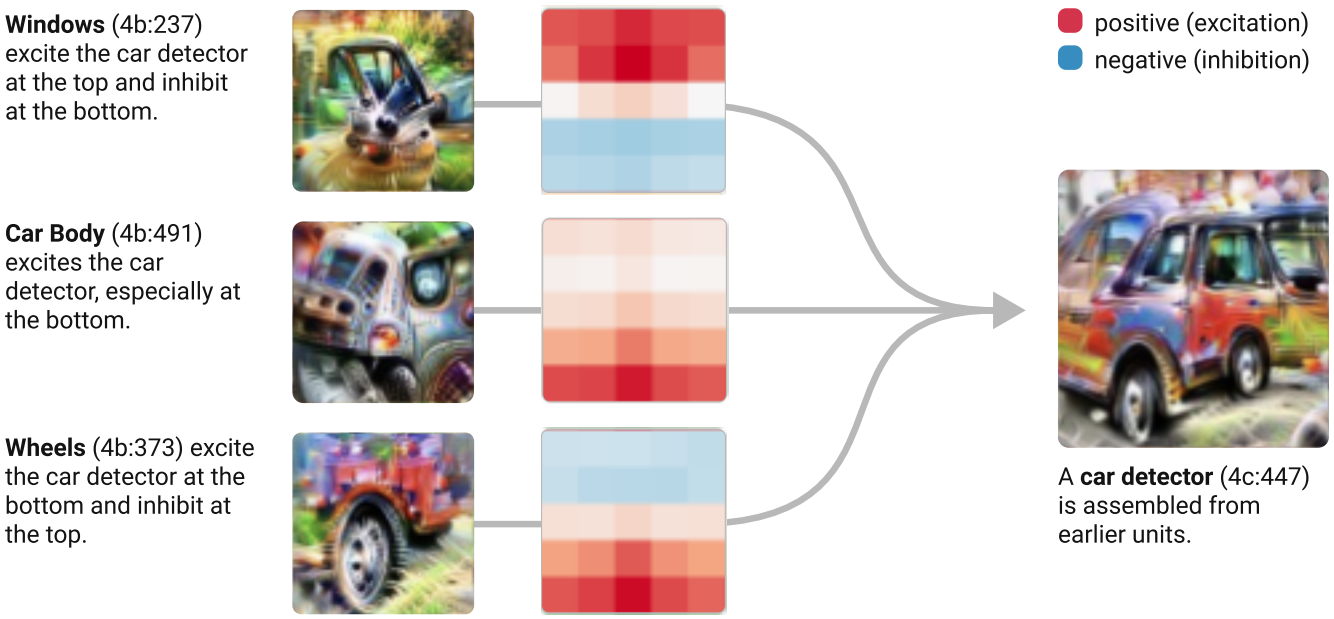
Zoom In: An Introduction to Circuits

By studying the connections between neurons, we can find meaningful algorithms in the weights of neural networks.



AUTHORS

Chris Olah — OpenAI
Nick Cammarata — OpenAI
Ludwig Schubert — OpenAI
Gabriel Goh — OpenAI
Michael Petrov — OpenAI
Shan Carter — OpenAI

PUBLISHED

March 10, 2020

DOI

10.23915/distill.00024.001

This article is part of the Circuits thread, an experimental format collecting invited short articles and critical commentary delving into the inner workings of neural networks.

← PREVIOUS ARTICLE NEXT ARTICLE →

Circuits Thread

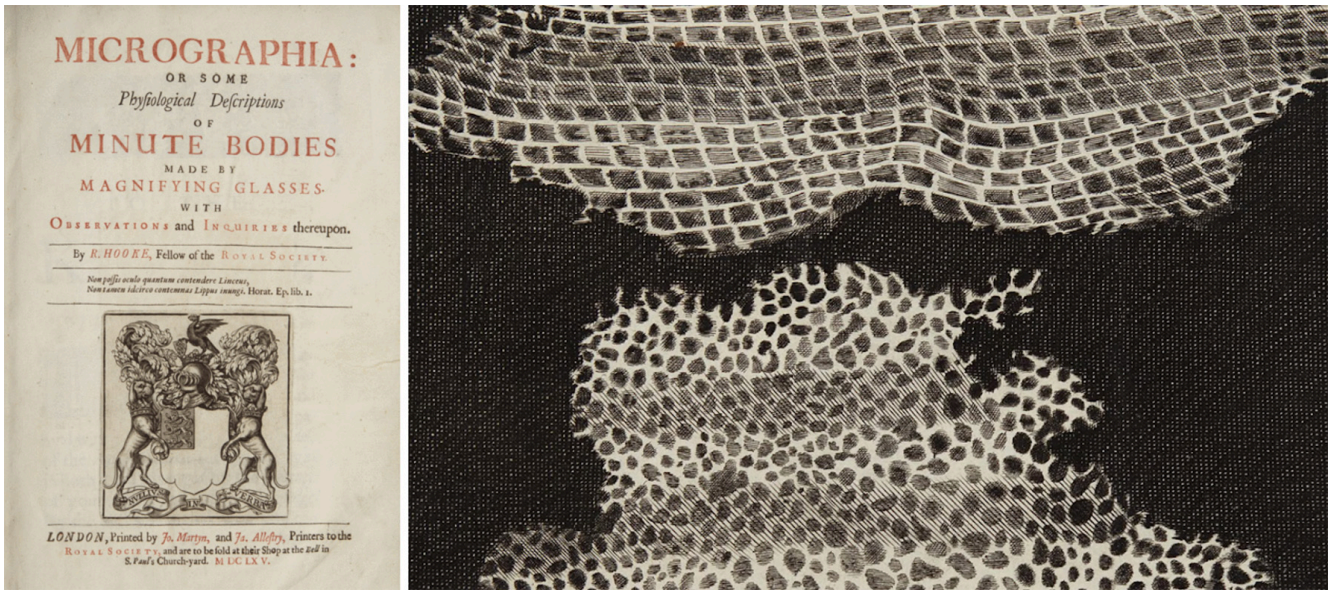
An Overview of Early Vision in InceptionV1

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.

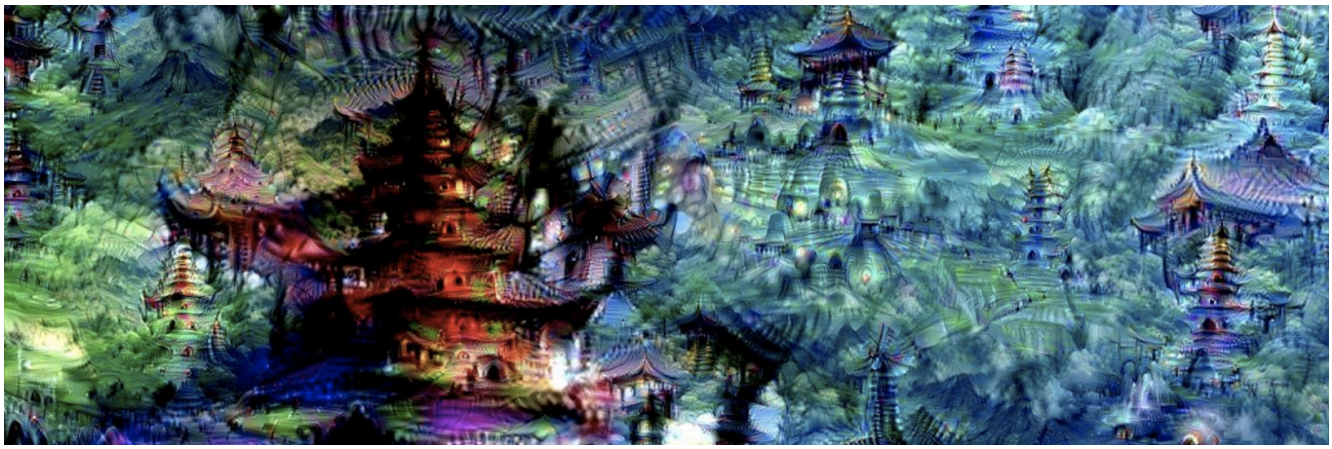


Hooke's Micrographia [1] revealed a rich microscopic world as seen through a microscope, including the initial discovery of cells.
Images from the National Library of Wales.

Just as the early microscope hinted at a new world of cells and microorganisms, visualizations of artificial neural networks have revealed tantalizing hints and glimpses of a rich inner world within our models (e.g. [2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13]). This has led us to wonder: Is it possible that deep learning is at a similar, albeit more modest, transition point?

Most work on interpretability aims to give simple explanations of an entire neural network's behavior. But what if we instead take an approach inspired by neuroscience or cellular biology—an approach of zooming in? What if we treated individual neurons, even individual weights, as being worthy of serious investigation? What if we were willing to spend thousands of hours tracing through every neuron and its connections? What kind of picture of neural networks would emerge?

In contrast to the typical picture of neural networks as a black box, we've been surprised how approachable the network is on this scale. Not only do neurons seem understandable (even ones that initially seemed inscrutable), but the "circuits" of connections between them seem to be meaningful algorithms corresponding to facts about the world. You can watch a circle detector be assembled from curves. You can see a dog head be assembled from eyes, snout, fur and tongue. You can observe how a car is composed from wheels and windows. You can even find circuits implementing simple logic: cases where the network implements AND, OR or XOR over high-level visual features.



Over the last few years, we've seen many incredible visualizations [2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 14, 12, 13] and analyses [15, 16, 17, 18] hinting at a rich world of internal features in modern neural networks. Above, we see a DeepDream [7] image, which sparked a great deal of excitement in this space.

This introductory essay offers a high-level overview of our thinking and some of the working principles that we've found useful in this line of research. In future articles, we and our collaborators will publish detailed explorations of this inner world.

But the truth is that we've only scratched the surface of understanding a single vision model. If these questions resonate with you, you are welcome to join us and our collaborators in the Circuits project, an open scientific collaboration hosted on the Distill slack.

Three Speculative Claims

One of the earliest articulations of something approaching modern cell theory was three claims by Theodor Schwann — who you may know for Schwann cells — in 1839:

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.

This translation/summarization of Schwann's claims can be found in many biology texts; we were unable to determine what the original source of the translation is. The image of Schwann's book is from the Deutsches Textarchiv.

