Orange
Innovation

CATS: Contrastive learning for Anomaly detection in Time Series

2024 IEEE Conference on Big Data – Washington DC, USA

December 15 – 18, 2024

Joël Roman Ky[1,2], Bertrand Mathieu[1], Abdelkader Lahmadi[2], Raouf Boutaba[3]

Orange Innovation Lannion[1]

Université de Lorraine, CNRS, LORIA[2]

David R. Cheriton School of Computer Science, University of Waterloo[3]

orange™

Outline

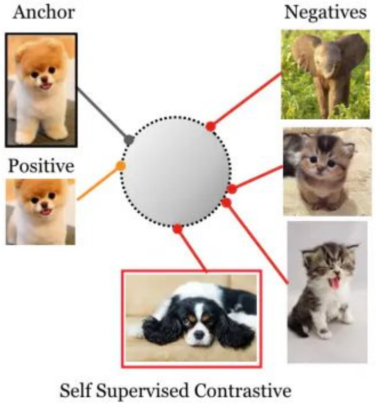
1. Context & Motivation

- Anomaly Detection (AD) is essential in a wide variety of applications.
  - AD reveals an importance for low-latency applications (Cloud Gaming or CloudVR) to be able to detect QoE deterioration as part of french ANR MOSAICO project.

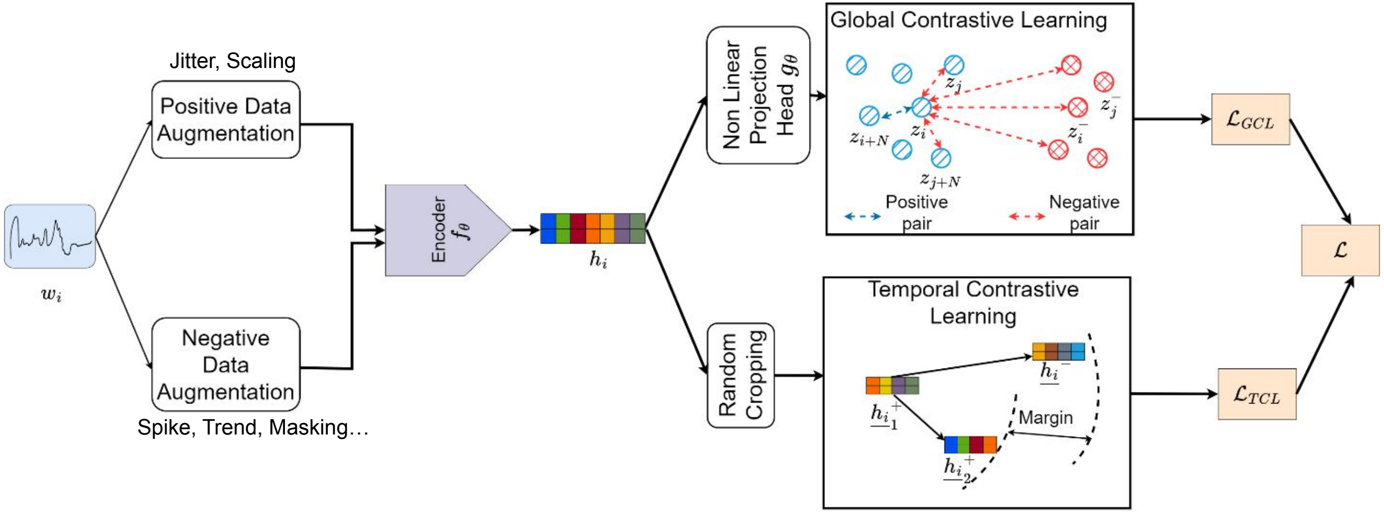
- Current unsupervised AD techniques for time series have some limitations.
  - Low performance
  - Impact of data contamination

