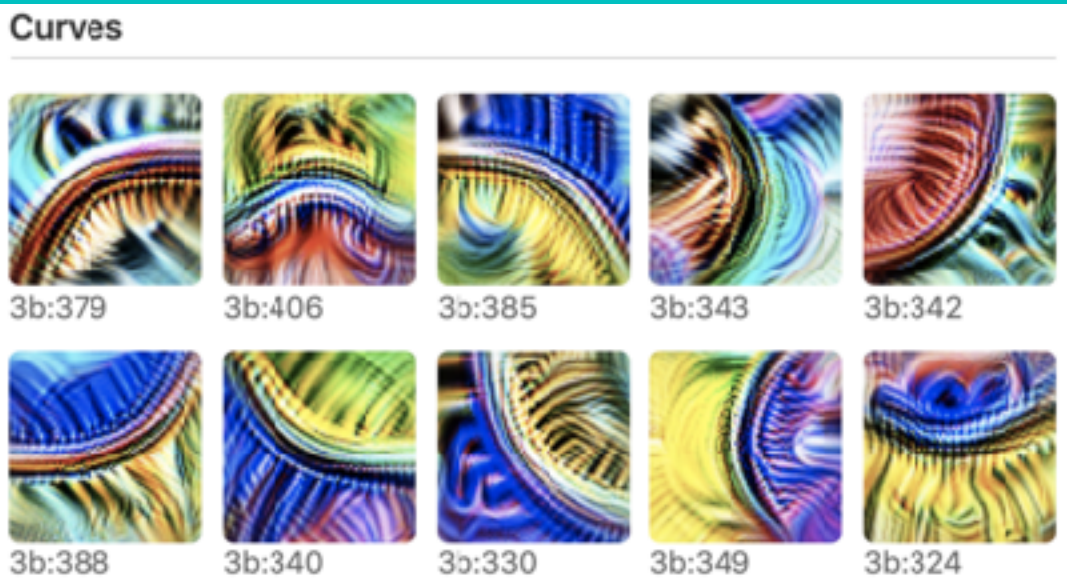
# Building Blocks: Features

- *Features are fundamental units of neural network. Features are how we describe what an activation in a network does.*

- They must be discovered, typically by:
  - Extensive visualization of excitations and filter weights (*forward analysis*)
  - Analysis of synthetic examples and dataset examples (*forward analysis*)
  - Through similarity to other features. e.g., rotations or scaling of a given feature (*parallel analysis*)
  - Through downstream features which *naturally* depend on the given feature working (*backward analysis*)
- With assumption of what **feature** is, a **circuit** can be implemented (even by hand) that nearly identically follows the assumed functionality



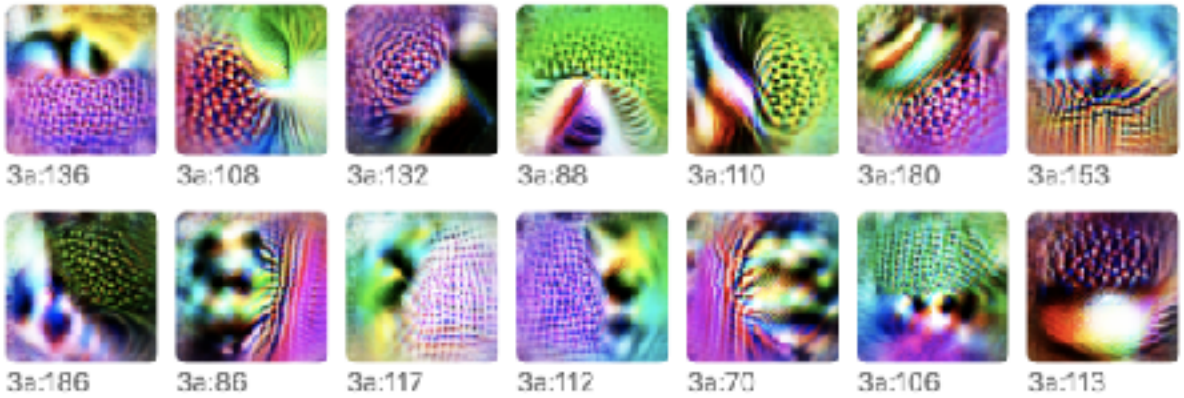
# Examples of Discovered Features

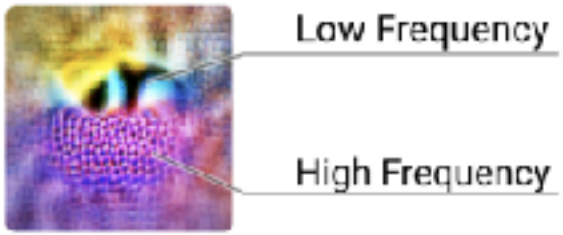



Hypothesized feature group (part of circuit)


