

Lecture Notes for **Neural Networks** **and Machine Learning**



CNN Circuits



Logistics and Agenda

- Logistics
 - Lab due very soon!
- Agenda
 - Last Time: Visualizing Convolutional Architectures
 - Student Paper Presentation: Augmentation Effectiveness
 - Today: Circuits in CNNs
 - If Time: Lab Town Hall



Student Paper Presentation

The Effectiveness of Data Augmentation in Image Classification using Deep Learning

Jesse Wang
Stanford University
450 Serra Mall
jwang11@stanford.edu

Luis Perez
Stanford University
450 Serra Mall
lperez11@stanford.edu

Abstract:

In this paper we introduce and compare multiple solutions to the problem of data augmentation in image classification. Previous work has demonstrated the effectiveness of data augmentation through image techniques, such as cropping, scaling, and flipping object images. By empirically comparing our own data augmentation method of the image for dataset, and comparing each data augmentation technique to them, this is the most accurate multi-dimensional technique in the traditional two-dimensional methods where the most experienced with GANs to generate images of different classes. Finally, we compare a few deep learning models to learn a representation that best improves the classifier, which are called neural augmentation. We discuss the inherent and shortcomings of our method in various datasets.

Data augmentation guided by super-labels helps [4], convolutional image augmentation [3], and has shown effectiveness in image classification [2].

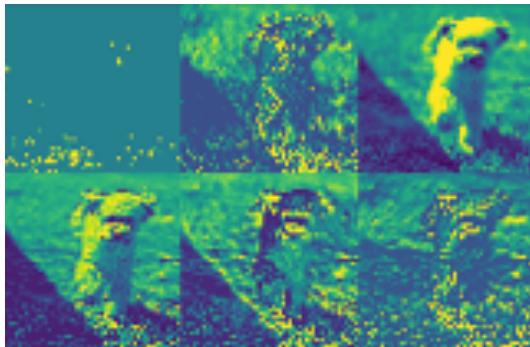
The motivation for this problem is look local and specific, specialized image and video classification tasks often have limited data. This is particularly true in the medical industry, where access to data is usually restricted due to privacy concerns. Importantly, it is much easier to classify convolutions [2] on datasets by this type of data. Therefore, we have developed which combines pre-trained networks with perturbation models. Convolutional neural networks in the AI industry often lack access to significant amounts of data. At the end of the day, we can only work with large training sets for most projects to access to reliable data, and as such, we explore the effectiveness of different data augmentation techniques in image classification tasks.

The datasets we examine are the famous digit MNIST data and CIFAR-10 [1]. They contain 60k training, 10k validation, and 10k test images of dimension 28x28x3. There is a total of 600 images per class with 100 distinct classes. MNIST consists of 60k handwritten digits in the training set and 10k in the test set, gray-scale with 10 classes with image dimensions of 28x28x1. To evaluate the effectiveness of augmentation techniques, we reduce our data to two classes and build convolutional neural net classifiers to correctly guess the class.

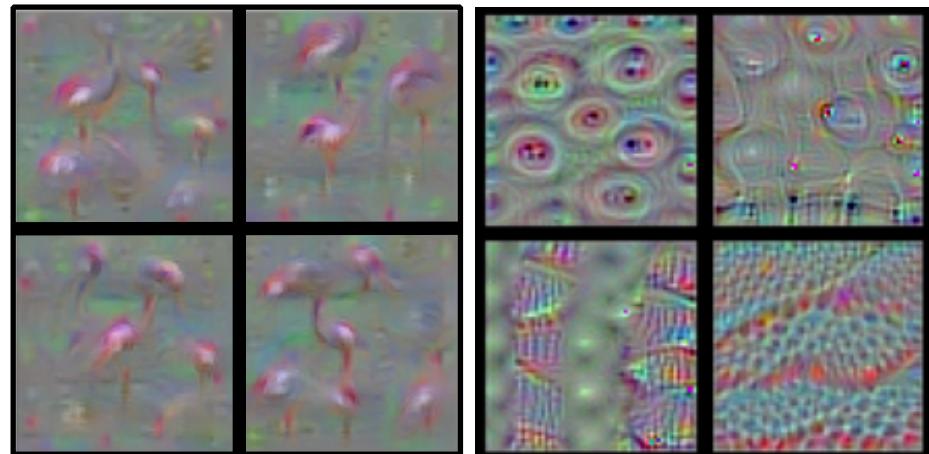
In particular, we will focus on even small test to perform a multi-classify classification. We will then present in our typical data augmentation techniques, and relate our methods. Then, we will make use of CycleGAN [4] to augment our data by translating styles from images in the dataset to a few predetermined images such as SegNetCityscapes or WinterSeason. Finally, we explore and propose a different kind of augmentation where we randomly rotate our data around its axis and slightly increase its standard augmentation noise, the convolutional image augmentation that been re-



Review: our visualization toolset



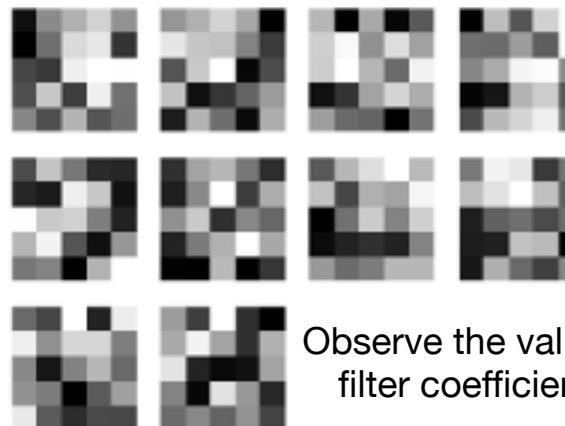
Visualize Activation
in response to input image



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

SCHWANN'S CLAIMS ABOUT CELLS

Claim 1

The cell is the unit of structure, physiology, and organization in living things.

Claim 2

The cell retains a dual existence as a distinct entity and a building block in the construction of organisms.

Claim 3

Cells form by free-cell formation, similar to the formation of crystals.