The first two of these claims are likely familiar, persisting in modern cellular theory. The third is likely not familiar, since it turned out to be horribly wrong.

We believe there's a lot of value in articulating a strong version of something one may believe to be true, even if it might be false like Schwann's third claim. In this spirit, we offer three claims about neural networks. They are intended both as empirical claims about the nature of neural networks, and also as normative claims about how it's useful to understand them.

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.

These claims are deliberately speculative. They also aren't totally novel: claims along the lines of (1) and (3) have been suggested before, as we'll discuss in more depth below.

But we believe these claims are important to consider because, if true, they could form the basis of a new "zoomed in" field of interpretability. In the following sections, we'll discuss each one individually and present some of the evidence that has led us to believe they might be true.

Claim 1: Features

> Features are the fundamental unit of neural networks. They correspond to directions. They can be rigorously studied and understood.

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.

Of course, being understandable doesn't mean being simple or easily understandable. Many neurons are initially mysterious and don't follow our a priori guesses of what features might exist! However, our experience is that there's usually a simple explanation behind these neurons, and that they're actually doing something quite natural. For example, we were initially confused by high-low frequency detectors (discussed below) but in retrospect, they are simple and elegant.

This introductory essay will only give an overview of a couple examples we think are illustrative, but it will be followed both by deep dives carefully characterizing individual features, and broad overviews sketching out all the features we understand to exist. We will take our examples from InceptionV1 [26] for now, but believe these claims hold generally and will discuss other models in the final section on universality.

Regardless of whether we're correct or mistaken about meaningful features, we believe this is an important question for the community to resolve. We hope that introducing several specific carefully explored examples of seemingly understandable features will help advance the dialogue.

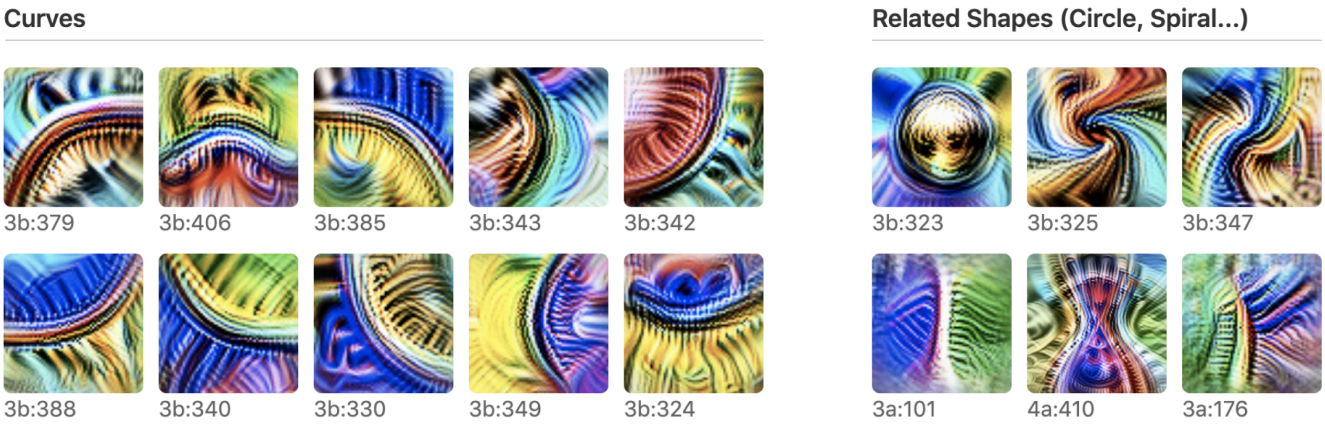
Example 1: Curve Detectors

Curve detecting neurons can be found in every non-trivial vision model we've carefully examined. These units are interesting because they straddle the boundary between features the community broadly agrees exist (e.g. edge detectors) and features for which there's significant skepticism (e.g. high-level features such as ears, automotives, and faces).

We'll focus on curve detectors in layer `mixed3b`, an early layer of InceptionV1. These units responded to curved lines and boundaries with a radius of around 60 pixels. They are also slightly additionally excited by perpendicular lines along the boundary of the curve, and prefer the two sides of the curve to be different colors.

Curve detectors are found in families of units, with each member of the family detecting the same curve feature in a different orientation. Together, they jointly span the full range of orientations.

It's important to distinguish curve detectors from other units which may seem superficially similar. In particular, there are many units which use curves to detect a curved sub-component (e.g. circles, spirals, S-curves, hourglass shape, 3d curvature, ...). There are also units which respond to curve related shapes like lines or sharp corners. We do not consider these units to be curve detectors.



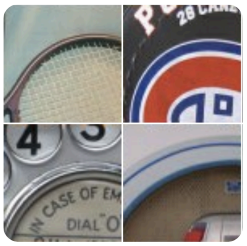
But are these "curve detectors" really detecting curves? We will be dedicating an entire later article to exploring this in depth, but the summary is that we think the evidence is quite strong.

We offer seven arguments, outlined below. It's worth noting that none of these arguments are curve specific: they're a useful, general toolkit for testing our understanding of other features as well. Several of these arguments — dataset examples, synthetic examples, and tuning curves — are classic methods from visual neuroscience (e.g. [27]). The last three arguments are based on circuits, which we'll discuss in the next section.

ARGUMENT 1: FEATURE VISUALIZATION

Optimizing the input to cause curve detectors to fire reliably produces curves. This establishes a causal link, since everything in the resulting image was added to cause the neuron to fire more.

You can learn more about feature visualization here.

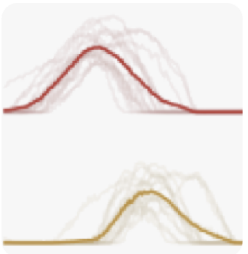


ARGUMENT 2: DATASET EXAMPLES

The ImageNet images that cause these neurons to strongly fire are reliably curves in the expected orientation. The images that cause them to fire moderately are generally less perfect curves or curves off orientation.

ARGUMENT 3: SYNTHETIC EXAMPLES

Curve detectors respond as expected to a range of synthetic curves images created with varying orientations, curvatures, and backgrounds. They fire only near the expected orientation, and do not fire strongly for straight lines or sharp corners.



ARGUMENT 4: JOINT TUNING

If we take dataset examples that cause a neuron to fire and rotate them, they gradually stop firing and the curve detectors in the next orientation begins firing. This shows that they detect rotated versions of the same thing. Together, they tile the full 360 degrees of potential orientations.

ARGUMENT 5: FEATURE IMPLEMENTATION (CIRCUIT-BASED ARGUMENT)

By looking at the circuit constructing the curve detectors, we can read a curve detection algorithm off of the weights. We also don't see anything suggestive of a second alternative cause of firing, although there are many smaller weights we don't understand the role of.

ARGUMENT 6: FEATURE USE (CIRCUIT-BASED ARGUMENT)

The downstream clients of curve detectors are features that naturally involve curves (e.g. circles, 3d curvature, spirals...). The curve detectors are used by these clients in the expected manner.

ARGUMENT 7: HANDWRITTEN CIRCUITS (CIRCUIT-BASED ARGUMENT)

Based on our understanding of how curve detectors are implemented, we can do a cleanroom reimplementation, hand setting all weights to reimplement curve detection. These weights are an understandable curve detection algorithm, and significantly mimic the original curve detectors.

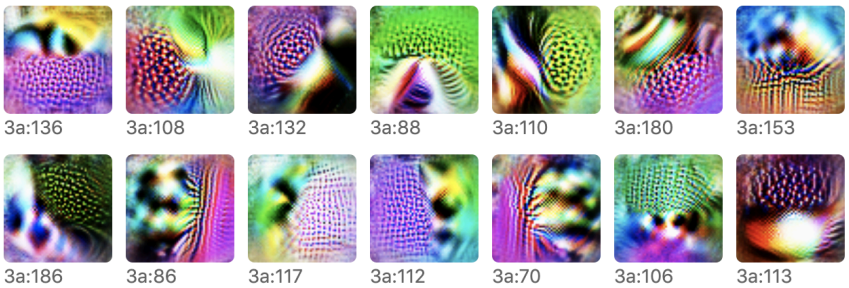
The above arguments don't fully exclude the possibility of some rare secondary case where curve detectors fire for a different kind of stimulus. But they do seem to establish that (1) curves cause these neurons to fire, (2) each unit responds to curves at different angular orientations, and (3) if there are other stimuli that cause them to fire those stimuli are rare or cause weaker activations. More generally, these arguments seem to meet the evidentiary standards we understand to be used in neuroscience, which has established traditions and institutional knowledge of how to evaluate such claims.

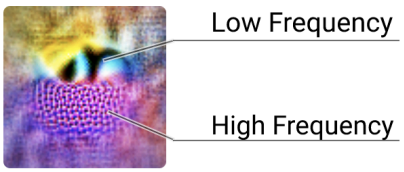
All of these arguments will be explored in detail in the later articles on curve detectors and curve detection circuits.

Example 2: High-Low Frequency Detectors

Curve detectors are an intuitive type of feature—the kind of feature one might guess exists in neural networks a priori. Given that they're present, it's not surprising we can understand them. But what about features that aren't intuitive? Can we also understand those? We believe so.

High-low frequency detectors are an example of a less intuitive type of feature. We find them in early vision, and once you understand what they're doing, they're quite simple. They look for low-frequency patterns on one side of their receptive field, and high-frequency patterns on the other side. Like curve detectors, high-low frequency detectors are found in families of features that look for the same thing in different orientations.





Why are high-low frequency detectors useful to the network? They seem to be one of several heuristics for detecting the boundaries of objects, especially when the background is out of focus. In a later article, we'll explore how they're used in the construction of sophisticated boundary detectors.

(One hope some researchers have for interpretability is that understanding models will be able to teach us better abstractions for thinking about the world [28]. High-low frequency detectors are, perhaps, an example of a small success in this: a natural, useful visual feature that we didn't anticipate in advance.)

All seven of the techniques we used to interrogate curve neurons can also be used to study high-low frequency neurons with some tweaking — for instance, rendering synthetic high-low frequency examples. Again we believe these arguments collectively provide strong support for the idea that these really are a family of high-low frequency contrast detectors.

Example 3: Pose-Invariant Dog Head Detector

Both curve detectors and high-low frequency detectors are low-level visual features, found in the early layers of InceptionV1. What about more complex, high-level features?

