3iA Côte d'Azur
Interdisciplinary Institute for Artificial Intelligence

—Postdoc Research Topic—
Neuro-Computational Insights on Machine Unlearning

Research axis of the 3IA: Axis 3 - AI for Computational Biology and Bio-inspired AI
Supervisor (3IA Chair): Emanuele Natale, Sophia Antipolis Laboratory for Computer Science, Signals and Systems (I3S), Sophia Antipolis
Hosting lab: I3S & INRIA UniCA

Apply by sending an email directly to the supervisor:
emanuele.natale@univ-cotedazur.fr

Primary discipline: Machine Learning
Secondary discipline: Neuroscience

Project Summary

This project proposes to explore how to equip artificial neural networks with the ability to efficiently unlearn, by applying neuroscientific insights from the same ability of mammal brains. In everyday life, disassociating memories is essential—letting us move on from fears, mistakes, or outdated beliefs. Similarly, forgetting is a challenge for artificial intelligence: once a machine learns something, it's hard to have it forget. This has unwanted implications when machines learn something wrong, private, copyrighted, or biased. The project aims to draw inspiration from how forgetting happens in biological brains, in order to advance the state of the art on how artificial neural networks can perform active unlearning more accurately, more efficiently, and more safely.

Scientific Description

Large Language Models (LLMs) are trained on massive text datasets—often in the order of terabytes—making it almost impossible to filter out undesirable or outdated information. When wrong, private, or copyrighted information is used for training, it often compromises the models with undesired associations that must be subsequently severed. Re-training the model may not be always viable, and it easily becomes necessary to prevent and remove particular associations within the model, without disrupting its integrity. Inevitably, one question arises: how can we make LLMs forget? This challenge is referred to as **Machine Unlearning**[1]. The problem is not dissimilar to that faced by biological organisms, which have evolved efficient strategies to update outdated or irrelevant information. In neuroscience research, a large body of literature has investigated the neural mechanisms that govern such memory unlearning processes.

[1] Thanh Tam Nguyen, Thanh Trung Huynh, Zhao Ren, Phi Le Nguyen, Alan Wee-Chung Liew, Hongzhi Yin, and Quoc Viet Hung Nguyen. 2025. A Survey of Machine Unlearning. ACM Trans. Intell. Syst. Technol. 2025. https://doi.org/10.1145/3749987

This project bridges computational neuroscience and machine learning to study the mechanisms of unlearning in artificial systems. Unlearning is crucial for biological organisms to adapt and remain flexible in dynamic environments, as well as for machines to optimize output integrity by shedding outdated or harmful associations. In this project we will draw analogies between memory dynamics in mammal brains and challenges in machine unlearning, exploring recent empirical ideas such as task vectors[2] and continual learning-unlearning frameworks[3], contextualizing them in the neuroscience of memory and aiming at providing rigorous theoretical grounds for them.
As for the neuroscience contextualization, it first aims at assessing how such unlearning strategies relate to the current state of the art on the understanding of forgetting in neuroscience, and whether such relationships and analogies can provide evidence either towards mechanism of memory erasure (disruption of an existing memory trace, referred as *reconsolidation update* in the field) or memory sidelearning (creation of a new memory trace that inhibits but does not delete the original one, referred as *extinction learning* in the field). Secondly, while the aforementioned recent strategies, among others, have been empirically validated, they lack a theoretical understanding that explains the mathematical principles underlying their efficacy. Providing a rigorous mathematical foundation for them can not only guide further improvements, but provide essential insights on the corresponding mechanisms in biological brains.

Positioning within Université Côte d'Azur

This interdisciplinary project is framed within a collaboration between 3IA Chair Emanuele Natale at I3S and INRIA d'Université Côte d'Azur and Bianca Silva at IPMC Sophia Antipolis. E. Natale and B. Silva are co-supervised a starting PhD project on building a biologically-grounded spiking neural network model of active forgetting, focusing on the neuroscientific aspects of unlearning. This postdoc project is complementing the latter direction by focusing on the machine learning aspects.

[2] G. Ilharco, M. T. Ribeiro, M. Wortsman, L. Schmidt, H. Hajishirzi, and A. Farhadi, "Editing models with task arithmetic," ICLR 2023, https://openreview.net/forum?id=6t0Kwf8-jrj
[3] Z. Huang et al., "Unified Gradient-Based Machine Unlearning with Remain Geometry Enhancement," Neurips 2024. https://openreview.net/forum?id=dheDf5EpBT