Memorization in Self-Supervised Learning Improves Downstream Generalization

Wenhao Wang*, Muhammad Ahmad Kaleem*, Adam Dziedzic*,
Michael Backes, Nicolas Papernot, Franziska Boenisch

CISPA Helmholtz Center for Information Security, University of Toronto, Vector Institute

Motivation

Memorization is a relevant concept to understand generalization, learning behavior, and privacy risks;

Self-supervised learning (SSL) Memorization was only empirically explored;

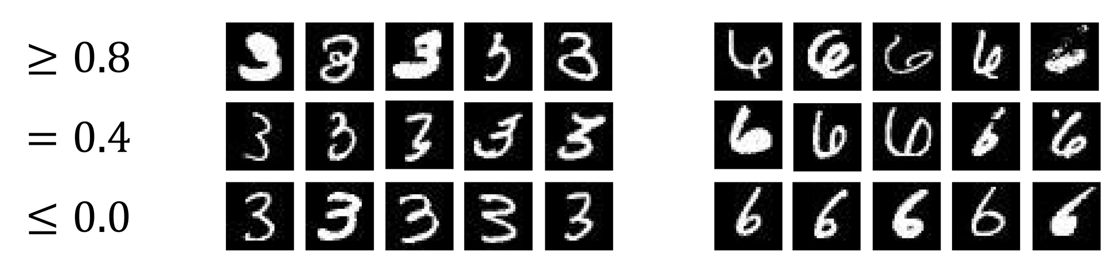
Formal definitions for memorization from supervised learning rely on labels and can not be applied;

Contributions

- Formal definition of memorization for SSL encoders (`SSLMem`): independent of SSL framework and training loss, operates on representations;
- Practical framework for experimentally approximating `SSLMem`;
- Extensive empirical evaluation of `SSLMem` on various SSL frameworks and datasets;

Summary of Findings

1. SSL memorizes especially atypical data points;



MNIST: class *3* and *6* for different levels of memorization (0: no memorization).

2. Highest memorized data points between different SSL frameworks align but differ significantly to highest memorized points in supervised learning;

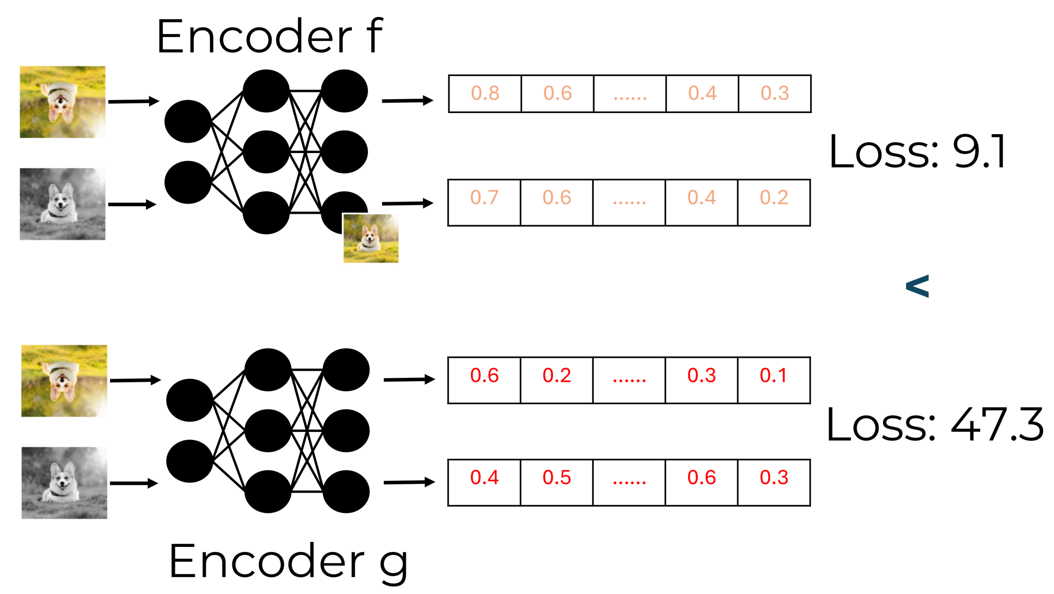
3. Memorization in the SSL encoder increases downstream generalization over different downstream data distributions and tasks;