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/>



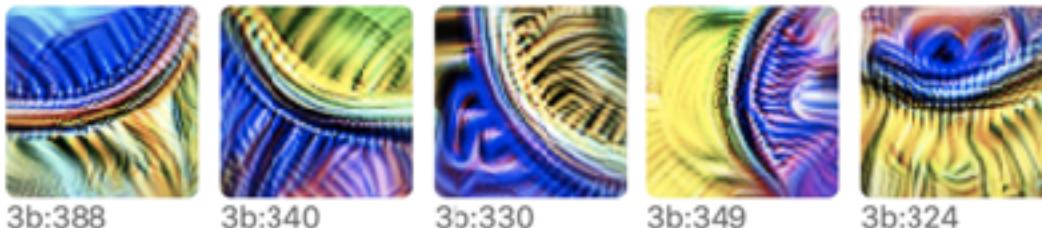
Building Blocks: Features

- *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



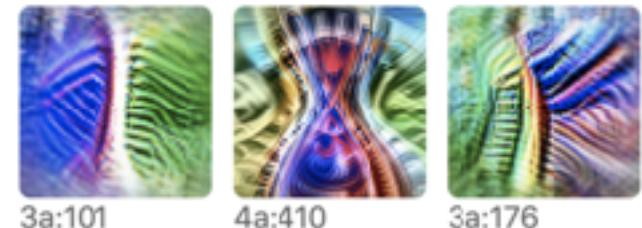
Examples of Discovered Features

Curves



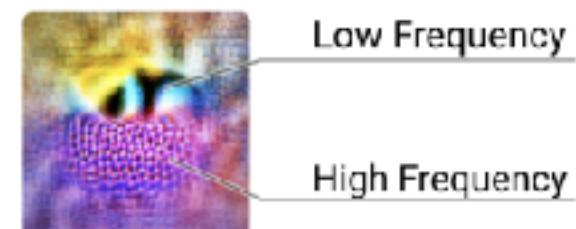
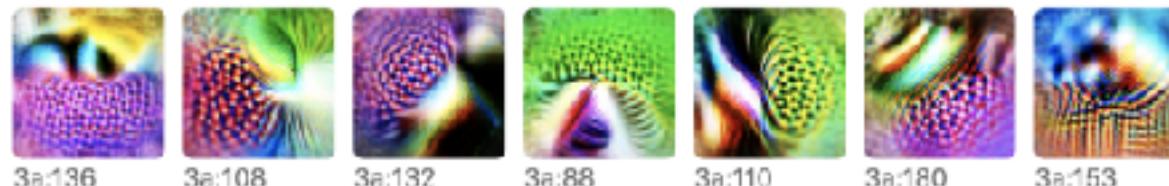
Hypothesized feature group (part of circuit)

Related Shapes (Circle, Spiral...)



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

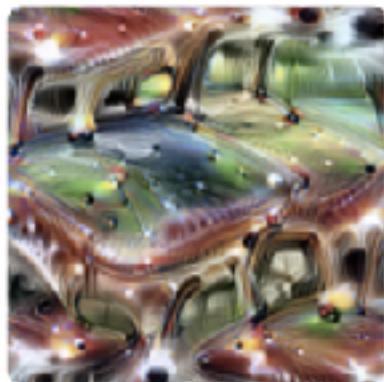


Neuron 4b:409



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].

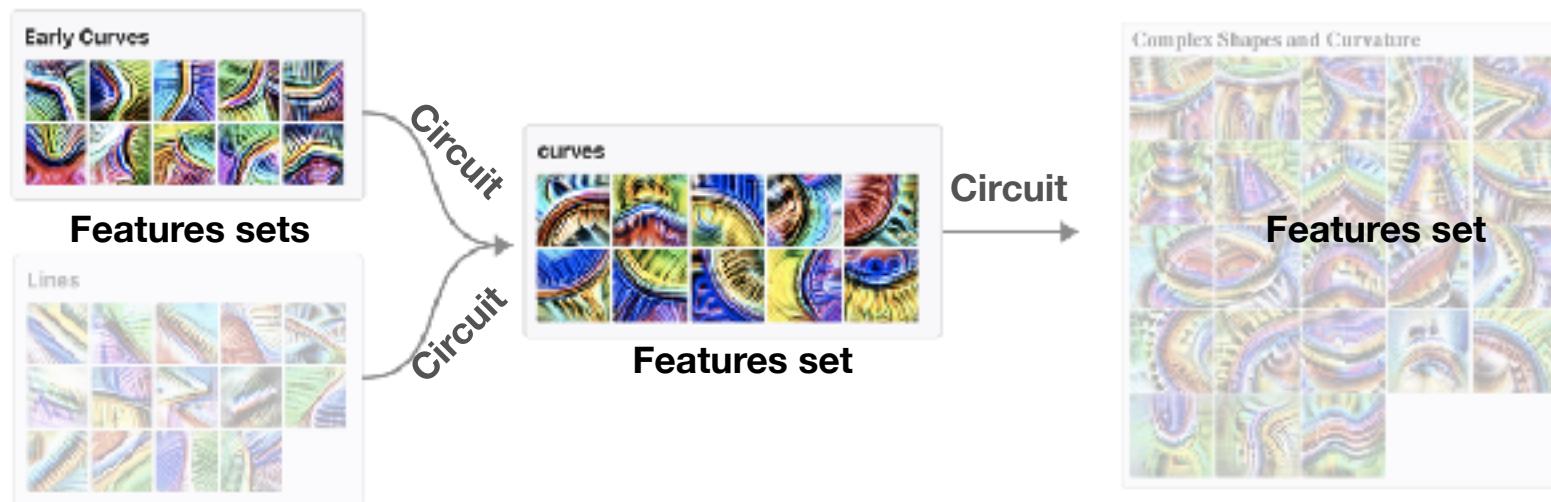
The existence of these neurons is likely one of the main criticism of network features.

