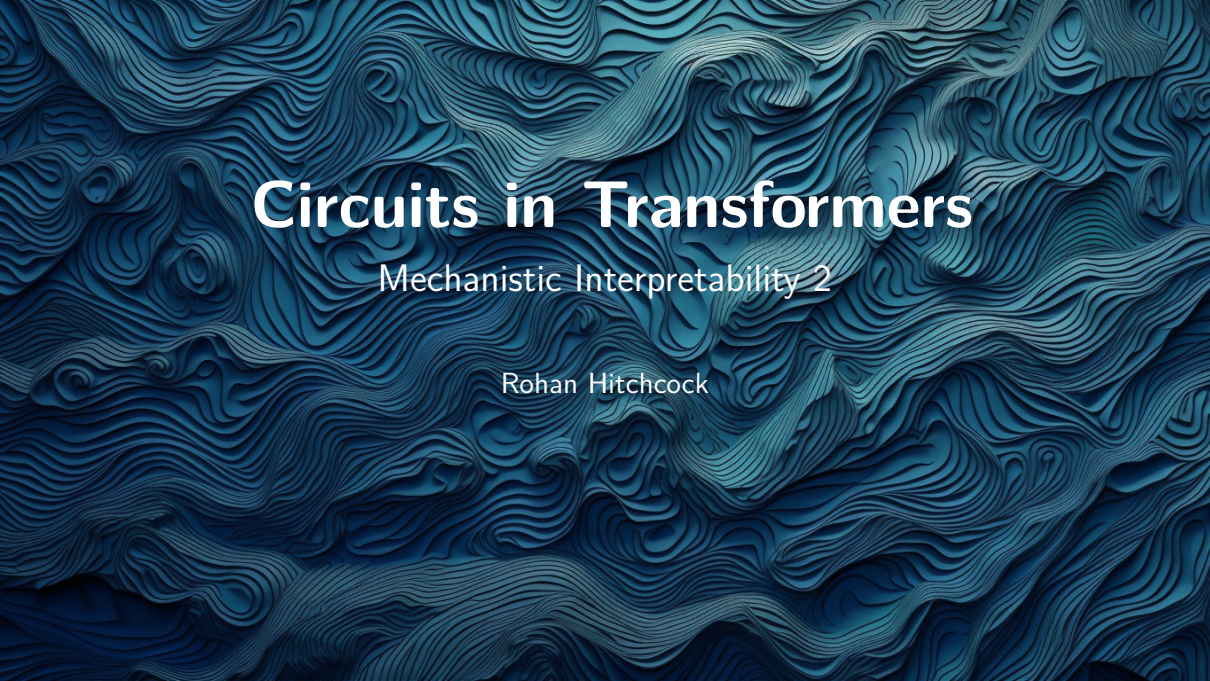

Circuits in Transformers
Mechanistic Interpretability 2
Rohan Hitchcock

Circuits and algorithms

A ***circuit*** is a part of a neural network which performs a certain task or algorithm.

- We define circuits by what they *do*, rather than how they are implemented.
- Circuits in transformers performing modular addition have been fully reverse engineered.
- Circuits in convolutional neural networks (e.g. for edge detection) have been studied (the term "circuits" was not used).

Much of mechanistic interpretability is dedicated to identifying the presence of certain circuits, and understanding how they work.