Downstream features

High to Low Frequency Transition: perhaps good at finding blurred versus area in focus



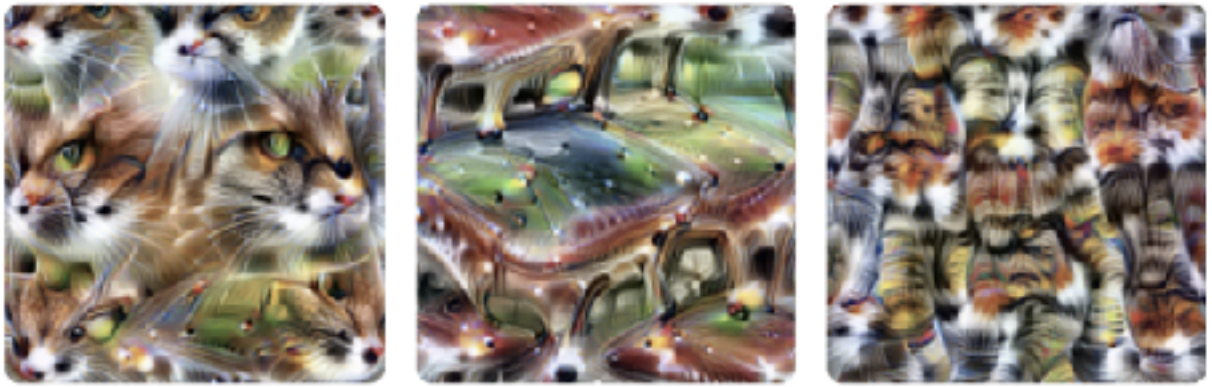
# More Examples: Higher Level Features

## Pose Invariant Dog-head Detection

Neuron 4b:409

Dataset examples for neuron 4b:409

## Polysemantic Neurons: things that become coupled…



4e:55 is a polysemantic neuron which responds to cat faces, fronts of cars, and cat legs. It was discussed in more depth in Feature Visualization [4].

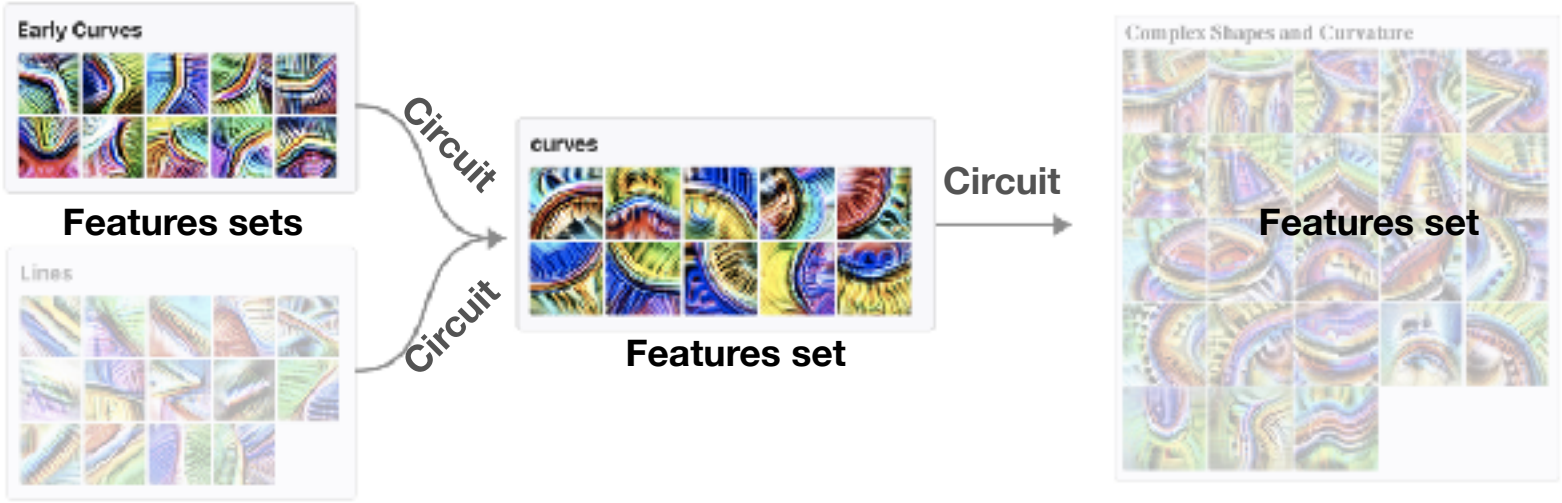
The existence of these neurons is likely one of the main criticism of network features.