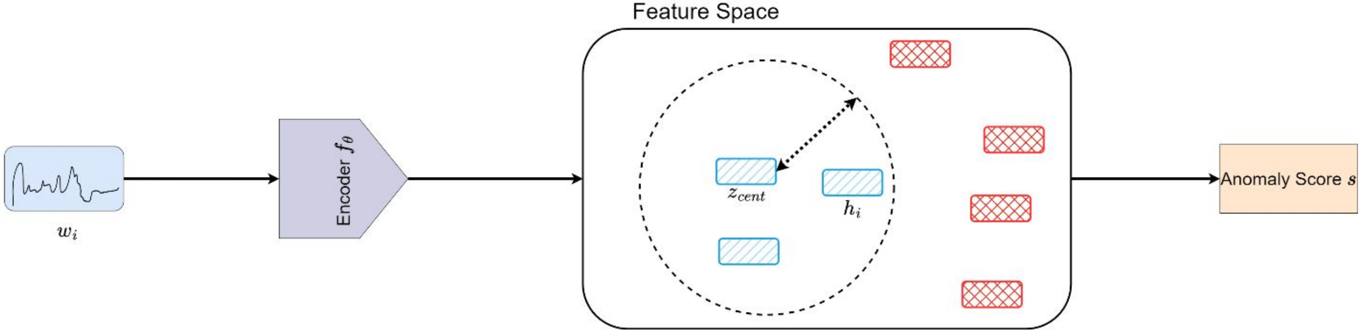
- Contrastive Learning (CL) proved its efficiency on many tasks on image, text and is now leveraged for time-series and network data.

- Contributions for time-series anomaly detection:
  - Use negative data augmentation techniques for time-series to be considered as anomalies (anomaly injection)
  - Consider temporal dependencies with a novel (Dynamic Time Warping) DTW-based temporal loss



2. CATS: Contrastive learning for Anomaly detection in Time Series



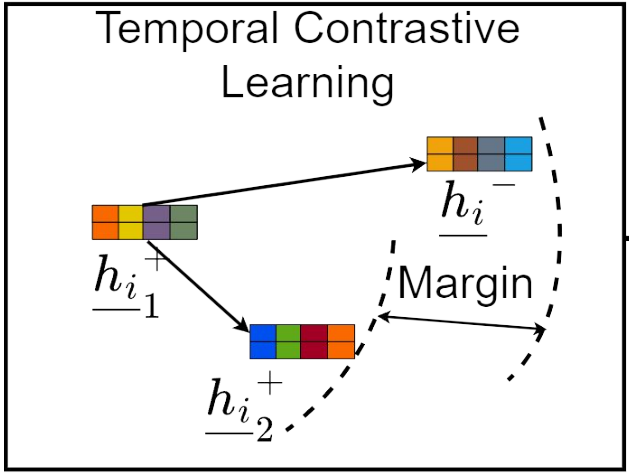


2-1. Temporal Contrastive Learning (TCL)

- **Dynamic Time Warping (DTW): a similarity measure between time series that seeks for the temporal alignment that minimizes Euclidean distance between aligned series.**
  - **However, DTW is not differentiable.**
  - **Soft-DTW was introduced using the soft-min operator to make DTW differentiable.**
- **TCL learns a temporal representation using a triplet loss with Soft-DTW and is defined as follows:**

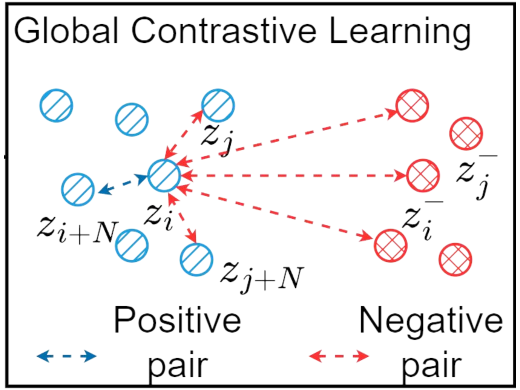
$$\mathrm{L}_{TCL} = \sum_{i=1}^{N} \max(d(h_i, h_i^+) - d(h_i, h_i^-) + m, 0)$$

$$d(h_i, h_j) = softDTW(h_i, h_j) - \frac{1}{2}(softDTW(h_i, h_i) + softDTW(h_j, h_j))$$



2-2. Global Contrastive Loss (GCL)