Let's consider this unit which we believe to be a pose-invariant dog detector. As with any neuron, we can create a feature visualization and collect dataset examples. If you look at the feature visualization, the geometry is... not possible, but very informative about what it's looking for and the dataset examples validate it.

Neuron 4b:409

Dataset examples for neuron 4b:409

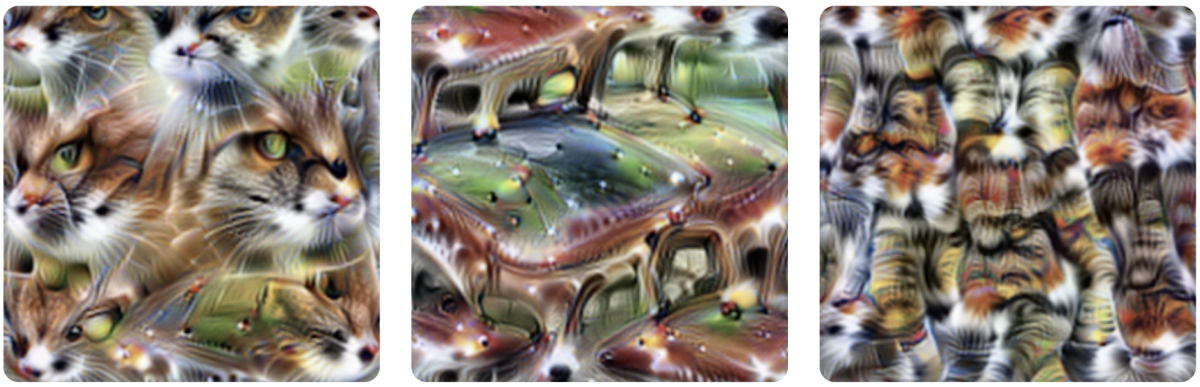
It's worth noting that the combination of feature visualization and dataset examples alone are already quite a strong argument. Feature visualization establishes a causal link, while dataset examples test the neuron's use in practice and whether there are a second type of stimuli that it reacts to. But we can bring all our other approaches to analyzing a neuron to bear again. For example, we can use a 3D model to generate synthetic dog head images from different angles.

At the same time, some of the approaches we've emphasized so far become a lot of effort for these higher-level, more abstract features. Thankfully, our circuit-based arguments — which we'll discuss more soon — will continue to be easy to apply, and give us really powerful tools for understanding and testing high-level features that don't require a lot of effort.

Polysemantic Neurons

This essay may be giving you an overly rosy picture: perhaps every neuron yields a nice, human-understandable concept if one seriously investigates it?

Alas, this is not the case. Neural networks often contain "polysemantic neurons" that respond to multiple unrelated inputs. For example, InceptionV1 contains one neuron that responds to cat faces, fronts of cars, and cat legs.



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].

To be clear, this neuron isn't responding to some commonality of cars and cat faces. Feature visualization shows us that it's looking for the eyes and whiskers of a cat, for furry legs, and for shiny fronts of cars — not some subtle shared feature.

We can still study such features, characterizing each different case they fire, and reason about their circuits to some extent. Despite this, polysemantic neurons are a major challenge for the circuits agenda, significantly limiting our ability to reason about neural networks. [4] Our hope is that it may be possible to resolve polysemantic neurons, perhaps by "unfolding" a network to turn polysemantic neurons into pure features, or training networks to not exhibit polysemanticity in the first place. This is essentially the problem studied in the literature of disentangling representations, although at present that literature tends to focus on known features in the latent spaces of generative models.

One natural question to ask is why do polysemantic neurons form? In the next section, we'll see that they seem to result from a phenomenon we call "superposition" where a circuit spreads a feature across many neurons, presumably to pack more features into the limited number of neurons it has available.

Claim 2: Circuits

> Features are connected by weights, forming circuits.
> These circuits can also be rigorously studied and understood.

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? To explore this, we find it helpful to study circuits: sub-graphs of the network, consisting a set of tightly linked features and the weights between them.

The remarkable thing is how tractable and meaningful these circuits seem to be as objects of study. When we began looking, we expected to find something quite messy. Instead, we've found beautiful rich structures, often with symmetry to them. Once you understand what features they're connecting together, the individual floating point number weights in your neural network become meaningful! *You can literally read meaningful algorithms off of the weights*.

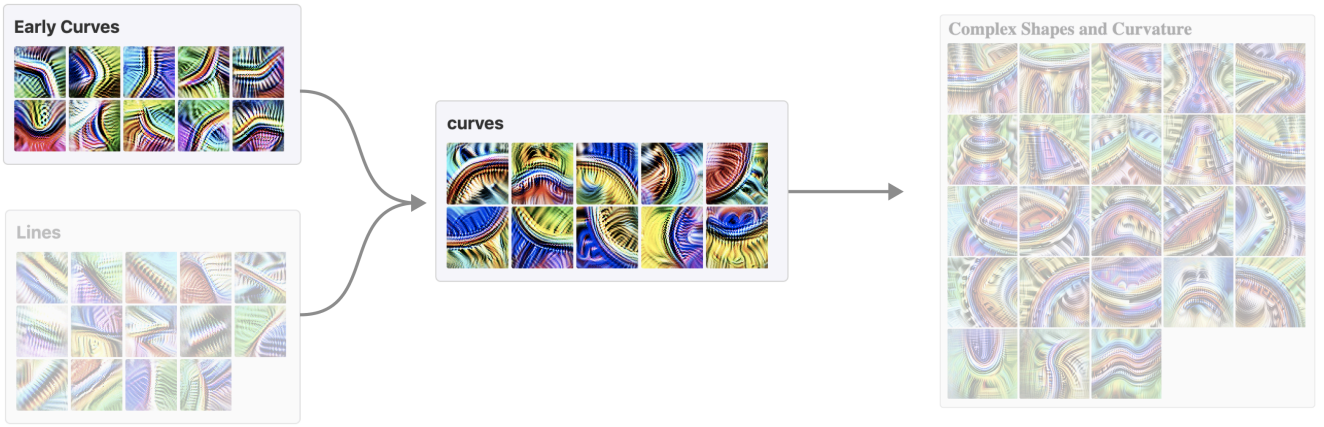
Let's consider some examples.

Circuit 1: Curve Detectors

In the previous section, we discussed curve detectors, a family of units detecting curves in different angular orientations. In this section, we'll explore how curve detectors are implemented from earlier features and connect to the rest of the model.

Curve detectors are primarily implemented from earlier, less sophisticated curve detectors and line detectors. These curve detectors are used in the next layer to create 3D geometry and complex shape detectors. Of course, there's a long tail of smaller connections to other features, but this seems to be the primary story.

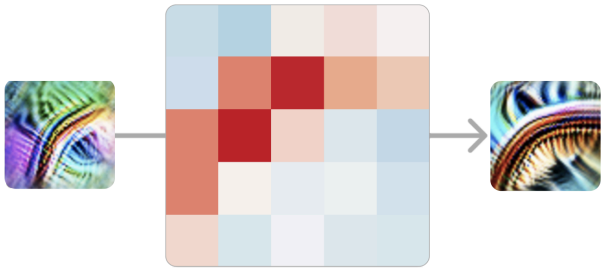
For this introduction, we'll focus on the interaction of the early curve detectors and our full curve detectors.



Let's focus even more and look at how a single early curve detector connects to a more sophisticated curve detector in the same orientation.

In this case, our model is implementing a 5x5 convolution, so the weights linking these two neurons are a 5x5 set of weights, which can be positive or negative. [5] A positive weight means that if the earlier neuron fires in that position, it excites the late neuron. Conversely a negative weight would mean that it inhibits it.

What we see are strong positive weights, arranged in the shape of the curve detector. We can think of this as meaning that, at each point along the curve, our curve detector is looking for a "tangent curve" using the earlier curve detector.



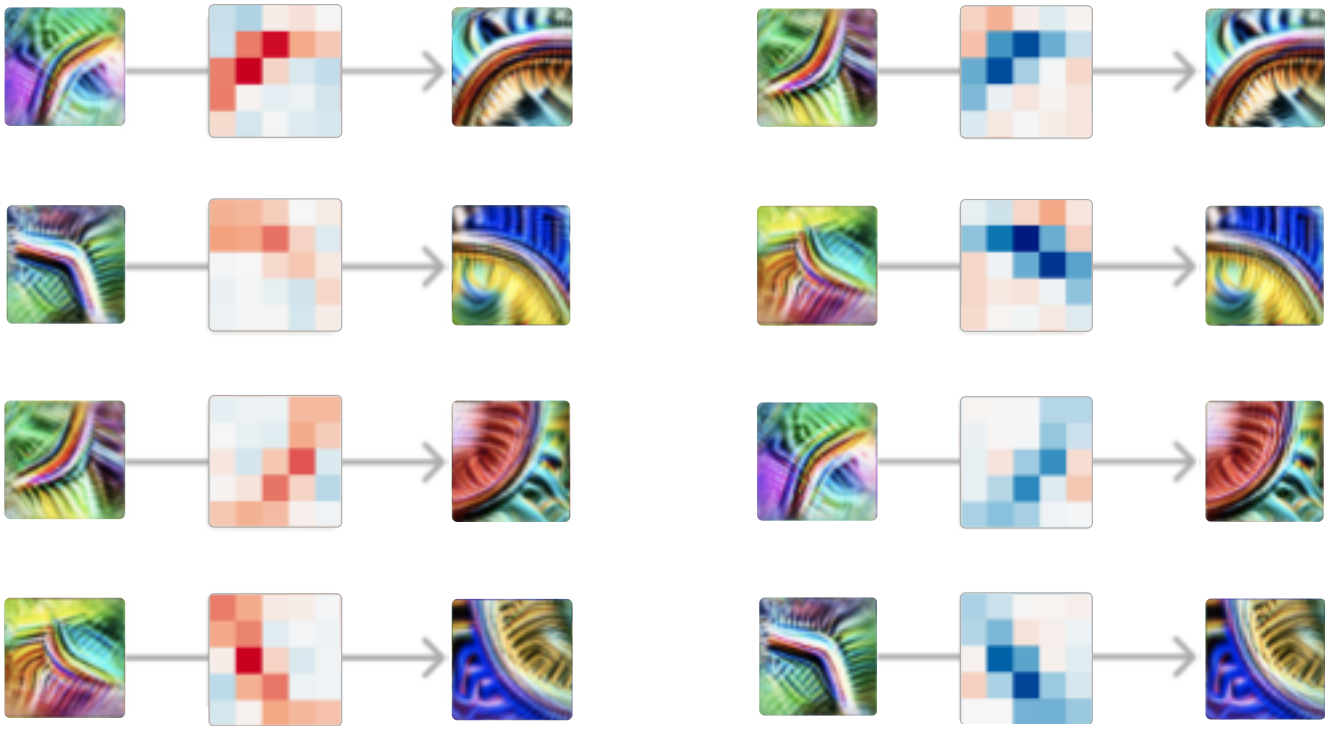
The raw weights between the early curve detector and late curve detector in the same orientation are a curve of **positive weights** surrounded by small **negative** or zero weights.

