A Trip Down Memory Lane: How Transformer Feed-Forward Layers Build Predictions in Language Models

Abstract:

The field of natural language processing is dominated by transformer-based language models (LMs). One of the core building blocks of these models is the feed-forward network (FFN) layers, which typically account for $>2/3$ of the network parameters. Yet, how these layers are being utilized by the model to build predictions is largely unknown. In this talk, I will share recent findings on the operation of FFN layers in LMs, and demonstrate their utility in real-world applications. First, I will show that FFN layers can be cast as human-interpretable key-value memories, and describe how the output from each layer can be viewed as a collection of updates to the model's output distribution. Then, I will demonstrate the utility of these findings in the context of (a) controlled language generation, where we reduce the toxicity of GPT2 by almost 50%, and (b) improving computation efficiency, through a simple rule for early exit, saving 20% of computation on average.