Online Self-Supervised Contrastive Learning for Time-Series Data

Abinav Chari
Georgia Institute of Technology
achari7@gatech.edu

Siddharth Pamidi
Georgia Institute of Technology
spamidi30@gatech.edu

1 Overview

Selfsupervised learning (SSL) has been shown to be a powerful method to learning representations in large-scale, unlabeled data. It has demonstrated success with all kinds of data, including time-series data. In particular, contrastive learning (CL) has proven to be a powerful means of learning representations through the use of positive and negative pairs of data. However, training these encoders is often very computationally expensive, and for data which follow a non-stationary distribution, requires constant retraining to maintain its efficacy.

In this paper, we explore the application of online updates to a contrastive encoder model based on streaming time-series data from a non-stationary distribution, aiming to compare the quality of representations of offline and online methods. We begin by introducing the offline contrastive model (SimCLR), then discuss the changes made to adapt it to an online setting.

2 Related Works

1. Arora, S., Khandeparkar, H., Khodak, M., Plevrakis, O., & Saunshi, N. (2019). A Theoretical Analysis of Contrastive Unsupervised Representation Learning. https://arxiv.org/abs/1902.09229.

   This paper provides a theoretical framework to analyze performance of offline representation learning algorithms based on sampling similar pairs of data from latent classes. It also proves a generalization bound for contrastive learning the typical classification task based on Rademacher complexity. This framework will be a great starting point for a theoretical analysis of online representation learning algorithms; by introducing a notion of regret to this existing framework, we can explore how online versions of these algorithms perform.

2. Bahroun, Y., & Soltoggio, A. (2017). Online representation learning with single and multi-layer hebbian networks for image classification. In Artificial Neural Networks and Machine Learning–ICANN 2017: 26th International Conference on Artificial Neural Networks, Alghero, Italy, September 11-14, 2017, Proceedings, Part I 26 (pp. 354-363). Springer International Publishing.

   This paper details online representation learning using Hebbian learning rules for Image classification. They use a single-layer hebbian network for feature extraction and then a multi-layer hebbian network for feature learning. This relates to our project because they make use of online learning by processing the data as it arrives rather than all at once. This setup is useful for our application of stock market prediction where the data will arrive sequentially rather than all at once or in batches.

Submitted to 34th Conference on Neural Information Processing Systems (NeurIPS 2020). Do not distribute.

3. Blaauwbroek, L., Olsak, M., Rute, J., Schaposnik Massolo, F.I., Piepenbrock, J. &; Pestun, V.. (2024). Graph2Tac: Online Representation Learning of Formal Math Concepts. *Proceedings of the 41st International Conference on Machine Learning*, *in Proceedings of Machine Learning Research* 235:4046-4076 Available from https://proceedings.mlr.press/v235/blaauwbroek24a.html.

   The paper details an online representation framework for formal math. Their goal is to learn representations of of formal math concepts through online learning. The authors use a graph based neural network approach to represent formal math concepts where nodes represent terms, propositions and edges represent relationships this helps the model learn representations required for proving theorems. It relates to our project because it they model supports online learning as the model improves as more proofs and concepts are introduced. Additionally this paper gives us insight to online learning in non-stationary data in this case the increasing growth of math concepts.

4. Hoi, S. C., Sahoo, D., Lu, J., & Zhao, P. (2021). Online learning: A comprehensive survey. Neurocomputing, 459, 249-289.

   This paper highlights the applications of Online learning which is beneficial to check if online learning is a valid strategy to approach tasks with non-distribution properties like the stock market. The paper outlines regret minimization, which is a common objective in online learning algorithms. Regret measures the difference between the cumulative loss of the algorithm and the loss of the best fixed decision in hindsight. The paper also goes into details of the algorithms we intend to use like Online Gradient Descent, Passive-Agressive algorithms, and Follow-The-Regularized-Leader (FTRL).

3 Problem Overview

Vanilla SSL for univariate time series has the following formulation: given a dataset $\mathcal{D} = \{x_i\}_{i=1,\dots,N}$ of N samples, where $x_i \in \mathbb{R}^d$. An SSL model aims to learn latent representations $z_i = f_\theta(x_i)$ for time series samples, where f is a function parametrized by θ. The goal is to produce representations which can be then applied to several different downstream tasks such as regression, classification, and clustering.

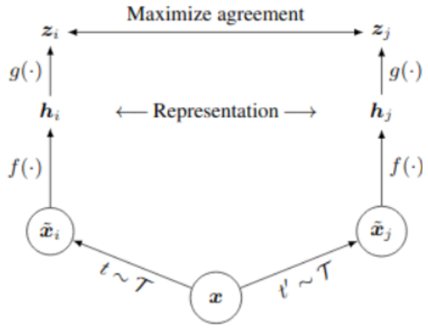
The novelty of this approach lies in the fact that labeled data is not required to learn the latent representations; rather, the SSL model learns to map key "features" in the input data to the new representation space, thus allowing for better performance in downstream tasks.

