Comparative Analysis of Contrastive Learning Methods Against Data Poisoning

Mohammad Akbarnezhad **Adithya Embar** **Dylan Gunn** **Anthony Holmes**

miles85@ucla.edu, aembar@ucla.edu, dylangunn@ucla.edu, anthonyjholmes1@ucla.edu
Department of Computer Science, UCLA, Los Angeles, CA, 90024.

Abstract

In recent years, the proliferation of data poisoning has emerged as a significant concern, casting a formidable shadow over the security of machine learning models. Over the past decade, we have witnessed a substantial increase in malicious poisoning attacks aimed at compromising the integrity of machine learning models. In our extensive experiments, we conducted a thorough evaluation of contrastive learning frameworks, specifically MoCo and SimCLR, using both clean and perturbed versions of the CIFAR-10 dataset with the intention of determining the most important aspects of a model that might contribute to its robustness against adversarial poisoning, specifically focusing on variations in data augmentation methods, encoders, and loss functions. Our findings, however, indicate that there is not a statistically significant difference in performance degradation between the two models. This suggests that the impact of these factors on contrastive learning may either be inconsequential or more intricate than initially presumed.

For reference, our code is available at : `https://github.com/anthonyjholmes1/com-sci-260D-Fall23`

Keywords: Contrastive Learning, Self-Supervised Learning, Data Poisoning, Robustness, Core-Set Selection, Poisoning Attacks.

1 Introduction

The advancements in contrastive learning techniques have propelled the capabilities of deep neural networks by fostering the extraction of rich and discriminative representations from unlabeled data. Despite their promise, the susceptibility of these models to poisoning attacks, particularly in the domain of contrastive learning, presents a formidable challenge. Poisoning attacks aim to corrupt the training data, introducing adversarial samples that can severely compromise the model's performance and reliability.

In the quest to fortify against these adversarial threats, this paper embarks on an exploration of the SAS (Subsets that maximize Augmentation Similarity to the full data) method as a core-set selection strategy within the realm of contrastive Semi-Supervised Learning (SSL). While SAS primarily aims to identify a subset of data points that alleviate computational burdens by maximizing similarity to the entire dataset under various augmentations, its performance under poisoned datasets in contrastive learning remains an uncharted territory.

The pivotal objective of this work is to evaluate the robustness and effectiveness of the SAS-selected core-set under the influence of poisoned data in contrastive SSL scenarios. Despite the valuable contributions of SAS in core-set selection for SSL, its behavior and adaptability when exposed to adversarial manipulations through poisoning attacks have yet to be thoroughly investigated in current research endeavors.

Preprint. Under review.

Through comprehensive empirical evaluations, we aim to shed light on the behavior of the SAS-selected subset when faced with poisoned datasets in contrastive learning environments. This investigation not only seeks to elucidate the response of the SAS method under adversarial scenarios but also endeavors to assess its resilience and suitability for practical deployment in scenarios susceptible to data poisoning.

In subsequent sections, this paper will delve into the fundamental principles of contrastive learning, outline the mechanisms behind poisoning attacks in this context, introduce the SAS core-set selection methodology, and present an extensive evaluation framework designed to examine the performance of SAS under the influence of poisoned datasets in contrastive SSL scenarios. The findings from this evaluation will contribute to a nuanced understanding of the robustness and effectiveness of the SAS method, addressing a critical gap in the current landscape of contrastive SSL research.

2 Related Work

Contrastive Learning Contrastive learning has emerged as a prominent paradigm in the domain of unsupervised representation learning, aiming to extract meaningful and discriminative features from raw data. The fundamental principle underlying contrastive learning revolves around the idea of learning representations by maximizing agreement between similar pairs of data samples while minimizing agreement between dissimilar pairs.

One of the foundational architectures in contrastive learning is the Siamese network, initially introduced for signature verification tasks. Siamese networks consist of twin networks sharing the same architecture and weights, where two input samples are processed through identical networks to generate embeddings. The objective of Siamese networks is to bring similar samples closer together in the embedding space while pushing dissimilar samples apart. This is typically achieved by utilizing contrastive loss functions, such as the contrastive loss introduced by Hadsell et al. in their seminal work Bertinetto et al. [2016]. The contrastive loss penalizes the model when embeddings of similar pairs are farther apart than a specified margin, and when embeddings of dissimilar pairs are closer than this margin.