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

This is true for every pair of early and full curve detectors in similar orientations. At every point along the curve, it detects the curve in a similar orientation. Similarly, curves in the opposite orientation are inhibitory at every point along the curve.

Curve detectors are **excited** by earlier detectors in **similar orientations**...

... and **inhibited** by earlier detectors in **opposing orientations**.



It's worth reflecting here that we're looking at neural network weights and they're meaningful.

And the structure gets richer the closer you look. For example, if you look at an early curve detector and full curve detector in similar, but not exactly the same orientation you can often see it have stronger positive weights on the side of the curve it is more aligned with.

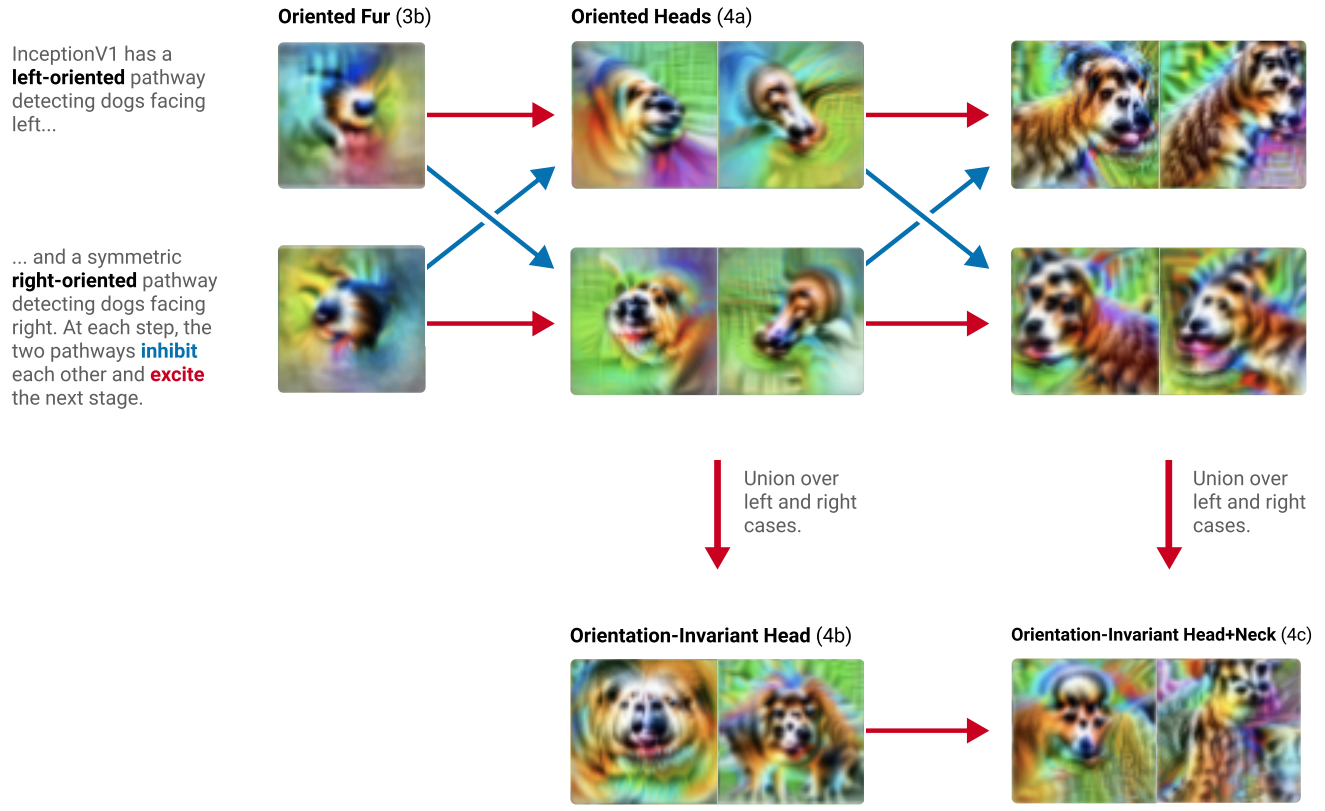
It's also worth noting how the weights rotate with the orientation of the curve detector. The symmetry of the problem is reflected as a symmetry in the weights. We call circuits with exhibiting this phenomenon an "equivariant circuit", and will discuss it in depth in a later article.

Circuit 2: Oriented Dog Head Detection

The curve detector circuit is a low-level circuit and only spans two layers. In this section, we'll discuss a higher-level circuit spanning across four layers. This circuit will also teach us about how neural networks implement sophisticated invariances.

Remember that a huge part of what an ImageNet model has to do is tell apart different animals. In particular, it has to distinguish between a hundred different species of dogs! And so, unsurprisingly, it develops a large number of neurons dedicated to recognizing dog related features, including heads.

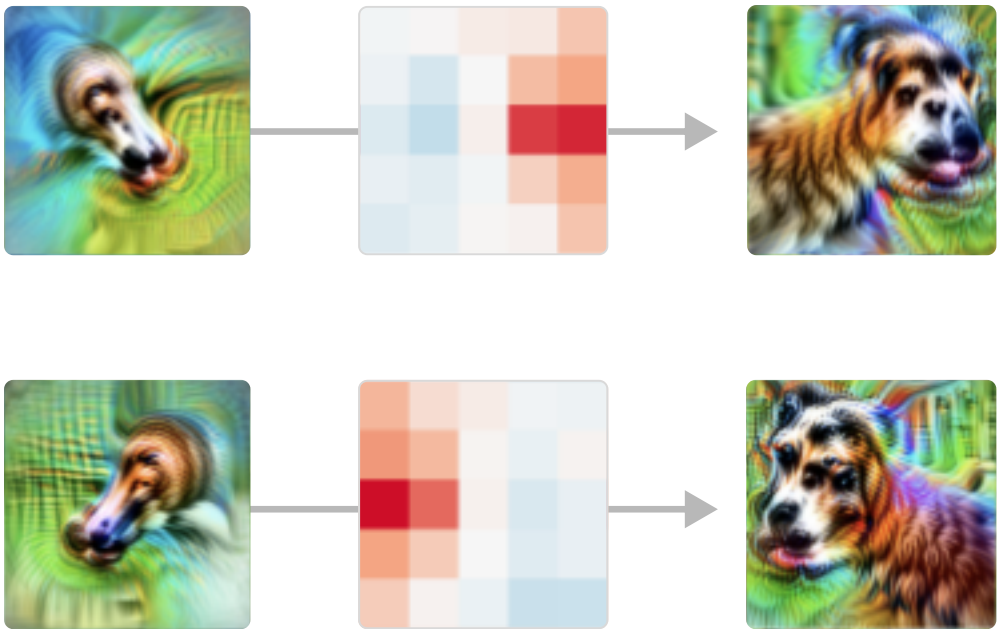
Within this "dog recognition" system, one circuit strikes us as particularly interesting: a collection of neurons that handle dog heads facing to the left and dog heads facing to the right. Over three layers, the network maintains two mirrored pathways, detecting analogous units facing to the left and to the right. At each step, these pathways try to inhibit each other, sharpening the contrast. Finally, it creates invariant neurons which respond to both pathways.



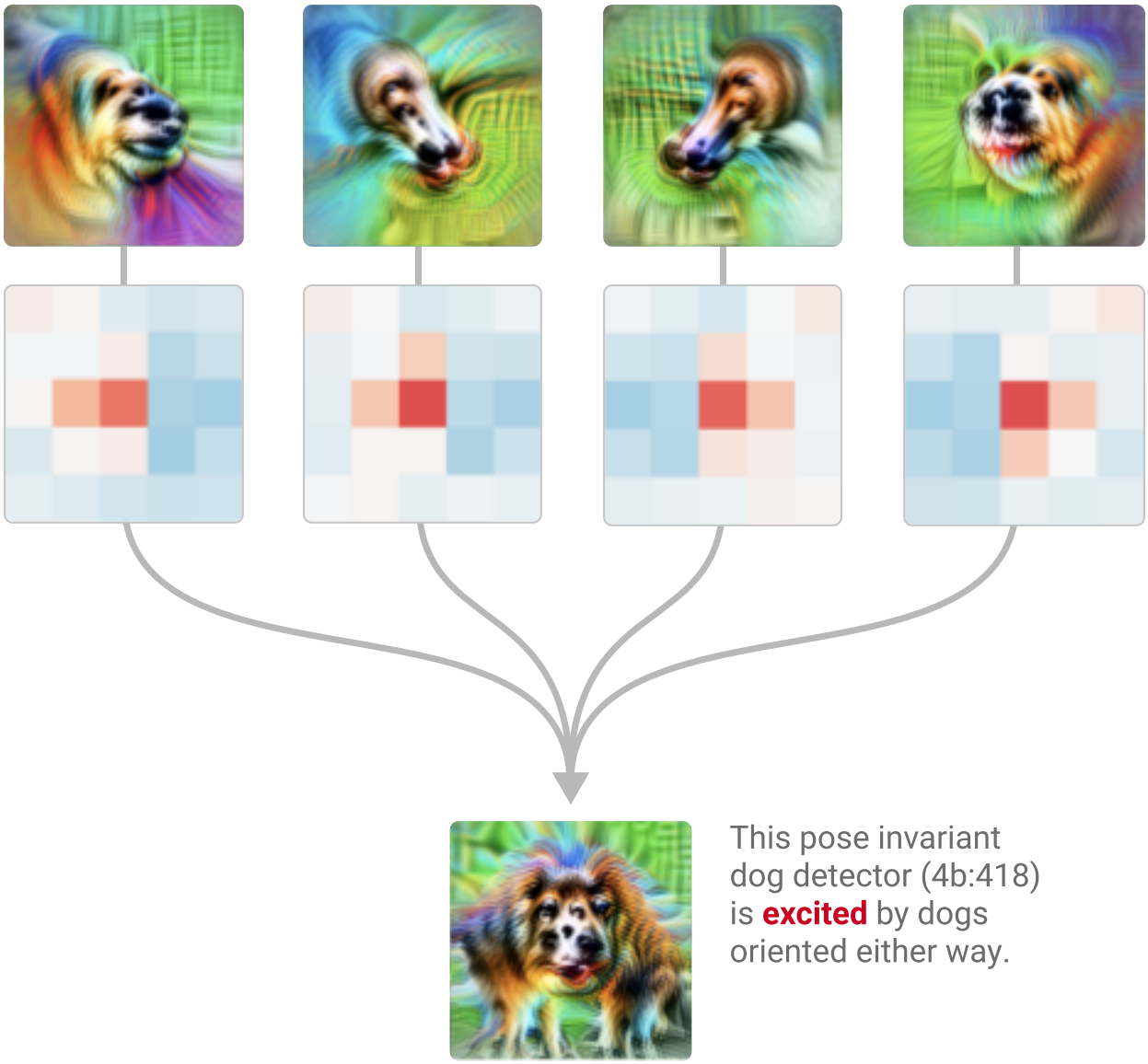
We call this pattern “unioning over cases”. The network separately detects two cases (left and right) and then takes a union over them to create invariant “multifaceted” [29] units. Note that, because the two pathways inhibit each other, this circuit actually has some XOR like properties.