However, in this paper, we also introduce a stream of inputs x_t, which is sampled from a non-stationary distribution. In other words, as the distribution moves, our original dataset will not properly represent the distribution. Typically, this requires retraining models on newer data, but this can often be very expensive. In this paper, we explore updating our latent representations based on these streamed data concurrently with inference to avoid expensive retraining.

3.1 Contrastive Learning

3.1.1 Pre-training

Contrastive learning aims to produce these latent representations with the property that similar, or positive, pairs are close together while dissimilar, or negative, pairs are far apart. As shown in the figure below, the samples are mapped to its learned latent representation, and then a contrastive loss function is applied, which minimizes the distance between positive pairs and maximizes the distance between negative pairs.



Positive pairs are typically generated through augmentations of input data, such as jittering, scaling, flipping, and masking. The idea is that, if the transformation is subtle enough, the augmented data should contain the same underlying features as the original, and thus should be similarly encoded in the latent space. Negative pairs are chosen to be any other sample and their augmentations. In this paper, pretraining is done using just jittering, which involves adding noise sampled from a Normal distribution to the original sample. The contrastive loss used in this paper is the NT-Xent loss, which is defined by

$$\mathcal{L}(z_i, z_i', z_j) = \frac{-1}{N} \sum_{i=1}^{N} log(\frac{\exp(sim(z_i, z_i'))}{\sum_{j=1, j \neq i}^{2N} \exp(sim(z_i, z_i'))})$$

where sim is the cosine similarity function and z_i, z_i', z_j are the latent representations of the sample, its augmentations/positive pairs, and negative pairs, respectively.

3.1.2 Fine-tuning

Fine-tuning to a specific downstream task requires some labeled dataset $\{(x_i, y_i)_{i=1,\ldots,N}\}$. The goal is to train a classifier g_ϕ parametrized by ϕ which minimizes the loss when comparing $g_\phi(z_i)$ and y_i. After fine-tuning, the encoder and classifier combined can be used for inference on a given input sample x_i by computing $g_\phi(f_\theta(x_i))$.

3.2 Online Contrastive Learning

Typically, once ϕ and θ are chosen through pretraining and fine-tuning respectively, they remain constant forever. However, our approach is to make constant updates to both each time inference is performed on a new input sample from the data stream.

The online learning problem is as follows: in each time step t, the learner picks an action, in this case, a set of parameters θ_t. Simultaneously, the environment selects an input sample x_t. The learner incurs a loss based on a predetermined function $\mathcal{L}(\theta_t, x_t)$, observes this loss, and updates θ for the next step.

Based on this problem statement, we can formulate a notion for regret:

$$R = \sum_{i=1}^{T} \mathcal{L}(\theta_i, x_i) - \min_{\theta} \sum_{i=1}^{T} \mathcal{L}(\theta, x_i)$$

Intuitively, this is the difference between the online algorithm's choices and the best fixed parameters in hindsight. If we substitute contrastive loss for the loss function in the regret formula, we could find an algorithm where the latent representations produced will approach the best contrastive loss in hindsight. In this paper, we utilize the Follow-the-Regularized-Leader (FTRL) update rule (McMahan 2011),

$$\theta_{t+1} = \arg\min_{\theta} (\eta R(\theta) + \theta \cdot \sum_{i=1}^{t} \nabla \mathcal{L}_i(\theta_i))$$

where R is some convex regularization function and $\mathcal{L}$ is the NT-Xent loss function defined above. Note that, normally, a batch of M datapoints is selected to compute NT-Xent loss: $M-1$ of these datapoints will be randomly sampled from the pretraining dataset and $\{x_i\}_{i=1,\ldots,t-1}$, with the last remaining sample being x_t.

4 Experiment

We conduct experiments with S&P 500 index price data. Stock prices are classic examples of a non-stationary process, which is why it is used in this study. Our dataset contains OHLCV and adjusted volume for every day from 1927 to 2020. Using this data, we construct rolling windows of 5 days of data, resulting in an input dimension of 30. The label for each window is a binary signal of whether the next price chronologically is larger or smaller. We split the training data into pre-training (1927-1980) and validation set (1981-2020). The pre-training dataset was shuffled before training while the validation set order was maintained in order to simulate the online environment. The finetune dataset is included as a subset of the pretraining set, which is partially labeled at a ratio of 0.1. We will use classification accuracy and other metrics as a proxy to compare the quality of representations.

Our implementation of the encoder mirrors the SimCLR paper (Chen et. al 2020), with 2-layer transformer blocks and a classifier with 2 fully-connected layers. The jittering augmentation used for positive pairs samples noise from a Standard Normal distribution.

