Lecture Notes for

Neural Networks and Machine Learning

CNN Circuits

Logistics and Agenda

- Logistics
  - Lab logistics!
- Agenda
  - Last Time: Visualizing Convolutional Architectures
  - Student Paper Presentation: Augmentation Effectiveness
  - Today: Circuits in CNNs
  - Next Time: Lab Town Hall

Student Paper Presentation

Transformer Interpretability Beyond Attention Visualization

Hila Chefer[1] Shir Gur[1] Lior Wolf[1,2]
[1]The School of Computer Science, Tel Aviv University
[2]Facebook AI Research (FAIR)

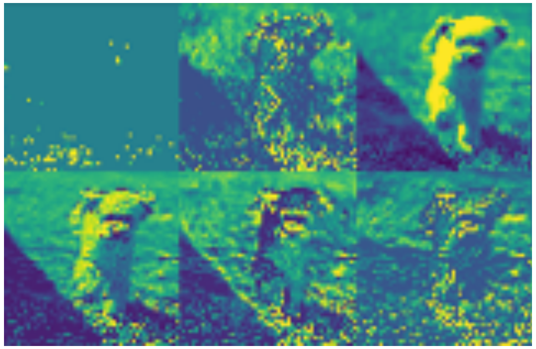
Abstract

Self-attention techniques, and specifically Transformers, are dominating the field of text processing and are becoming increasingly popular in computer vision classification tasks. In order to visualize the parts of the image that led to a certain classification, existing methods either rely on the obtained attention maps, or employ heuristic propagation along the attention graph. In this work, we propose a novel way to compute relevancy for Transformer networks. The method assigns local relevance based on the deep Taylor decomposition principle and then propagates these relevancy scores through the layers. This propagation involves attention layers and skip connections, which challenge existing methods. Our solution is based on a specific formulation that is shown to maintain the total relevancy across layers. We benchmark our method on very recent visual Transformer networks, as well as on a text classification problem, and demonstrate a clear advantage over the existing explainability methods. Our code is available at: https://github.com/hila-chefer/Transformer-Explainability

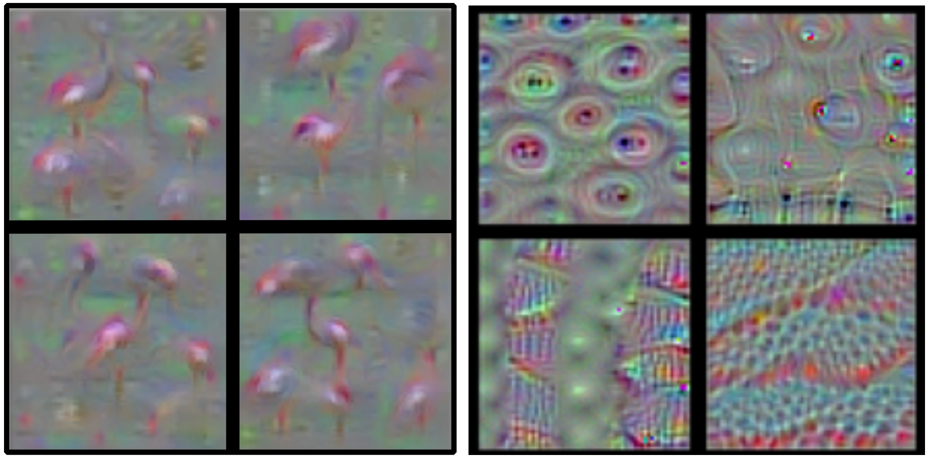
be associated with a patch [11, 4]. A common practice when trying to visualize Transformer models is, therefore, to consider these attentions as a relevancy score [39, 41, 4]. This is usually done for a single attention layer. Another option is to combine multiple layers. Simply averaging the attentions obtained for each token, would lead to blurring of the signal and would not consider the different roles of the layers: deeper layers are more semantic, but each token accumulates additional context each time self-attention is applied. The rollout method [1] is an alternative, which reassigns all attention scores by considering the pairwise attentions and assuming that attentions are combined linearly into subsequent contexts. The method seems to improve results over the utilization of a single attention layer. However, as we show, by relying on simplistic assumptions, irrelevant tokens often become highlighted.

