Computing Reviews
Association for Computing Machinery
ThinkLoud

Review

Can AI learn to forget?
Greengard S. Communications of the ACM 65(4): 9-11, 2022. Type: Article

Date Reviewed: 08/08/22

Nowadays, we can assume readers of *Computing Reviews* are familiar with the ideas behind machine learning, where neural networks are trained with large training sets so that they “learn” to recognize patterns within said dataset; some of those patterns are readily identifiable by a human observer, but some might be deep, very subtle, and impossible for us to understand. Neural networks usually keep learning (and thus adjusting their behaviors) from further data they receive in production. This short article deals with the problem termed as “machine unlearning”: how can we ensure all traces of a given problematic case can be permanently erased from a neural network?

This issue is relevant for many reasons; Greengard quotes legal requests derived from laws such as the European General Data Protection Regulation (GDPR) or the California Consumer Privacy Act (CCPA), where individuals might exercise the “right to be forgotten” and demand that service providers delete their records from a dataset. A record might need to be removed because it is too much an outlier in some aspects to the rest of the items, so it gets “memorized,” leading to potentially pathological patterns. It has even been shown that the specific data for given records can be retrieved by specially crafting queries to the neural network. Such a removal could be carried out by removing the user from the training database and running again the training process, but this requires all source data to be retained and an expensive and time-consuming training process to be carried out for each requested deletion.

Greengard refers to a promising machine unlearning method called SISA (sharded, isolated, sliced, and aggregated): the training data can be divided into disjoint sets, each of them much quicker to re-train, and then combined together; removing specific records from one of them still requires the whole shard to be processed, but yields a significant speed improvement over the whole process. He also presents model checkpointing, where the learner builds and stores many different discrete models, each of them excluding certain data points.

This article, easy and pleasant to read, brings forward a simple to state problem that is likely to open a wide array of research directions in the next few years. It can be brought into discussion for advanced undergraduate-level courses, or as a starting point for graduate studies.

Reviewer: Gunnar Wolf Review #: CR147482