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/>

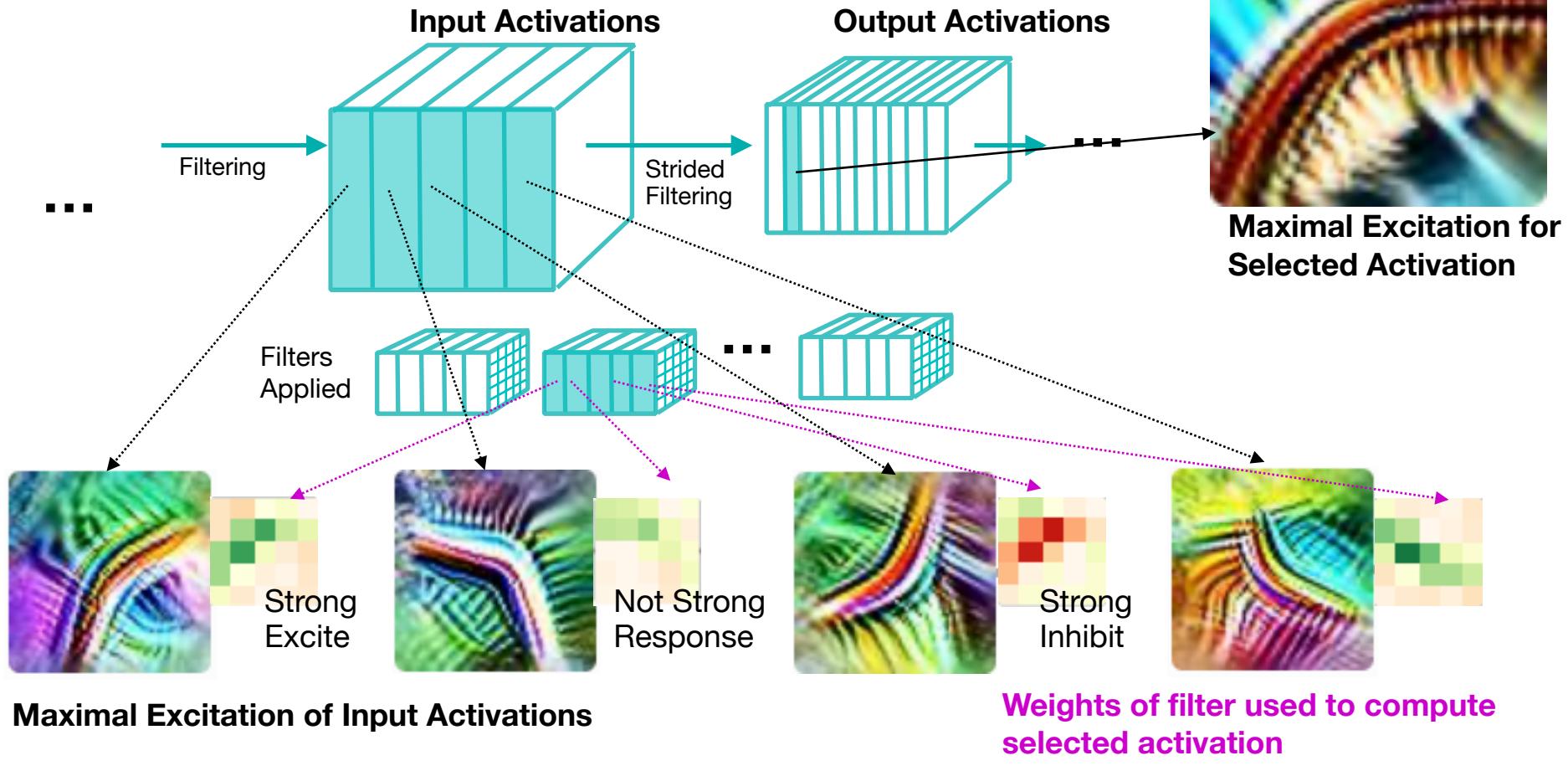
29



What weights comprise a circuit?

Structure of Each Tensor:

Channels x Rows x Columns

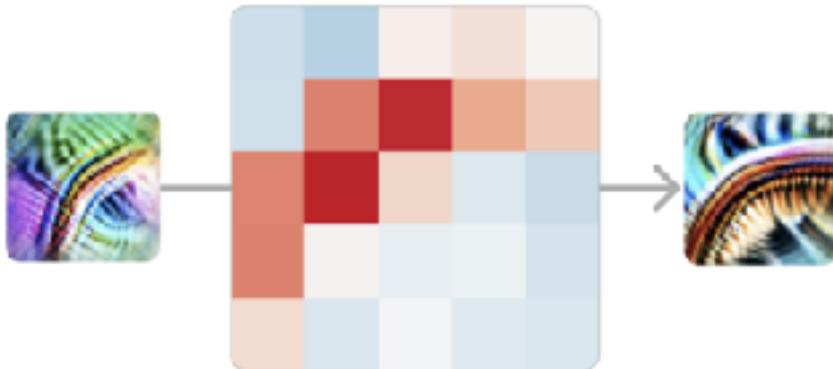


<https://distill.pub/2020/circuits/curve-circuits/>



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.