In this work, we follow the line of work that assigns relevancy and propagates it, such that the sum of relevancy is maintained throughout the layers [26]. While the application of such methods to Transformers has been attempted [40], this was done in a partial way that does not propagate attention throughout all layers.

Review: our visualization toolset

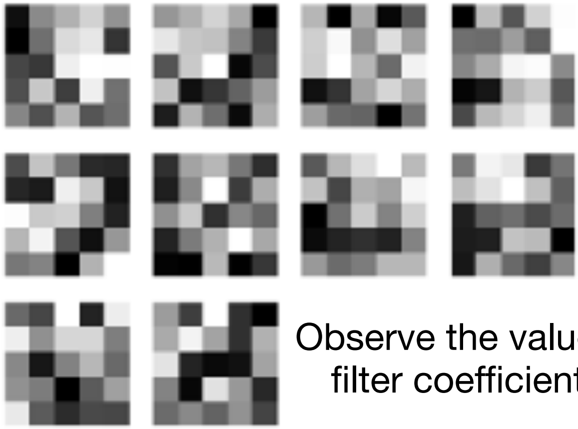


Visualize Activation
in response to input image



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.

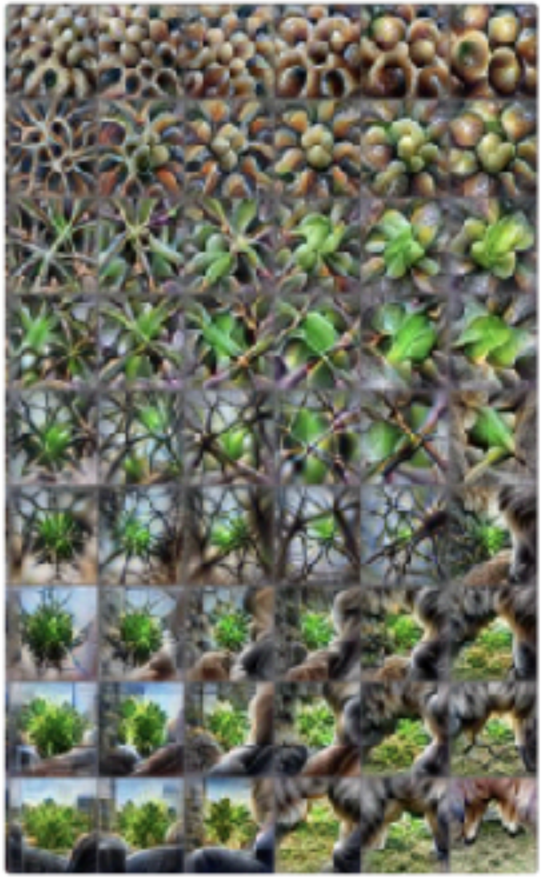
Claim 3

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.

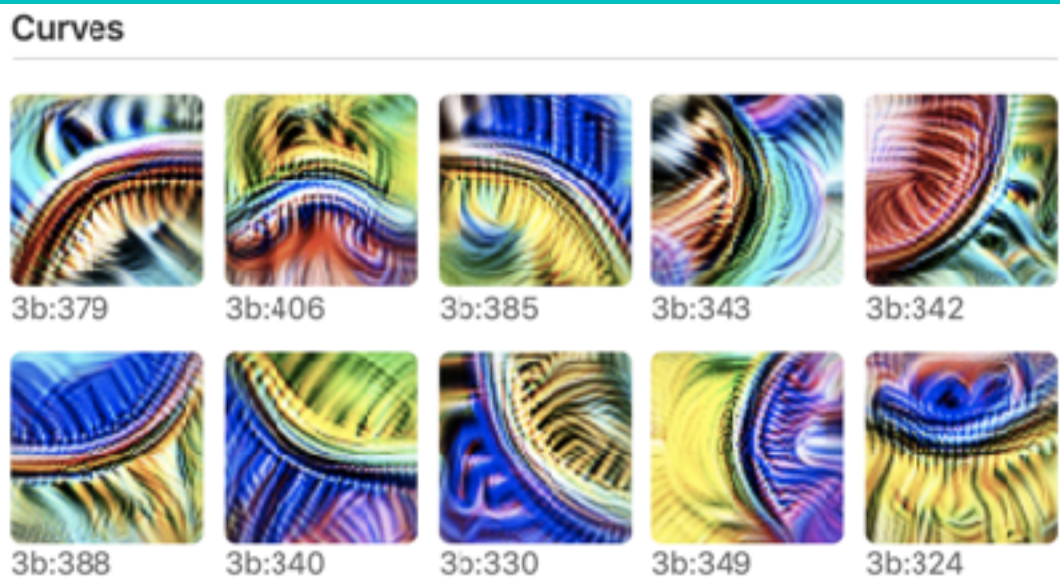
https://distill.pub/2020/circuits/zoom-in/

