Understanding and acceleration of grokking phenomena in learning arithmetic operations via Kolmogorov-Arnold representation

Yeachan Park

Korea Institute for Advanced Study

We propose novel methodologies aimed at accelerating the grokking phenomenon, which refers to the rapid increment of test accuracy after a long period of overfitting as reported by Power et al. (2022). Focusing on the grokking phenomenon that arises in learning arithmetic binary operations via the transformer model, we begin with a discussion on data augmentation in the case of commutative binary operations. To further accelerate, we elucidate arithmetic operations through the lens of the Kolmogorov-Arnold (KA) representation theorem, revealing its correspondence to the transformer architecture: embedding, decoder block, and classifier. Observing the shared structure between KA representations associated with binary operations, we suggest various transfer learning mechanisms that expedite grokking. This interpretation is substantiated through a series of rigorous experiments. In addition, our approach is successful in learning two nonstandard arithmetic tasks: composition of operations and a system of equations. Furthermore, we reveal that the model is capable of learning arithmetic operations using a limited number of tokens under embedding transfer, which is supported by a set of experiments as well.

[1] Power, Alethea, et al. "Grokking: Generalization beyond overfitting on small algorithmic datasets." arXiv preprint arXiv:2201.02177 (2022).