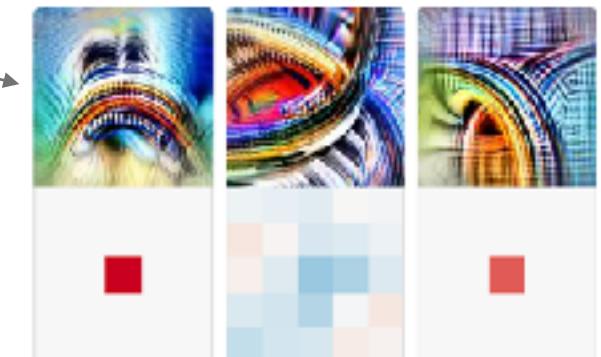
Superposition of Early Curves



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

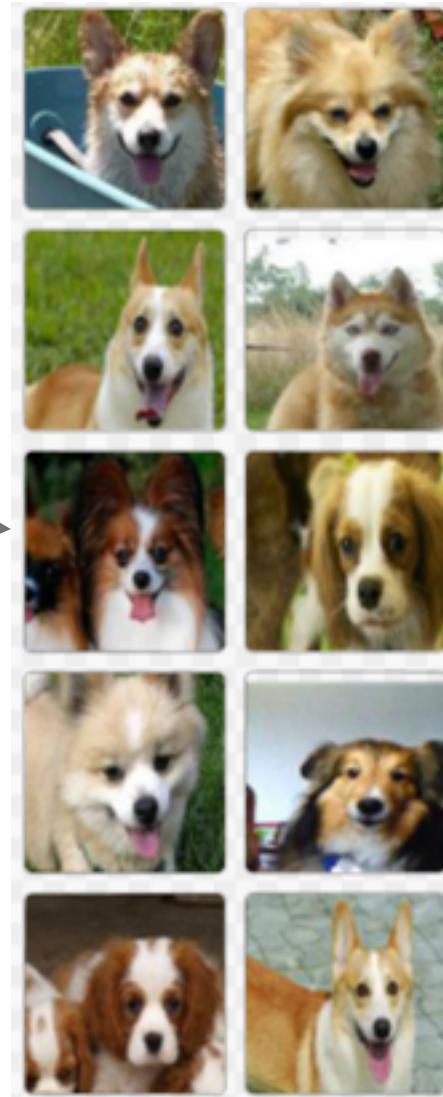
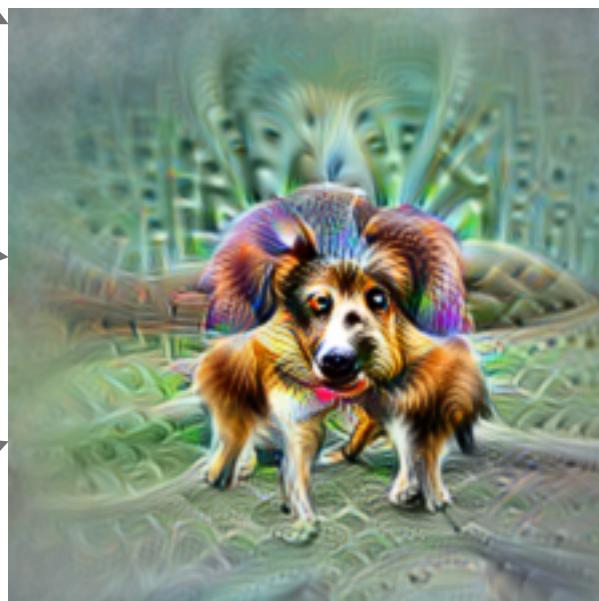
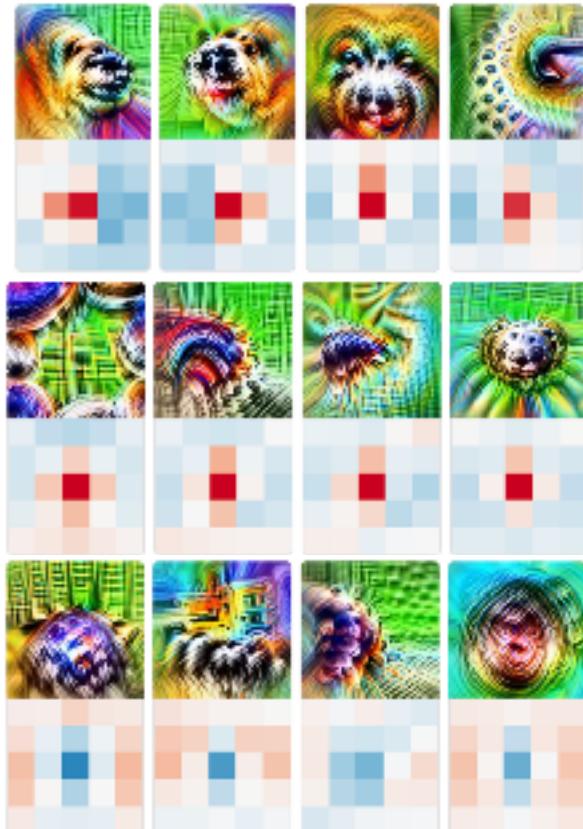


