Observing and Validating Induction heads in SOLU-8l-old

Brian Muhia

Interpretability Hackathon Report

Apart Research

PIs: Esben Kran, Neel Nanda, Fazl Barez

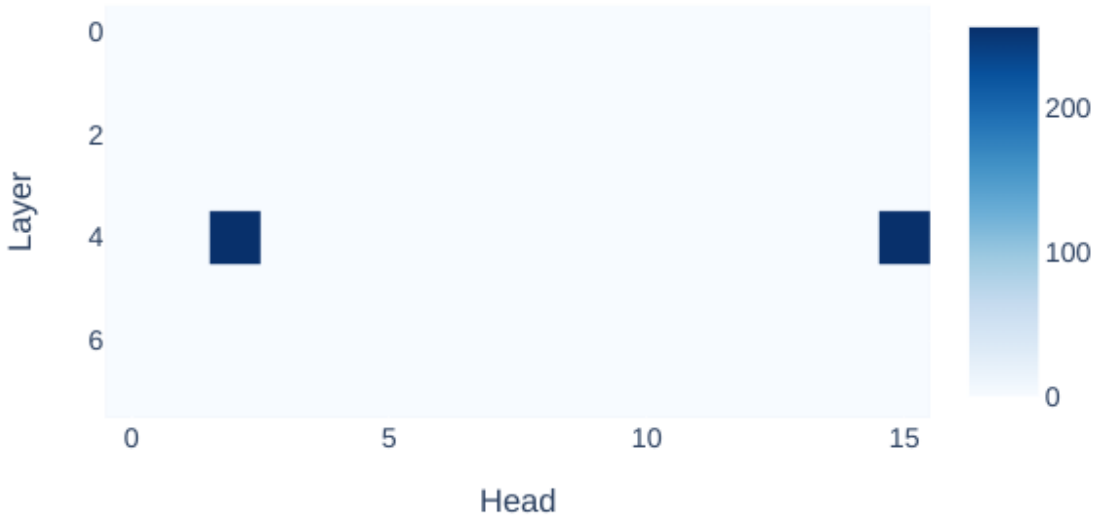
Date: 13th November, 2022

Abstract

We look for induction heads by feeding in a random sequence of tokens repeated twice and looking for heads that attend from a second copy of a token to the token just after the first copy. This is to test the generality of methods for detecting induction heads by looking at attention scores on a linear scale, which leaves open questions about whether the method itself is reproducible. We make observations about the mean attention score used to determine if an attention head is an induction head in **SoLU-8l-old** compared to **GPT2-small.** This is a reproducible project, linked here https://github.com/poppingtonic/transformer-visualization/blob/main/SOLU-8l-old-Observations.ipynb

Observing and Validating Induction heads in SOLU-8l-old

The SoLU (Softmax Linear Unit) transformer is an architecture designed for mechanistic interpretability. In this report, we search for induction heads and compute attention patterns for a random sequence of tokens that is repeated. We then find the heads with the highest values, ablate them and observe the effect on the same random repeated sequence. Compared to GPT-2 small, investigated here, we observe that with a max attention score of 0.6 we obtain 2 attention heads that when pruned give a drastic increase in loss from -6.09 to -8.3. For a max attention score of 0.5 one more attention head is obtained, increasing the loss a bit more to 9.74. This leaves us with questions on these thresholds. Is a sweep necessary to understand how this value changes with model scale? A second observation is on which attention heads mostly attend to the previous token. We run a simple test and identify the layer 3 head 8. An investigation on the effects of pruning this head is left for future work, as is an investigation into the value of the induction score changes with model scale. A more general, or less confusing, method for detecting induction heads is needed.



A threshold of 0.5 gives us 3 heads and drops performance in the following way:

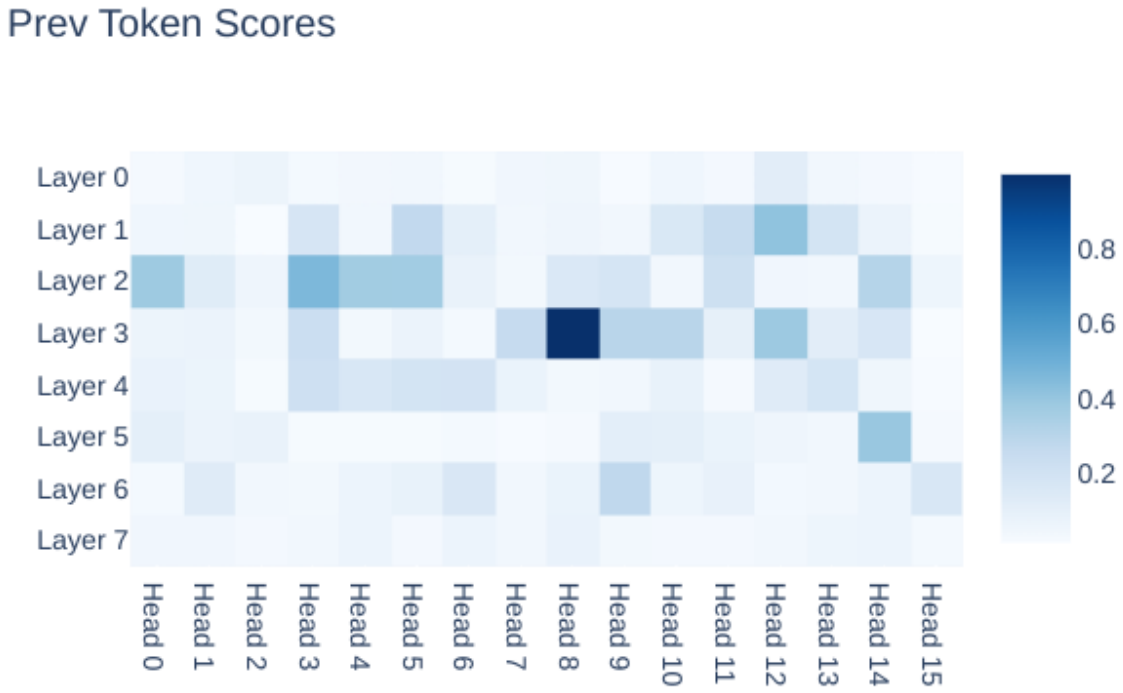
> Original loss on repeated sequence: -6.096261024475098

> Loss on repeated sequence without induction heads: -9.745206832885742

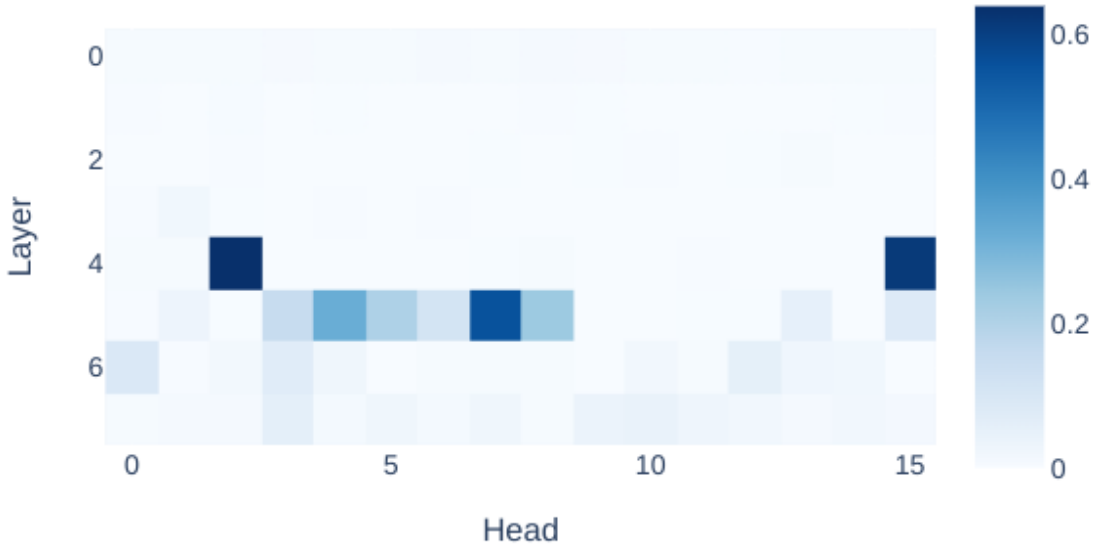
A threshold of 0.6 gives us two heads, with the following loss:

> Original loss on repeated sequence: -6.096261024475098

> Loss on repeated sequence without induction heads: -8.345271110534668



We identify Head 8, Layer 3 as mostly responsible for identifying the previous token. Future work: Find the effect of pruning that head



It appears that the max score for the random tokens is around 0.6. Above that threshold there appear to be two attention heads `L4H2` with score 0.6402724, and `L4H15` with score 0.6168649. A third attention head is slightly less activated, `L5H7`, with score 0.5587158. Is this an induction head, or some other kind of attention head?

References

Elhage et. al. (2022). Softmax Linear Units *https://transformer-circuits.pub/2022/solu/index.html*. Anthropic.

Nanda (2022). EasyTransformer https://github.com/neelnanda-io/Easy-Transformer