**Why do these exist?**



# From Features to Circuits

- *Features are connected by weights, forming circuits*
- *"All neurons in our network are formed from linear combinations of neurons in the previous layer, followed by ReLU. If we can understand the features in both layers, shouldn't we also be able to understand the connections between them?"*
- *"Once you understand what features they're connecting together… You can literally read meaningful algorithms off of the weights."*

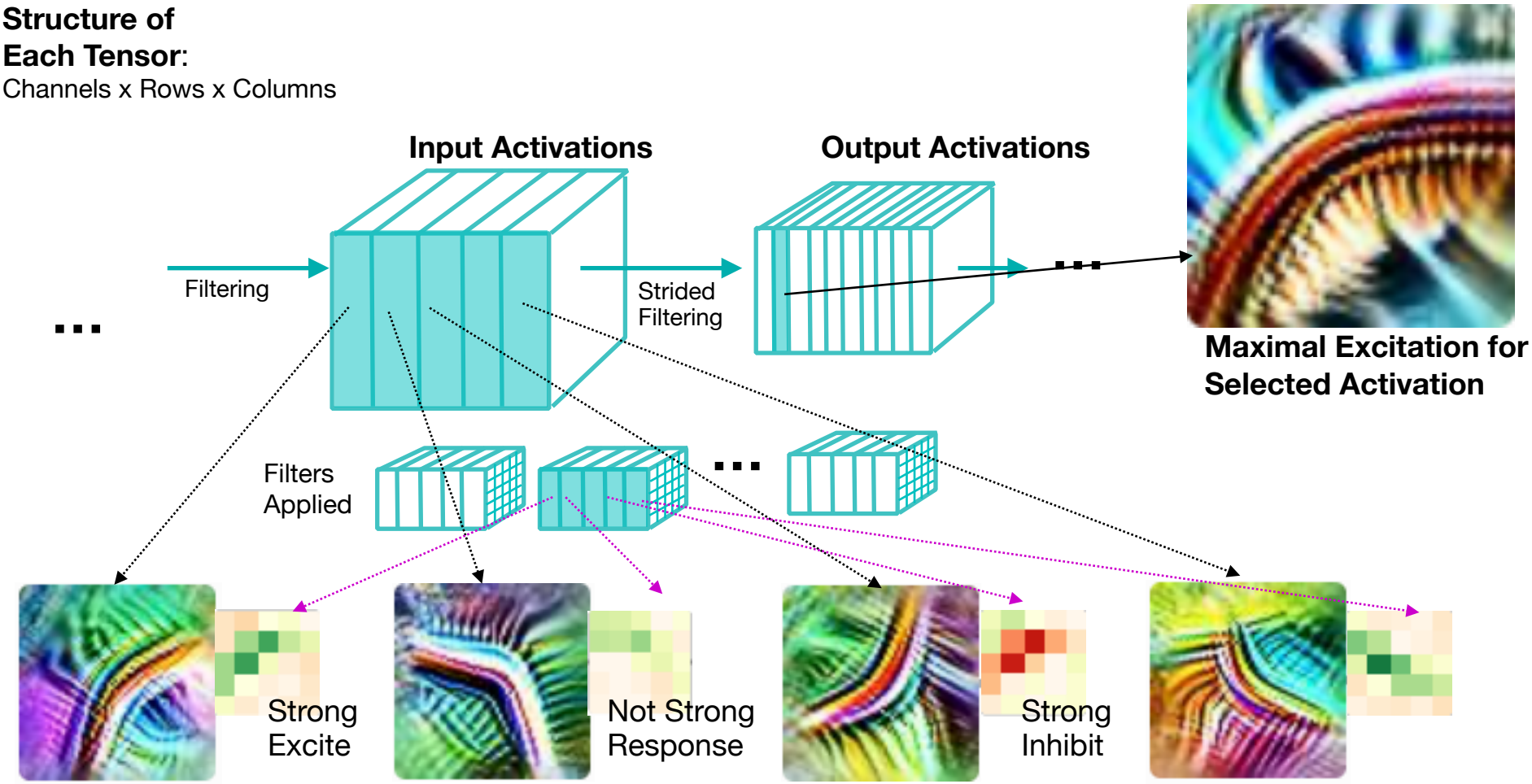