Another influential development in contrastive learning is InfoNCE (Information Noise Contrastive Estimation), proposed by Oord et al. [2018]. InfoNCE is rooted in the framework of self-supervised learning, where the objective is to learn useful representations from unlabeled data. InfoNCE formulates the contrastive objective based on the mutual information between different views of the same instance. By maximizing the agreement between representations derived from different augmentations of the same data sample while minimizing the agreement between representations from different samples, InfoNCE facilitates the learning of robust and informative representations.

Recent advancements in contrastive learning have witnessed various extensions and improvements to the core principles. Approaches such as Momentum Contrast (MoCo) He et al. [2020], SimCLR (Simple Contrastive Learning Representation) Chen et al. [2020], and BYOL (Bootstrap Your Own Latent) Grill et al. [2020] have demonstrated substantial performance gains by refining the mechanisms of generating positive and negative pairs, designing more effective augmentation strategies, or introducing novel pretext tasks.

Mechanisms Behind Poisoning Attacks in Contrastive Learning Contrastive learning models are vulnerable to poisoning attacks, where adversaries strategically inject malicious samples into the training data to manipulate the model's behavior. Understanding the mechanisms of these poisoning attacks is crucial in fortifying models against such adversarial threats. Adversarial perturbations play a pivotal role in poisoning attacks against contrastive learning models. Adversaries craft perturbations that are imperceptible to the human eye but can substantially alter the learned representations by the model Goodfellow et al. [2014]. These perturbations, when added to training data, aim to deceive the model during the learning process, causing it to misclassify or generalize poorly.

Poisoning attacks in contrastive learning can also manipulate the embedding space and decision boundaries of the model. By strategically placing poisoned instances in the training data, adversaries aim to shift the representations of specific classes or alter the decision boundaries, leading to misclassifications or biases in learned representations Kim et al. [2020]. Another critical aspect is the transferability of poisoning attacks across different models or tasks. Adversarial examples generated to poison a specific contrastive learning model might generalize and cause disruptions in other models, making them susceptible to similar attacks Papernot et al. [2016]. Understanding this transferability is essential for devising robust defense strategies.

Poisoning attacks can significantly impact the learned representations and generalization capabilities of contrastive learning models. The introduction of poisoned instances during training can corrupt the learned representations, leading to compromised generalization performance and reduced model robustness against unseen data Muñoz-González et al. [2017]. Adversaries continually evolve attack strategies to evade existing defense mechanisms. Robust defense against poisoning attacks in contrastive learning requires a comprehensive understanding of these evolving adversarial techniques and the development of proactive defense strategies Biggio et al. [2013].

SAS Core-Set Selection in Contrastive SSL Finding examples that significantly contribute to contrastive Semi-Supervised Learning (SSL) is notably challenging compared to core-set selection in supervised learning. Methods in supervised learning rely on loss or confidence of predictions, requiring labeled data. Contrastingly, SSL lacks labeled data, making it difficult to identify crucial examples.

SAS Joshi and Mirzasoleiman [2023] addresses this challenge by maximizing alignment between augmented views within a class and minimizing dissimilarity between views of different classes. These examples, pivotal for contrastive SSL, pull together instances within a class and maintain learned class representation centers. Surprisingly, Joshi and Mirzasoleiman [2023] observed that examples vital for contrastive SSL have less impact on supervised learning. Therefore, they concluded that high-confidence, low-forgetting-score examples can be safely excluded from supervised learning without affecting accuracy, while difficult-to-learn examples crucial in supervised learning might hinder contrastive SSL performance.

Extensive evaluations by Joshi and Mirzasoleiman [2023] demonstrated SAS's efficacy across various datasets (e.g., CIFAR, STL10, TinyImageNet) and contrastive learning methods (e.g., SimCLR, BYOL). SAS subsets consistently outperform random subsets by over 3% in downstream performance, efficiently extracting subsets critical for SSL early in training or using smaller proxy models. Nevertheless, a significant inquiry remains unaddressed is the performance assessment of the SAS-selected subset under a poisoned dataset is yet to be explored. This gap underscores the necessity to scrutinize the resilience and efficiency of the SAS-selected subset when exposed to poisoned data, representing a crucial aspect that has not been adequately investigated in current research.

