Circuits can help explain transformer language models’ linguistic abilities

Learning to Agree: How Language Models Implement Subject-Verb Agreement

Michael Hanna
m.w.hanna@uva.nl
University of Amsterdam
Institute for Logic, Language, and Computation
Amsterdam, Netherlands

Language Models

NLP relies on pre-trained language models (LMs), neural models that predict the next word given a context. LMs possess linguistic abilities, like subject-verb agreement (SVA):

The dog that

p(barks) = 0.1
p(run) = 0.05
p(bark) = 0.00...

Interpretability and Circuits

We want an explanation of SVA that is:

• low-level: at the attention head/MLP level
• causal: we can prove it works
• comprehensive: from inputs to outputs

We thus search for a circuit: a minimal computational subgraph of our LM that suffices to perform SVA. How to find one? To start, we visualize the comp. graph of a toy LM.

We then ablate edges, replacing one activation (MLP1->MHA2) with another input’s.

Attention head 6.4 clearly transmits number information to MLPs 8-10 and the logits. We apply the logit lens [4] to head 6.4, and find it boosts words that agree with the subject:

• are
• were
• sont
• hebben

These words agree with the example’s plural subject across languages; sont and hebben are plural-form verbs in French and Dutch.

Key Takeaways

• Circuits provide low-level explanations of model behavior at the sub-layer level.
• Zooming into LMs yields clearer insights, potentially even algorithmic explanations.
• Next time you study LM representations, ask where the info in the representations comes from. Why / how do LMs create it?

The Transformer Architecture

How LMs Learn SVA

I want to understand how LMs’ processing changes during training. Do circuits only change with performance? Or are they dynamic even when performance flatlines? I conducted a behavioral evaluation of Pythia-160m’s SVA abilities.

Data and Metric

Our SVA dataset is a pre-existing dataset [3] of sentences with challenging constructions, e.g., center embedding. We run ACDC on same-structure subsets of this. We measure model behavior thus. Let $x_i$ be a sentence, and $A_i$, $D_i$ the sets of tokens that agree / disagree with its subject. For each $x_i$ in our dataset, we measure:

$$\sum_{x_i} p(a_i|x_i) - \sum_{x_i} p(d_i|x_i)$$

Learning occurs between steps 100 and 10,000; elsewhere, performance is static.

References