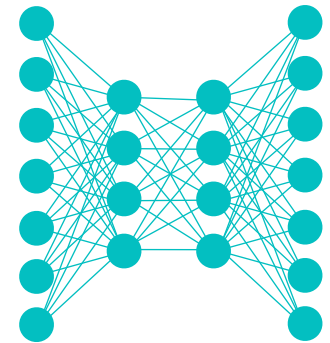


Lecture Notes for **Neural Networks and Machine Learning**



CNN Circuits
Continued

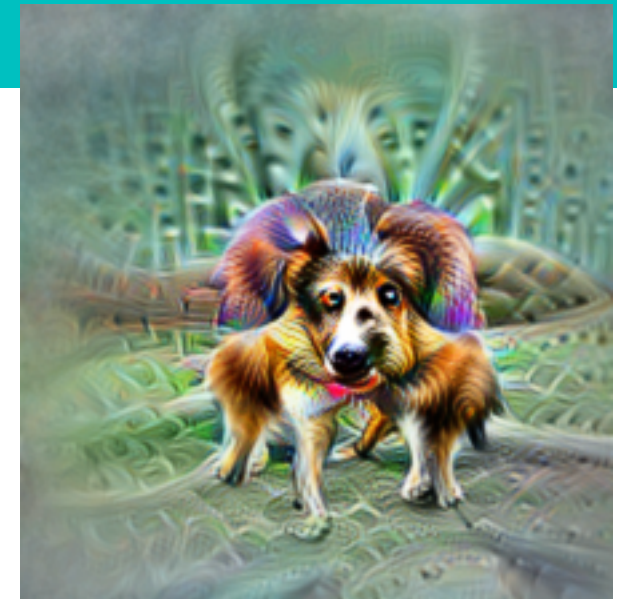
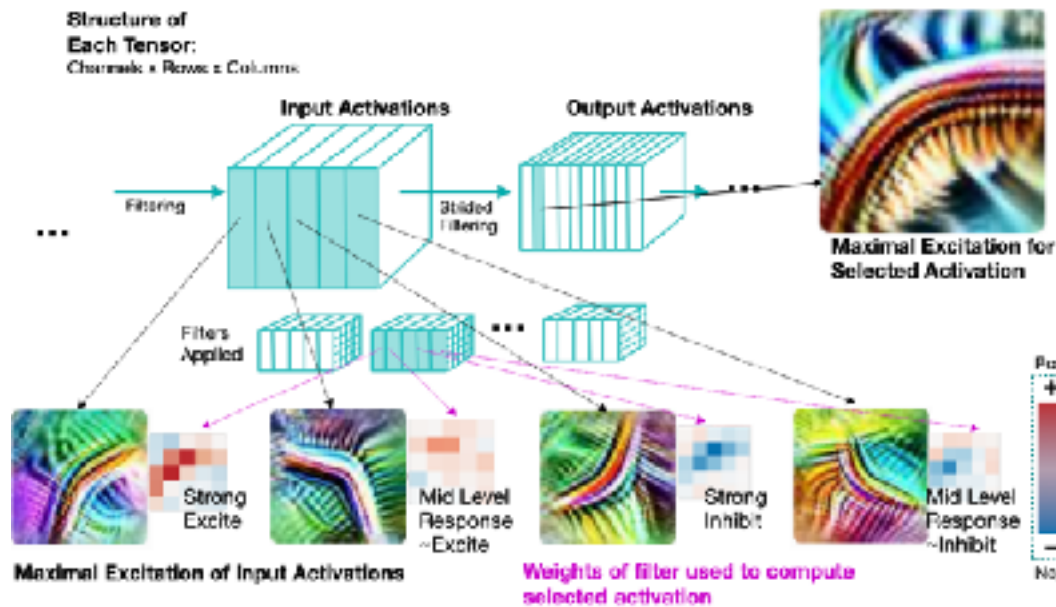


Logistics and Agenda

- Logistics
 - Grading Update
 - Lab logistics!
- Agenda
 - Last Time: Circuits in CNNs
 - Continued Circuits
 - Student Paper Presentation
 - Lab Three Town Hall
 - Next Time: Fully Convolutional Networks



