Making a Machine Learning Model Forget About You

By Martin Anderson

First published ***August 11th, 2021*** *at:*
https://www.unite.ai/making-a-machine-learning-model-forget-about-you-forsaken-forgetting/
Web-archived version

Removing a particular piece of data that contributed to a machine learning model is like trying to remove the second spoonful of sugar from a cup of coffee. The data, by this time, has already become intrinsically linked to many other neurons inside the model. If a data point represents 'defining' data that was involved in the earliest, high-dimensional part of the training, then removing it can radically redefine how the model functions, or even require that it be re-trained at some expenditure of time and money.

Nonetheless, in Europe at least, Article 17 of the General Data Protection Regulation Act (GDPR) requires that companies remove such user data on request. Since the act was formulated on the understanding that this erasure would be no more than a database 'drop' query, the legislation destined to emerge from the Draft EU Artificial Intelligence Act will effectively copy and paste the spirit of GDPR into laws that apply to trained AI systems rather than tabular data.

Further legislation is being considered around the world that will entitle individuals to request deletion of their data from machine learning systems, while the California Consumer Privacy Act (CCPA) of 2018 already provides this right to state residents.

Why It Matters

When a dataset is trained into an actionable machine learning model, the characteristics of that data become generalized and abstract, because the model is designed to infer principles and *broad trends* from the data, eventually producing an algorithm that will be useful in analyzing specific and non-generalized data.

However, techniques such as model inversion have revealed the possibility of re-identifying the contributing data that underlies the final, abstracted algorithm, while membership inference attacks are also capable of exposing source data, including sensitive data that may only have been permitted to be included in a dataset on the understanding of anonymity.

Escalating interest in this pursuit does not need to rely on grass-roots privacy activism: as the machine learning sector commercializes over the next ten years, and nations come under pressure to end the current *laissez faire* culture over the use of screen scraping for dataset generation, there will be a growing commercial incentive for IP-enforcing organizations (and IP trolls) to decode and review the data that has contributed to proprietary and high-earning classification, inference and generative AI frameworks.

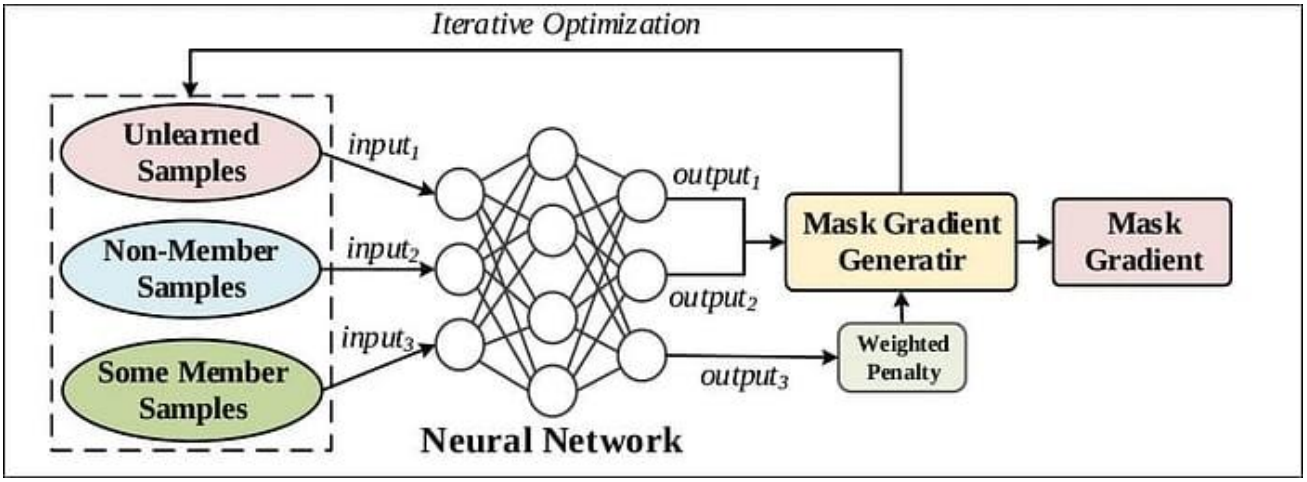
Inducing Amnesia in Machine Learning Models

Therefore we are left with the challenge of getting the sugar out of the coffee. It's a problem that has been vexing researchers in recent years: in 2021 the EU-supported paper *A Comparative Study on the Privacy Risks of Face Recognition Libraries* found that several popular face recognition algorithms were capable of enabling sex or race based discrimination in re-identification attacks; in 2015 research out of Columbia University proposed a 'machine unlearning' method based on updating a number of summations within the data; and in 2019 Stanford researchers offered novel deletion algorithms for K-means clustering implementations.

Now a research consortium from China and the US has published new work that introduces a uniform metric for evaluating the success of data deletion approaches, together with a new 'unlearning' method called Forsaken, which the researchers claim is capable of achieving a more than 90% forgetting rate, with only a 5% accuracy loss in the overall performance of the model.