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 and backward 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

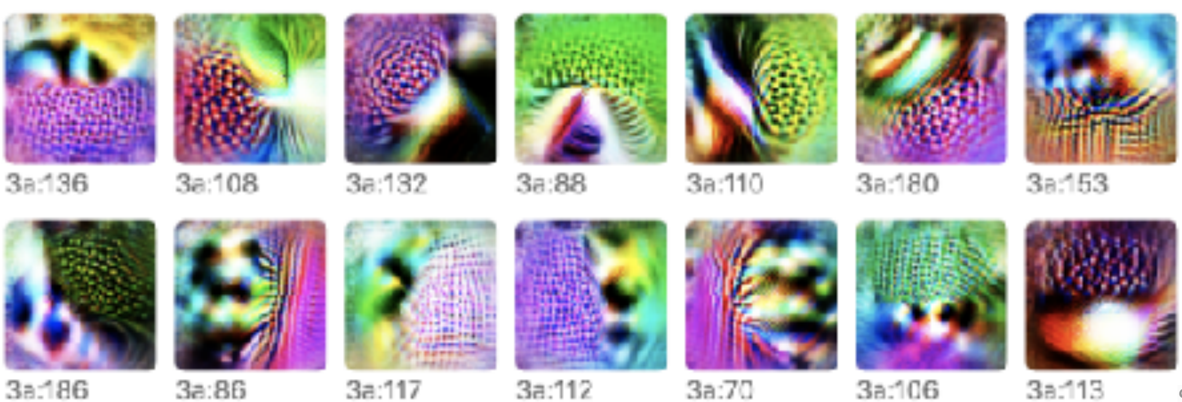
Examples of Discovered Features

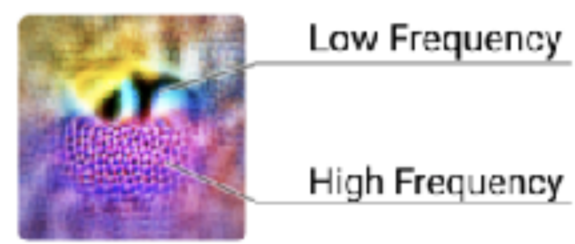


Hypothesized feature group (part of circuit)

Downstream features

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





Shubert, et al., “Hi-Lo Freq. Detectors”, 2021.