Last Time



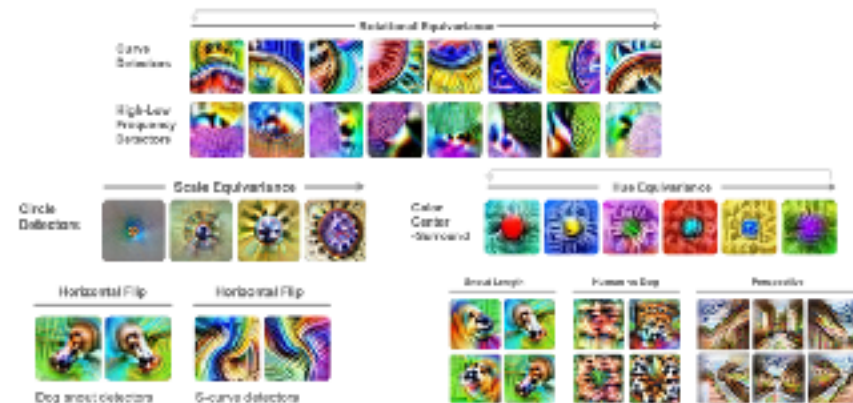
Many features that are part of a circuit are clearly designed for rotation, hue, and other invariance



Neuron 4b:409

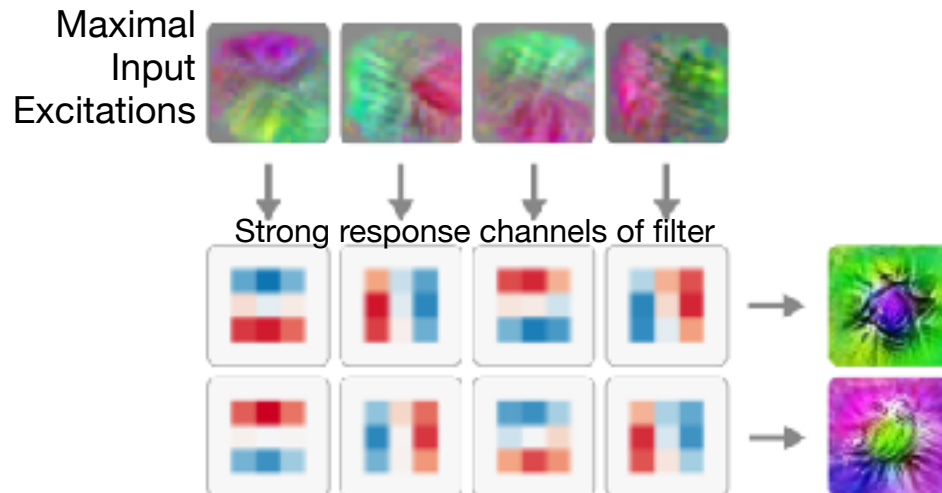
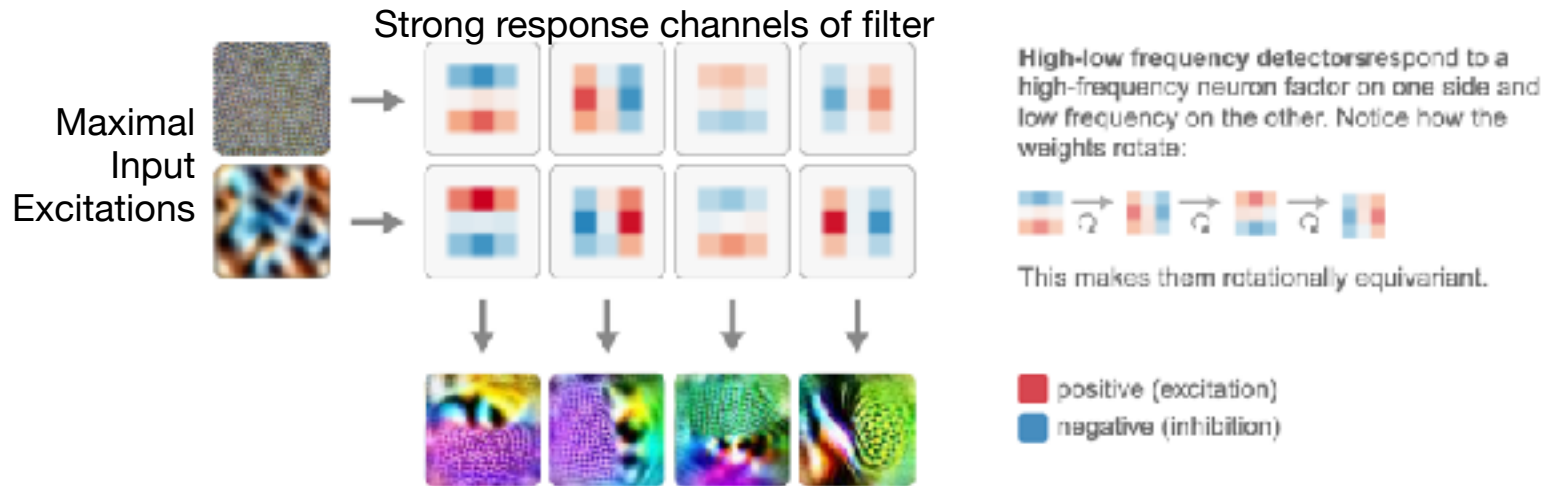