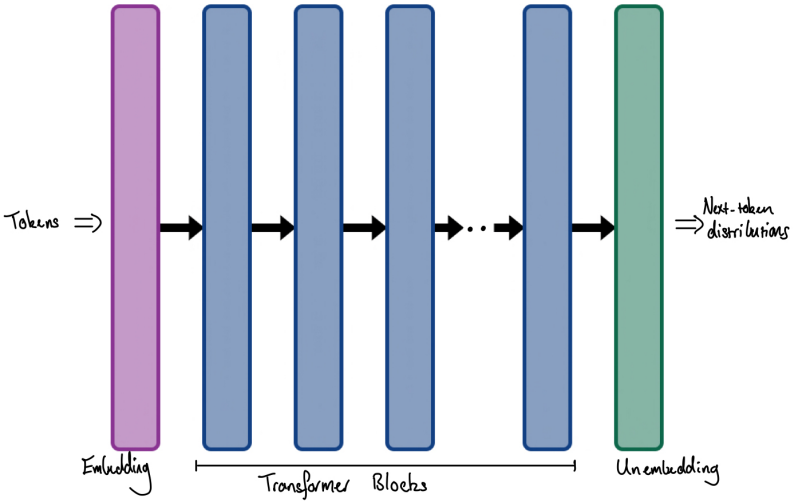
Overview of talk

Goal: Give an overview of how circuits in transformers are studied, using paper "In-context Learning and Induction Heads" Olsson et. al. (2022) as a case study.

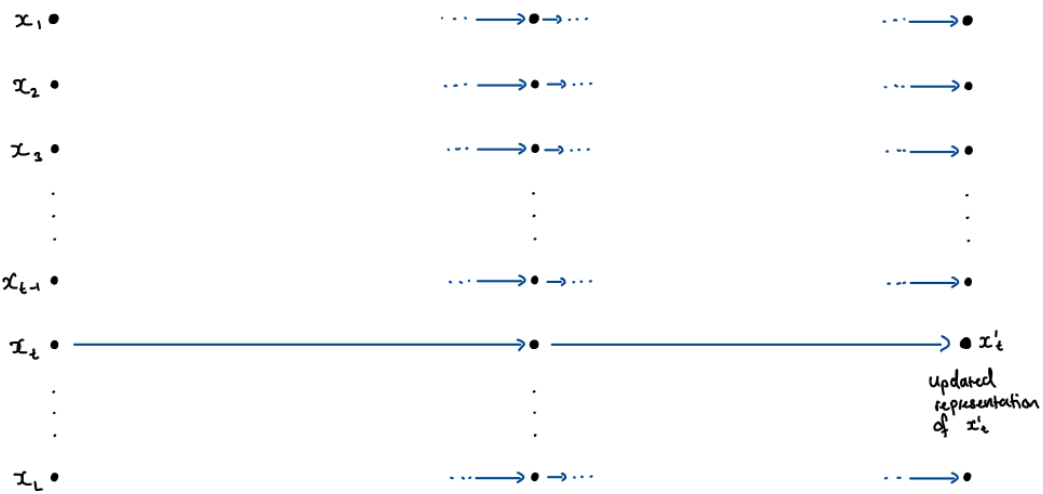
Outline

- Brief review of transformer architecture.
- In-context learning
- Discussion of the paper "In-context Learning and Induction Heads" Olsson et. al. (2022)
- A framework for studying circuits