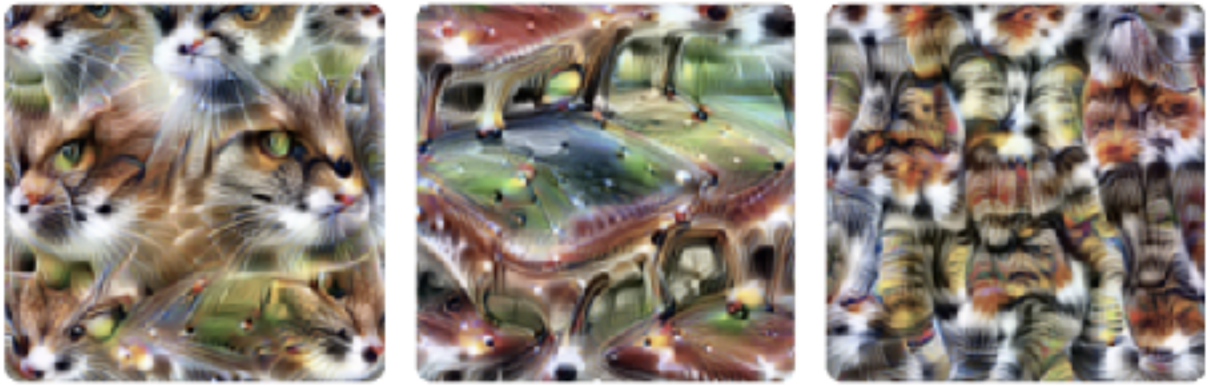
More Examples: Higher Level Features

Pose Invariant Dog-head Detection

Neuron 4b:409

Dataset examples for neuron 4b:409

Polysemantic Neurons: things that become coupled…



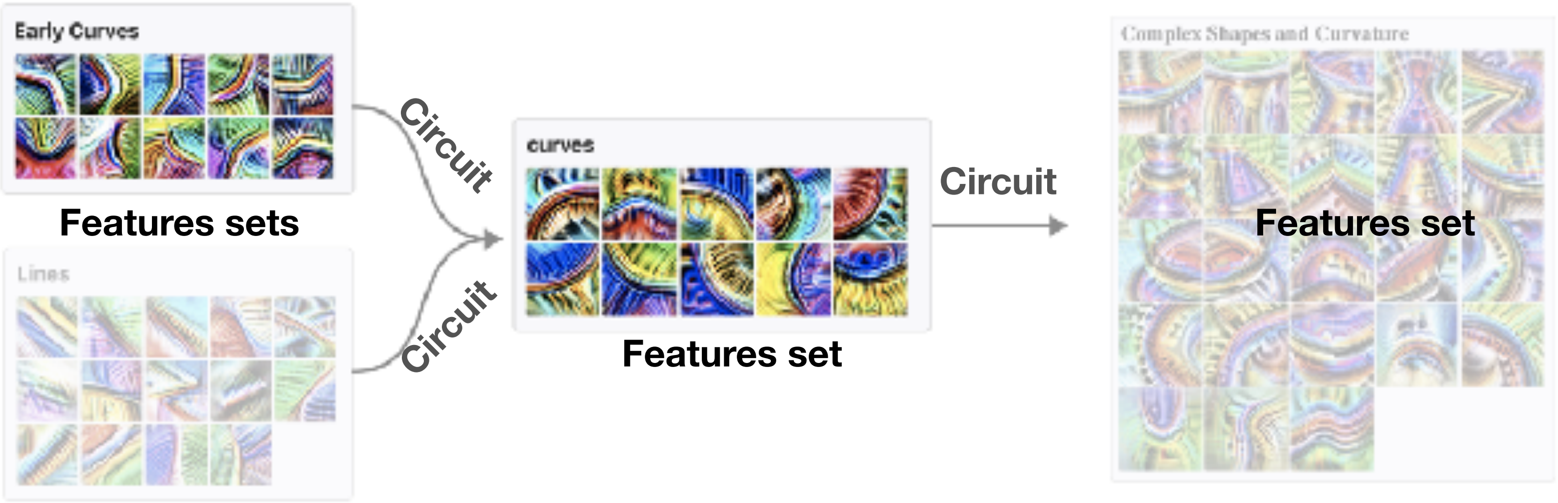
4e:55 is a polysemantic neuron which responds to cat faces, fronts of cars, and cat legs. It was discussed in more depth in Feature Visualization [4].

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

Why do these exist?

