

Memorization in Self-Supervised Learning

Muhammad Ahmad Kaleem

Joint work with Wenhao Wang, Adam Dziedzic, Franziska Boenisch, Nicolas Papernot

University of Toronto

October 18, 2023

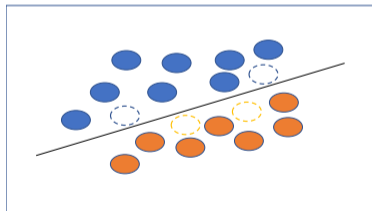
Table of Contents

- 1 Background
- 2 Prior Work and Motivation
- 3 Proposed Method
- 4 Experimental Results
- 5 Next Steps

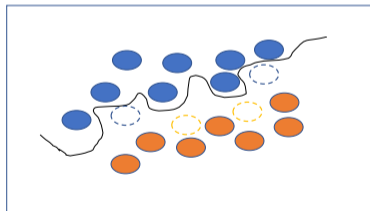
Background

Memorization

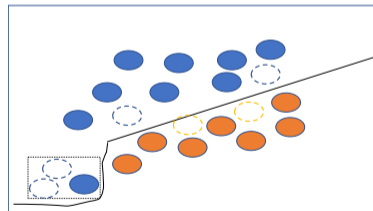
- Important property of learning algorithms and neural networks
- Memorized datapoints have a large impact on the output of a learning algorithm: source of privacy leakage
- Overparametrized deep neural networks can easily memorize training datapoints
- Memorization generally not favourable but required in certain cases for good generalization (Feldman, 2020)



Low memorization



High memorization



Memorization of outliers

Self-Supervised Learning (SSL)

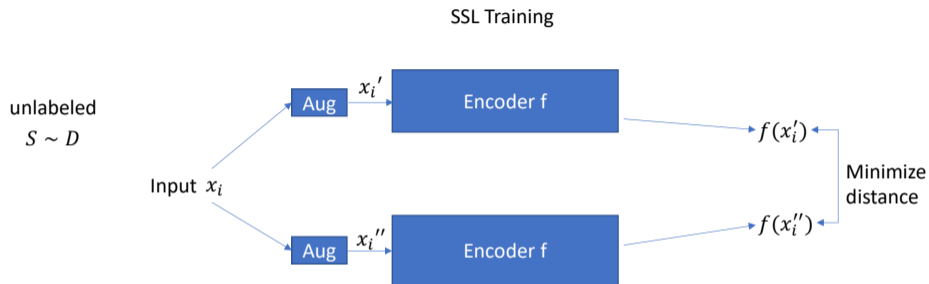
- Learning paradigm for unsupervised representation learning
- Main objective to learn implicit structures in input data so representations are a useful encoding
- Common form is contrastive learning: representations so that similar inputs have similar representations, dissimilar ones have dissimilar representations
- Training relies on the use of augmentations e.g. cropping, rotation, blurring to achieve this goal
- Trained encoders can be used for different types of downstream tasks e.g. classification, segmentation

Model of SSL

- Unlabeled dataset $S = \{x_i\}_{i=1}^m$, encoder $f : \mathbb{R}^n \rightarrow \mathbb{R}^d$
- Model data distribution \mathcal{D} as being composed of K latent classes: $\Gamma_1, \dots, \Gamma_K$
- Set of possible augmentations Aug
 - For each point x_i , define an augmentation set $\text{Aug}(x_i) = \{a(x_i) | a \in \text{Aug}\}$
- During training, SSL methods directly or indirectly minimize the distance between representations of augmentations of an input (alignment)
- Alignment loss for a single input x_i :

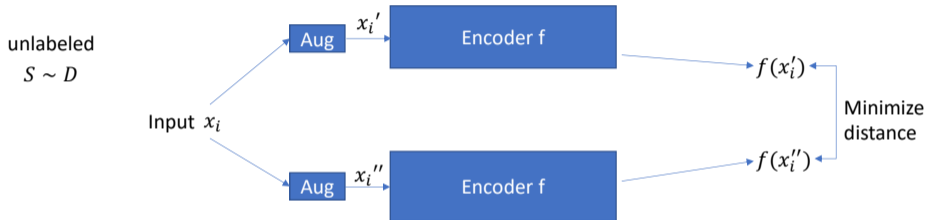
$$\mathcal{L}_{\text{align}}(f, x_i) = \mathbb{E}_{x_i', x_i'' \sim \text{Aug}(x_i)} [d(f(x_i'), f(x_i''))]$$