- **GCL learn representations at the instance level using the NT-Xent loss while considering more negative pairs.**
  - **NT-Xent loss consider two views of same instance as positive and view of different instances as negative.**
  - **GCL also include the views generated through negative data augmentation.**
  - **Consequently, instead of contrasting one positive pair and N-1 negative pairs in NT-Xent =, GCL contrasts one pair and 2N-1 negative pairs.**



$$L_{GCL} = -\frac{1}{2N} \sum_{i \in B_a \cup B^+} \log \frac{\exp(sim(z_i, z_i^+)/\tau)}{\sum_{j \in B \text{ and } j \neq i} \exp(sim(z_i, z_j)/\tau)}$$

$$sim(z_i, z_j) = \frac{z_i^T z_j}{\|z_j\| \|z_i\|}$$

$$B = \{B_a, B^+, B^-\}$$

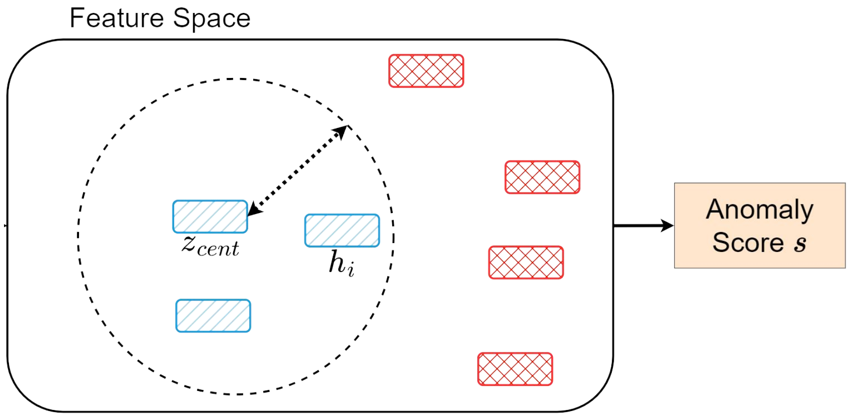
2-3. Anomaly detection

- **After training, we assume that the encoder has learn sufficient information to be efficient for our downstream task (AD).**

- **Anomaly can be identified using a simple anomaly score computed as follows:**

$$s(w_t) = D(f_\theta(w_t), z_{cent})$$

$$z_{cent} = \frac{1}{N_{train}} \sum h_i$$



3. Experimental results

- **Datasets:**
  - Cloud Gaming QoE/QoS datasets: (STD, GFN, XC)
  - Benchmark datasets (SMD, SMAP, MSL)
- **Evaluations metrics:**
  - F1-score
  - AUPR
  - MCC (Matthews Coefficient Correlation)
- **Competing solutions:**
  - iForest
  - Deep-SVDD, AE, USAD
  - SimCLR, SimSiam, TS2Vec
- **Experiments:**
  - Performance comparison
  - Ablation studies
  - Data contamination
  - Hyper-parameters influence

3-1. Experimental results: Performance

TABLE II: Performance comparison on the datasets. Mean and standard deviation computed over all entities for benchmark datasets and over five runs for case-study datasets. Bold values indicate best results.