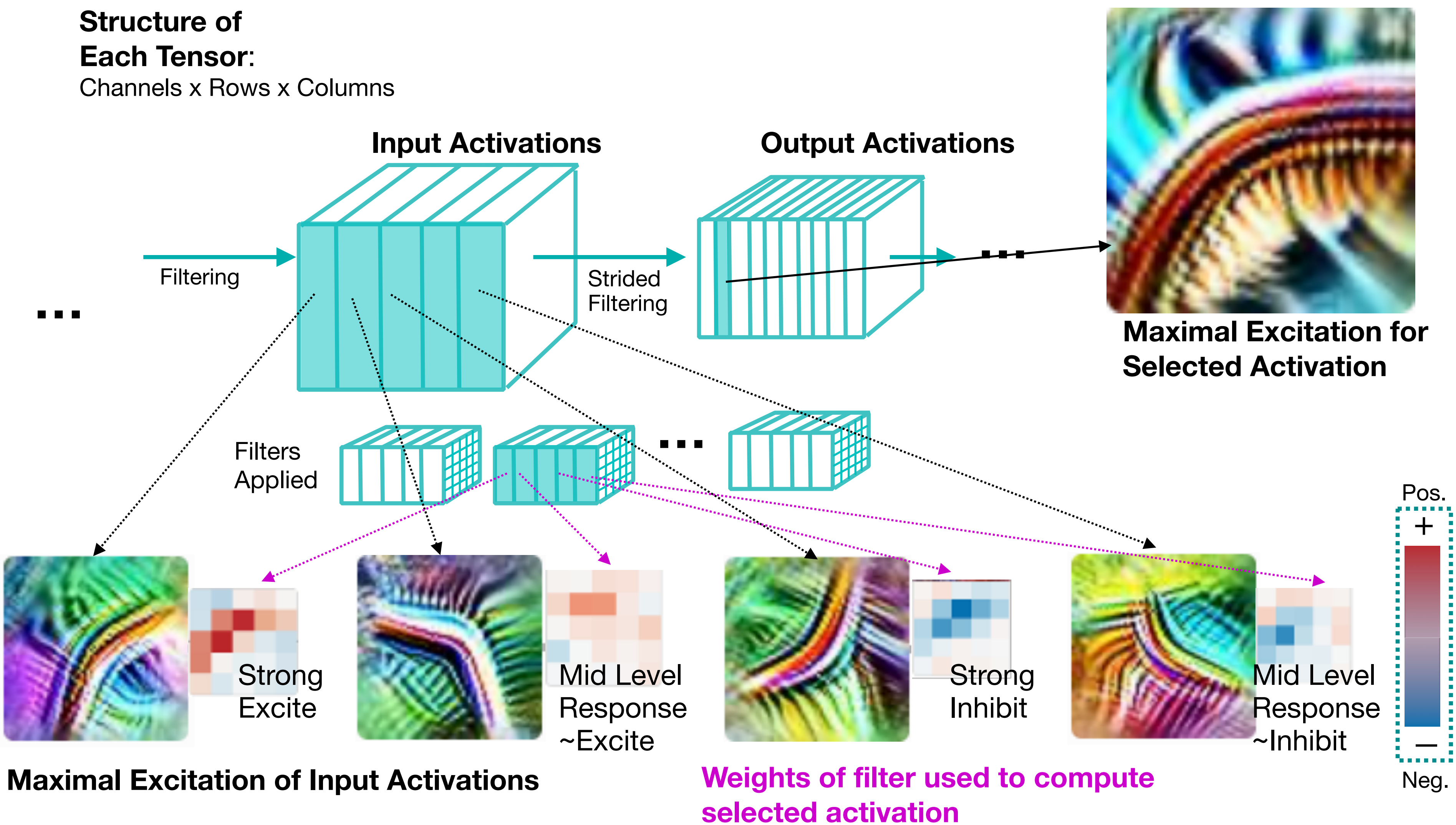
From Features to Circuits

- *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?



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

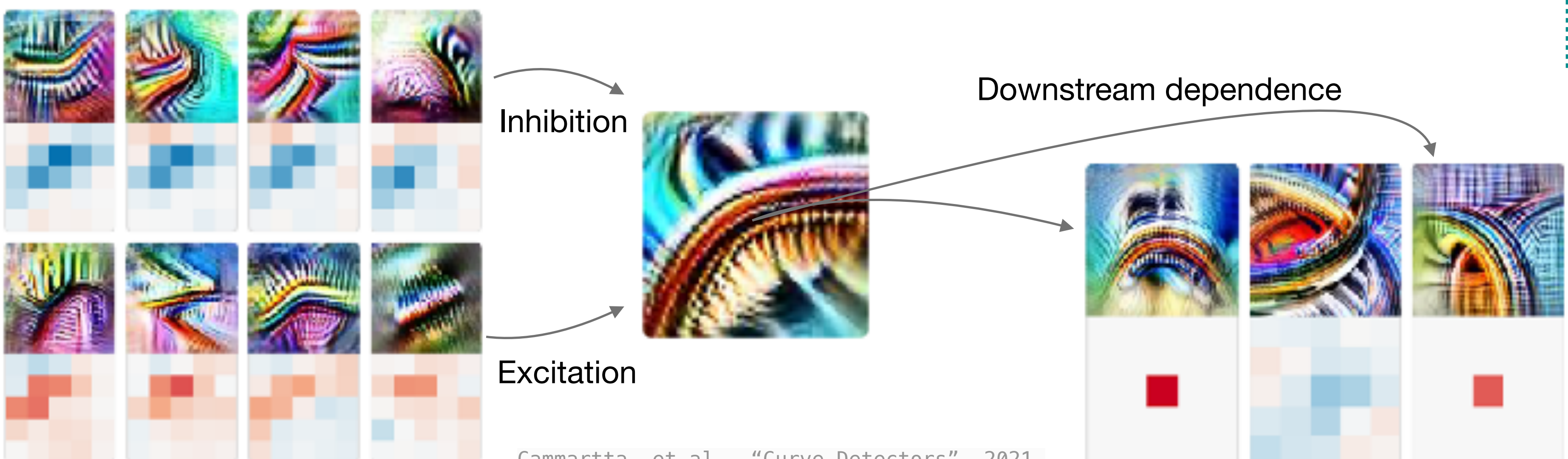
Example: Circuit for Better Curve Detection