Model of SSL



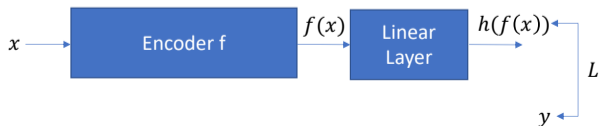
Model of SSL

SSL Training



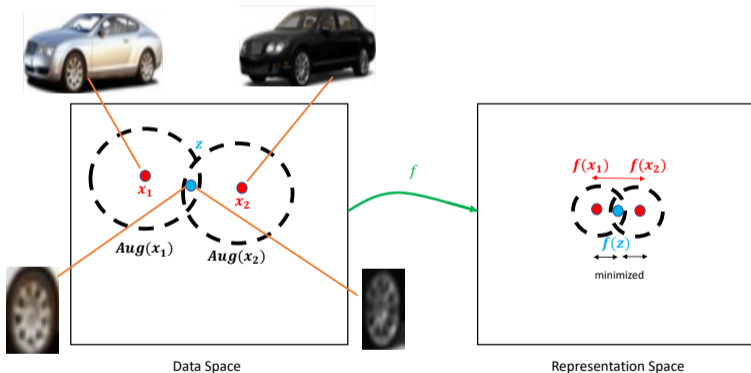
Downstream Task

labeled $S_{down} \sim D$



Augmentations

- Key intuition: Similar datapoints often have overlapping augmentation sets
 - Minimizing alignment within an augmentation set indirectly leads to minimizing distance between representations of similar images (triangle inequality)



Prior Work and Motivation

Memorization in Supervised Learning

- Standard definition based on leave-one-out approach
- Consider training two models f and g on dataset S with learning algorithm \mathcal{A}
- g is trained without a specific datapoint x
- Large difference between predictions of f and g on x indicates memorization since high impact on model
- Definition focuses on label memorization

$$m(x) = \Pr_{f \sim \mathcal{A}(S)} [f(x) = y] - \Pr_{g \sim \mathcal{A}(S \setminus x)} [g(x) = y]$$

- Note: Probability computed over possible outcomes of \mathcal{A}

Memorization in SSL

- Fundamentally different setting due to lack of labels
 - Existing definition or methods based around it do not carry over directly
- Recent work (Meehan et al., 2023) has started exploring memorization in SSL
 - Method based on correlations between representation of the crop of an image and representations of images from the same class
 - Strong assumptions: requires access to labeled data from same distribution
 - Relies on a particular augmentation in SSL (cropping) - does not carry over to SSL algorithms in general
 - Does not provide a score of memorization, only a binary result
- Main motivation: Propose a unified definition of memorization in SSL

Proposed Method

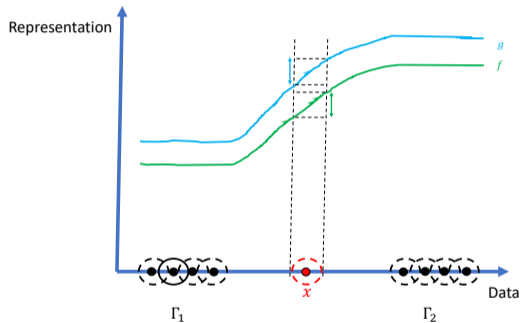
Definition

- Based on leave-one-out approach
- As alternative to labels, use alignment loss due to its importance in SSL
- Compare alignment loss of encoders f and g on x
- Larger difference signifies higher impact on training: higher memorization score

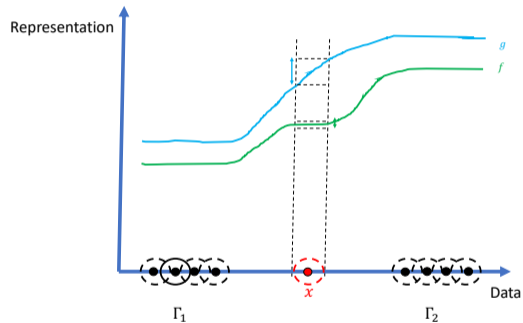
$$m(x) = \mathbb{E}_{g \sim \mathcal{A}(S \setminus x)} \mathbb{E}_{x', x'' \sim \text{Aug}(x)} [d(g(x'), g(x''))] - \mathbb{E}_{f \sim \mathcal{A}(S)} \mathbb{E}_{x', x'' \sim \text{Aug}(x)} [d(f(x'), f(x''))].$$