Poisoning Data and Semi-Supervised Learning Resiliency tests carried out by poisoning the unlabeled data part of a dataset and training semi-supervised models on them found that a modification to the dataset as little as 0.1% has the capacity to affect the classification power of the final model. In fact, Carlini claims that more accurate models are more vulnerable to poisoning attacks, dismissing the possibility that improving the models themselves could solve the underlying issue Carlini [2021].

3 Problem Formulation

Clustering and classification methods, especially those employing supervised learning methods, are especially vulnerable to poisoning. By introducing a small subset of permuted or altered data, the accuracy of the entire resultant model trained on that data can be massively affected. Even on the scale of datasets used to train models that do work like spam identification in Gmail and malware detection in crowdsourced antivirus software, a targeted mass attack via submission of false negatives is enough to impact these models' classifications Constantin [2021].

Because contrastive learning methods make use of identified subsets that maximize expected augmentation similarity, the only way that a poisoned dataset can affect its performance is by affecting those learned minimal subsets; essentially, the permutations have to be drastic enough to affect the subsets that are chosen to maximize augmentation similarity in the feature space. The resultant subset is more eloquently described as follows:

$$S_k \quad = \quad argmin_{S \in V, |S| \leq r_k} \sum_{i \in V_k \S_k} \sum_{j \in S_k} d_{i,j}, \quad d_{i,j} =< f(x_i), f(x_j) >$$

Theoretically, this should provide some robustness to dataset corruption and perturbation. While poisoning other types of data, like incorporating uncommon misspellings into a sentiment analysis-performing model, can cause massive reductions in performance, contrastive learning should not be so sensitive. For example, Google's toxicity detector can be nearly completely overridden by slightly changing some words; phrases that rank 90% in "toxicity" easily are reduced to 10-15% toxicity with the addition of basic perturbations Hossein Hosseini and Poovendran [2017].

In our investigation, we aim to assess the robustness of contrastive learning frameworks, including notable methodologies such as MoCo, SimCLR, and CMC. To comprehensively evaluate their resilience against data poisoning, we will explore various key aspects identified in the literature. These critical facets encompass strategies such as Data Augmentation, Parallel Augmentation, Architectural choices, Loss Functions, and Data Modalities. For the scope of our study, we will exclusively focus on two widely utilized datasets, namely CIFAR-10 and CIFAR-10-P, both of which are image-based. As all of our experiments share the same data modality, the consideration of Data Modality variability becomes unnecessary in this particular context. CIFAR-10 was selected as our baseline dataset, providing a consistent reference point with the same model configuration. This choice allows us to establish a solid foundation for performance assessment, which we can then use to validate the outcomes when applied to the CIFAR-10P dataset.

Crucial to our research is the pursuit of a deeper understanding of the significance behind the choices made in data augmentation, parallel augmentation, architectural decisions, and loss function selection within the context of contrastive learning. We seek to elucidate why these choices have a bearing on the robustness of these frameworks.

Additionally, we will investigate whether there exists any discernible correlation between the choice of dataset and any of these key aspects. This examination will provide insights into whether certain datasets are more susceptible to the influence of these factors, shedding light on potential dataset-specific nuances that impact the performance of contrastive learning frameworks.

4 Method

To establish a fair basis for comparison, we made use of publicly-available MoCo and SimCLR models. We attempted to standardize our environment as much as possible to reduce the impact of confounding variables by only using implementations of these models in Torchvision; if the original papers' implementations were done on another framework, we only made use of alternative implementations that had proven equivalent performance on the same datasets, using accuracy as our performance measure, to the original models.

Each of the models was trained for 50 epochs on both CIFAR-10 and CIFAR-10-P datasets. It is important to note that the CIFAR-10-P dataset is a perturbed collection of the original CIFAR-10 dataset. While it is not optimally poisoned and is not targeted on any specific classes, it still functions similarly to an indiscriminate attack on a dataset, working to reduce the overall accuracy of the final model. This is done by applying transformations on the original images, including but not limited to: brightness changes, adding gaussian blur and noise, adding motion blur, layered effects, and more common and easy augmentations like rotation, scaling, and shearing of images.

The use of only 50 epochs for training each of these models may affect the final conclusions of our investigation. In many cases, standardized comparison of final accuracies between contrastive learning models is done on a scale of 200 to 500 epochs, at which point iterative gain is rather inconsequential and potentially causing the model to overfit to its training data. Our restriction on computational resources, being done on only a single GPU and in a time-constrained environment (Google Colab requires constant interaction, or will otherwise close the open instances), made it such that it was unreasonable to attempt anything more. This means that we will lack the tail behavior of training these models for more epochs, potentially causing us to make premature assumptions about the training behavior of these models for both unperturbed and perturbed data. It is also due to limited computational power that these models were trained on CIFAR-10 datasets, as the more commonly used ImageNet datasets are far too large and vastly reduce the maximum batch size that fits in our GPU memory.

