Cybersecurity & AI

Freie Universität Berlin

An Investigation into the Performance of Non-Contrastive Self-Supervised Learning Methods for Network Intrusion Detection

Hamed Fard, Tobias Schalau, Gerhard Wunder
Cybersecurity and AI Research Group,
Freie Universität Berlin

27.08.2024

Motivation

- Network intrusion detection (NID): reliance on supervised learning algorithms in the past two decades
  - Limited to detecting only known anomalies
  - Labeled data requirement: costly and time-consuming
- Inspired by recent success of Self-Supervised Learning (SSL) in computer vision:
  - Rising interest in adapting this paradigm for NID
- Prior research: mainly delved into contrastive SSL
- Efficacy of non-contrastive methods in conjunction with encoder architectures and augmentation methods remains unclear
  - Single non-contrastive model outperforms all baselines
  - Preference towards specific augmentation strategies
  - No systematic comparison of non-contrastive SSL methods similar to computer vision

Main Contribution

- Performance comparison of five non-contrastive SSL models with three encoder architectures and six augmentation methods
- Ninety experiments conducted on two network intrusion detection datasets: UNSW-NB15 and 5G-NIDD
- For each self-supervised model: combination of encoder architecture and augmentation method yielding highest average precision, recall, F1-score, and AUCROC reported
- Best-performing combinations compared to two unsupervised baselines: DeepSVDD and Autoencoder

Brief Recap on SSL

- SSL leverages patterns and structure in data to learn meaningful representations without labeled data
- SSL with joint-embedding architecture: learning representations invariant to various distortions
  - Seek to generate similar embeddings for different augmented views of the same sample
- Contrastive: define positive and negative sample pairs through data augmentation
  - Aims to bring positive pairs' embeddings closer while pushing negative pairs' embeddings further apart
  - Requires comparing each sample with many others to work effectively
- Non-Contrastive: require no negative samples, differing primarily in how they avoid representation collapse
  - Architectural modification
  - Loss function design

Augmentation Methods

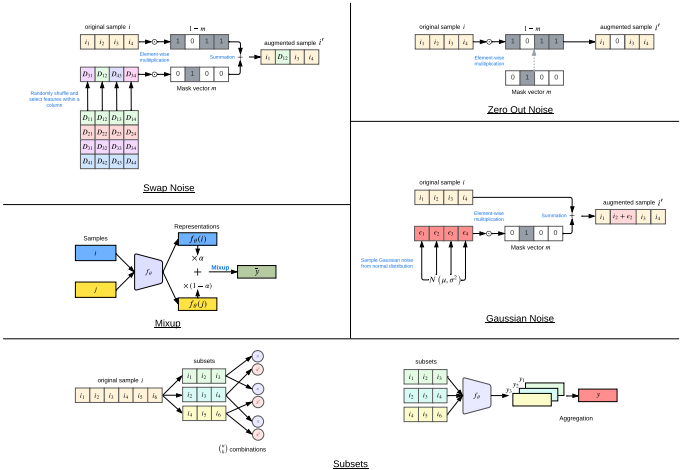

Figure: Visualisation of different augmentation methods.

Random Shuffle Augmentation

Algorithm 1 Pseudocode of Fisher-Yates inspired *Random Shuffle* augmentation method.

Require: $i_j = \{f_j^{(1)}, f_j^{(2)}, \ldots, f_j^{(d_D)}\}$ ▷ network traffic sample i_j composed of d_D features
  for $k = d_D - 1$ to 0 **do**
    $p \leftarrow random\_integer(0, k)$
    $i_j^{(p)}, i_j^{(k)} = i_j^{(k)}, i_j^{(p)}$
  end for

Encoders

- MLP:
  - Four fully connected layers (input layer, two hidden layers, output layer)
  - Batch normalization and ReLU applied after first three layers
  - First layer input dimension equals number of input features; remaining layers set to 256

- CNN:

Table: Architecture of the CNN encoder. *conv* is a convolutional layer followed by a ReLU activation function, and pooling represents a pooling layer. For each layer, the kernel size, number of filters (only for convolutional layers), input shape, and output shape are given for an example network traffic sample with 196 features.

Layer	Kernel	Filter	Input	Output
conv1	1×2	32	$1 \times 196 \times 1$	$1 \times 195 \times 32$
conv2	1×2	64	$1 \times 195 \times 32$	$1 \times 194 \times 64$
conv3	1×2	128	$1 \times 194 \times 64$	$1 \times 193 \times 128$
pooling	1×3	$-$	$1 \times 193 \times 128$	$1 \times 64 \times 128$
conv4	1×2	256	$1 \times 64 \times 128$	$1 \times 63 \times 256$
pooling	1×2	$-$	$1 \times 63 \times 256$	$1 \times 31 \times 256$
conv5	1×2	512	$1 \times 31 \times 256$	$1 \times 30 \times 512$
pooling	1×4	$-$	$1 \times 30 \times 512$	$1 \times 7 \times 512$