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

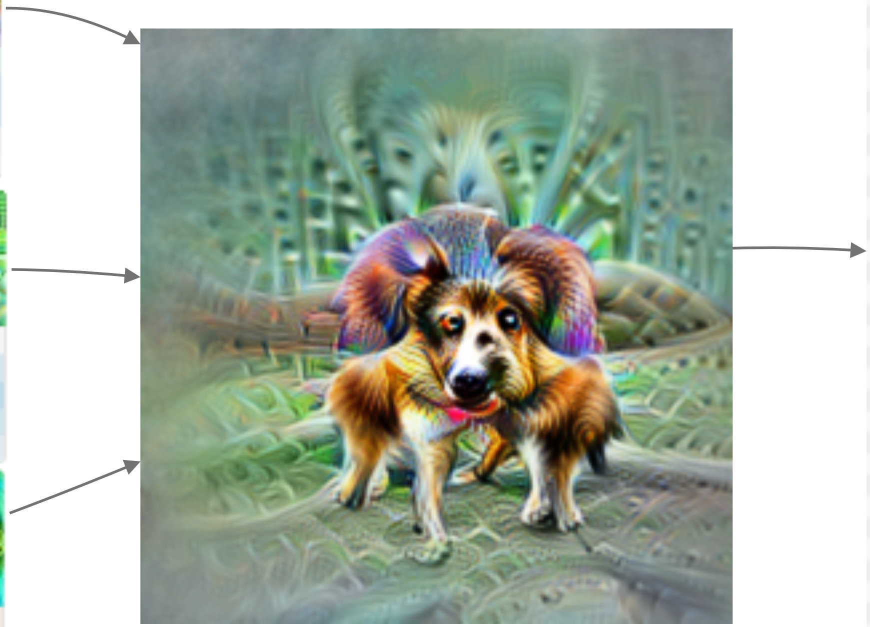
This can be interpreted as looking for "tangent curves" at each point along the curve.



Cammartta, et al., "Curve Detectors", 2021.