https://microscope.openai.com/models/inceptionv1/



# What weights comprise a circuit?

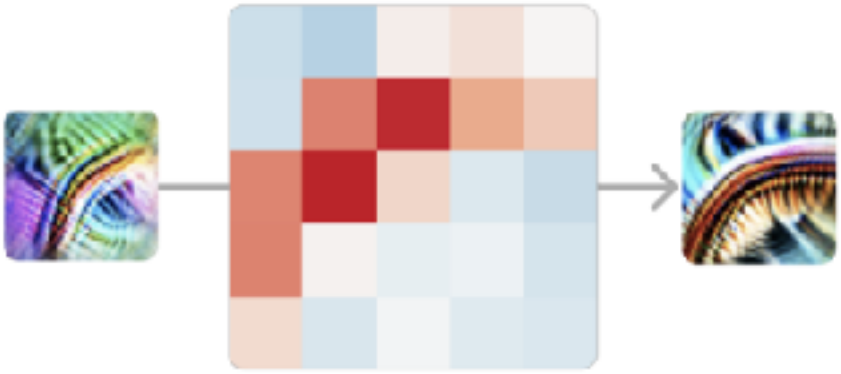
**Structure of Each Tensor:**
Channels x Rows x Columns




https://distill.pub/2020/circuits/curve-circuits/

# Example: Circuit for Better Curve Detection

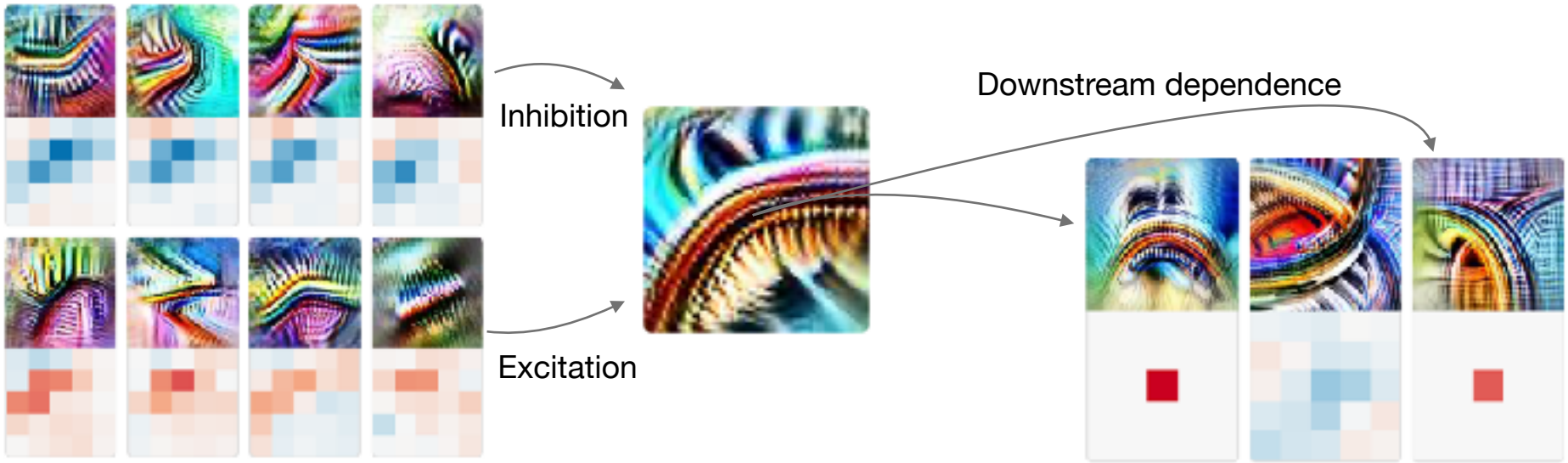
Visualize 5x5 Conv Filter to next Feature



The raw weights between the early curve detector and late curve detector in the same orientation are a curve of **positive weights** surrounded by small **negative** or zero weights.