Dataset examples for neuron 4b:409



Review of Equivariant Circuits

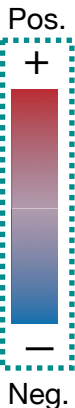
- Possible to reveal patterns of circuits via sets of weights



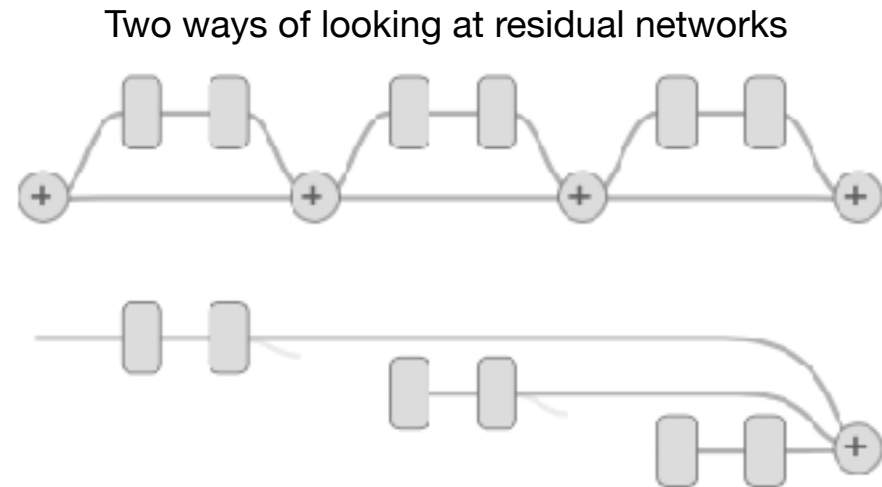
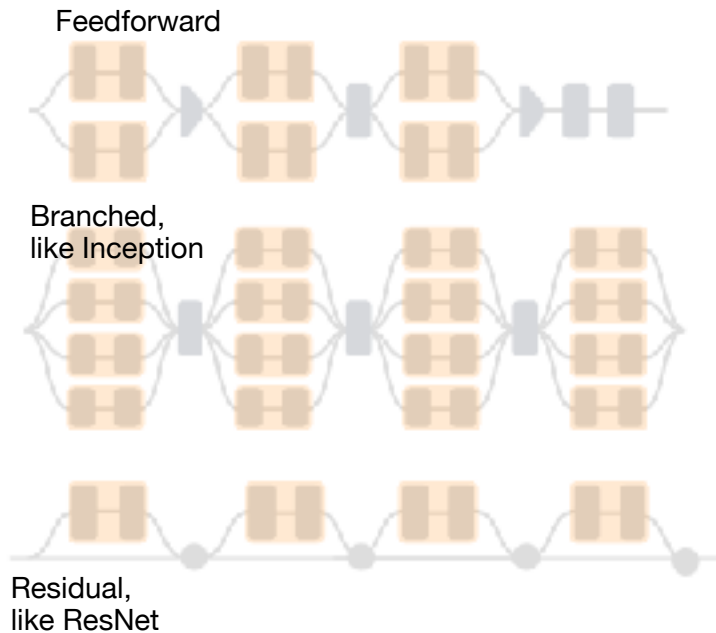
Rotational equivariance can be turned into invariance with the transpose of an invariant \rightarrow equivariant circuit.