To ensure the reproducibility of our work, we have made our research repository openly accessible on GitHub `https://github.com/anthonyjholmes1/com-sci-260D-Fall23`. In this repository, we have provided the Google Colab files used to run our experiments. Setup in individual Google Colab environments allowed easy parallelization of computation and comparison of results. All that is required for reproduction of our tests is for the individual Python notebook files to be imported into either a local or Colab instance and run. Our code and its specific imports rely on the existing Colab functionality; some packages may not be available on a local machine without additional installation.

All tests were run on instances containing single NVIDIA T4 GPUs, via Google Colab.

5 Experiments

We trained and tested two contrastive learning approaches against clean and perturbed imaged based datasets found online (CIFAR-10 and CIFAR-10-P). The contrastive learning methods we used are MoCo and SimCLR. With the help of online implementations, we formulated our experiments and altered them to our specific use-cases. We conducted our experiments with 50 epochs in order to enable us to have multiple passes on the data, preventing any error that may have come about by single poorly-performing training instances. At this point, we started to see diminishing returns in performance gain per epoch and felt it was an the appropriate number, especially given clearly distinct differences in test accuracy at this point. This was also set as a constant to enable comparative analysis even between different contrastive learning models.

After a few iterations of each instance, we got the following results:

MoCo, CIFAR-10: 72.4% accuracy

SimCLR, CIFAR-10: 74.8% accuracy

MoCo, CIFAR-10-P: 42.3% accuracy

SimCLR, CIFAR-10-P: 42.9% accuracy

Interestingly, when comparing the accuracies between the two models on both the CIFAR-10 and CIFAR-10-P datasets, we don't notice a significant difference. Both perform similarly overall, with the different models, for each dataset, having potentially statistically insignificant differences.

This may be compounded by the fact that we are lacking insight into the performance of MoCo and SimCLR on these datasets for more epochs (ideally 200 or more), but to assume that this might be impacting our observations would be a gross extrapolation.

Overall, we see that each model drops in accuracy about 30%. When we calculate the percent decrease using the clean results as a normalizer, we conclude a 41.6% degradation of performance on MoCo and a 42.6% degradation on SimCLR.

6 Conclusion

In our experiments with clean datasets, SimCLR consistently outperformed MoCo, showcasing a significantly higher accuracy rate of 74.8% compared to MoCo's 72.4%. This suggests that SimCLR, in this specific setting, demonstrates superior performance, at least for low-epoch training. When subjected to perturbed datasets, both MoCo and SimCLR experienced a notable decrease in accuracy, reflecting the vulnerability of these frameworks to data poisoning. While SimCLR maintained a higher average accuracy of 42.9% in this scenario, and MoCo followed closely with an average of 42.3%, it is evident that both methods are susceptible to perturbations. Relative to their performance on the clean datasets, MoCo was affected by 41.6% and SimCLR by 42.6%.

Additionally, it is noteworthy that when we compared the models against each other, considering various factors such as data augmentation strategies, encoders, and loss functions, we observed that neither SimCLR nor MoCo consistently outperformed the other by a significant margin. This suggests that the single choice of any these factors may not be the sole determinant of superior performance in the face of poisoned data, and the interplay of these elements in contrastive learning frameworks may be more complex and nuanced than initially anticipated. This intriguing finding calls for further exploration into the methods by which contrastive learning models may become more robust to both corruption and adversarial perturbation. Interestingly, these findings are at least consistent with Carlini's conjecture that more accurate models are more vulnerable to poisoning attacks.

Our results contribute valuable insights into the broader context of contrastive learning research. They emphasize the need for further investigation into the robustness of contrastive learning frameworks when confronted with perturbed data, a critical consideration in real-world applications. Additionally, our findings underscore the relevance of selecting appropriate contrastive learning methods based on the specific problem and dataset, highlighting SimCLR as a strong candidate in certain image-based scenarios. However, at least amongst the models tested, there is no recommendation to be made for robustness on all datasets.