Superposition of Early Curves

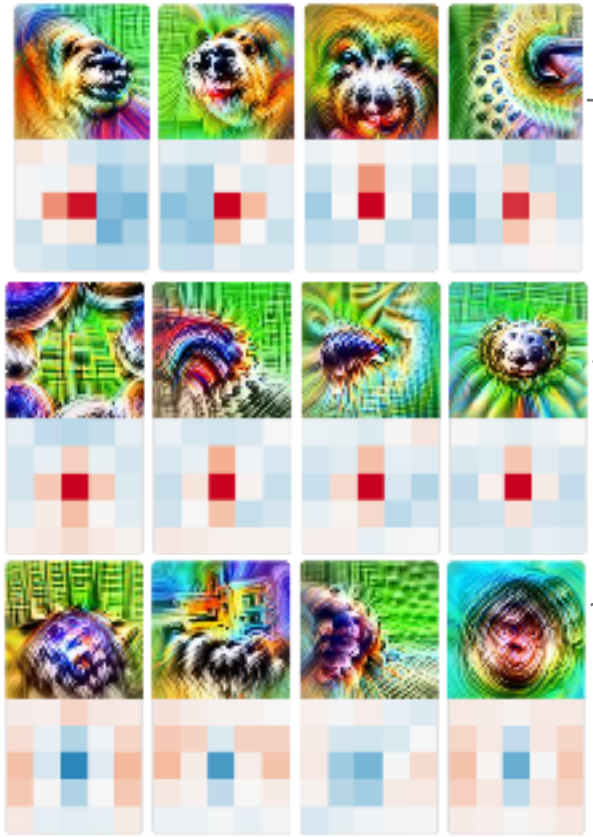
This can be interpreted as looking for "tangent curves" at each point along the curve.

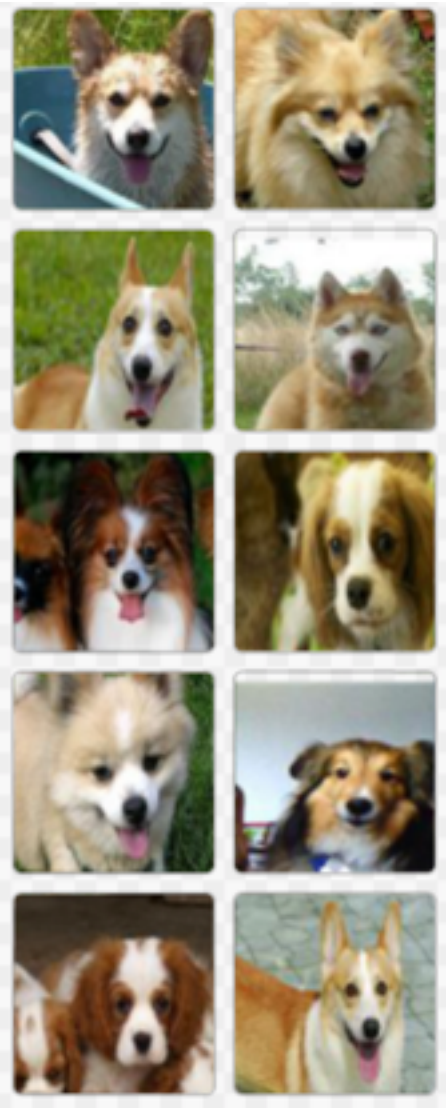





# Another Example: Dog head

Compact Circuit Visualization





This example is also **polysemantic** due to the "**espresso maker**" class also being excited by this…



# Equivariance from Circuits

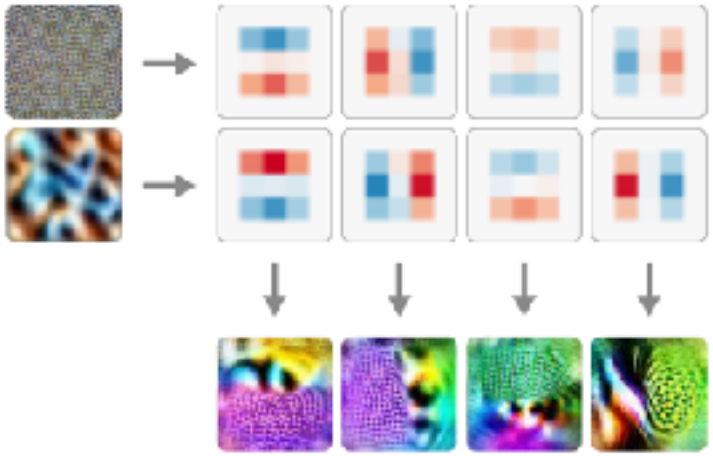
- Many features that are part of a circuit are clearly designed for rotation, hue, and other invariance


https://distill.pub/2020/circuits/equivariance/



# Equivariant circuits: a Motif

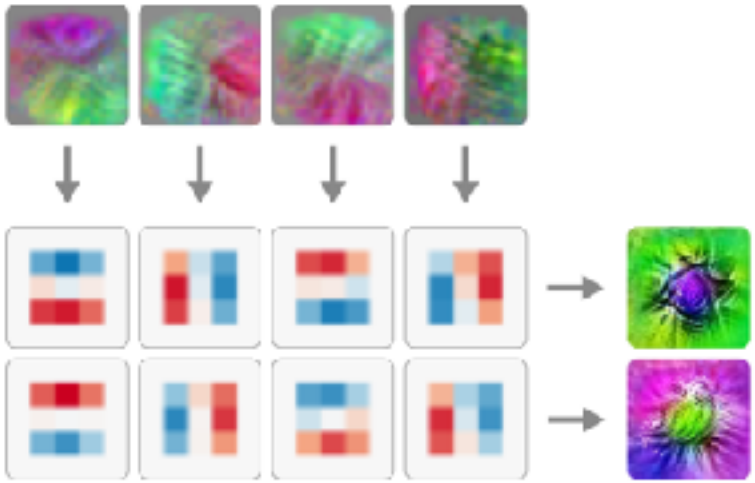
- Possible to reveal patterns of circuits via sets of weights