Notation

Symbol	Explanation
$\mathcal{A}$	SSL learning algorithm
$S = \{x_i\}_{i=1}^m$	Training dataset
$S' = S \setminus x$	Reference dataset
$\mathbf{Aug}(x)$	Augmentation set
$f : \mathbb{R}^n \to \mathbb{R}^d$	Encoder trained on S
$g : \mathbb{R}^n \to \mathbb{R}^d$	Reference encoder trained on $S' = S \setminus \{x_j\}$
S_S	training data shared between encoders f and g
S_C	candidate set, training data for f only

Intuition of SSLMem

- SSL frameworks optimize for *representation alignment, i.e.,* two augmentations of the same data point should have close representations;
- Quantify memorization of data point x by comparing representation alignment of encoder f trained with x and reference encoder g trained without x;
- Intuition: x is memorized more the more f's representation alignment is better than g's;
- If f's and g's representation alignment is close, x does not influence f's behavior significantly (no memorization);



Formalizing SSLMem

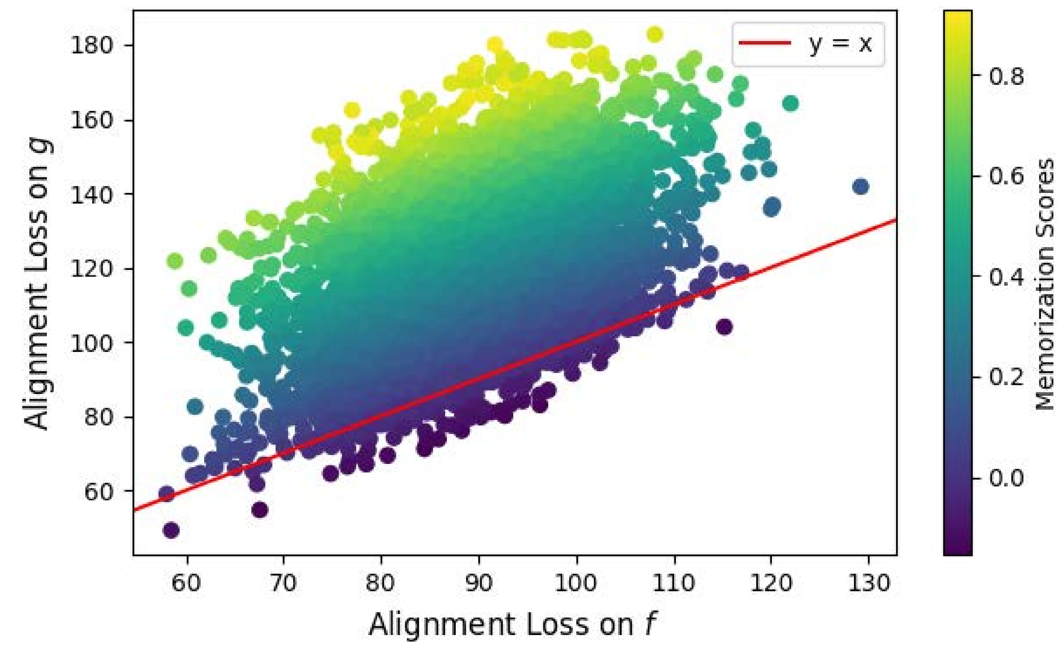
Definition *alignment loss* to quantify representation alignment for metric d, *e.g.,* the ℓ_2 distance:

$$\mathcal{L}_{\text{align}}(f, x) = \mathop{\mathbb{E}}_{x', x'' \sim \text{Aug}(x)} [d\,(f(x'), f(x''))]$$

Our Definition of Memorization Score:

$$\mathcal{H}_{\text{align}}(f, x, S) = \mathop{\mathbb{E}}_{f \sim \mathcal{A}(S)} \mathcal{L}_{\text{align}}(f, x)$$