Downstream dependence



Another Example: Dog head

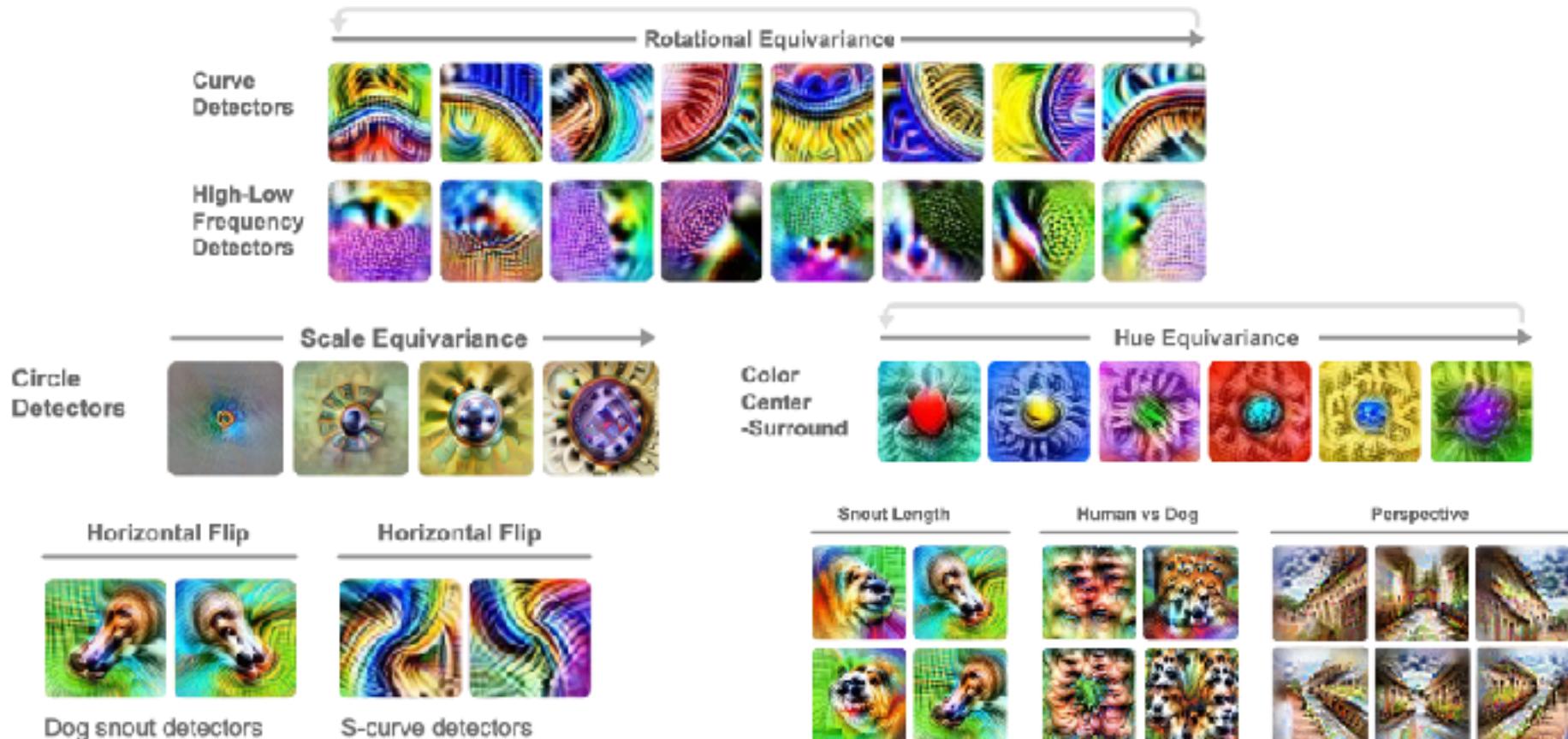
Compact Circuit Visualization



This example is also **polysemantic** due to the “**espresso maker**” class also being excited by this...

Equivariant Circuits

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



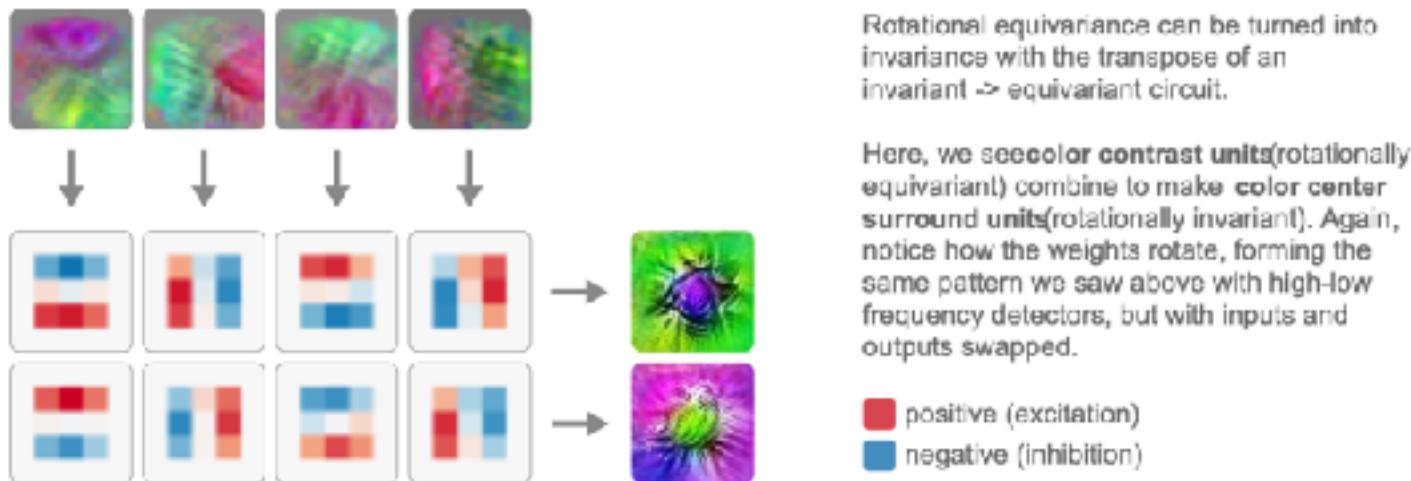
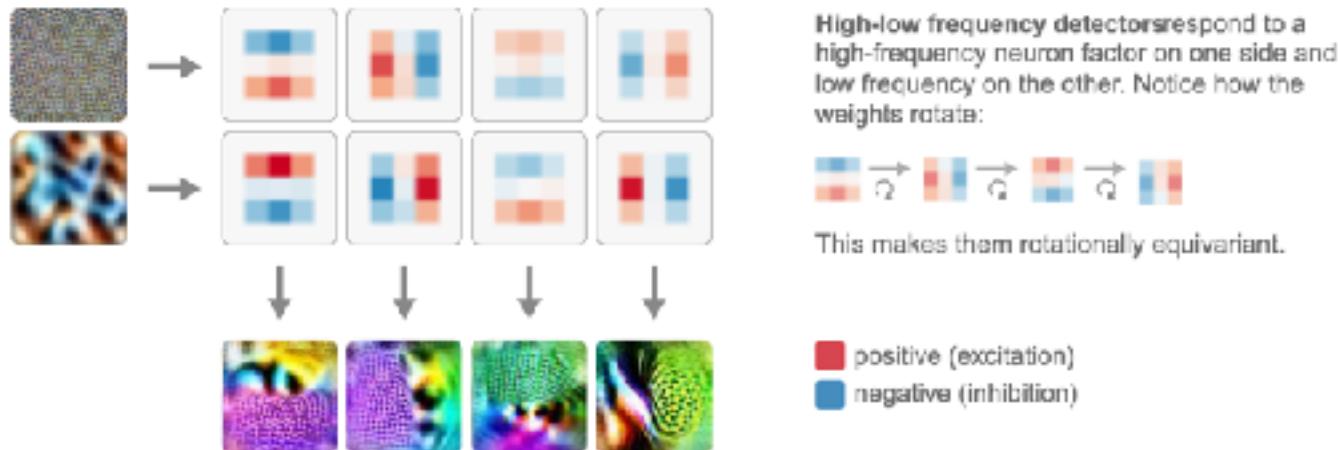
<https://distill.pub/2020/circuits/equivariance/>

