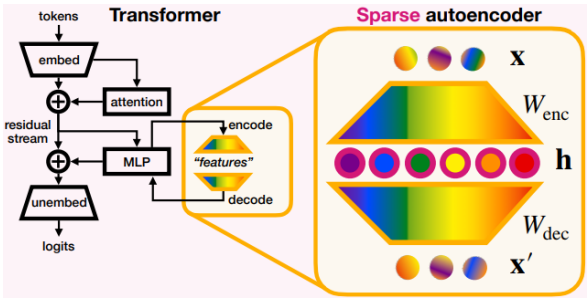
ETH
Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

Distributed Computing

Prof. R. Wattenhofer

Sparse Autoencoders for Sensitive Feature Removal in Federated Learning



Federated learning is a machine-learning approach where multiple devices or organizations collaboratively train a shared model without sharing their raw data. Each participant trains the model locally and sends only information about the model to a central server for aggregation. A key challenge in this area is effectively preserving privacy of sensitive data from the training set. Several strategies are currently used in federated or distributed settings. A common one is parameter averaging, where model weights or gradients are aggregated into a single global model. Another approach uses ensemble methods, where participating models remain separate, and their outputs are combined without merging their parameters into a shared space. A third approach relies on knowledge distillation, where each model's predictions on a shared public or proxy dataset are used to train a new unified model. However, none of these methods do guarantee by nature the privacy of sensitive training data.

In this project, we will focus on work within the subfield of model distillation. Originally, model distillation is done by simply using the outputs of the models, but in recent years, approaches [1] have been presented that not only use the predictions of the models but also leverage the hidden representations of some layers to enforce even greater similarity between the distillation model and the original models. However, outputs as well as hidden representations can include sensitive data from the training data. For example, when using "The email from John Doe is " as input for an Large Language Model (LLM), the hidden representations can include the representation of John Doe's private email, effectively leaking private data from the LLM's learning data.

The goal of this work is to take inspiration from the field of **mechanistic interpretability** to approach the **privacy** issue. Recent work[1] in mechanistic interpretability has proposed sparse autoencoders (SAE) [2] to identify features in hidden representations. SAE are utilized in work like [3, 4] to detect features which result in the leakage of sensitive information. In [3], this is used to prevent the generation an of sensitive data by the model. But, similar to [4], we are more interested in using SAEs to create shareable representations (and outputs) that do not contain sensitive information. We also aim to investigate how using such representations affects the performance of the distilled model, compared tosharing the full hidden representations and output for model distillation.

We also note that [3] remove sensitive data in a supervised manner, meaning they train their method to remove features and directions associated with email addresses. [4] remove sensitive data by first manually identifying sensitive features within a sparse autoencoder.

[1] https://transformer-circuits.pub/2023/monosemantic-features/index. html

Therefore, another step would be to explore how to transition to a non-supervised and non-manual method for removing sensitive features. This would involve both theoretical and empirical determination of what sensitive features in SAE are and how to remove them, without significantly harming distillation performance. An early idea in this direction would be to remove all features that are active in only x% of the private training data points, while also intervening on the remaining feature values in a way that ensures information related to only x% of the private training data cannot be reconstructed.

Interested? Contact us for more information

- Denis Sutter: densutter@ethz.ch, ETZ G97
- Dr. Antonio Di Maio: adimaio@ethz.ch, ETZ G97

Last update: December 2025

References

[1] Gustavo Aguilar et al. "Knowledge Distillation from Internal Representations". In: *CoRR* abs/1910.03723 (2019). arXiv: 1910.03723. URL: http://arxiv.org/abs/1910.03723.

[2] Hoagy Cunningham et al. *Sparse Autoencoders Find Highly Interpretable Features in Language Models*. 2023. arXiv: 2309.08600 [cs.LG]. URL: https://arxiv.org/abs/2309.08600.

[3] Ahmed Frikha et al. *PrivacyScalpel: Enhancing LLM Privacy via Interpretable Feature Intervention with Sparse Autoencoders*. 2025. arXiv: 2503.11232 [cs.LG]. URL: https://arxiv.org/abs/2503.11232.

[4] Joris Heemskerk et al. "Privacy-Aware Traffic Re-Identification with Interpretable Sparse Autoencoders". In: *The 37th Benelux Conference on Artificial Intelligence and the 34th Belgian Dutch Conference on Machine Learning*. 2025. URL: https://openreview.net/forum?id=GVlD0M7o5a.