This circuit is striking because the network could have easily done something much less sophisticated. It could easily create invariant neurons by not caring very much about where the eyes, fur and snout went, and just looking for a jumble of them together. But instead, the network has learned to carve apart the left and right cases and handle them separately. We’re somewhat surprised that gradient descent could learn to do this! [6]

But this summary of the circuit is only scratching the surface of what is going on. Every connection between neurons is a convolution, so we can also look at where an input neuron excites the the next one. And the models tends to be doing what you might have optimistically hoped. For example, consider these “head with neck” units. The head is only detected on the correct side:



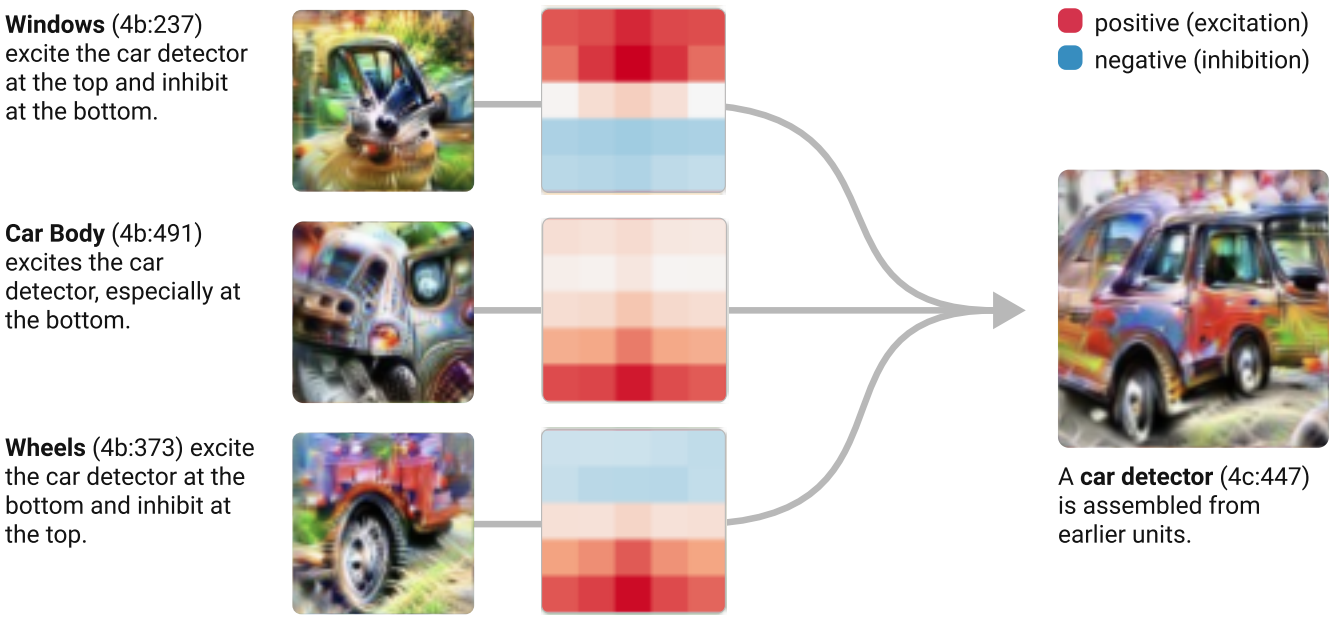
The union step is also interesting to look at the details of. The network doesn't indiscriminately respond to the heads into the two orientations: the regions of excitation extend from the center in different directions depending on orientation, allowing snouts to converge in to the same point.



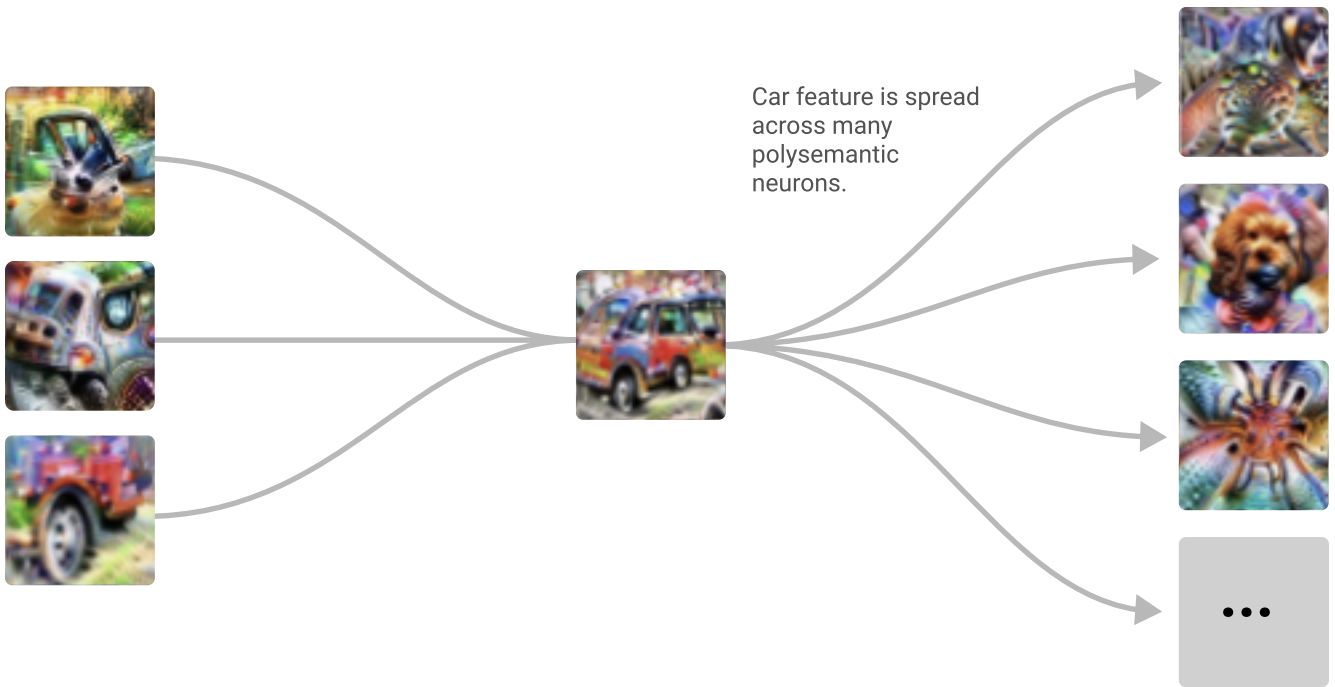
There's a lot more to say about this circuit, so we plan to return to it in a future article and analyze it in depth, including testing our theory of the circuit by editing the weights.

Circuit 3: Cars in Superposition

In `mixed4c`, a mid-late layer of InceptionV1, there is a car detecting neuron. Using features from the previous layers, it looks for wheels at the bottom of its convolutional window, and windows at the top.



But then the model does something surprising. Rather than create another pure car detector at the next layer, it spreads its car feature over a number of neurons that seem to primarily be doing something else—in particular, dog detectors.



This circuit suggests that polysemantic neurons are, in some sense, deliberate. That is, you could imagine a world where the process of detecting cars and dogs was deeply intertwined in the model for some reason, and as a result polysemantic neurons were difficult to avoid. But what we're seeing here is that the model had a "pure neuron" and then mixed it up with other features.

We call this phenomenon superposition.

Why would it do such a thing? We believe superposition allows the model to use fewer neurons, conserving them for more important tasks. As long as cars and dogs don't co-occur, the model can accurately retrieve the dog feature in a later layer, allowing it to store the feature without dedicating a neuron. [7]

Circuit Motifs

As we've studied circuits throughout InceptionV1 and other models, we've seen the same abstract patterns over and over. Equivariance, as we saw with the curve detectors. Unioning over cases, as we saw with the pose-invariant dog head detector. Superposition, as we saw with the car detector.

In biology, a circuit motif [30] is a recurring pattern in complex graphs like transcription networks or biological neural networks. Motifs are helpful because understanding one motif can give researchers leverage on all graphs where it occurs.

We think it's quite likely that studying motifs will be important in understanding the circuits of artificial neural networks. In the long run, it may be more important than the study of individual circuits. At the same time, we expect investigations of motifs to be well served by us first building up a solid foundation of well understood circuits first.