33

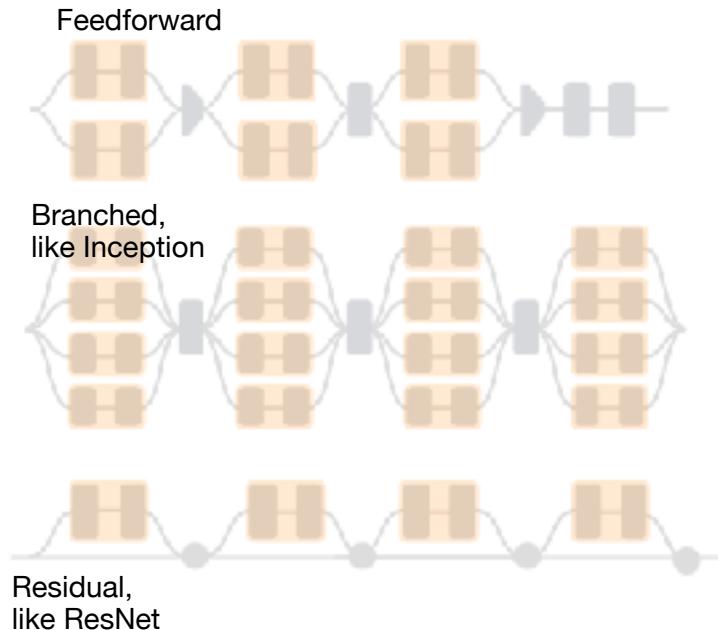


Equivariant circuits: a Motif

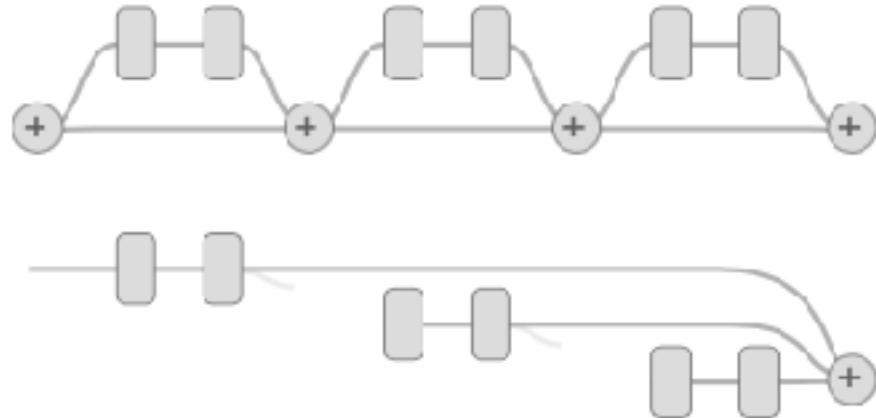
- Possible to reveal patterns of circuits via sets of weights



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



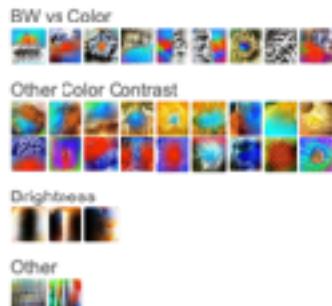
<https://distill.pub/2020/circuits/branch-specialization/>



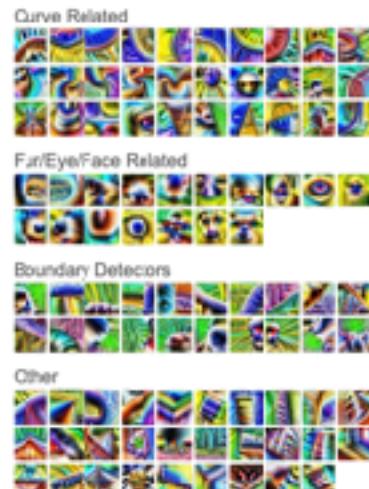
Branch Specialization



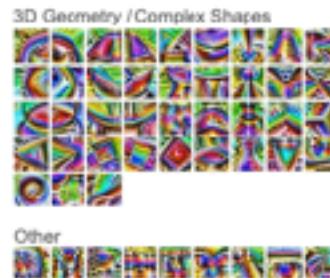
mixed3a_5x5: The 5x5 branch of mixed3a, a relatively early layer, is specialized on color detection, and especially black-and-white vs. color detection.



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

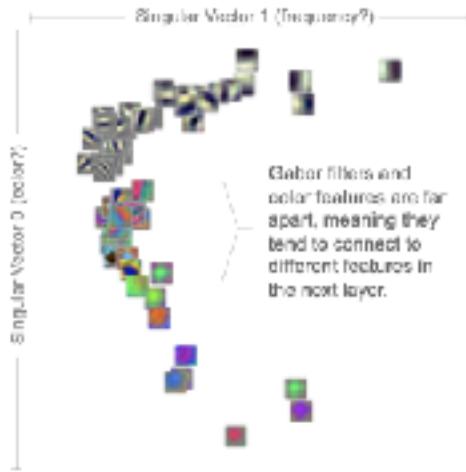


Investigating Major Variation via SVD

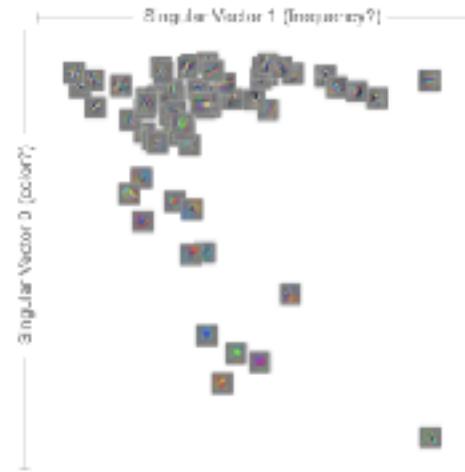
InceptionV1 (tf-slim version) trained on ImageNet.

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 $[W]$.



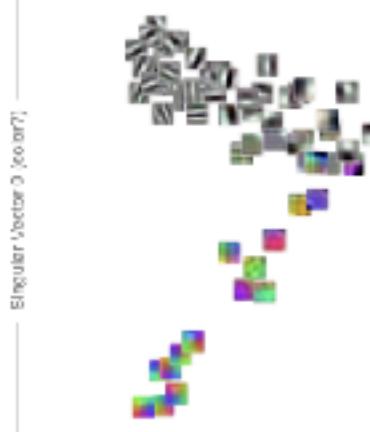
Neurons in the second convolutional layer organized by the right singular vectors of $[W]$.



