Title: Machine unlearning

URL: https://en.wikipedia.org/wiki/Machine_unlearning

PageID: 78565600

Categories: Category:Machine learning, Category:Machine learning stubs

-----

Machine unlearning is a branch of machine learning focused on removing specific undesired element, such as private data, wrong or manipulated training data, outdated information, copyrighted material, harmful content, dangerous abilities, or misinformation, without needing to rebuild models from the ground up.

Large language models, like the ones powering ChatGPT , may be asked not just to remove specific elements but also to unlearn a "concept," "fact," or "knowledge," which aren't easily linked to specific examples. New terms such as "model editing," "concept editing," and "knowledge unlearning" have emerged to describe this process. [1]

History

Early research efforts were largely motivated by Article 17 of the GDPR , the European Union's privacy regulation commonly known as the "right to be forgotten" (RTBF), introduced in 2014. [2]

Present

The GDPR did not anticipate that the development of large language models would make data erasure a complex task. This issue has since led to research on "machine unlearning," with a growing focus on removing copyrighted material, harmful content, dangerous capabilities, and misinformation. Just as early experiences in humans shape later ones, some concepts are more fundamental and harder to unlearn. A piece of knowledge may be so deeply embedded in the model’s knowledge graph that unlearning it could cause internal contradictions, requiring adjustments to other parts of the graph to resolve them. [citation needed] .

Researchers have now also started studying unlearning in the context of removing incorrect or adversarially manipulated training data such as systematically biased labels or poisoning attacks. [3]

References

This machine learning -related article is a stub . You can help Wikipedia by expanding it .