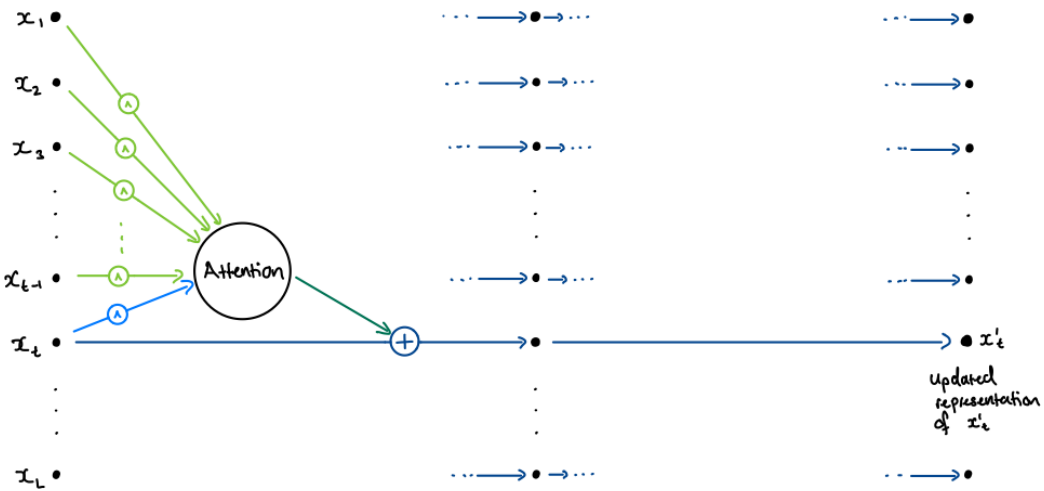
Transformers, briefly



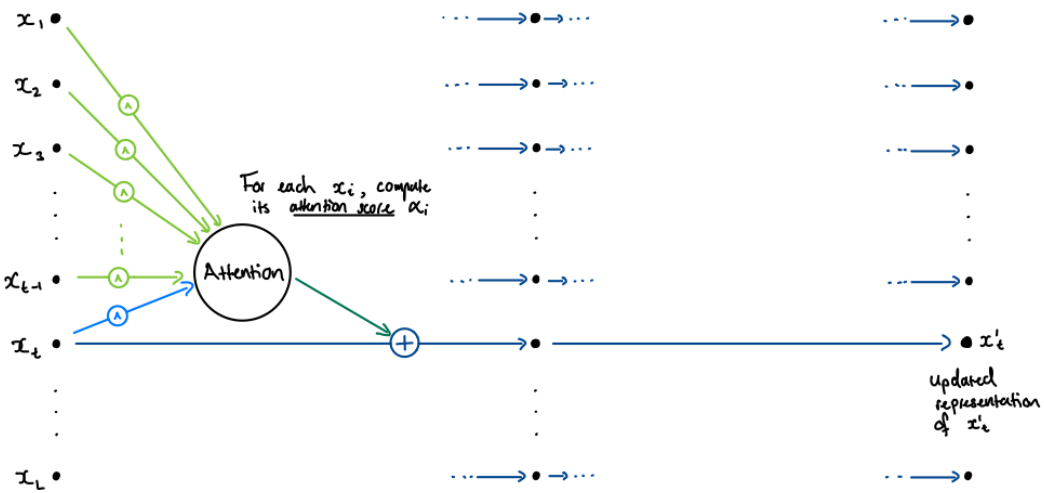
Transformers: a transformer block



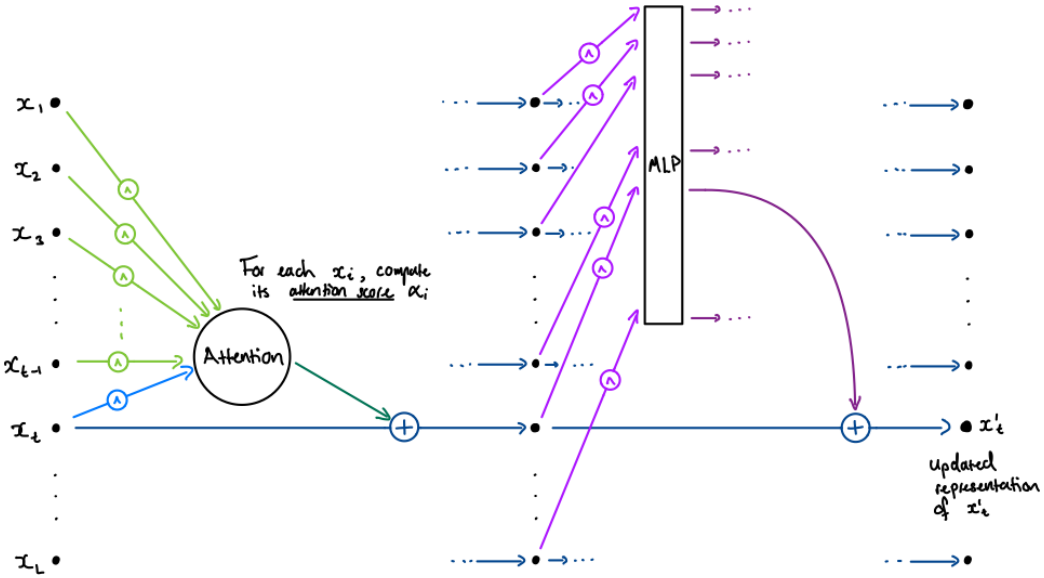
Transformers: a transformer block



Transformers: a transformer block



Transformers: a transformer block



In-context learning

Definition

In-context learning is the phenomenon where transformer language models get better at predicting tokens given a longer prompt (aka context). In other words, they learn from context.