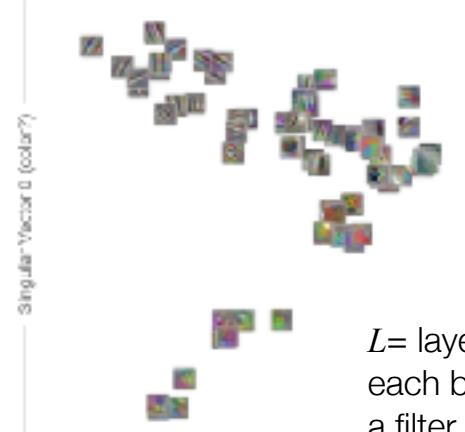
InceptionV1 trained on Places365

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 1 (frequency?)



- Singular Value Decomposition (SVD) decomposes a matrix into three elements

- $W = U\Sigma V^T$
- U is eig-vec of WW^T
 V is eig-vec of W^TW
- U and V are orthogonal such that
 $UUT=I \quad VVT=I$
- Σ is a diagonal matrix of the singular values
- These values characterize the variability in a matrix

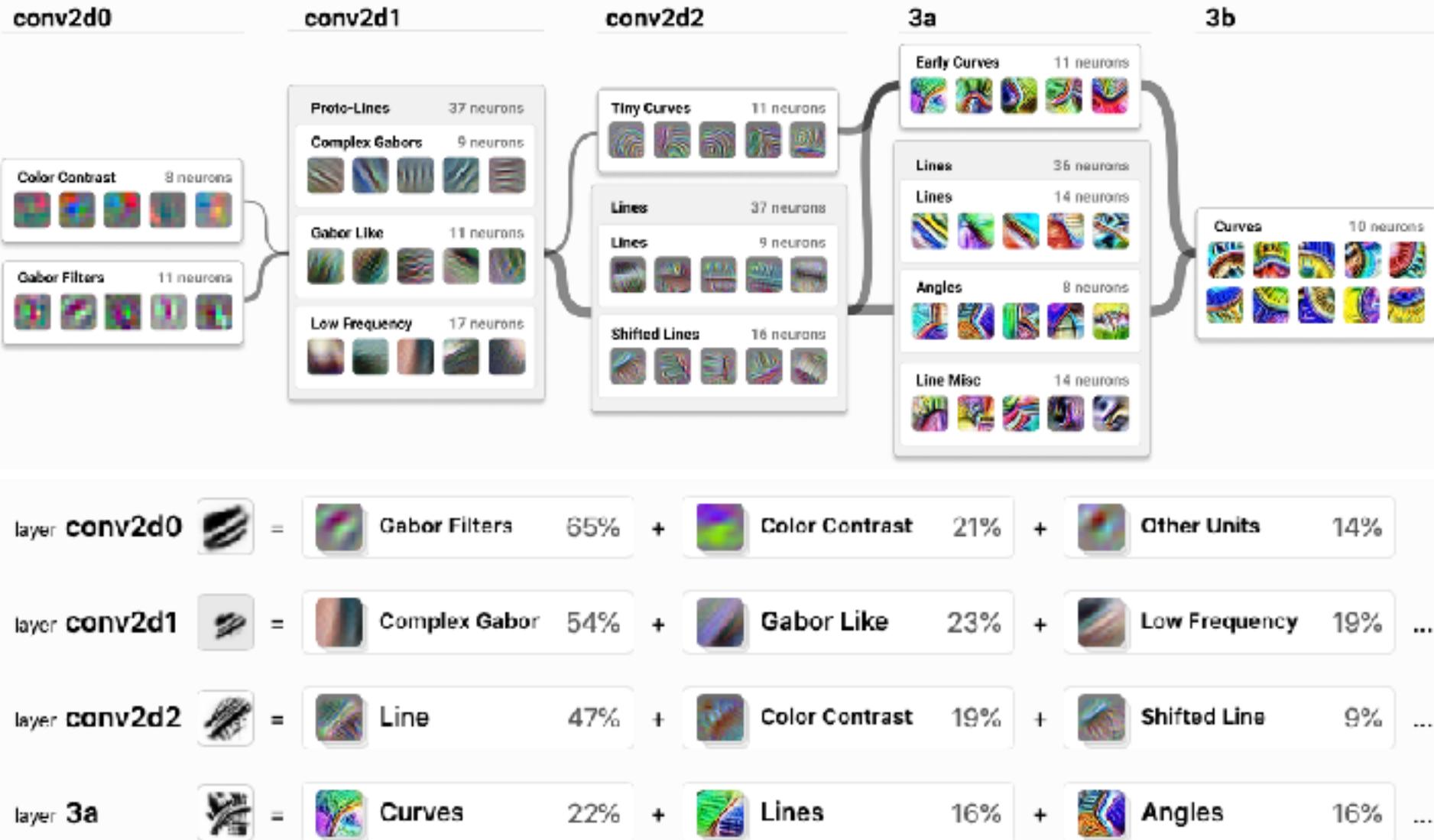
$$SVD(|\mathbf{W}_f^{(L)}|)$$

$$SVD(|\mathbf{W}_f^{(L+1)}|)$$

$L =$ layer
each block is
a filter, f , in layer



Neural Nets: Directed Graph of Circuits



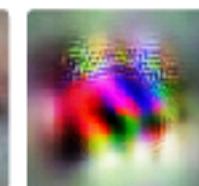
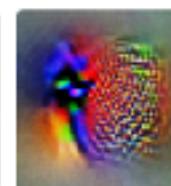
Universality of Circuits

- Analogous features and circuits form across models and tasks

Curve detectors

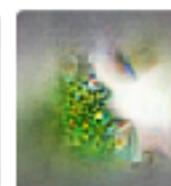


High-Low Frequency detectors



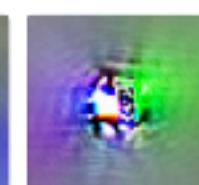
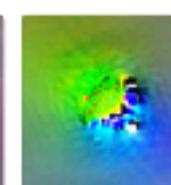
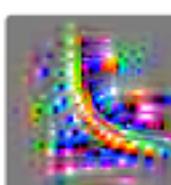
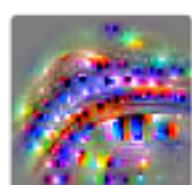
ALEXNET

