Title: Grokking (machine learning)

URL: https://en.wikipedia.org/wiki/Grokking_(machine_learning)

PageID: 77080272

Categories: Category:Machine learning, Category:Phenomena

In machine learning (ML), grokking , or delayed generalization , is a phenomenon observed in some settings where a model abruptly transitions from overfitting (performing well only on training data) to generalizing (performing well on both training and test data), after many training iterations with little or no improvement on the held-out data. This contrasts with what is typically observed in machine learning, where generalization occurs gradually alongside improved performance on training data.

Etymology

Grokking was introduced in January 2022 by OpenAI researchers who were studying generalization on small datasets. It is derived from the word grok coined by Robert Heinlein in his novel Stranger in a Strange Land . In ML research, "grokking" is not used as a synonym for "generalization"; rather, it names a sometimes-observed delayed■generalization training phenomenon in which training and held■out performance do not improve in tandem, and in which held■out performance rises abruptly later. Authors also analyze the "grokking time", the epoch or step at which this transition occurs in those scenarios.

Interpretations

Grokking can be understood as a phase transition during the training process. In particular, recent work has shown that grokking may be due to a complexity phase transition in the model during training. While grokking has been thought of as largely a phenomenon of relatively shallow models, grokking has been observed in deep neural networks and non-neural models and is the subject of active research.

One potential explanation is that the weight decay (a component of the loss function that penalizes higher values of the neural network parameters, also called regularization) slightly favors the general solution that involves lower weight values, but that is also harder to find. According to Neel Nanda, the process of learning the general solution may be gradual, even though the transition to the general solution occurs more suddenly later.

Recent theories have hypothesized that grokking occurs when neural networks transition from a "lazy training" regime where the weights do not deviate far from initialization, to a "rich" regime where weights abruptly begin to move in task-relevant directions. Follow-up empirical and theoretical work has accumulated evidence in support of this perspective, and it offers a unifying view of earlier work as the transition from lazy to rich training dynamics is known to arise from properties of adaptive optimizers, weight decay, initial parameter weight norm, and more. This perspective is complementary to a unifying "pattern learning speeds" framework that links grokking and double descent ; within this view, delayed generalization can arise across training time ("epoch■wise") or across model size ("model■wise"), and the authors report "model■wise grokking".

References