		Models	IForest	Deep-SVDD	AE	USAD	SimCLR	SimSiam	TS2Vec	CATS
Benchmark datasets	SMD	AUC	77.10 (± 11.9)	75.31 (± 14.5)	81.33 (± 13.2)	81.08 (± 12.5)	80.83 (± 14.7)	77.26 (± 14.9)	74.25 (± 16.6)	**82.21** (± 14.3)
		F1	29.88 (± 20.6)	34.75 (± 21.5)	46.00 (± 24.3)	46.62 (± 26.3)	46.51 (± 25.7)	41.82 (± 25.3)	43.18 (± 25.9)	**50.65** (± 23.6)
		MCC	29.62 (± 20.8)	36.25 (± 22.0)	47.00 (± 24.1)	47.98 (± 25.1)	48.06 (± 24.2)	43.24 (± 24.9)	44.95 (± 24.7)	**50.85** (± 23.6)
	MSL	AUC	56.94 (± 14.1)	61.38 (± 17.1)	62.30 (± 16.1)	63.31 (± 14.3)	61.09 (± 15.4)	62.07 (± 14.3)	63.95 (± 15.0)	**64.98** (± 15.7)
		F1	21.24 (± 21.4)	27.93 (± 25.6)	26.02 (± 22.9)	27.16 (± 23.0)	25.72 (± 23.1)	23.78 (± 23.2)	28.43 (± 24.5)	**29.15** (± 24.2)
		MCC	11.09 (± 21.8)	19.24 (± 29.2)	16.49 (± 24.4)	17.33 (± 24.8)	16.30 (± 25.1)	14.11 (± 24.0)	19.86 (± 24.8)	**20.14** (± 27.8)
	SMAP	AUC	56.98 (± 17.3)	62.52 (± 19.1)	**64.30** (± 19.6)	61.11 (± 19.4)	63.99 (± 17.7)	62.12 (± 17.1)	61.42 (± 20.3)	64.07 (± 18.6)
		F1	22.80 (± 27.2)	29.20 (± 33.0)	28.93 (± 33.5)	**30.10** (± 33.1)	28.23 (± 32.2)	27.46 (± 33.2)	28.26 (± 33.2)	29.07 (± 29.07)
		MCC	11.38 (± 29.0)	23.93 (± 33.1)	23.96 (± 34.0)	23.66 (± 34.9)	22.52 (± 32.2)	21.44 (± 32.5)	23.5 (± 32.8)	**24.28** (± 32.7)
Case-study datasets	STD	AUC	74.57 (± 1.63)	91.19 (± 1.08)	96.04 (± 0.27)	96.09 (± 0.08)	95.78 (± 0.39)	75.65 (± 11.3)	95.63 (± 1.94)	**97.93** (± 0.13)
		F1	75.79 (± 1.42)	87.18 (± 1.24)	90.35 (± 0.51)	90.02 (± 0.24)	90.15 (± 0.52)	74.21 (± 9.22)	92.83 (± 1.92)	**94.06** (± 0.45)
		MCC	39.56 (± 3.66)	71.83 (± 2.77)	78.93 (± 1.14)	77.89 (± 0.36)	78.48 (± 1.17)	39.31 (± 19.8)	84.33 (± 4.12)	**86.72** (± 0.88)
	GFN	AUC	61.97 (± 0.87)	71.78 (± 3.41)	74.05 (± 0.84)	74.84 (± 0.42)	78.50 (± 1.95)	67.07 (± 3.25)	74.91 (± 4.32)	**84.35** (± 1.23)
		F1	74.12 (± 0.71)	75.51 (± 2.11)	74.05 (± 0.84)	77.80 (± 0.38)	81.20 (± 2.61)	74.25 (± 2.93)	76.76 (± 2.71)	**82.88** (± 0.96)
		MCC	17.07 (± 1.27)	24.26 (± 6.56)	28.08 (± 0.14)	31.40 (± 1.22)	37.46 (± 3.87)	17.86 (± 6.27)	28.19 (± 8.39)	**47.27** (± 1.49)
	XC	AUC	78.71 (± 1.13)	67.32 (± 6.52)	89.18 (± 2.31)	89.97 (± 0.26)	85.81 (± 3.17)	83.35 (± 10.6)	**96.96** (± 1.36)	96.10 (± 0.41)
		F1	63.33 (± 1.18)	50.83 (± 7.69)	75.94 (± 3.30)	77.59 (± 0.58)	70.58 (± 3.45)	69.09 (± 13.4)	**89.60** (± 2.03)	86.69 (± 0.83)
		MCC	43.42 (± 2.43)	27.40 (± 11.4)	63.95 (± 4.63)	65.35 (± 0.72)	56.59 (± 4.89)	52.30 (± 21.3)	**84.07** (± 2.94)	79.67 (± 0.11)

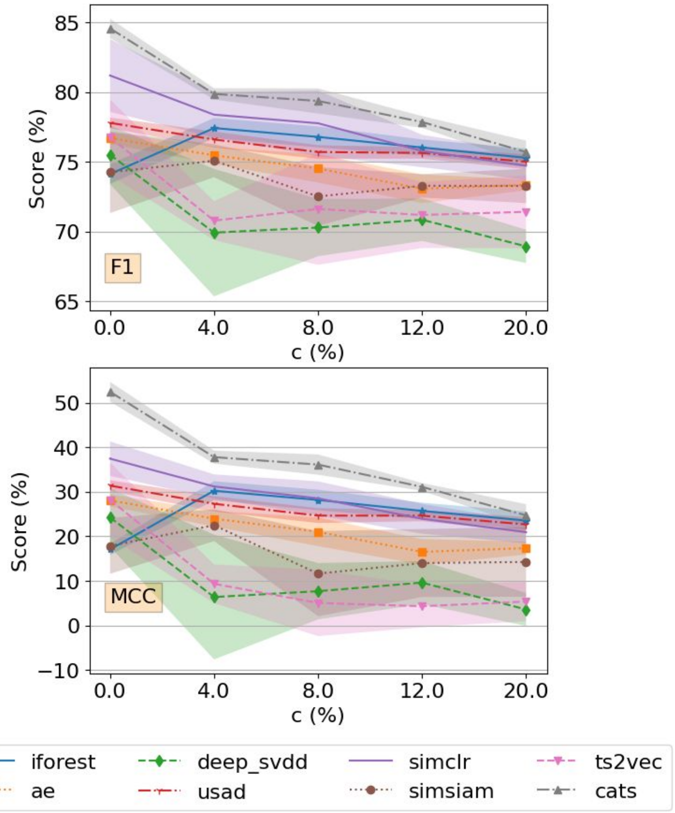
3-2. Experimental results: Ablation study