**High-low frequency detectors**respond 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)



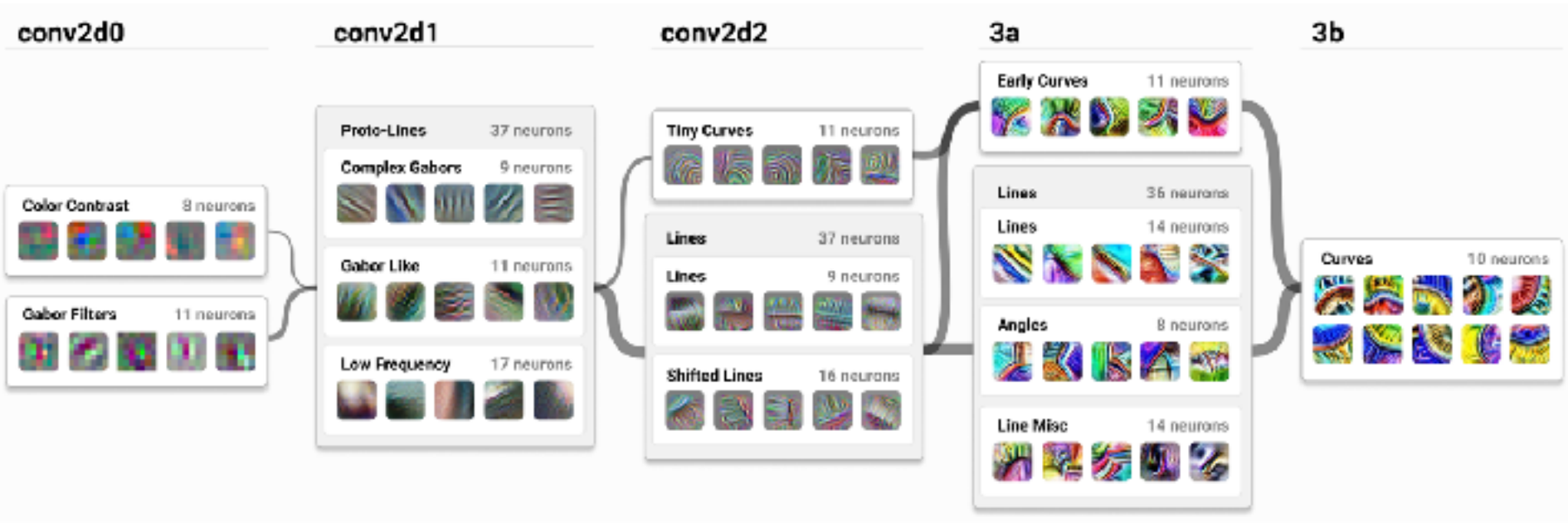
Rotational equivariance can be turned into invariance with the transpose of an invariant -> equivariant circuit.

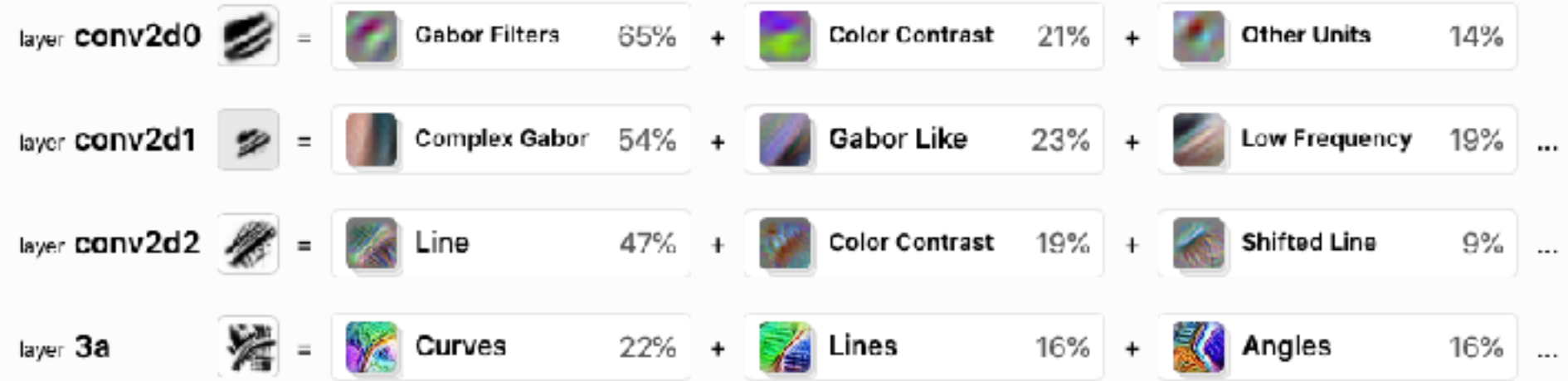
Here, we see**color contrast units**(rotationally equivariant) combine to make **color center surround units**(rotationally invariant). Again, notice how the weights rotate, forming the same pattern we saw above with high-low frequency detectors, but with inputs and outputs swapped.