Krizhevsky et al. [34]



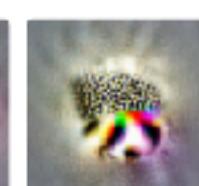
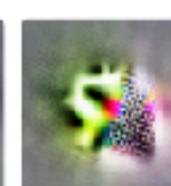
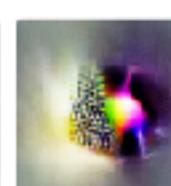
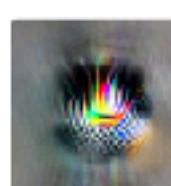
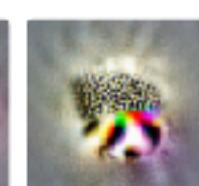
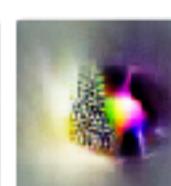
INCEPTIONV1

Szegedy et al. [26]



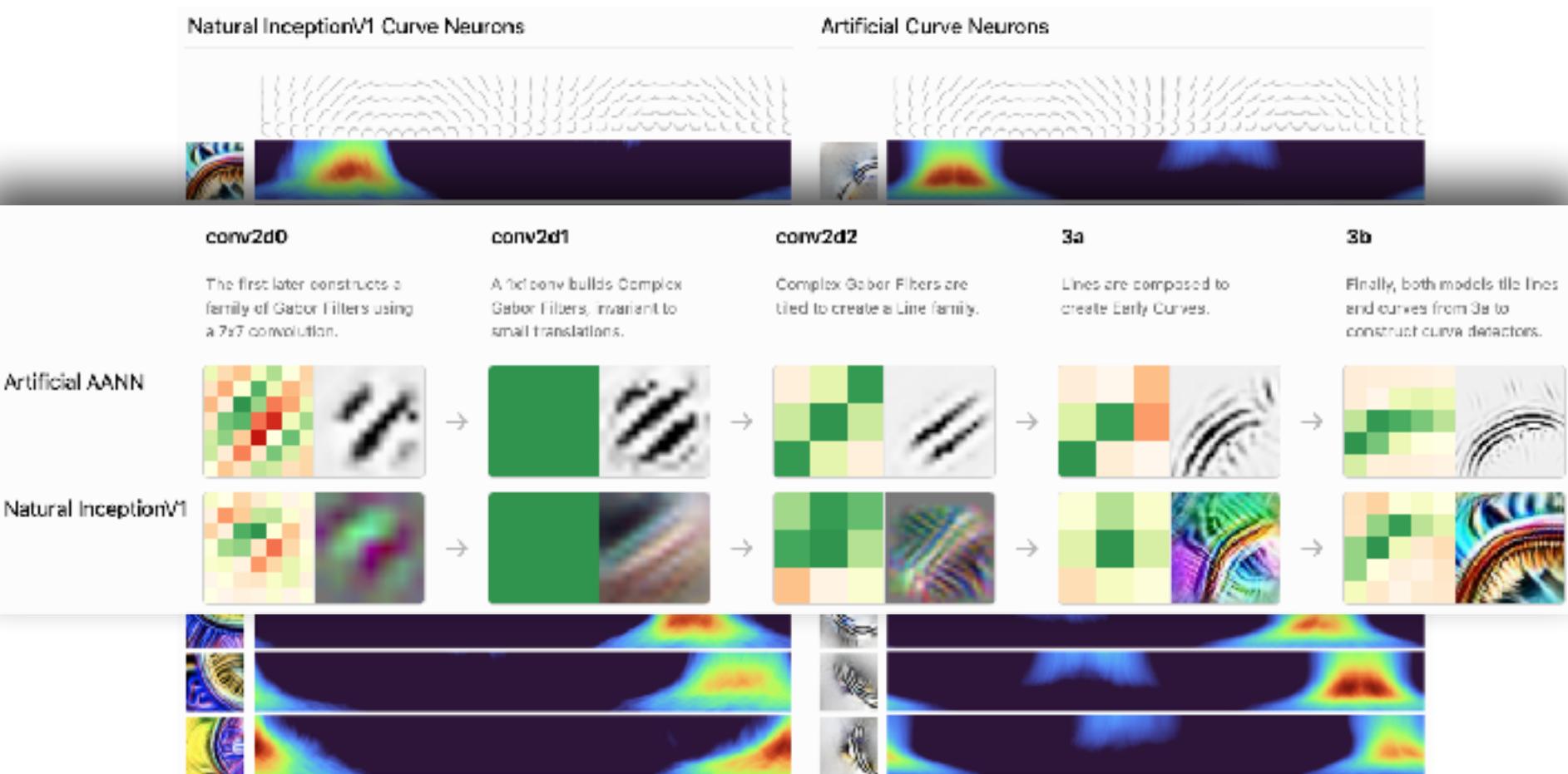
VGG19

Simonyan et al. [35]



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 Two Town Hall



Tamás Görbe @TamasGorbe · 8h

student: how do i become a grad.student?

me: here *hands them a nabla ∇ *

∇ student

@TamasGorbe



Figure for Circuits Lab



Structure of Each Tensor:

Channels x Rows x Columns

Input Activations

Filtering

Rows

5. Look at the input activations for each channel of the multi-channel filter, see which are the most influential.

Multi-channel Filters

1. Choose a network and middle layer to analyze. This example shows activations from a VGG layer with 512 channels.

2. Choose Output Activation of Interest

Output Activations

A diagram showing a stack of vertical bars of varying heights, representing a column of data. The bars are colored in a gradient from light blue at the bottom to dark blue at the top. A pink arrow points to the top bar, which has diagonal hatching. To the left of the stack is a green arrow pointing right, and to the right is a grey arrow pointing left.

3. Find multi-channel filter that is responsible for selected activation

4. Look at each channel of the multi-channel filter. Each channel can be thought of as a filter applied to the activations in the previous layer.

Lecture Notes for Neural Networks and Machine Learning

CNN Circuits



Next Time:
Fully Convolutional Learning
Reading: Chollet 5.4