Encoders

- Feature Tokenizer Transformer (FT-T):
  - Processes numerical/categorical features, stacks embeddings to form final input featue embedding
  - Set embedding dimension of the feature tokenizer to 32,
  - 4 Multi-Head Self-Attention heads,
  - 4 Transformer layers
  - Transformer encoder output flattened for further computations
  - Added dropout value of 0.1 to attention and feedforward sub-layers

SSL Models

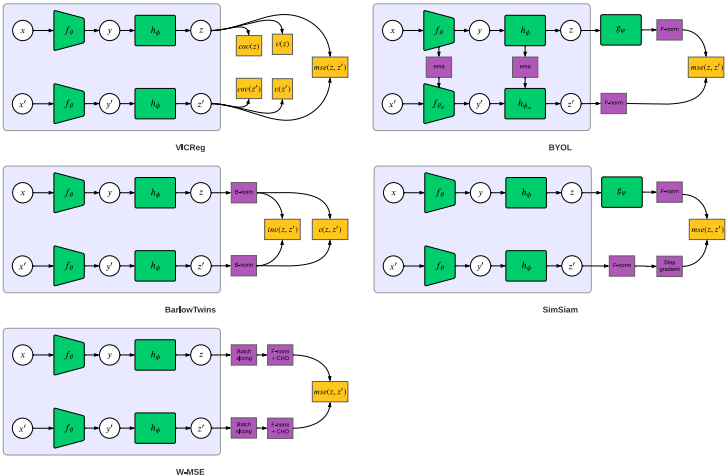

Figure: Comparison of non-contrastive SSL models: architecture and loss function.

Classifier

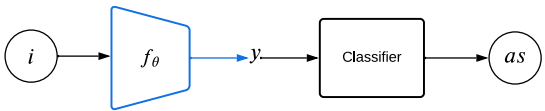

Figure: K-means classifier generates an anomaly score (as) for a network traffic sample i.

- After training, only the encoder f_θ is retained; all other model parts discarded
- Encoder weights frozen; simple classifier trained on top frozen data representation

Evaluation Protocol and Metrics

- Training set consists exclusively of normal data
- Test set includes both normal data and anomalies
- Positive class consistently defined as the anomalous class
- Performance metrics include:
  - AUROC (threshold-independent)
  - Precision
  - Recall (detection rate)
  - F1-Score
- Each combination of augmentation method, encoder architecture, and SSL model as a distinct experiment, executed in 10 runs to account for statistical uncertainty
- Mean and standard deviation calculated over all runs

Hyperparameter Optimization

- Three sets of variable hyperparameters: model-specific, augmentation, and general training (learning rate, epochs)
- Hyperparameter optimization conducted per dataset, model, augmentation, and encoder using Tune
- ADAM optimizer consistently used for optimization in all runs
- Initial learning rate set to 1×10^{-4} with a maximum of 200 epochs
- Optimal model and augmentation parameters determined using BayesOptSearch with 200 trials; ; followed by grid search on learning rates

Datasets

- UNSW-NB15 (well-known in NID community)
- 5G-NIDD:
  - Generation of benign traffic by actual mobile devices in the network
  - Benign traffic: HTTP, HTTPS, SSH, and SFTP
  - Two attack categories:
  - Port Scan (including SYN Scan, TCP Connect Scan, UDP Scan)
  - Dos/DDoS (covering ICMP flood, UDP flood, SYN flood, HTTP flood, Slow rate DoS - Slowloris, Slow rate DoS - Torshammer)

Preprocessing

- Datasets cleaned by: removing NaN values, dropping duplicated features and samples
- Normalized feature values
- Categorical features one-hot encoded

Table: General information on the datasets after preprocessing,

Dataset	Number of Samples	Number of Features	Attack Ratio
UNSW-NB15	154 098	196	0.4437
5G-NIDD	1 215 655	58	0.6072

Results

Table: Comparison of non-contrastive SSL models with the highest average performance metrics (all with standard deviation) on the UNSW-NB15 dataset.

Model	ENC	AUG	Precision	Recall	F1-Score	AUROC
BYOL	FT-T	GN	0.720 ± 0.021	0.776 ± 0.024	0.747 ± 0.022	0.704 ± 0.037
SimSiam	FT-T	ZON	0.762 ± 0.049	$\mathbf{0.823 \pm 0.053}$	0.791 ± 0.051	0.762 ± 0.048
VICReg	MLP	S	$\mathbf{0.788 \pm 0.059}$	0.810 ± 0.062	$\mathbf{0.798 \pm 0.056}$	$\mathbf{0.786 \pm 0.071}$
BarlowTwins	MLP	S	0.783 ± 0.066	0.809 ± 0.053	0.795 ± 0.056	0.764 ± 0.105
W-MSE	CNN	M	0.763 ± 0.031	0.806 ± 0.019	0.784 ± 0.018	0.756 ± 0.043