positive (excitation)
negative (inhibition)



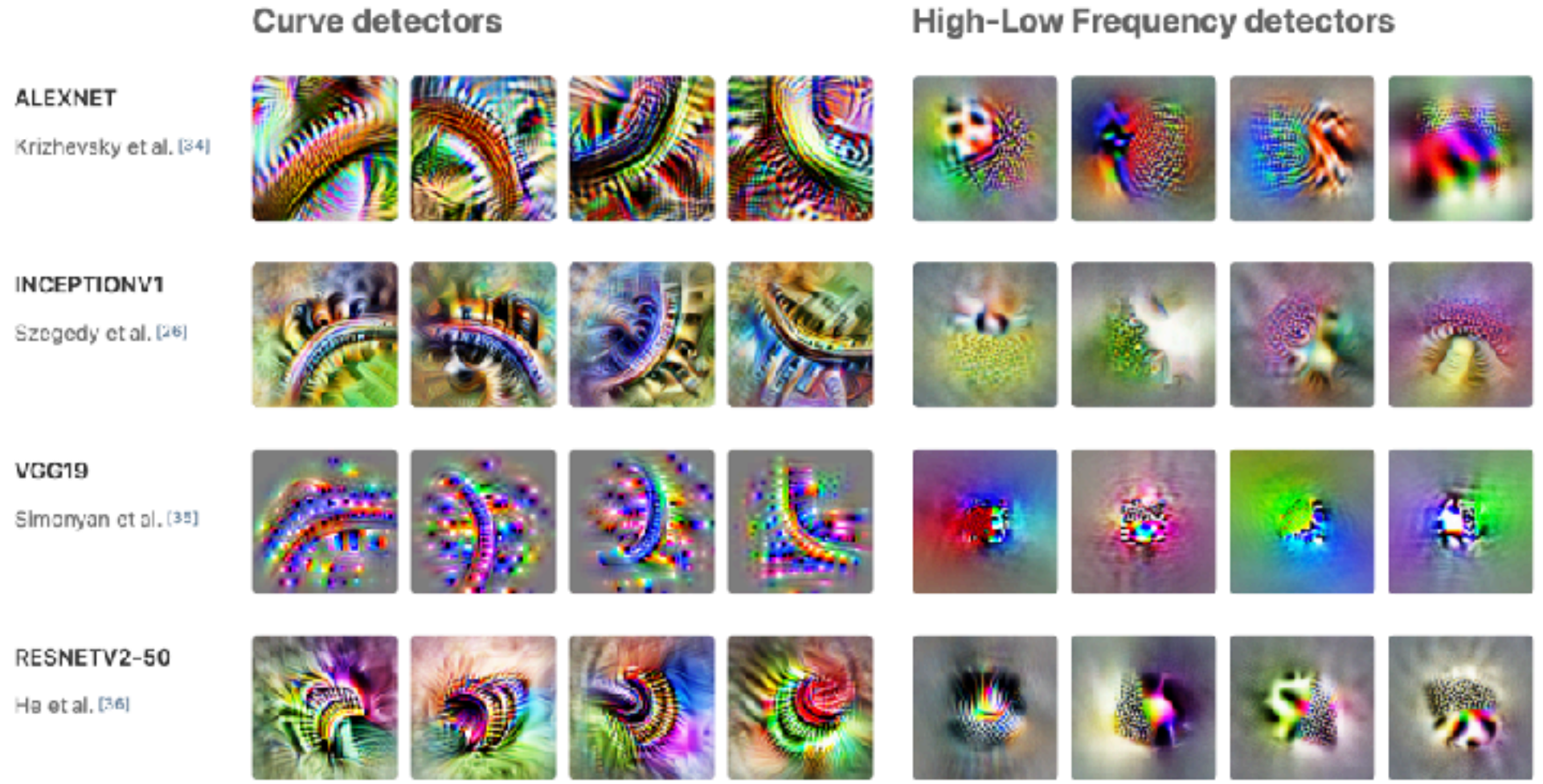
# Neural Nets: Directed Graph of Circuits