Claim 3: Universality

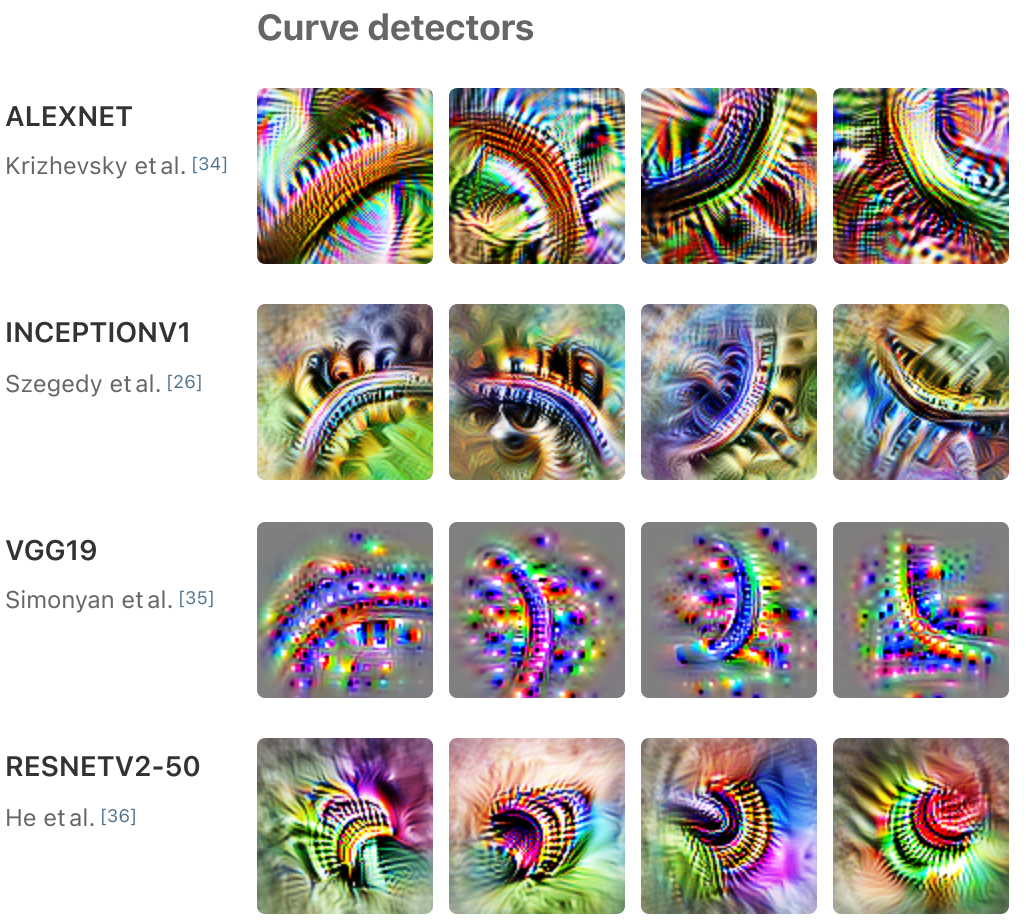
> Analogous features and circuits form across models and tasks.

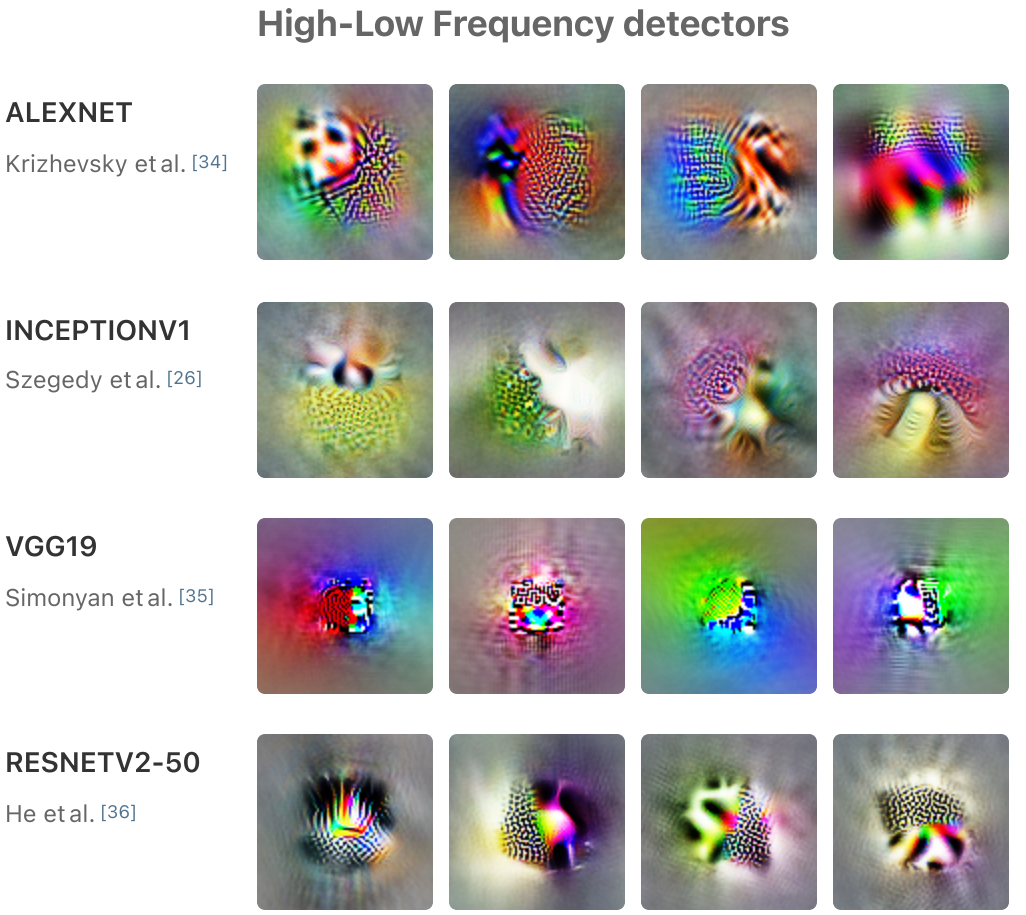
It's a widely accepted fact that the first layer of vision models trained on natural images will learn Gabor filters. Once you accept that there are meaningful features in later layers, would it really be surprising for the same features to also form in layers beyond the first one? And once you believe there are analogous features in multiple layers, wouldn't it be natural for them to connect in the same ways?

Universality (or "convergent learning") of features has been suggested before. Prior work has shown that different neural networks can develop highly correlated neurons [31] and that they learn similar representations at hidden layers [32, 33]. This work seems highly suggestive, but there are alternative explanations to analogous features forming. For example, one could imagine two features—such as a fur texture detector and a sophisticated dog body detector—being highly correlated despite being importantly different features. Adopting the meaningful feature-skeptic perspective, it doesn't seem definitive.

Ideally, one would like to characterize several features and then rigorously demonstrate that those features—and not just correlated ones—are forming across many models. Then, to further establish that analogous circuits form, one would want to find analogous features over several layers of multiple models and show that the same weight structure forms between them in each model.

Unfortunately, the only evidence we can offer today is anecdotal: we simply have not yet invested enough in the comparative study of features and circuits to give confident answers. With that said, we have observed that a couple low-level features seem to form across a variety of vision model architectures (including AlexNet, InceptionV1, InceptionV3, and residual networks) and in models trained on Places365 instead of ImageNet. We've also observed them repeatedly form in vanilla conv nets trained from scratch on ImageNet.





These results have led us to suspect that the universality hypothesis is likely true, but further work will be needed to understand if the apparent universality of some low-level vision features is the exception or the rule.

If it turns out that the universality hypothesis is broadly true in neural networks, it will be tempting to speculate: might biological neural networks also learn similar features? Researchers working at the intersection of neuroscience and deep learning have already shown that the units in artificial vision models can be useful for modeling biological neurons [37, 38, 39]. And some of the features we've discovered in artificial neural networks, such as curve detectors, are also believed to exist in biological neural networks (e.g. [40, 41]). This seems like significant cause for optimism. [8]

Focusing on the study of circuits, is universality really necessary? Unlike the first two claims, it wouldn't be completely fatal to circuits research if this claim turned out to be false. But it does greatly inform what kind of research makes sense. We introduced circuits as a kind of "cellular biology of deep learning." But imagine a world where every species had cells with a completely different set of organelles and proteins. Would it still make sense to study cells in general, or would we limit ourselves to the narrow study of a few kinds of particularly important species of cells? Similarly, imagine the study of anatomy in a world where every species of animal had a completely unrelated anatomy: would we seriously study anything other than humans and a couple domestic animals?

In the same way, the universality hypothesis determines what form of circuits research makes sense. If it was true in the strongest sense, one could imagine a kind of "periodic table of visual features" which we observe and catalogue across models. On the other hand, if it was mostly false, we would need to focus on a handful of models of particular societal importance and hope they stop changing every year. There might also be in between worlds, where some lessons transfer between models but others need to be learned from scratch.

Interpretability as a Natural Science

The Structure of Scientific Revolutions by Thomas Kuhn [42] is a classic text on the history and sociology of science. In it, Kuhn distinguishes between "normal science" in which a scientific community has a paradigm, and "extraordinary science" in which a community lacks a paradigm, either because it never had one or because it was weakened by crisis. It's worth noting that "extraordinary science" is not a desirable state: it's a period where researchers struggle to be productive.

Kuhn's description of pre-paradigmatic fields feel eerily reminiscent of interpretability today. [9] There isn't consensus on what the objects of study are, what methods we should use to answer them, or how to evaluate research results. To quote a recent interview with Ian Goodfellow: "For interpretability, I don't think we even have the right definitions." [43]

One particularly challenging aspect of being in a pre-paradigmatic field is that there isn't a shared sense of how to evaluate work in interpretability. There are two common proposals for dealing with this, drawing on the standards of adjacent fields. Some researchers, especially those with a deep learning background, want an "interpretability benchmark" which can evaluate how effective an interpretability method is. Other researchers with an HCI background may wish to evaluate interpretability methods through user studies.

But interpretability could also borrow from a third paradigm: natural science. In this view, neural networks are an object of empirical investigation, perhaps similar to an organism in biology. Such work would try to make empirical claims about a given network, which could be held to the standard of falsifiability.

