Thread: *Circuits*

What can we learn if we invest heavily in reverse engineering a single neural network?

PUBLISHED
March 10, 2020

DOI
10.23915/distill.00024

In the original narrative of deep learning, each neuron builds progressively more abstract, meaningful features by composing features in the preceding layer. In recent years, there's been some skepticism of this view, but what happens if you take it really seriously?

InceptionV1 is a classic vision model with around 10,000 unique neurons — a large number, but still on a scale that a group effort could attack. What if you simply go through the model, neuron by neuron, trying to understand each one and the connections between them? The circuits collaboration aims to find out.

Articles & Comments

The natural unit of publication for investigating circuits seems to be short papers on individual circuits or small families of features. Compared to normal machine learning papers, this is a small and unusual topic for a paper.

To facilitate exploration of this direction, Distill is inviting a "thread" of short articles on circuits, interspersed with critical commentary by experts in adjacent fields. The thread will be a living document, with new articles added over time, organized through an open slack channel (#circuits in the Distill slack). Content in this thread should be seen as early stage exploratory research.

Articles and comments are presented below in chronological order:

Zoom In: An Introduction to Circuits

AUTHORS
Chris Olah, Nick Cammarata, Ludwig Schubert, Gabriel Goh, Michael Petrov, Shan Carter

AFFILIATIONS
OpenAI

Does it make sense to treat individual neurons and the connections between them as a serious object of study? This essay proposes three claims which, if true, might justify serious inquiry into them: the existence of meaningful features, the existence of meaningful circuits between features, and the universality of those features and circuits.

It also discuses historical successes of science "zooming in," whether we should be concerned about this research being qualitative, and approaches to rigorous investigation.

→ READ FULL ARTICLE

An Overview of Early Vision in InceptionV1

AUTHORS

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

AFFILIATIONS

OpenAI

An overview of all the neurons in the first five layers of InceptionV1, organized into a taxonomy of "neuron groups." This article sets the stage for future deep dives into particular aspects of early vision.

→ READ FULL ARTICLE

Curve Detectors

AUTHORS

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

AFFILIATIONS

OpenAI

Every vision model we've explored in detail contains neurons which detect curves. Curve detectors is the first in a series of three articles exploring this neuron family in detail.

→ READ FULL ARTICLE

Naturally Occurring Equivariance in Neural Networks

AUTHORS

Chris Olah, Nick Cammarata, Chelsea Voss, Ludwig Schubert, Gabriel Goh

AFFILIATIONS

OpenAI

Neural networks naturally learn many transformed copies of the same feature, connected by symmetric weights.

→ READ FULL ARTICLE

High-Low Frequency Detectors

AUTHORS

Ludwig Schubert, Chelsea Voss, Nick Cammarata, Gabriel Goh, Chris Olah

AFFILIATIONS

OpenAI

A family of early-vision neurons reacting to directional transitions from high to low spatial frequency.

→ READ FULL ARTICLE

Curve Circuits

AUTHORS

Nick Cammarata, Gabriel Goh, Shan Carter, Chelsea Voss, Ludwig Schubert, Chris Olah

AFFILIATIONS

OpenAI

We reverse engineer a non-trivial learned algorithm from the weights of a neural network and use its core ideas to craft an artificial artificial neural network from scratch that reimplements it.

→ READ FULL ARTICLE

Visualizing Weights

AUTHORS
Chelsea Voss, Nick Cammarata, Gabriel Goh, Michael Petrov, Ludwig Schubert, Ben Egan, Swee Kiat Lim, Chris Olah

AFFILIATIONS
OpenAI, Mount Royal University, Stanford University

We present techniques for visualizing, contextualizing, and understanding neural network weights.

→ READ FULL ARTICLE

Branch Specialization

AUTHORS
Chelsea Voss, Gabriel Goh, Nick Cammarata, Michael Petrov, Ludwig Schubert, Chris Olah

AFFILIATIONS
OpenAI

When a neural network layer is divided into multiple branches, neurons self-organize into coherent groupings.

→ READ FULL ARTICLE

Weight Banding

AUTHORS
Michael Petrov, Chelsea Voss, Ludwig Schubert, Nick Cammarata, Gabriel Goh, Chris Olah

AFFILIATIONS
OpenAI

Weights in the final layer of common visual models appear as horizontal bands. We investigate how and why.

→ READ FULL ARTICLE

THIS IS A LIVING DOCUMENT

Expect more articles on this topic, along with critical comments from experts.

Get Involved

The Circuits thread is open to articles exploring individual features, circuits, and their organization within neural networks. Critical commentary and discussion of existing articles is also welcome. The thread is organized through the open `#circuits` channel on the Distill slack. Articles can be suggested there, and will be included at the discretion of previous authors in the thread, or in the case of disagreement by an uninvolved editor.

If you would like get involved but don't know where to start, small projects may be available if you ask in the channel.

About the Thread Format

Part of Distill's mandate is to experiment with new forms of scientific publishing. We believe that that reconciling faster and more continuous approaches to publication with review and discussion is an important open problem in scientific publishing.

Threads are collections of short articles, experiments, and critical commentary around a narrow or unusual research topic, along with a slack channel for real time discussion and collaboration. They are intended to be earlier stage than a full Distill paper, and allow for more fluid publishing, feedback and discussion. We also hope they'll allow for wider participation. Think of a cross between a Twitter thread, an academic workshop, and a book of collected essays.

Threads are very much an experiment. We think it's possible they're a great format, and also possible they're terrible. We plan to trial two such threads and then re-evaluate our thought on the format.

Citation Information

If you wish to cite this thread as a whole, citation information can be found below. The author order is all participants in the thread in alphabetical order. Since this is a living document, the citation may add additional authors as it evolves. You can also cite individual articles using the citation information provided at the bottom of the corresponding article.

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

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

BibTeX citation

```
@article{cammarata2020thread:,
  author = {Cammarata, Nick and Carter, Shan and Goh, Gabriel and Olah, Chris and Petrov, Michael and
Schubert, Ludwig and Voss, Chelsea and Egan, Ben and Lim, Swee Kiat},
  title = {Thread: Circuits},
  journal = {Distill},
  year = {2020},
  note = {https://distill.pub/2020/circuits},
  doi = {10.23915/distill.00024}
}
```