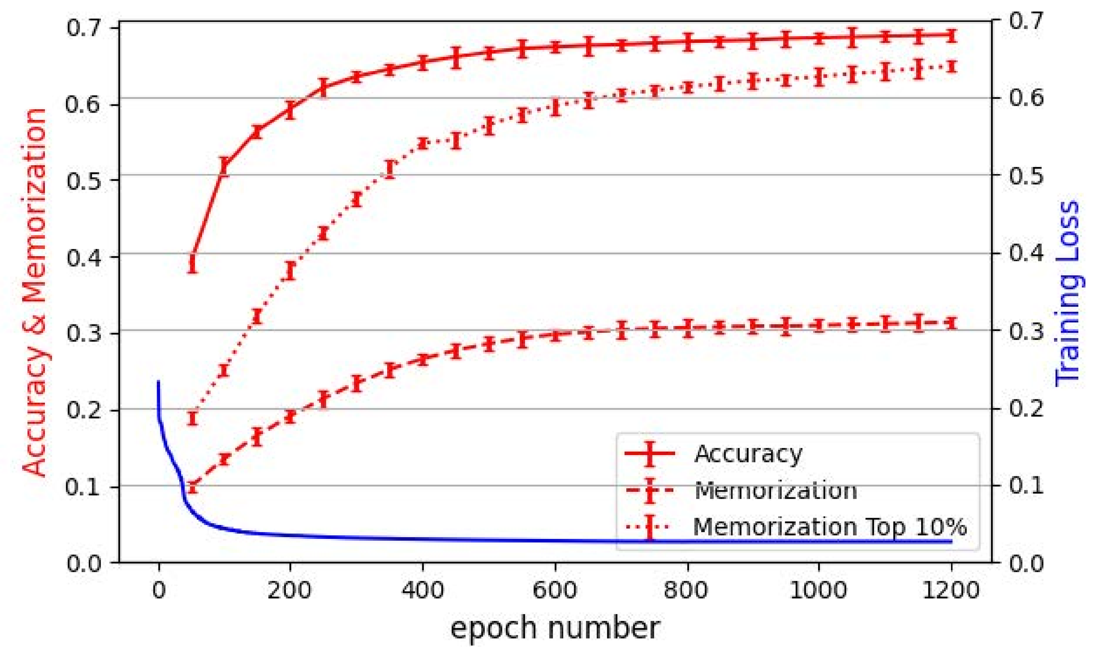
$$\texttt{SSLMem}(g, f, x, S', S) = \mathcal{H}_{\text{align}}(g, x, S') - \mathcal{H}_{\text{align}}(f, x, S)$$



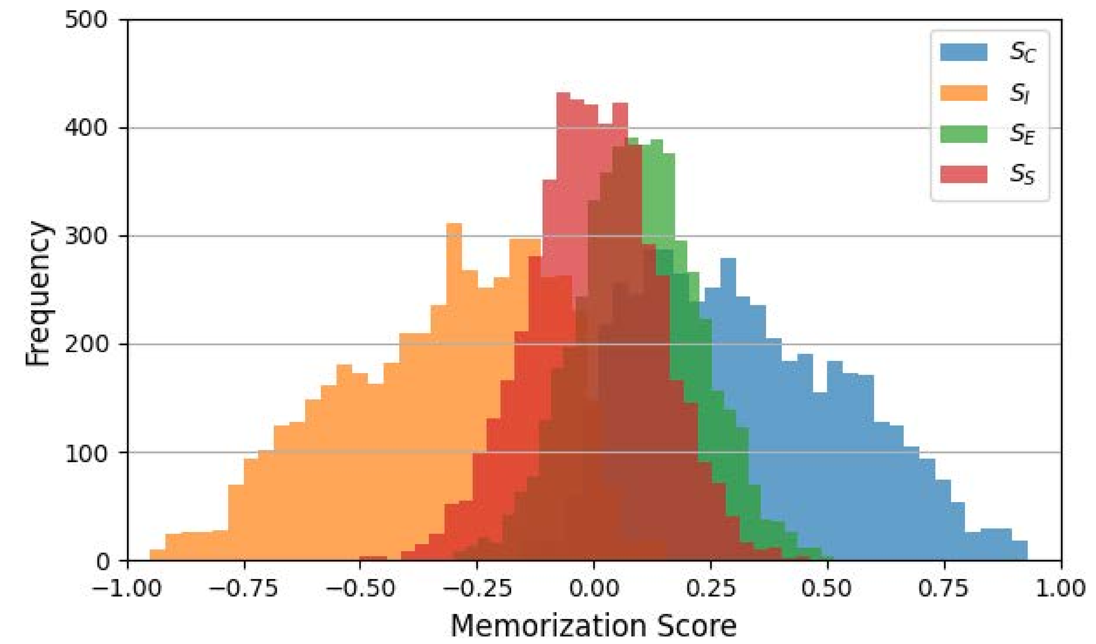
Encoder alignment loss vs. SSL memorization.