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)



Branch Specialization



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



Branch Specialization

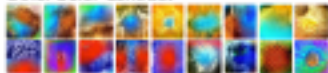


mixed3a_5x1: The 5x1 branch of mixed3a, a relatively early layer, is specialized on color detection, and especially black-and-white vs. color detection.

B/W vs Color



Other Color Contrast



Brightness



Other

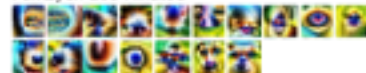


mixed3b_5x5: This branch contains all 30 of the curve-related features for this layer (all curves, double curves, circles, spirals, S-shape and more features, etc). It also contains a disproportionate number of boundary, eye, and fur detectors, many of which share sub-components with curves.

Curve Related



Fur/Eye/Face Related



Boundary Detectors

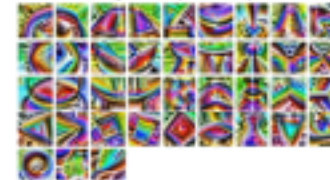


Other



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.

3D Geometry / Complex Shapes



Other

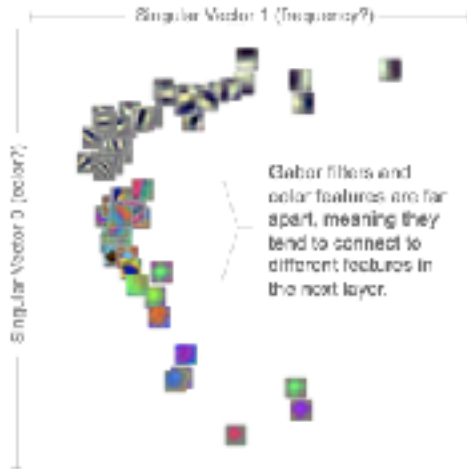


Motifs appear in branches. Similar clusters of operations can be found across different architectures



Investigating Connection Clusters via SVD

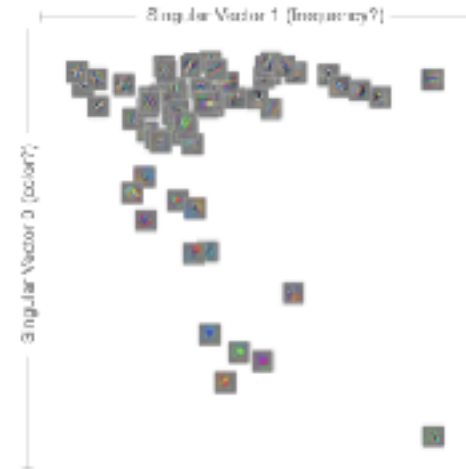
Neurons in the first convolutional layer organized by the left singular vectors of $[W]$.



InceptionV1 (tf-slim version) trained on ImageNet.

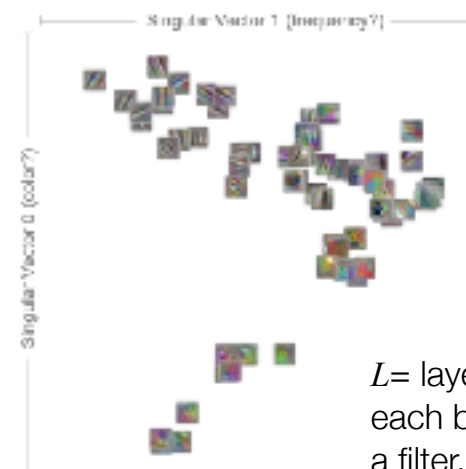
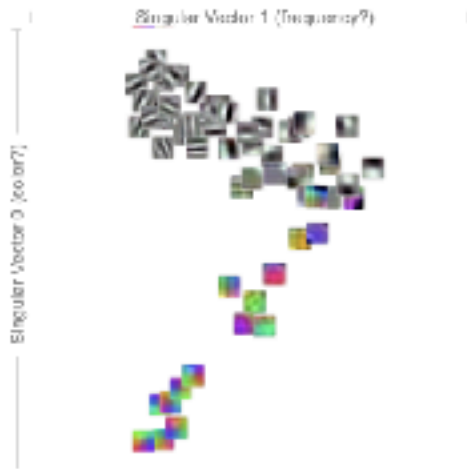
The first singular vector separates color and black and white, meaning that's the largest dimension of variation in which neurons connect to which in the next layer.