- Identified in GPT2/GPT3 where it was called *few-shot learning*
- If a model is doing in-context learning then the loss at each token decreases with increasing token index.

"In-context Learning and Induction Heads", Olsson et al. (2022)

Summary of paper:

Presents evidence that a type of circuit called an ***induction head*** is a mechanism for in-context learning in transformer models:

1. During training formation of induction heads appears to coincide with the model gaining the ability to do in-context learning.
2. Adjusting the model to promote induction heads leads to in-context learning.
3. Disrupting induction heads during test degrades in-context learning.
4. Induction heads can do more complex things in some contexts.
5. Can reverse-engineer induction heads in small models.

I didn't work on this paper and I'm not affiliated with the authors in any way.

Induction heads

Definition

An ***induction head*** is an attention head which exhibits a type of completion algorithm behaviour: for any tokens A and B and sequence of the form $\ldots AB \ldots A$ an induction head influences the next token prediction to be B.

Measuring induction heads

Definition

Given an attention head, its ***prefix matching score*** is computed as follows:

1. Generate a sequence of 25 random tokens and repeat it 4 times.
2. Average the attention assigned to the token following the current token in earlier repeats:

Measuring in-context learning

Definition

The ***in-context learning score*** (ILS) of a model is the average loss at the 500^{th} token minus the average loss of the 50^{th} token, over 512-token test sequences.

- A model with a *more negative* ILS is *better* at in-context learning.

Model details

In this talk we focus on results from two classes of models:

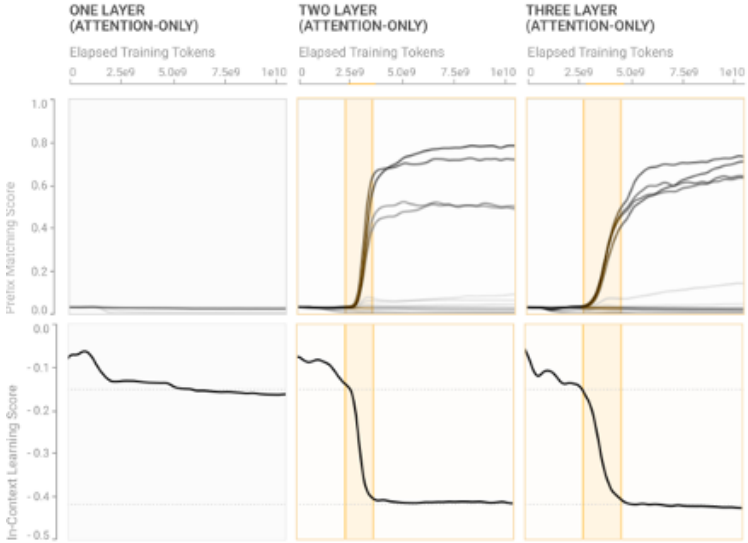
- **Small models:** Attention-only (no MLP) transformers with 1, 2 and 3 transformer blocks.
- **Large models:** "Full" transformers with 13M to 13B parameters (4 to 40 blocks).

Note: we don't expect induction heads to form in 1 layer models.