Insights into SSLMem

We train an MAE SSL encoder based on VIT-tiny using CIFAR10.



Memorization is not just an effect of increasing/ decreasing accuracy: while loss and accuracy stagnate after a few hundred epochs, memorization increases.



The encoders exhibit memorization indicated by significantly higher scores for S_C (candidates used to train only f) compared to S_S (shared training set for f and g).

Retained Points	CIFAR10	CIFAR100	STL10
25k (full encoder)	63.3%±0.92%	61.1%±1.14%	61.6%±0.83%
24k (most memorized)	**64.4%±1.03%**	**61.3±0.98%**	**61.7±1.18%**
22k (most memorized)	**63.8%±0.76%**	**61.8±1.24%**	**62.4±1.05%**
20k (most memorized)	63.2%±1.07%	60.8%±0.68%	61.1±1.05%
16k (most memorized)	61.8%±1.11%	58.4%±0.91%	59.9±0.89%
12k (most memorized)	59.7%±0.74 %	55.6%±1.32%	55.2±1.24%

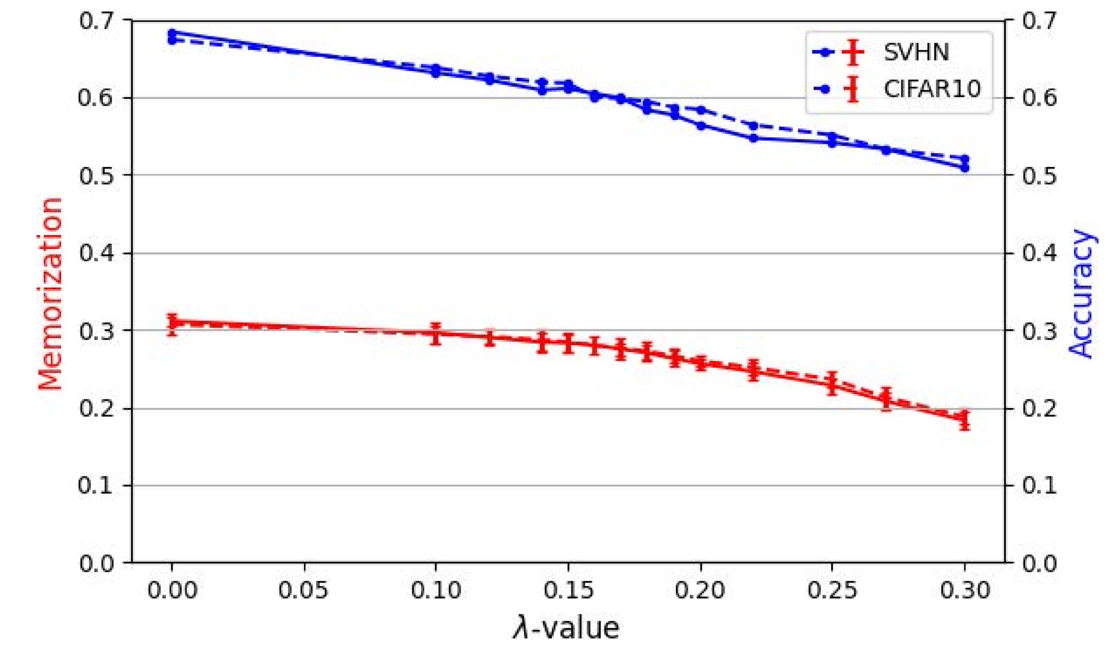
CoreSet Selection: Training only on the most memorized data points yields same performance.