Why don't we see more of this kind of evaluation of work in interpretability and visualization? [10] Especially given that there's so much adjacent ML work which does adopt this frame! One reason might be that it's very difficult to make robustly true statements about the behavior of a neural network as a whole. They're incredibly complicated objects. It's also hard to formalize what the interesting empirical statements about them would, exactly, be. And so we often get standards of evaluations more targeted at whether an interpretability method is useful rather than whether we're learning true statements.

Circuits sidestep these challenges by focusing on tiny subgraphs of a neural network for which rigorous empirical investigation is tractable. They're very much falsifiable: for example, if you understand a circuit, you should be able to predict what will change if you edit the weights. In fact, for small enough circuits, statements about their behavior become questions of mathematical reasoning. Of course, the cost of this rigor is that statements about circuits are much smaller in scope than overall model behavior. But it seems like, with sufficient effort, statements about model behavior could be broken down into statements about circuits. If so, perhaps circuits could act as a kind of epistemic foundation for interpretability.

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.

This article is part of the Circuits thread, a collection of short articles and commentary by an open scientific collaboration delving into the inner workings of neural networks.

← PREVIOUS ARTICLE
Circuits Thread

NEXT ARTICLE →
An Overview of Early Vision in InceptionV1

Glossary

This essay introduces some new terminology, and also uses some terminology which isn't common. To help, we provide the following glossary:

Circuit - A subgraph of a neural network. Nodes correspond to neurons or directions (linear combinations of neurons). Two nodes have an edge between them if they are in adjacent layers. The edges have weights which are the weights between those neurons (or $n_1 W n_2^T$ if the nodes are linear combinations). For convolutional layers, the weights are 2D matrices representing the weights for different relative positions of the layers.

Circuit Motif - A recurring, abstract pattern found in circuits, such as equivariance or unioning over cases. Inspired by the use of circuit motifs in systems biology [30].

Client Neuron or **Client Feature** - A neuron in a later layer which relies on a particular earlier neuron. For example, a circle detector is a client of curve detectors.

Direction - A linear combination of neurons in a layer. Equivalently, a vector in the representation of a layer. A direction can be an individual neuron (which is a basis direction of the vector space). For intuition about directions as an object, see Building Blocks [44] (in particular, the section titled "What Does the Network See?") and Activation Atlases [13].

Downstream / Upstream - In a later layer / In an earlier layer.

Equivariance - For equivariance in the context of circuits (eg. equivariant features, equivariant circuits), see the circuit article on equivarinace. For the more general idea of equivariance in mathematics, see the wikipedia equivariant map article.

Family - A set of features found in one layer which detect small variations of the same thing. For example, curve detectors exist in a family detecting curves in different orientations.

Feature - A scalar function of the input. In this essay, neural network features are directions, and often simply individual neurons. We claim such features in neural networks are typically meaningful features which can be rigorously studied (Claim 1).

Meaningful Feature - A feature that genuinely responds to an articulable property of the input, such as the presence of a curve or a floppy ear. Meaningful features may still be noisy or imperfect.

Polysemantic Feature - A feature that responds to multiple unrelated latent variables, such as the cat/car neuron [4]. This can be seen a special case of a "multifaceted features" [29] which responds to multiple different cases, but include both "real" multi-faceted features such as the pose-invariant dog head or polysemantic neurons. Contrast with pure.

Pure Feature - A feature which responds to only a single latent variable. Contrast with polysemantic.

Universal Feature - A feature which reliably forms across different models and tasks.

Representation - The vector space formed by the activations of all neurons in a layer, with vectors of the form *(activation of neuron 1, activation of neuron 2, ...)*. A representation can be thought of as the collection of all features that exist in a layer. For intuition about representations in vision models, see Activation Atlases [13].

Author Contributions

Writing: The text of this essay was primarily written by Christopher Olah, drawing extensively on the research and thinking of the entire Clarity team. Nick Cammarata was deeply involved in developing the framing and revising the final text.

Research: This essay articulates themes that developed as a result of several people's research into how neural networks implement features. Chris began initial attempts to understand the mechanistic implementations of neurons in terms of their weights in 2018, and developed several tools that enabled this line of work. This work was extended by Gabriel Goh, who discovered the first of what we now call motifs (using negative weights for specialization), in addition to describing the mechanisms behind several neurons. At this point, Nick Cammarata took up this line of research to characterize much larger and deeper circuits, greatly expand the number of neurons we understood mechanistically, and performed detailed, rigorous characterizations of curve detectors. Nick also introduced the connection to Systems biology. Ludwig Schubert performed detailed analysis of high-low frequency detectors. Chris gave research advice and mentorship throughout.

Infrastructure: Michael Petrov, Shan Carter, Ludwig and Nick built a variety of infrastructural tools which made our research possible.

Historical Note

The ideas in this introductory essay were previously presented as a keynote talk by Chris Olah at VISxAI 2019. It was also informally presented at MILA, the Vector Institute, the Redwood Center for Neuroscience, and a private workshop.

Acknowledgments

All our work understanding InceptionV1 is indebted to Alex Mordvintsev, whose early explorations of vision models paved pathways we still follow. We're deeply grateful to Nick Barry, and Sophia Sanborn for their deep engagement on potential connections between our work and neuroscience, and to Tom McGrath who pointed out the similarities between Kuhn's "pre-paradigmatic science" and the state interpretability as a field to us. The careful comments and criticism of Brice Menard were also invaluable in sharpening this essay.

In addition to Nick and Sophia's deep engagement, we're more generally appreciative of the neuroscience community's engagement with us, especially in sharing hard-won lessons about methodological weaknesses in our work. In particular, we appreciate Brian Wandell pushing us in 2019 on not using tuning curves and the importance of families of neurons, which we think has made our work much stronger. We're also very grateful for the comments and support of Mareike Grotheer, Natalia Bilenko, Bruno Olshausen, Michael Eickenberg, Charles Frye, Philip Sabes, Paul Merolla, James Redd, Thong-Wei Koh, and Ivan Alvarez. We think we have a lot to learn from the neuroscience community and are excited to continue doing so.

One of the privileges of working on circuits has been the open collaboration and feedback in the Distill slack's #circuits channel. We've especially appreciated the detailed feedback we received from Stefan Sietzen, Shahab Bakhtiari, and Flora Liu (Stefan has additionally run with many of these ideas, and we're excited to see his work in future articles in this thread!).

We benefitted greatly from the comments of many people on meta-science and framing questions around this essay, but especially appreciated the comments of Arvind Satyanarayan, Miles Brundage, Amanda Askell, Aaron Courville, and Martin Wattenberg. We're grateful to Taco Cohen, Tess Smidt, and Sara Sabour for their extremely helpful comments on equivariance. We're grateful to Nikita Obidin, Nick Barry, and Chelsea Voss for helpful conversation and references about systems biology and circuit motifs. (Nikita and Nick initially introduced Nick Cammarata to circuit motifs.) Finally, we're grateful for the institutional support of OpenAI, and for the support and comments of all our colleagues and friends across institutions, including Dario Amodei, Daniela Amodei, Jonathan Uesato, Laura Ball, Katarina Slama, Alethea Power, Jacob Hilton, Jacob Steinhardt, Tom Brown, Preetum Nakkiran, Ilya Sutskever, Ryan Lowe, Erin McCloskey, Eli Chen, Fred Hohman, Jason Yosinski, Pallavi Koppol, Reiichrio Nakano, Sam McCandlish, Daniel Dewey, Anna Goldie, Jochen Görtler, Hendrik Strobelt, Ravi Chunduru, Tom White, Roger Grosse, David Duvenaud, Daniel Burkhardt, Janelle Tam, Jeff Clune, Christian Szegedy, Alec Radford, Alex Ray, Evan Hubinger, Scott Gray, Augustus Odena, Mikhial Pavlov, Daniel Filan, Jascha Sohl-Dickstein and Kris Sankaran.

Footnotes

1. By "direction" we mean a linear combination of neurons in a layer. You can think of this as a direction vector in the vector space of activations of neurons in a given layer. Often, we find it most helpful to talk about individual neurons, but we'll see that there are some cases where other combinations are a more useful way to analyze networks — especially when neurons are "polysemantic." (See the glossary for a detailed definition.) [↩]

2. A "circuit" is a computational subgraph of a neural network. It consists of a set of features, and the weighted edges that go between them in the original network. Often, we study quite small circuits — say with less than a dozen features — but they can also be much larger. (See the glossary for a detailed definition.) [↩]

3. The community disagreement on meaningful features is hard to pin down, and only partially expressed in the literature. Foundational descriptions of deep learning often describe neural networks as detecting a hierarchy of meaningful features [25], and a number of papers have been written demonstrating seemingly meaningful features in different domains domains [15, 2, 16, 17, 4, 18]. At the same time, a more skeptical parallel literature has developed suggesting that neural networks primarily or only focus on texture, local structure, or imperceptible patterns [20, 21, 22, 23], that meaningful features, when they exist, are less important than uninterpretable ones [24] and that seemingly interpretable neurons may be misunderstood [19]. Although many of these papers express a highly nuanced view, that isn't always how they've been understood. A number of media articles have been written embracing strong versions of these views, and we anecdotally find that the belief that neural networks don't understand anything more than texture is quite common. Finally, people often have trouble articulating their exact views, because they don't have clear language for articulating nuances between "a texture detector highly correlated with an object" and "an object detector." [↩]

4. Why are polysemantic neurons so challenging? If one neuron with five different meanings connects to another neuron with five different meanings, that's effectively 25 connections that can't be considered individually. [↩]

5. Many of the neurons discussed in this article, including curve detectors, live in branches of InceptionV1 that are structured as a 1x1 convolution that reduce the number of channels to a small bottleneck followed by a 3x3 or 5x5 convolution. The weights we present in this essay are the multiplied out version of the 1x1 and larger conv weights. We think it's often useful to view this as a single low-rank weight matrix, but this technically does ignore one ReLU non-linearity. [↩]

6. To be clear, there are also more direct pathways by which various constituents of heads influence these later head detectors, without going through the left and right pathways [↩]

7. Fundamentally, this is a property of the geometry of high-dimensional spaces, which only allow for n orthogonal vectors, but exponentially many almost orthogonal vectors. [↩]

8. One particularly exciting possibility might be if artificial neural networks could predict features which were previously unknown but could then be found in biology. (Some neuroscientists we have spoken to have suggested that high-low

frequency detectors might be a candidate for this.) If such a prediction could be made, it would be extremely strong evidence for the universality hypothesis. [↩]

9. We were introduced to Kuhn's work and this connection by conversations with Tom McGrath at DeepMind [↩]

10. To be clear, we do see researchers who take more of this natural science approach, especially in earlier interpretability research. It just seems less common right now. [↩]

References

1. Micrographia: or Some Physiological Descriptions of Minute Bodies Made by Magnifying Glasses. With Observations and Inquiries Thereupon [link]
   Hooke, R., 1666. The Royal Society. DOI: 10.5962/bhl.title.904

2. Visualizing and understanding recurrent networks [PDF]
   Karpathy, A., Johnson, J. and Fei-Fei, L., 2015. arXiv preprint arXiv:1506.02078.