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 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 $UU^T = I$ $VV^T = I$
- 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

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

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



Neural Nets: Directed Graph of Circuits



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



Universality of Circuits

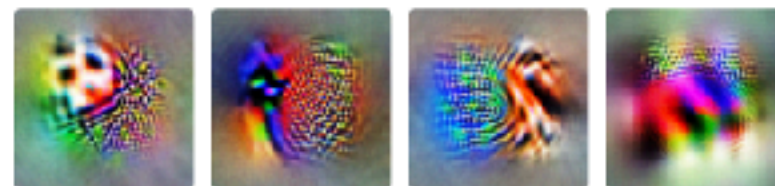
- Analogous features and circuits form across models and tasks

Curve detectors

High-Low Frequency detectors

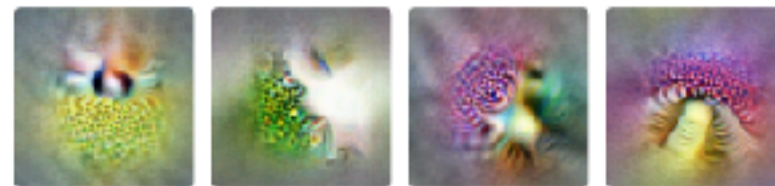
ALEXNET

Krizhevsky et al. [34]



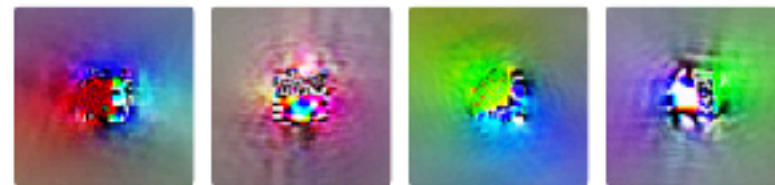
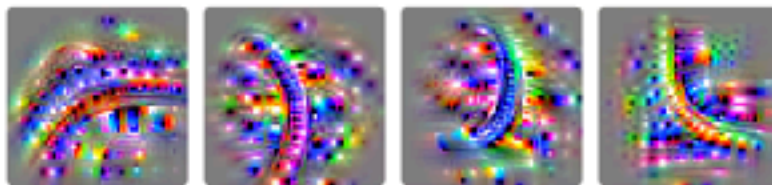
INCEPTIONV1

Szegedy et al. [26]



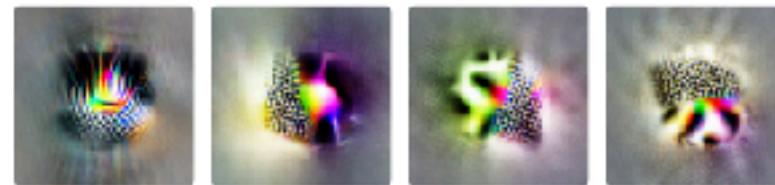
VGG19

Simonyan et al. [35]