While our study provides valuable insights, it is not without limitations. First, our experiments were conducted on a specific subset of datasets, and the generalizability of our findings to other domains may vary. Future research should extend these investigations to diverse datasets and problem settings. Furthermore, we acknowledge that our study did not delve deeply into the mechanisms causing the observed performance disparities. Future research avenues should focus on understanding the underlying reasons behind these differences, potentially involving architectural and hyperparameter analyses. It would also be useful to repeat these experiments for more epochs, to see if there is a divergence that might be statistically significant once the models have better fitted.

In conclusion, our research contributes to the ongoing discourse on contrastive learning robustness, offering a foundation for further exploration in diverse contexts and datasets. The pursuit of robust machine learning models remains an imperative task, and our findings illuminate important considerations in this endeavor.

7 Contributions

Mohammad Akbarnezhad

1. Created content for the presentation
2. Wrote the Introduction section
3. Wrote the Related Work section

Adithya Embar

1. Created content for the presentation
2. Presented work
3. Wrote the Experiments section

Dylan Gunn

1. Wrote the project code
2. Created content for the presentation
3. Presented work
4. Contributed to the Related Work section
5. Wrote the Methodology section
6. Contributed to the Experiments section
7. Contributed to the Conclusion section
8. General final editing

Anthony Holmes

1. Created content for the presentation
2. Presented work
3. Wrote the Abstract
4. Wrote the Problem Formulation section
5. Contributed to the Methodology section
6. Contributed to the Experiments section
7. Wrote the Conclusion section
8. Wrote the project code

References

Luca Bertinetto, Jack Valmadre, Joao F Henriques, Andrea Vedaldi, and Philip HS Torr. Fully-convolutional siamese networks for object tracking. In *Computer Vision–ECCV 2016 Workshops: Amsterdam, The Netherlands, October 8-10 and 15-16, 2016, Proceedings, Part II 14*, pages 850–865. Springer, 2016.

Battista Biggio, Igino Corona, Davide Maiorca, Blaine Nelson, Nedim Šrndić, Pavel Laskov, Giorgio Giacinto, and Fabio Roli. Evasion attacks against machine learning at test time. In *Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2013, Prague, Czech Republic, September 23-27, 2013, Proceedings, Part III 13*, pages 387–402. Springer, 2013.

Nicholas Carlini. Poisoning the unlabeled dataset of Semi-Supervised learning. In *30th USENIX Security Symposium (USENIX Security 21)*, pages 1577–1592. USENIX Association, August 2021. ISBN 978-1-939133-24-3. URL `https://www.usenix.org/conference/usenixsecurity21/presentation/carlini-poisoning`.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. *arXiv preprint arXiv:2002.05709*, 2020.

Lucian Constantin. How data poisoning attacks corrupt machine learning models. *CSO Online*, 2021.

Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014.

Jean-Bastien Grill, Florian Strub, Florent Altche, Corentin Tallec, Pierre H Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Daniel Guo, and Mohammad Gheshlaghi Azar. Bootstrap your own latent: A new approach to self-supervised learning. *arXiv preprint arXiv:2006.07733*, 2020.

Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9729–9738, 2020.

Baosen Zhang Hossein Hosseini, Sreeram Kannan and Radha Poovendran. Deceiving google's perspective api built for detecting toxic comments. *arXiv preprint arXiv:1702.08138*, 2017. URL `https://arxiv.org/pdf/1702.08138.pdf`.

Siddharth Joshi and Baharan Mirzasoleiman. Data-efficient contrastive self-supervised learning: Most beneficial examples for supervised learning contribute the least. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett, editors, *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 15356–15370. PMLR, 23–29 Jul 2023. URL `https://proceedings.mlr.press/v202/joshi23b.html`.

Minseon Kim, Jihoon Tack, and Sung Ju Hwang. Adversarial self-supervised contrastive learning. *Advances in Neural Information Processing Systems*, 33:2983–2994, 2020.

Luis Muñoz-González, Battista Biggio, Ambra Demontis, Andrea Paudice, Vasin Wongrassamee, Emil C Lupu, and Fabio Roli. Towards poisoning of deep learning algorithms with back-gradient optimization. In *Proceedings of the 10th ACM workshop on artificial intelligence and security*, pages 27–38, 2017.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

Nicolas Papernot, Patrick McDaniel, and Ian Goodfellow. Transferability in machine learning: from phenomena to black-box attacks using adversarial samples. *arXiv preprint arXiv:1605.07277*, 2016.