3. Visualizing higher-layer features of a deep network [PDF]
   Erhan, D., Bengio, Y., Courville, A. and Vincent, P., 2009. University of Montreal, Vol 1341, pp. 3.

4. Feature Visualization [link]
   Olah, C., Mordvintsev, A. and Schubert, L., 2017. Distill. DOI: 10.23915/distill.00007

5. Deep inside convolutional networks: Visualising image classification models and saliency maps [PDF]
   Simonyan, K., Vedaldi, A. and Zisserman, A., 2013. arXiv preprint arXiv:1312.6034.

6. Deep neural networks are easily fooled: High confidence predictions for unrecognizable images [PDF]
   Nguyen, A., Yosinski, J. and Clune, J., 2015. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 427--436. DOI: 10.1109/cvpr.2015.7298640

7. Inceptionism: Going deeper into neural networks [HTML]
   Mordvintsev, A., Olah, C. and Tyka, M., 2015. Google Research Blog.

8. Plug & play generative networks: Conditional iterative generation of images in latent space [PDF]
   Nguyen, A., Clune, J., Bengio, Y., Dosovitskiy, A. and Yosinski, J., 2016. arXiv preprint arXiv:1612.00005.

9. Visualizing and understanding convolutional networks [PDF]
   Zeiler, M.D. and Fergus, R., 2014. European conference on computer vision, pp. 818--833.

10. Interpretable Explanations of Black Boxes by Meaningful Perturbation [PDF]
    Fong, R. and Vedaldi, A., 2017. arXiv preprint arXiv:1704.03296.

11. PatternNet and PatternLRP--Improving the interpretability of neural networks [PDF]
    Kindermans, P., Schutt, K.T., Alber, M., Muller, K. and Dahne, S., 2017. arXiv preprint arXiv:1705.05598. DOI: 10.1007/978-3-319-10590-1_53

12. Visualizing and Measuring the Geometry of BERT [PDF]
    Reif, E., Yuan, A., Wattenberg, M., Viegas, F.B., Coenen, A., Pearce, A. and Kim, B., 2019. Advances in Neural Information Processing Systems, pp. 8592--8600.

13. Activation atlas [link]
    Carter, S., Armstrong, Z., Schubert, L., Johnson, I. and Olah, C., 2019. Distill, Vol 4(3), pp. e15. DOI: 10.23915/distill.00015

14. Summit: Scaling Deep Learning Interpretability by Visualizing Activation and Attribution Summarizations [PDF]
    Hohman, F., Park, H., Robinson, C. and Chau, D.H.P., 2019. IEEE Transactions on Visualization and Computer Graphics, Vol 26(1), pp. 1096--1106. IEEE.

15. Distributed representations of words and phrases and their compositionality [PDF]
Mikolov, T., Sutskever, I., Chen, K., Corrado, G.S. and Dean, J., 2013. Advances in neural information processing systems, pp. 3111--3119.

16. Learning to generate reviews and discovering sentiment [PDF]
Radford, A., Jozefowicz, R. and Sutskever, I., 2017. arXiv preprint arXiv:1704.01444.

17. Object detectors emerge in deep scene cnns [PDF]
Zhou, B., Khosla, A., Lapedriza, A., Oliva, A. and Torralba, A., 2014. arXiv preprint arXiv:1412.6856.

18. Network Dissection: Quantifying Interpretability of Deep Visual Representations [PDF]
Bau, D., Zhou, B., Khosla, A., Oliva, A. and Torralba, A., 2017. Computer Vision and Pattern Recognition.

19. On Interpretability and Feature Representations: An Analysis of the Sentiment Neuron
Donnelly, J. and Roegiest, A., 2019. European Conference on Information Retrieval, pp. 795--802.

20. Measuring the tendency of CNNs to Learn Surface Statistical Regularities [PDF]
Jo, J. and Bengio, Y., 2017. arXiv preprint arXiv:1711.11561.

21. ImageNet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness [PDF]
Geirhos, R., Rubisch, P., Michaelis, C., Bethge, M., Wichmann, F.A. and Brendel, W., 2018. arXiv preprint arXiv:1811.12231.

22. Approximating cnns with bag-of-local-features models works surprisingly well on imagenet [PDF]
Brendel, W. and Bethge, M., 2019. arXiv preprint arXiv:1904.00760.

23. Adversarial examples are not bugs, they are features [PDF]
Ilyas, A., Santurkar, S., Tsipras, D., Engstrom, L., Tran, B. and Madry, A., 2019. Advances in Neural Information Processing Systems, pp. 125--136.

24. On the importance of single directions for generalization [PDF]
Morcos, A.S., Barrett, D.G., Rabinowitz, N.C. and Botvinick, M., 2018. arXiv preprint arXiv:1803.06959.

25. Deep learning [PDF]
LeCun, Y., Bengio, Y. and Hinton, G., 2015. nature, Vol 521(7553), pp. 436--444. Nature Publishing Group.

26. Going deeper with convolutions [PDF]
Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., Rabinovich, A. and others,, 2015. DOI: 10.1109/cvpr.2015.7298594

27. Receptive fields, binocular interaction and functional architecture in the cat's visual cortex
Hubel, D.H. and Wiesel, T.N., 1962. The Journal of physiology, Vol 160(1), pp. 106--154. Wiley Online Library.

28. Using Artificial Intelligence to Augment Human Intelligence [link]
Carter, S. and Nielsen, M., 2017. Distill. DOI: 10.23915/distill.00009

29. Multifaceted feature visualization: Uncovering the different types of features learned by each neuron in deep neural networks [PDF]
Nguyen, A., Yosinski, J. and Clune, J., 2016. arXiv preprint arXiv:1602.03616.

30. An introduction to systems biology: design principles of biological circuits
Alon, U., 2019. CRC press. DOI: 10.1201/9781420011432

31. Convergent learning: Do different neural networks learn the same representations? [PDF]
Li, Y., Yosinski, J., Clune, J., Lipson, H. and Hopcroft, J.E., 2015. FE@ NIPS, pp. 196--212.

32. SVCCA: Singular Vector Canonical Correlation Analysis for Deep Learning Dynamics and Interpretability [PDF]
Raghu, M., Gilmer, J., Yosinski, J. and Sohl-Dickstein, J., 2017. Advances in Neural Information Processing Systems 30, pp. 6078--6087. Curran Associates, Inc.

33. Similarity of neural network representations revisited [PDF]
Kornblith, S., Norouzi, M., Lee, H. and Hinton, G., 2019. arXiv preprint arXiv:1905.00414.

34. ImageNet Classification with Deep Convolutional Neural Networks [PDF]
Krizhevsky, A., Sutskever, I. and Hinton, G.E., 2012. Advances in Neural Information Processing Systems 25, pp. 1097--1105. Curran Associates, Inc.

35. Very Deep Convolutional Networks for Large-Scale Image Recognition [PDF]
Simonyan, K. and Zisserman, A., 2014. CoRR, Vol abs/1409.1556.

36. Deep Residual Learning for Image Recognition [PDF]
He, K., Zhang, X., Ren, S. and Sun, J., 2015. CoRR, Vol abs/1512.03385.

37. Performance-optimized hierarchical models predict neural responses in higher visual cortex
Yamins, D.L., Hong, H., Cadieu, C.F., Solomon, E.A., Seibert, D. and DiCarlo, J.J., 2014. Proceedings of the National Academy of Sciences, Vol 111(23), pp. 8619--8624. National Acad Sciences.

38. Deep neural networks reveal a gradient in the complexity of neural representations across the ventral stream
Gu{\c{c}}lu, U. and van Gerven, M.A., 2015. Journal of Neuroscience, Vol 35(27), pp. 10005--10014. Soc Neuroscience.

39. Seeing it all: Convolutional network layers map the function of the human visual system
Eickenberg, M., Gramfort, A., Varoquaux, G. and Thirion, B., 2017. NeuroImage, Vol 152, pp. 184--194. Elsevier.

40. Discrete neural clusters encode orientation, curvature and corners in macaque V4 [link]
Jiang, R., Li, M. and Tang, S., 2019. bioRxiv. Cold Spring Harbor Laboratory. DOI: 10.1101/808907

41. Shape representation in area V4: position-specific tuning for boundary conformation
Pasupathy, A. and Connor, C.E., 2001. Journal of neurophysiology, Vol 86(5), pp. 2505--2519. American Physiological Society Bethesda, MD.

42. The structure of scientific revolutions
Kuhn, T.S., 1962. University of Chicago press. DOI: 10.7208/chicago/9780226458106.001.0001

43. Ian Goodfellow: Generative Adversarial Networks [link]
Goodfellow, I. and Fridman, L., 2019. Artificial Intelligence Podcast.

44. The Building Blocks of Interpretability [link]
Olah, C., Satyanarayan, A., Johnson, I., Carter, S., Schubert, L., Ye, K. and Mordvintsev, A., 2018. Distill. DOI: 10.23915/distill.00010

Updates and Corrections

If you see mistakes or want to suggest changes, please create an issue on GitHub.

Reuse

Citation

For attribution in academic contexts, please cite this work as

```
Olah, et al., "Zoom In: An Introduction to Circuits", Distill, 2020.
```

BibTeX citation

```
@article{olah2020zoom,
  author = {Olah, Chris and Cammarata, Nick and Schubert, Ludwig and Goh, Gabriel and Petrov,
Michael and Carter, Shan},
  title = {Zoom In: An Introduction to Circuits},
  journal = {Distill},
  year = {2020},
  note = {https://distill.pub/2020/circuits/zoom-in},
  doi = {10.23915/distill.00024.001}
}
```