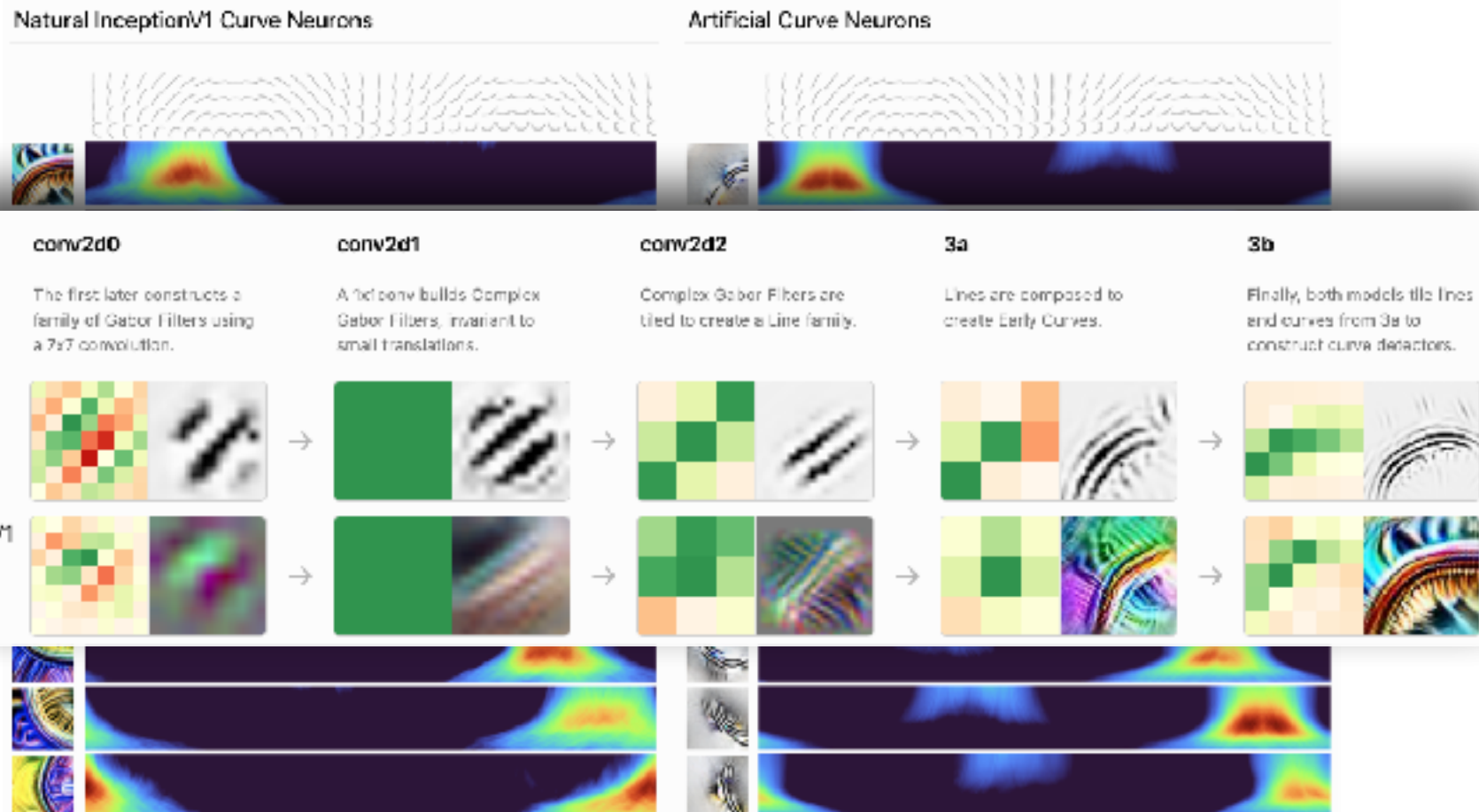
RESNETV2-50

He et al. [36]



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/>



Student Paper Presentation

Group Normalization

Yuxin Wu

Kaiming He

Facebook AI Research (FAIR)

Abstract Batch Normalization (BN) is a milestone technique in the development of deep learning, enabling various networks to train. However, normalizing along the batch dimension introduces problems — BN's error increases rapidly when the batch size becomes smaller, caused by inaccurate batch statistics estimation. This limits BN's usage for training larger models and transferring features to computer vision tasks including detection, segmentation, and video, which require small batches constrained by memory consumption. In this paper, we present Group Normalization (GN) as a simple alternative to BN. GN divides the channels into groups and computes within each group the mean and variance for normalization. GN's computation is independent of batch sizes, and its accuracy is stable in a wide range of batch sizes. On ResNet-50 trained in ImageNet, GN has 10.5% lower error than its BN counterpart when using a batch size of 2; when using typical batch sizes, GN is comparably good with BN and outperforms other normalization variants. Moreover, GN can be naturally transferred from pre-training to fine-tuning. GN can outperform its BN-based counterparts for object detection and segmentation in COCO, and for video classification in Kinetics, showing that GN can effectively replace the powerful BN in a variety of tasks. GN can be easily implemented by a few lines of code.



Lab Three Town Hall



Tamás Görbe @TamasGorbe · 8h
student: how do i become a grad.student?