Results

- BYOL shows lower performance on UNSW-NB15 dataset compared to other models
- Gaussian Noise augmentation yields best result for BYOL despite being previously deemed unsuitable
- Random Shuffle augmentation consistently underperforms across models and is excluded from results
- Swap Noise augmentation, used with VICReg and MLP encoder, also excluded due to poor performance
- VICReg achieves highest average precision, F1-Score, and AUCROC with Subsets augmentation
- Barlow Twins achieves best performance metrics with MLP encoder and Subsets augmentation
- SimSiam combined with Zero Out Noise and FT-Transformer achieves highest detection rate

Results

Table: Comparison of non-contrastive SSL models with the highest average performance metrics (all with standard deviation) on the 5G-NIDD dataset.

Model	ENC	AUG	Precision	Recall	F1-Score	AUROC
BYOL	CNN	SN	0.867 ± 0.015	0.911 ± 0.017	0.888 ± 0.013	0.775 ± 0.017
SimSiam	CNN	S	0.841 ± 0.070	0.877 ± 0.070	0.858 ± 0.069	0.724 ± 0.134
VICReg	CNN	GN	$\mathbf{0.932 \pm 0.010}$	$\mathbf{0.961 \pm 0.026}$	$\mathbf{0.946 \pm 0.009}$	0.908 ± 0.005
BarlowTwins	MLP	M	0.916 ± 0.012	0.909 ± 0.012	0.912 ± 0.009	$\mathbf{0.925 \pm 0.009}$
W-MSE	MLP	M	0.836 ± 0.054	0.891 ± 0.057	0.863 ± 0.056	0.756 ± 0.077

Results

- VICReg with CNN encoder and Gaussian Noise achieved highest average precision, recall, and F1-Score among non-contrastive SSL models, showcasing Gaussian Noise viability
- Exclusion of Gaussian Noise in [31] possibly due to insufficient hyperparameter optimization or suboptimal combinations with other augmentations
- Random Shuffle consistently underperformed and is excluded from Table
- Barlow Twins with MLP encoder and Mixup augmentation achieved highest AUCROC
- Mixup with MLP encoder achieved highest metrics for W-MSE on this dataset

Results

Table: Comparison of non-contrastive SSL models with the highest average performance metrics (all with standard deviation) on the UNSW-NB15 dataset against unsupervised models on the same dataset.

Model	ENC	AUG	Precision	Recall	F1-Score	AUROC
SimSiam	FT-T	ZON	0.762 ± 0.049	0.823 ± 0.053	0.791 ± 0.051	0.762 ± 0.048
VICReg	MLP	S	$\mathbf{0.788 \pm 0.059}$	0.810 ± 0.062	0.798 ± 0.056	0.786 ± 0.071
DeepSVDD	—	—	0.683 ± 0.021	0.735 ± 0.025	0.708 ± 0.023	0.656 ± 0.047
AE	—	—	0.786 ± 0.013	$\mathbf{0.837 \pm 0.029}$	$\mathbf{0.811 \pm 0.018}$	$\mathbf{0.793 \pm 0.024}$

Results

Table: Comparison of non-contrastive SSL models with the highest average performance metrics (all with standard deviation) on the 5G-NIDD dataset against unsupervised models on the same dataset.

Model	ENC	AUG	Precision	Recall	F1-Score	AUROC
VICReg	CNN	GN	0.932 ± 0.010	0.961 ± 0.026	0.946 ± 0.009	0.908 ± 0.005
BarlowTwins	MLP	M	0.916 ± 0.012	0.909 ± 0.012	0.912 ± 0.009	0.925 ± 0.009
DeepSVDD	—	—	0.895 ± 0.060	0.937 ± 0.055	0.915 ± 0.057	0.865 ± 0.117
AE	—	—	$\mathbf{0.939 \pm 0.027}$	$\mathbf{0.965 \pm 0.018}$	$\mathbf{0.951 \pm 0.020}$	$\mathbf{0.932 \pm 0.020}$

Results

- In the UNSW-NB15 dataset, VICReg achieves highest average precision; AE consistently outperforms non-contrastive SSL models on other metrics
- On the 5G-NIDD dataset, AE attains highest average performance metrics
- DeepSVDD shows less favorable results
- Lack of hyperparameter tuning details in prior studies may lead to misleading confidence in non-contrastive SSL model comparisons

Limitations and Outlook

- Experiments highlight the importance of augmentation methods but reveal two drawbacks:
  - Not tailored for NID
  - May violate domain constraints and generate unrealistic samples
- Limitation: feature selection/elimination methods could significantly reduce processing time
- Future research: designing NID-specific augmentation methods that satisfy domain constraints for SSL methods
- Future work: explore improved distance metrics like Mahalanobis distance for K-means detector
- Incorporate advanced unsupervised detectors (Isolation Forest, OCSVM) to close the performance gap between non-contrastive SSL models and reconstruction-based approaches

THANK YOU