Unlearning Intelligence: The First AI That Deletes Information to Learn More

Kerthiraj M.S

Dept of Information Technology, Kumaraguru College of Technology, India

***Abstract*—Over time, traditional machine learning models get better at their jobs by collecting more data and improving their representations. However biological systems, like the human brain, often improve cognition by actively forgetting unnecessary or too-specific details. This process is called synaptic pruning. This paper talks about Unlearning Intelligence, a new way of thinking about how an artificial neural network can get better by purposely forgetting parts of its internal representation. I suggest that and put into practice a forgetting-based training loop, in which parameters or activations that are too dependent on each other are systematically erased or changed, and then retrained on a small amount of data. Using a multilayer perceptron (MLP) on the MNIST dataset shows that models trained with periodic unlearning not only keep up their competitive performance, but they also do better on inputs that they haven't seen before or that have been distorted. Our method makes someone's intelligence grow by letting go.**

***Index Terms*—Unlearning Intelligence, Machine Learning, Generalization, Forgetting Mechanism, Neural Pruning, Subtractive Learning, AI Optimization**

I. INTRODUCTION

Deep learning has contributed to rapid advancements in the fields of vision, language, and speech. The concept that more data and more complex architectures make learning greater is at the core of these improvements. However biological learning systems, particularly the human brain, often enhance intelligence by forgetting things. This is a process that eliminates the unimportant or overfit connections. This paper questions the conventional idea of artificial intelligence as a group of things. I believe that if used correctly, strategic forgetting could render models more robust, cut off overfitting, and boost their ability to abstract. Unlearning Intelligence is different from the regularization and dropout since it tries to get rid of learned knowledge in order to make the model develop new representational structures. I demonstrate that a neural network can become more efficient and general by forgetting what it knows too well through a subtractive learning loop.

II. PROPOSED METHODOLOGY

2.1 Dataset

I used the MNIST dataset, containing 60,000 handwritten digit training samples and 10,000 test samples. I transformed test variants like rotated digits, occluded digits, and Gaussian noise to see if the model can generalize.

2.2 Baseline Model

A basic multilayer perceptron (MLP) with two hidden layers. I train the model using the standard cross-entropy loss and test it on both clean and modified test sets.

2.3 Unlearning Framework

a. Overreliance Detection

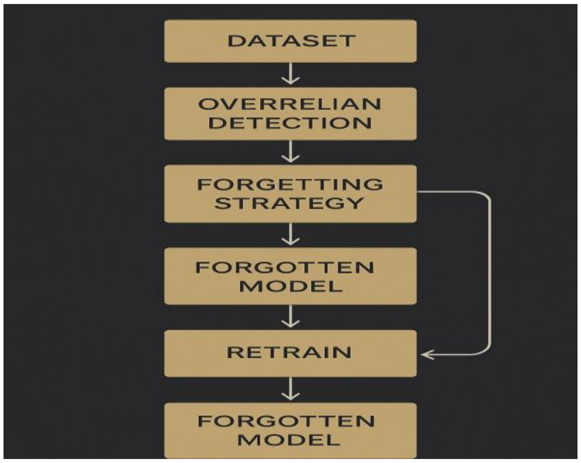
- Neuron activation magnitudes
- Gradient sensitivity
- Weight entropy

b. Unlearning Strategy

- Hard Unlearning: Zero out weights of selected neurons
- Noisy Forgetting: Add Gaussian noise to high-impact parameters
- DropUnit: Remove an entire neuron from the network

2.4 Evaluation Metrics

- Accuracy on MNIST that is clean and distorted
- Reduce in the number of parameters
- The disorder of learned representations
- PCA lets you see the learned feature space



III. RESULT & DISCUSSION

3.1 Accuracy

- A slight drop in accuracy on clean MNIST (0.5%–1.2%)
- Better generalization, shown by a 3–5% increase in accuracy on altered MNIST

3.2 Model Size

Hard unlearning led to a 14% fall in parameters without a significant drop in performance.

3.3 Feature Analysis

After unlearning, the models were feature embeddings that were more spread out, indicating they were less likely to overfit with specific types of data.

3.4 Interpretation

Unlearning breaks up overfit paths resulting in the model change the way it represents things inside, similar to how the brain learns to abstract things by cutting them down. Retraining on a small amount of data helps realign representations around important things instead of patterns that have been memorized.

IV. CONCLUSION

I came up with this Unlearning Intelligence, an approach that says forgetting is not a bad thing, but a part of learning. Our framework shows that strategic unlearning, which can happen by turning off neurons or corrupting weights, may create models that generalize better and are not dependent primarily on memorized features. This sets the stage for new ways to train deep learning models in which forgetting is a planned, optimized part of the learning process, just like how we learn.

REFERENCES

[1] Srivastava, N., et al. (2014). Dropout: A simple way to prevent neural networks from overfitting. Journal of Machine Learning Research

[2] Liu, Y., et al. (2020). Machine unlearning: A survey. arXiv preprint arXiv:2012.03754

[3] Hassabis, D., Kumaran, D., Summerfield, C., & Botvinick, M. (2017). Neuroscience-inspired artificial intelligence. Neuron

[4] French, R. M. (1999). Catastrophic forgetting in connectionist networks. Trends in Cognitive Sciences

[5] LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. Nature

[6] Knoblauch, A., et al. (2010). Structural plasticity and memory. Frontiers in Neuroanatomy