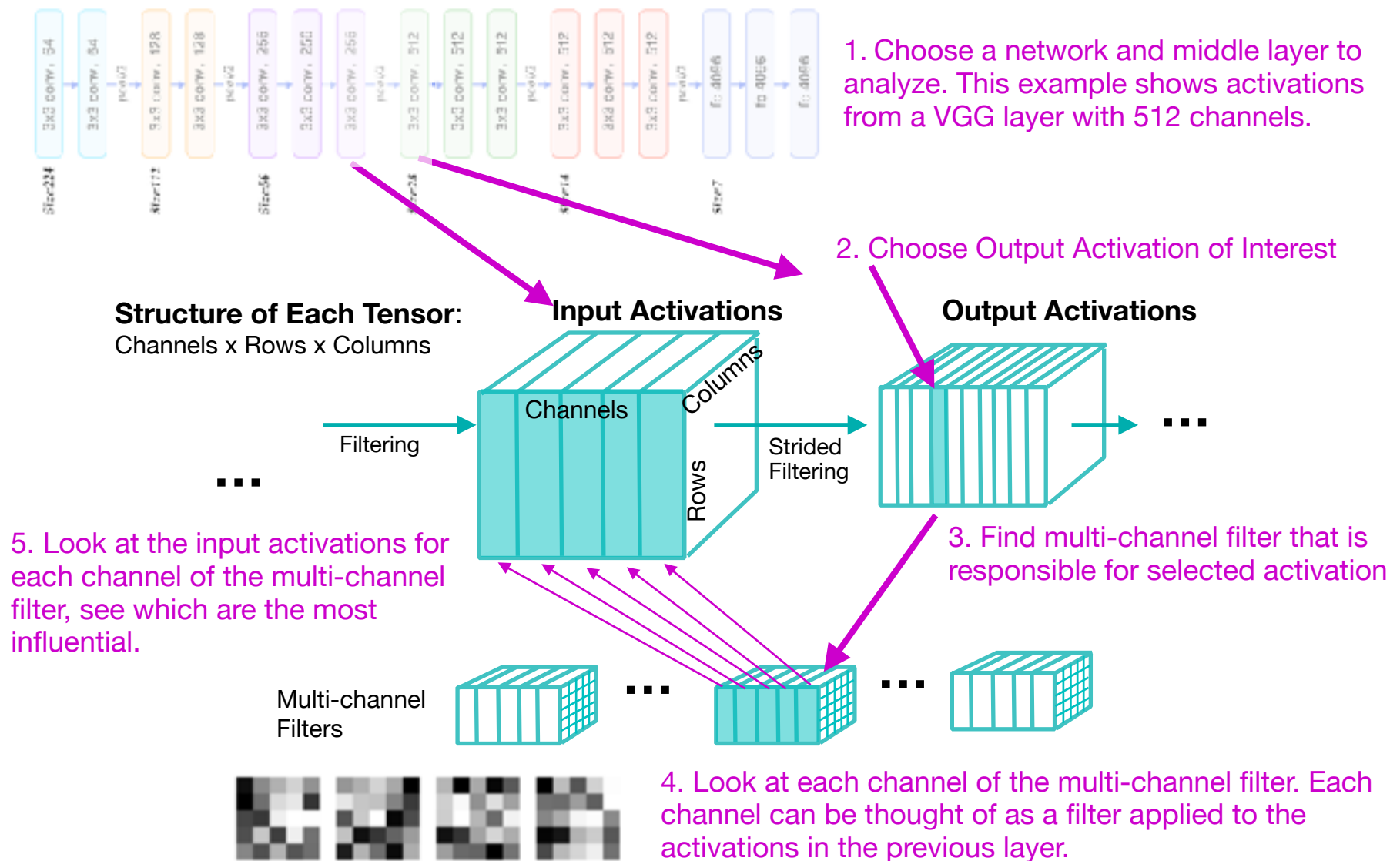
me: here *hands them a nabla ∇ *

∇ student

@TamasGorbe



Figure for Circuits Lab



Lecture Notes for Neural Networks and Machine Learning

CNN Circuits



Next Time:
Fully Convolutional Learning
Reading: Chollet 5.4

