Locating and Editing Factual Associations in GPT

Kevin Meng* MIT CSAIL
David Bau* Northeastern University
Alex Andonian MIT CSAIL
Yonatan Belinkov† Technion – IIT

Abstract

We analyze the storage and recall of factual associations in autoregressive transformer language models, finding evidence that these associations correspond to localized, directly-editable computations. We first develop a causal intervention for identifying neuron *activations* that are decisive in a model's factual predictions. This reveals a distinct set of steps in middle-layer feed-forward modules that mediate factual predictions while processing subject tokens. To test our hypothesis that these computations correspond to factual association recall, we modify feed-forward *weights* to update specific factual associations using Rank-One Model Editing (ROME). We find that ROME is effective on a standard zero-shot relation extraction (zsRE) model-editing task. We also evaluate ROME on a new dataset of difficult counterfactual assertions, on which it simultaneously maintains both specificity and generalization, whereas other methods sacrifice one or another. Our results confirm an important role for mid-layer feed-forward modules in storing factual associations and suggest that direct manipulation of computational mechanisms may be a feasible approach for model editing. The code, dataset, visualizations, and an interactive demo notebook are available at `https://rome.baulab.info/`.

1 Introduction

Where does a large language model store its facts? In this paper, we report evidence that factual associations in GPT correspond to a localized computation that can be directly edited.

Large language models can predict factual statements about the world (Petroni et al., 2019; Jiang et al., 2020; Roberts et al., 2020). For example, given the prefix "*The Space Needle is located in the city of*," GPT will reliably predict the true answer: "*Seattle*" (Figure 1a). Factual knowledge has been observed to emerge in both autoregressive GPT models (Radford et al., 2019; Brown et al., 2020) and masked BERT models (Devlin et al., 2019).

In this paper, we investigate how such factual associations are stored within GPT-like autoregressive transformer models. Although many of the largest neural networks in use today are autoregressive, the way that they store knowledge remains under-explored. Some research has been done for masked models (Petroni et al., 2019; Jiang et al., 2020; Elazar et al., 2021a; Geva et al., 2021; Dai et al., 2022; De Cao et al., 2021), but GPT has architectural differences such as unidirectional attention and generation capabilities that provide an opportunity for new insights.

We use two approaches. First, we trace the causal effects of hidden state activations within GPT using causal mediation analysis (Pearl, 2001; Vig et al., 2020b) to identify the specific modules that mediate recall of a fact about a subject (Figure 1). Our analysis reveals that feedforward MLPs at a range of middle layers are decisive when processing the last token of the subject name (Figures 1b,2b,3).

Second, we test this finding in model weights by introducing a Rank-One Model Editing method (ROME) to alter the parameters that determine a feedfoward layer's behavior at the decisive token.

*Equal contribution. Correspondence to `mengk@mit.edu`, `davidbau@northeastern.edu`.
†Supported by the Viterbi Fellowship in the Center for Computer Engineering at the Technion.