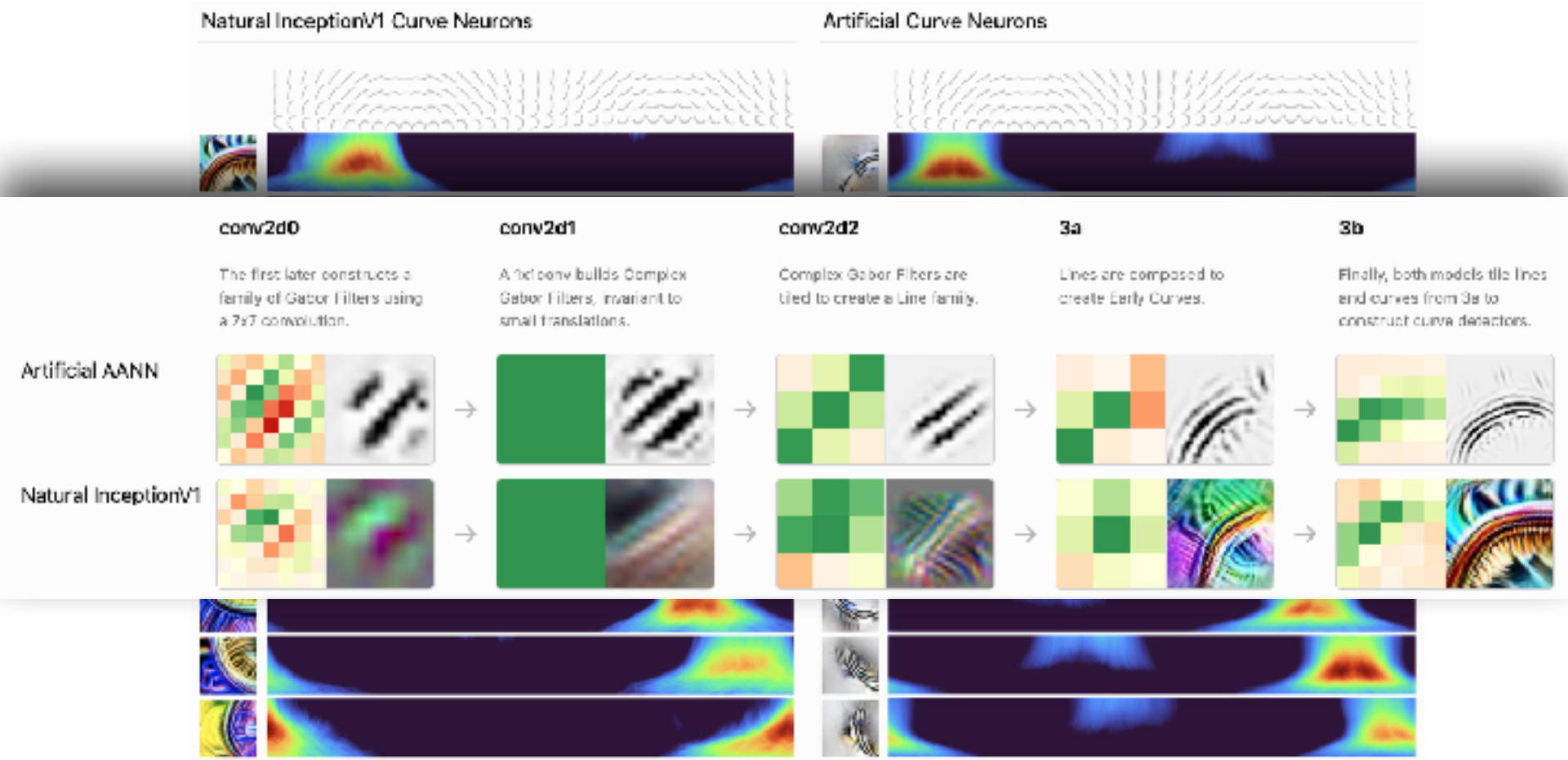
# Universality of Circuits

- 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



# 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/



# Lab One Town Hall


Lecture Notes for

# Neural Networks and Machine Learning

## CNN Circuits

**Next Time:**

Fully Convolutional Learning

**Reading:** Chollet 5.4