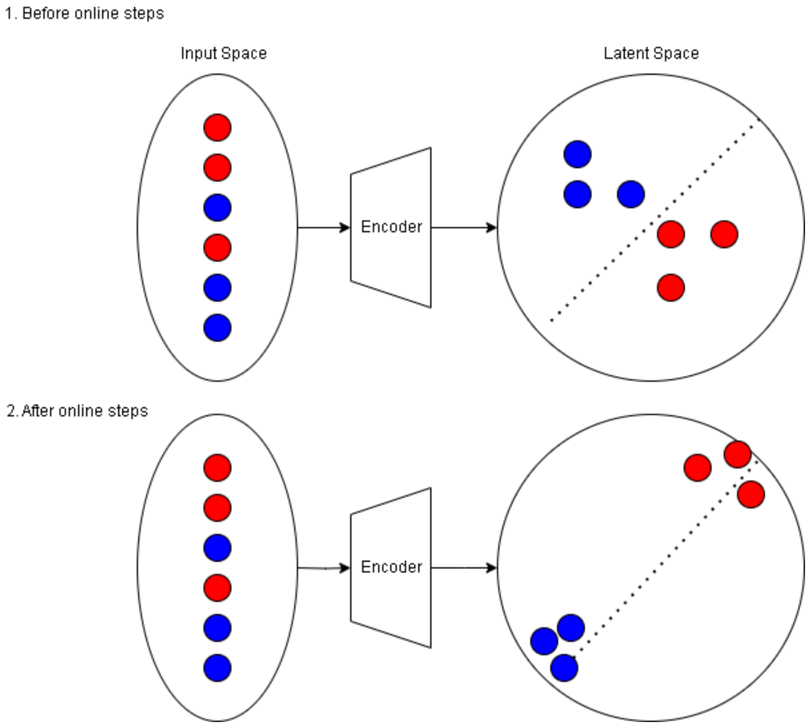
The pre-training lasted for 200 epochs, and we chose the model which achieved the lowest loss when tuning hyperparameters.

Table 1: Performance comparison between SimCLR and online SimCLR (O-SimCLR)

Model	Accuracy	Precision	Recall	F1 Score
SimCLR	0.7328	0.7394	0.7315	0.7364
O-SimCLR	0.7296	0.7322	0.7280	0.7301

5 Discussion

As shown in Table 1, O-SimCLR has about the same performance as SimCLR. THis would suggest that O-SimCLR is producing representations with similar or worse quality to SimCLR. However, this may not be the case; since we are using the classification metrics as a proxy, it might not be the latent representations which are suffering from non-stationarity, but the classifier itself. Furthermore, if the classifier is not able to adapt to the change in latent representation, the accuracy of the classifier might be lower even when the contrastive properties of the updated latent space are improved, as shown in Figure 2.



Although the encoder better separates the samples with different patterns, the classifier isn't fine-tuned to this change in representation, resulting in worse classification. Thus, these results could suggest that the one-step online updates for the classifier are not able to keep up with the change in representation space.

6 Future Work

This paper provides some insight into a very expansive and practical problem, but limitations in our approach give room for future exploration.

Firstly, due to time constraints, we only tested on one dataset. For a better analysis on the efficacy of online learned representations, testing on other datasets could be beneficial. Problems like anomaly detection or other financial modeling would be logical next steps for time series contrastive learning.

Another possible future direction would be to explore how generative-based representation learning can be adapted to a similar online setting, such as VAEs, and how they compare to online contrastive learning frameworks.

Lastly, building on online generative representation learning, one particular problem that would have numerous practical uses is in language models; language models encode text into a high-dimensional latent vector whose direction holds features of the original text. However, especially when dealing with Internet data, words can often change meaning or use over time, such as with slang and pop culture. Thus, being able to update representations of words in an online fashion (without retraining) could be an interesting extension of this, especially with the high cost of retraining large language models.

7 Conclusion

In this paper, we presented a novel way to combat chronological non-stationarity in input data for contrastive representation learning. Our approach of making gradient-based weight updates to the encoder during the evaluation process was inspired by a regret-based analysis on the contrastive loss minimization procedure, which involved using the FTRL algorithm to produce a sublinear regret solution. Our findings give a starting point for further research into online representation learning.

References

[1] Alexander, J.A. & Mozer, M.C. (1995) Template-based algorithms for connectionist rule extraction. In G. Tesauro, D.S. Touretzky and T.K. Leen (eds.), *Advances in Neural Information Processing Systems 7*, pp. 609–616. Cambridge, MA: MIT Press.

[2] Bower, J.M. & Beeman, D. (1995) *The Book of GENESIS: Exploring Realistic Neural Models with the GEneral NEural SImulation System.* New York: TELOS/Springer–Verlag.

[3] Chen, T. et. al (2020). A Simple Framework for Contrastive Learning of Visual Representations. *https://arxiv.org/abs/2002.05709*

[4] Hasselmo, M.E., Schnell, E. & Barkai, E. (1995) Dynamics of learning and recall at excitatory recurrent synapses and cholinergic modulation in rat hippocampal region CA3. *Journal of Neuroscience* **15**(7):5249-5262.

[5] McMahan, B. (2011) Follow-the-Regularized-Leader and Mirror Descent: Equivalence Theorems and L1 Regularization. *Volume 15 of JMLR: W&CP*