The paper is called *Learn to Forget: Machine Unlearning via Neuron Masking*, and features researchers from China and Berkeley.

Neuron masking, the principle behind Forsaken, uses a mask gradient generator as a filter for the removal of specific data from a model, effectively updating it rather than forcing it to be retrained either from scratch or from a snapshot that occurred prior to the inclusion of the data (in the case of streaming-based models that are continuously updated).



The architecture of the mask gradient generator. Source: https://arxiv.org/pdf/2003.10933.pdf

Biological Origins

The researchers state that this approach was inspired by the biological process of 'active forgetting', where the user takes strident action to erase all engram cells for a particular memory by manipulation of a special type of dopamine.

Forsaken continuously evokes a mask gradient that replicates this action, with safeguards to slow down or halt this process in order to avoid catastrophic forgetting of non-target data.

The advantages of the system are that it is applicable to many kinds of existing neural networks, whereas recent similar work has enjoyed success largely in computer vision networks; and that it does not interfere with model training procedures, but rather acts as an adjunct, without requiring that the core architecture be altered or the data retrained.

Restricting The Effect

Deletion of contributed data can have a potentially deleterious effect on the functionality of a machine learning algorithm. To avoid this, the researchers have exploited norm regularization, a feature of normal neural network training that is commonly used to avoid overtraining. The particular implementation chosen is designed to ensure that Forsaken does not fail to converge in training.

To establish a usable dispersal of data, the researchers used out-of-distribution (OOD) data (i.e., data not included in the actual dataset, mimicking 'sensitive' data in the actual dataset) to calibrate the way that the algorithm should behave.

Testing On Datasets

The method was tested over eight standard datasets and in general achieved close-to or higher forgetting rates than full retraining, with very little impact on model accuracy.

TABLE III: Forgetting rate of `Forsaken` on different datasets for OOD data unlearning

Dataset	Parameter Size	BT (BF)	Forgetting Rate (Average / Variance)				
			Full Retraining	`Forsaken`	SMU [9]	SISA [10]	SISA-DP [11]
C10.S.	14.77M	168 (32)	93.45 / 0.34%	97.62 / 0.44%	86.91 / 4.89%	91.67 / 0.79%	94.05 / 1.21%
C10.T.	14.77M	176 (24)	94.89 / 0.28%	98.29 / 0.69%	93.75 / 5.94%	96.03 / 0.63%	97.16 / 0.73%
C100.T.	14.77M	117 (83)	96.58 / 0.41%	95.73 / 1.32%	71.79 / 4.13%	86.32 / 2.89%	89.74 / 1.59%
I.C.	2.62M	195 (5)	94.36 / 0.16%	98.46 / 0.79%	43.59 / 6.33%	95.89 / 0.41%	96.41 / 1.14%
Reuters (35-11)	5.26M	192 (8)	94.79 / 0.13%	96.35 / 0.82%	81.25 / 2.51%	95.32 / 0.76%	93.75 / 0.87%
News (15-5)	0.25M	162 (38)	95.06 / 0.23%	97.53 / 0.59%	86.93 / 4.47%	92.26 / 1.09%	94.44 / 1.81%

TABLE IV: Forgetting rate of `Forsaken` on different datasets for ID data unlearning

Dataset	Parameter Size	BT (BF)	Forgetting Rate (Average / Variance)				
			Full Retraining	`Forsaken`	SMU [9]	SISA [10]	SISA-DP [11]
CIFAR10	14.77M	167 (33)	85.32 / 0.09%	88.63 / 2.29%	85.03 / 5.77%	86.23 / 0.41%	88.02 / 0.43%
CIFAR100	14.77M	191 (9)	94.24 / 0.04%	87.96 / 1.63%	79.06 / 5.05%	84.29 / 1.52%	88.48 / 1.46%
IMDB	2.62M	151 (49)	84.77 / 0.11%	94.04 / 2.19%	62.25 / 6.13%	87.42 / 0.71%	90.72 / 0.81%
Reuters	5.26M	160 (40)	83.13 / 0.09%	86.25 / 1.29%	69.38 / 5.33%	81.25 / 0.29%	84.38 / 0.76%
News	0.25M	172 (28)	85.47 / 0.71%	90.12 / 1.84%	86.86 / 6.67%	84.31 / 0.37%	87.21 / 0.84%

It seems impossible that full retraining on an edited dataset could actually do worse than any other method, since the target data is entirely absent. However, the model has by this time abstracted various features of the deleted data in a 'holographic' fashion, in the way (by analogy) that a drop of ink redefines the utility of a glass of water.

In effect, the weights of the model have already been influenced by the excised data, and the only way to entirely remove its influence is to retrain the model from absolute zero, rather than the far speedier approach of retraining the weighted model on an edited dataset.