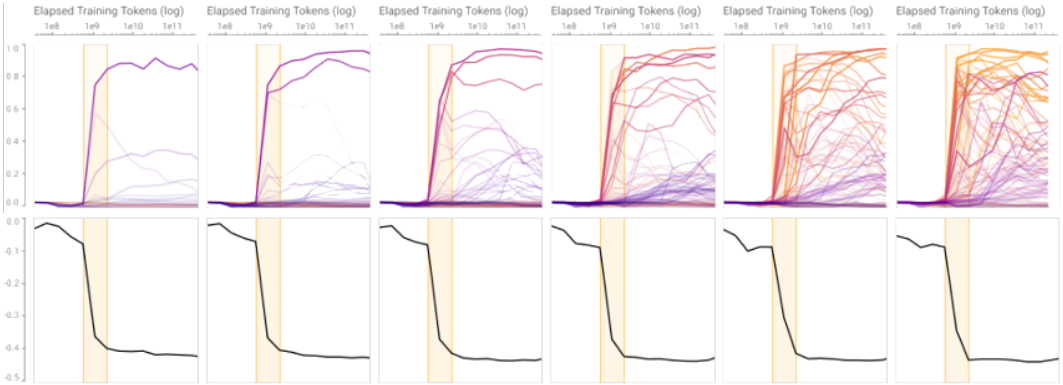
Results: small models

PMS: (induction heads)

ILS: (in-context learning)
More negative = more in-context learning



Results: large models



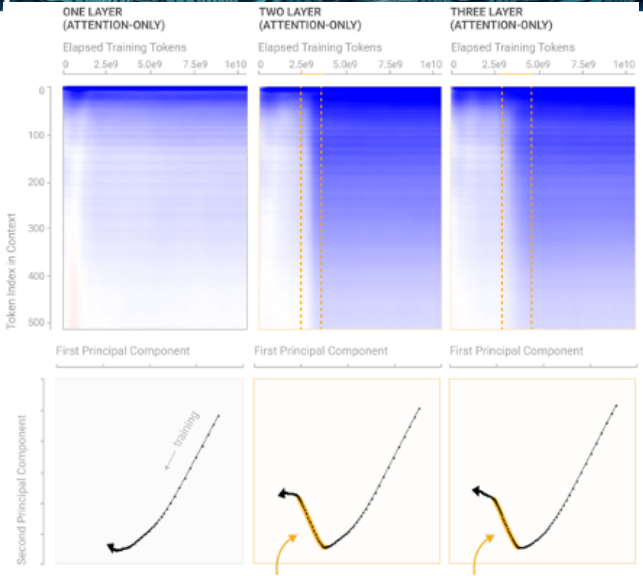
PMS (top), **ILS** (bottom)
Models from 13M parameters (left) to 13B parameters (right).

Figs from Olsson et al., "In-context Learning and Induction Heads", Transfomer Circuits Thread 2022

“Phase transtion” shows up in other measurements too

Change in loss per log-token-index

Principal components of *per-token loss



Figs from Olsson et al., “In-context Learning and Induction Heads”, Transfomer Circuits Thread 2022

Induction heads can do other things

Attention heads which score highly as induction heads are observed doing other things:

- Translation
- A synthetic classification task

Some attention heads appear to implement a more general induction algorithm.

Other methods

The paper did several other things:

- Changing architecture to make it easier to form induction head circuits
- Disrupting the induction heads at test time
- Reverse engineer how induction heads work in smaller models

A framework for studying circuits

In summary:

- Identify an interesting property or phenomenon in neural networks (in-context learning).
- Find a way to measure it (in-context learning score)
- Identify an algorithm associated to your property (copying)
- Find a way to measure whether that algorithm is occurring (prefix-matching score)

How does this connect with SLT?

- Does circuit formation correspond with genuine phase transitions?
- Do phase transitions always arise due to circuit formation?
- Can SLT give us more principled ways of measuring circuit formation?

References and further reading

- C. Olsson et al., "In-context Learning and Induction Heads," Transformer Circuits Thread, 2022.
- Modular addition circuits: N. Nanda, L. Chan, T. Lieberum, J. Smith, and J. Steinhardt, "Progress measures for grokking via mechanistic interpretability." `https://arxiv.org/abs/2301.05217`, 2023
- More on transformers: M. Phuong and M. Hutter, "Formal Algorithms for Transformers," `http://arxiv.org/abs/2207.09238`, 2022