Intuition

- Consider 1 dimensional data, representation space



Datapoint with low memorization



Datapoint with high memorization

Experimental Results

Examples of Memorized Datapoints

- With a trained encoder, memorization scores were estimated for all training datapoints
- Samples ranked by memorization scores

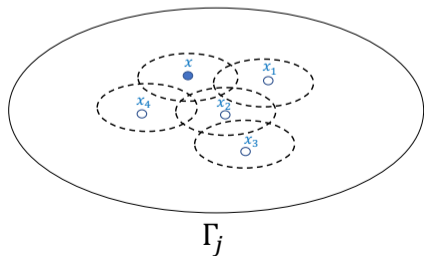


Examples of datapoints by memorization score, MNIST class 3 and 6.

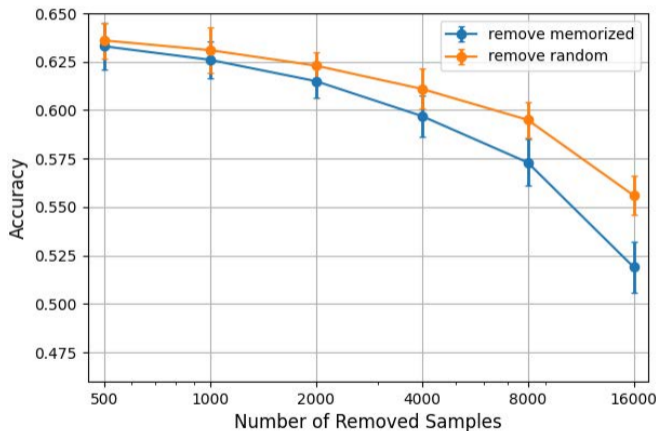
- As expected, atypical examples generally have higher memorization scores
- Observation: Many datapoints with high memorization scores across different SSL methods and datasets

Outlier Datapoints, Generalization

- Hypothesis: memorization of datapoints from outlier subpopulations helps reduce generalization error (similar to supervised learning)
- Two ways to define generalization error of encoder: focusing on generalization on downstream tasks (Huang et al., 2023)
- Consider an outlier latent class Γ_j with a single datapoint x in training dataset S
- Memorization may help in achieving lower alignment in region around x and thus encourage representations of points in Γ_j close to $f(x)$: better generalization of encoder on Γ_j



Outlier Datapoints, Generalization (cont'd)







Effect of removing memorized vs random datapoints on downstream accuracy

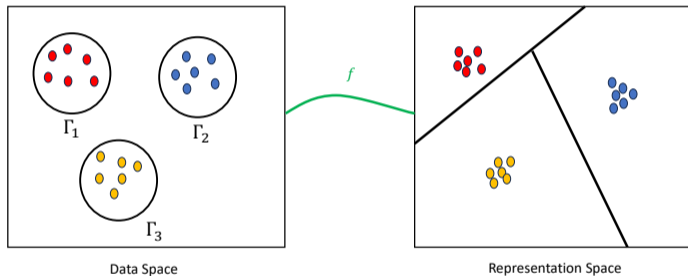
Next Steps

- Theoretical analysis for relationships between memorization and generalization similar to Feldman, 2020
- Considering alternative gradient based definitions of memorization similar to supervised setting (Zielinski et al., 2020)
- More practical estimators of leave-one-out definition
- Applications beyond vision based SSL methods

Bibliography I

-  Feldman, Vitaly (2020). “Does learning require memorization? a short tale about a long tail”. In: *Proceedings of the 52nd Annual ACM SIGACT Symposium on Theory of Computing*, pp. 954–959.
-  Huang, Weiran, Mingyang Yi, Xuyang Zhao, and Zihao Jiang (2023). “Towards the Generalization of Contrastive Self-Supervised Learning”. In: *The Eleventh International Conference on Learning Representations*. URL: <https://openreview.net/forum?id=XDJwuEYHhme>.
-  Meehan, Casey, Florian Bordes, Pascal Vincent, Kamalika Chaudhuri, and Chuan Guo (2023). “Do SSL Models Have Déjà Vu? A Case of Unintended Memorization in Self-supervised Learning”. In: *arXiv e-prints*, arXiv–2304.
-  Zielinski, Piotr, Shankar Krishnan, and Satrajit Chatterjee (2020). “Weak and strong gradient directions: Explaining memorization, generalization, and hardness of examples at scale”. In: *arXiv preprint arXiv:2003.07422*.

Augmentations (cont'd)



Linear Separability of representations makes downstream tasks easier