ε	`SSLMem`	Acc. (%)
∞	0.307 ± 0.013	69.40% ± 1.12%
20	0.182 ± 0.009	54.22% ± 0.98%
8	0.107 ± 0.012	33.66% ± 1.76%

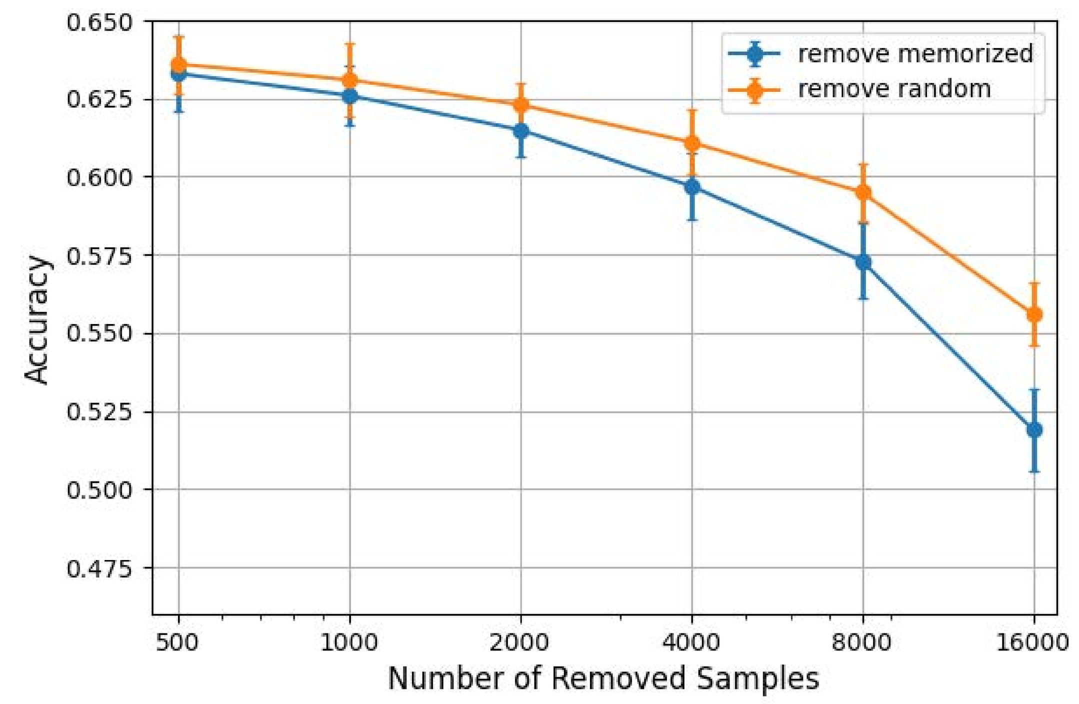
Effect of differential privacy.

Evaluation of SSLMem

We assess the influence of memorization on downstream generalization.



Limiting memorization harms downstream accuracy.



Removal of memorized data points harms accuracy over all downstream tasks more than the removal of random data points.

	Without removing	Removing 10000		Removing 20000	
		Memorized	Random	Memorized	Random
mIoU	45.4	44.8	45.1	43.8	44.4
Acc. (%)	69.89%±0.84%	68.33% ± 0.92%	68.91%±0.77%	66.51%±1.03%	67.58% ±0.82%

Evaluating the effect of memorization on a semantic segmentation downstream task.

Conclusions

- We introduce `SSLMem`, a formal definition for memorization in self-supervised learning.
- `SSLMem` generalizes across different encoder architectures and SSL training frameworks, and is independent of any downstream task and label.
- We show that encoders require memorization to generalize well to downstream tasks.