Another Example: Dog head

Compact Circuit Visualization



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

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

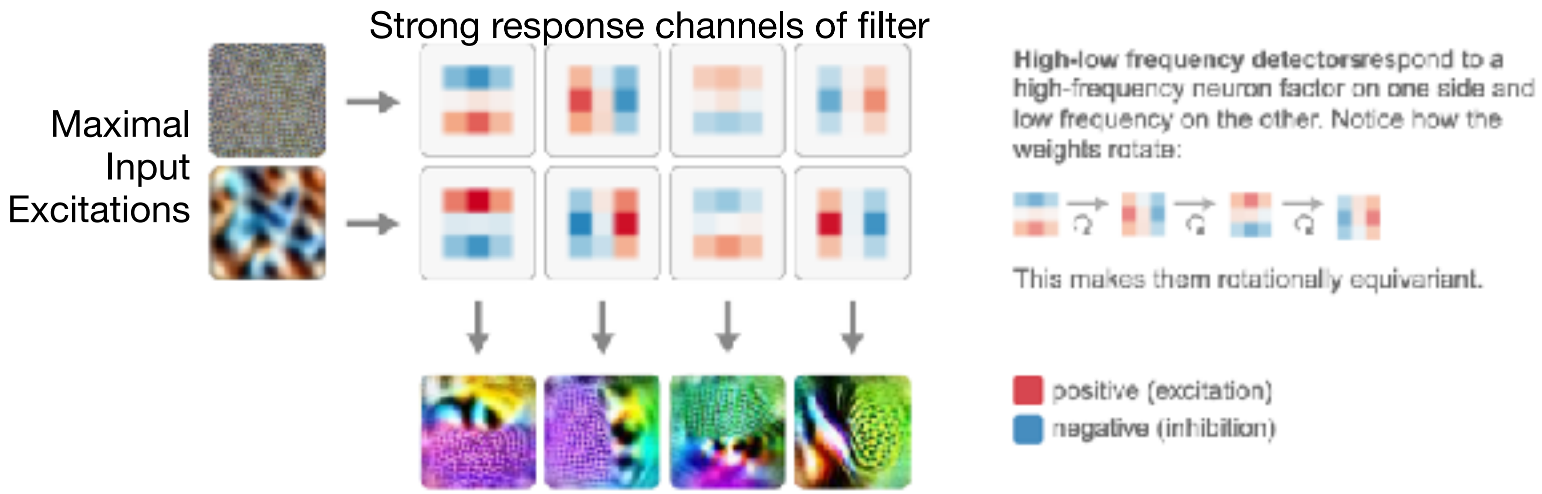
Equivariant Circuits

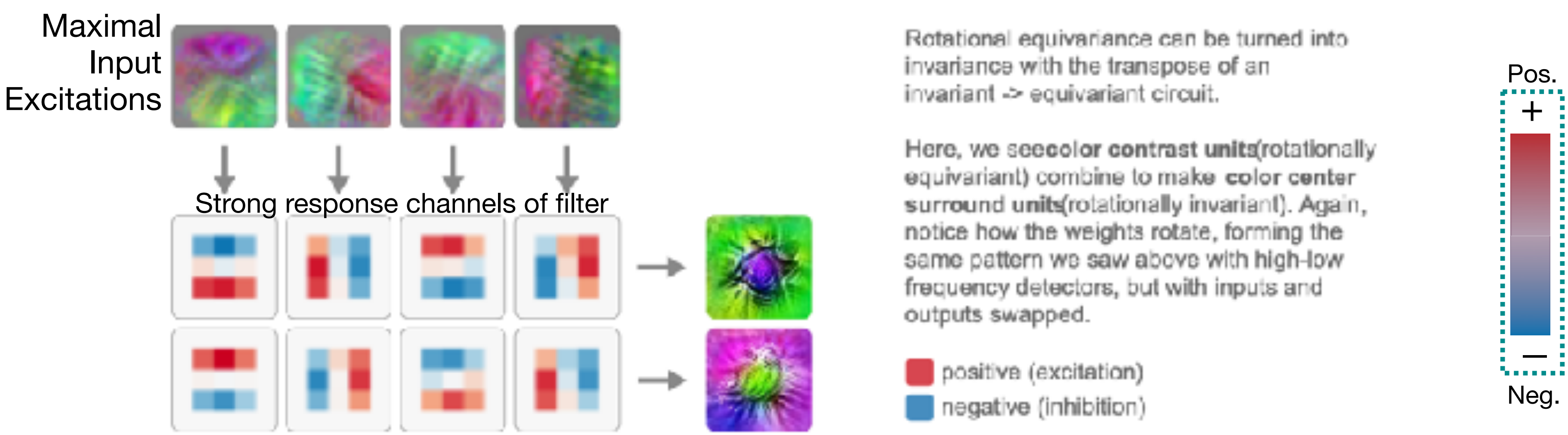
- Many features that are part of a circuit are clearly designed for rotation, hue, and other invariance

Olah, et al., "Naturally Occurring Equivariance in NN", 2021. https://distill.pub/2020/circuits/equivariance/

Equivariant circuits: a Motif

- Possible to reveal patterns of circuits via sets of weights





Olah, et al., "Naturally Occurring Equivariance in NN", 2021.

Lecture Notes for

Neural Networks and Machine Learning

CNN Visualization

Next Time:
CNN Circuits

Reading: OpenAI Circuits