Contrastive Unlearning in Hyperbolic space

December 31, 2024

Emanuele Leone

Abstract

This study investigates contrastive unlearning within hyperbolic spaces. By utilizing hyperbolic geometry to reposition unlearning samples, I demonstrate the potential effectiveness of this approach. While the results are promising, several computational challenges emerge, indicating opportunities for future research aimed at enhancing the efficiency of hyperbolic contrastive unlearning. Project link: https://github.com/FFMasterSlave/Progetto-DL.

1. Introduction

Machine unlearning is crucial for data privacy, especially with GDPR's "right to be forgotten," which demands removing specific data from machine learning models without degrading performance on other tasks. Traditional unlearning methods, like retraining, are costly and time-intensive. *Contrastive unlearning* offers a solution for classification problem by leveraging contrastive learning principles. It actively "forgets" data by adjusting embeddings: pushing unlearned data away from its class and pulling it toward other classes, effectively erasing its influence while preserving model utility. As a geometric approach to unlearning, it opens possibilities of experimentation with new data representation spaces that are gaining popularity in the latest years, such as *Hyperbolic spaces*. These prove effective in capturing complex, hierarchical data. Testing hyperbolic embeddings in contrastive unlearning could enhance accuracy and robustness over traditional Euclidean approaches, making unlearning more effective.

2. Related work

Machine unlearning research focuses on removing the influence of specific data from models while retaining overall model performance. Approaches generally fall into *exact unlearning*, which ensures complete data removal through costly retraining, and *approximate unlearning*, which minimizes data influence using efficient, often heuristic methods. While exact unlearning methods are effective, their reliance on retraining can make them impractical at scale. Approximate methods offer computational advantages but often compromise on completeness or model utility. Recent advancements in approximate unlearning involve directly modifying the model's representation space to remove data influence more effectively. Many methods, however, primarily focus on either unlearning samples or preserving the remaining dataset without explicitly optimizing both in the latent space. Some algorithms, like NegGrad (Golatkar et al., 2020), apply gradient reversal on unlearning samples but can alter class boundaries, impacting utility. Others, such as SCRUB (Kurmanji et al., 2023), rely on iterative adjustments across the full dataset, which is computationally costly. The Contrastive Unlearning approach of (kyu Lee et al., 2024) addresses these gaps by using representation learning to target both unlearning effectiveness and model preservation. It contrasts unlearning samples with both same-class and different-class samples, pushing unlearning data away from its original class and toward other classes in the embedding space. By balancing these contrastive forces, this approach achieves efficient, high-fidelity unlearning while maintaining model performance on remaining samples.

3. Method

The novelty of contrastive unlearning is in utilizing geometric properties of the latent representation space to achieve unlearning. Specifically, (kyu Lee et al., 2024) hypothesize that, after training, a model produces geometrically similar embeddings for samples of the same class and more distant embeddings for samples of different classes, even without explicit representation learning techniques. Building on this, contrastive unlearning modifies the representation space by contrasting each unlearning sample with remaining samples: (1) pushing embeddings of unlearning samples away from remaining samples of the same class (positive pairs) and (2) pulling them closer to samples of different classes (negative pairs). Through this process, unlearning samples are repositioned centrally among

Email: Emanuele Leone <leone.1859441@studenti.uniroma1.it>.

Deep Learning and Applied AI 2024, Sapienza University of Rome, 2nd semester a.y. 2023/2024.

remaining samples in the latent space. To show the validity of the approach, two scenarios are used: *single-class* unlearning and *random-sample* unlearning. In the *single-class unlearning* scenario, the goal is to eliminate the influence of an entire class by adjusting the embeddings of its samples to remove their contribution to model predictions. This procedure involves pushing these class embeddings away from their original positions and towards different classes, effectively achieving "forgetting" for this class. Evaluations focus on the model's performance on remaining classes, as well as the extent to which the chosen class samples no longer influence the model, measured by reduced classification accuracy on the unlearned class. For *random-sample* unlearning, a randomly selected batch of samples is chosen to be "forgotten." The contrastive unlearning approach pushes these samples away from their original class embeddings and pulls them towards embeddings of other classes, following a similar procedure as in *single-class* unlearning. To extend this approach, I implemented it within a hyperbolic Poincaré ball model. This required adapting the original model to work with spaces of negative curvature and remoduling the contrastive loss function to use hyperbolic distances, leveraging hyperbolic space's ability to capture hierarchical relationships. Here below the unlearning loss functions for both *single-class (SC)* and *random-sample (RS)* scenarios:

$$L_{RS} = \sum_{x_i \in X^u} -\frac{1}{|N_z(x_i)|} \sum_{z_a \in N_z} \log \frac{\exp(z_i \cdot z_a/\tau)}{\sum_{z_p \in P_z(x_i)} \exp(z_i \cdot z_p/\tau)} \tag{1}$$

$$L_{SC} = \sum_{x_i \in X^u} -\frac{1}{|N_z(x_i)|} \sum_{z_a \in N_z} \log \frac{\exp(z_i \cdot z_a/\tau)}{|N_z(x_i)|} \tag{2}$$

And the total loss of the Unlearning algorithm for both scenarios:

$$L = \lambda_{UL} L_{UL} + \lambda_{CE} L_{CE}(F(X^r), Y^r) \tag{3}$$

You can find the details about these formulas on the original article of (kyu Lee et al., 2024) or the Jupyter notebook at https://github.com/FFMasterSlave/Progetto-DL.

4. Experimental results

The experiments conducted in the article assess the contrastive unlearning algorithm's performance on both single-class and random-sample unlearning scenarios. The datasets used for testing are the CIFAR-10 dataset (subset of 10 natural classes form the CIFAR-100 dataset) and SVHN (The Street View House Numbers dataset). Following the work of (kyu Lee et al., 2024), i used the accuracy metric for validating the efficency of the unlearning procedure. Although this parameter can give an idea of the effectiveness of the unlearning procedure, the random nature of the sampling of which samples to compare with the data to be forgotten, make standard accuracy-based metrics less meaningful, as different batches may vary significantly in their impact on the overall model. Consequently, the primary focus here is on observing shifts in the data distribution within the latent space rather than on changes in classification accuracy alone. This focus on spatial distribution enables a clearer understanding of how unlearning samples affect the representation space, regardless of batch composition randomness. Figure 3 is an example of data distribution after the unlearning algorithm:

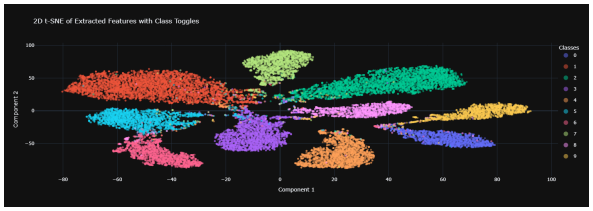

Figure 1. Data representation after Euclidean Contrastive Unlearning

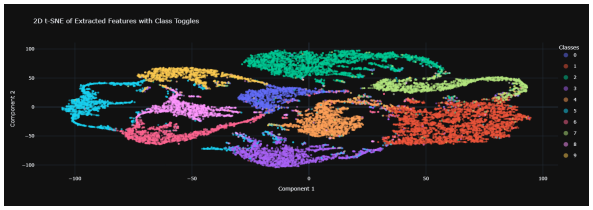

Figure 2. Data representation after Hyperbolic Contrastive Unlearning

Figure 3. Example of data distribution after the unlearning procedure of class 5 (cian color)

Some accuracy results can be found at the Github page.

5. Discussion and conclusions

Contrastive unlearning shows promise in effectively unlearning data while maintaining model performance. Experiments suggest hyperbolic spaces, though slower than Euclidean ones, offer structured representations that help separate complex data clusters, aiding unlearning. Despite its potential, computational challenges remain. This study lays the groundwork for exploring non-Euclidean geometries to enhance unlearning in machine learning, especially in improving computational efficiency.

References

Golatkar, A., Achille, A., and Soatto, S. Eternal sunshine of the spotless net: Selective forgetting in deep

networks, 2020. URL https://arxiv.org/abs/1911.04933.

Kurmanji, M., Triantafillou, P., Hayes, J., and Triantafillou, E. Towards unbounded machine unlearning, 2023. URL https://arxiv.org/abs/2302.09880.

kyu Lee, H., Zhang, Q., Yang, C., Lou, J., and Xiong, L. Contrastive unlearning: A contrastive approach to machine unlearning, 2024. URL https://arxiv.org/abs/2401.10458.