- **Impact of each loss components**

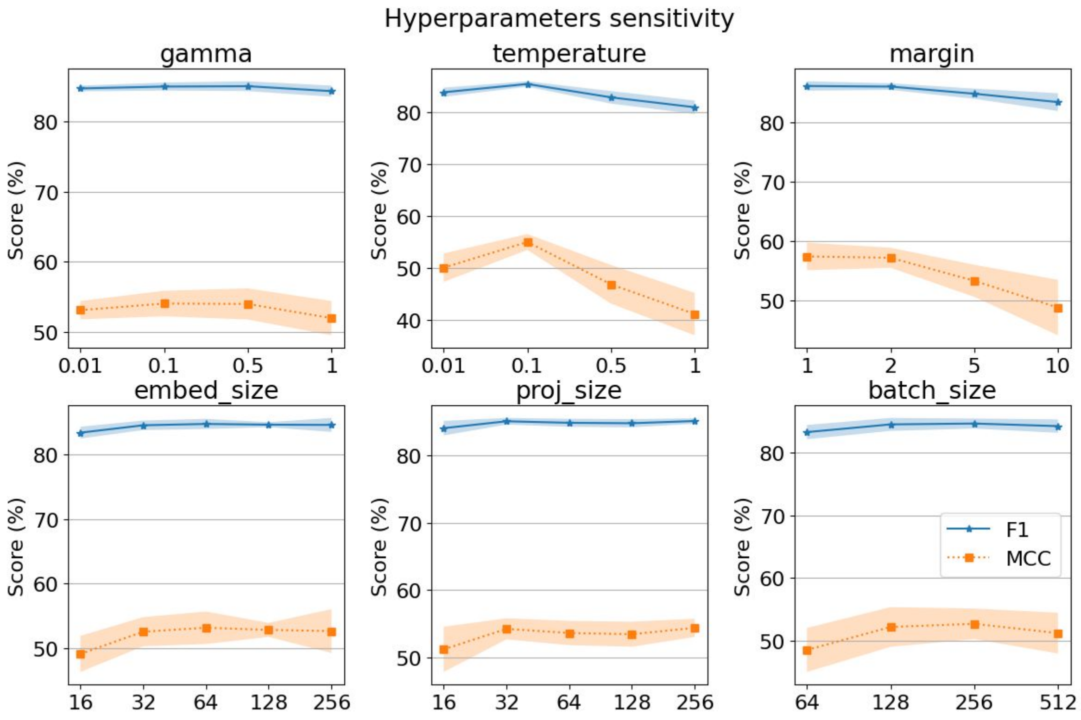
Table 4. Ablation study on loss components.

	GFN		XC	
Loss	F1	MCC	F1	MCC
$\mathcal{L}_{NTXent}$	$81.20_{(\pm 2.61)}$	$37.46_{(\pm 3.87)}$	$70.58_{(\pm 3.45)}$	$56.59_{(\pm 4.89)}$
$\mathcal{L}_{GCL}$	$82.52_{(\pm 1.77)}$	$40.73_{(\pm 2.32)}$	$85.68_{(\pm 2.52)}$	$78.31_{(\pm 3.67)}$
$\mathcal{L}_{TCL}$	$79.93_{(\pm 2.69)}$	$38.12_{(\pm 8.32)}$	$76.57_{(\pm 5.68)}$	$65.71_{(\pm 8.34)}$
$\mathcal{L}_{w/o-crop}$	$80.44_{(\pm 2.01)}$	$33.45_{(\pm 1.96)}$	$85.68_{(\pm 2.52)}$	$78.31_{(\pm 3.67)}$
$\mathcal{L}_{GCL} + \mathcal{L}_{TCL}$	**82.88**$_{(\pm 0.96)}$	**47.27**$_{(\pm 1.49)}$	**86.69**$_{(\pm 0.83)}$	**79.67**$_{(\pm 0.11)}$

3-3. Experimental results: Data contamination



3-4. Experimental results: Hyperparameters impact



4. Conclusion

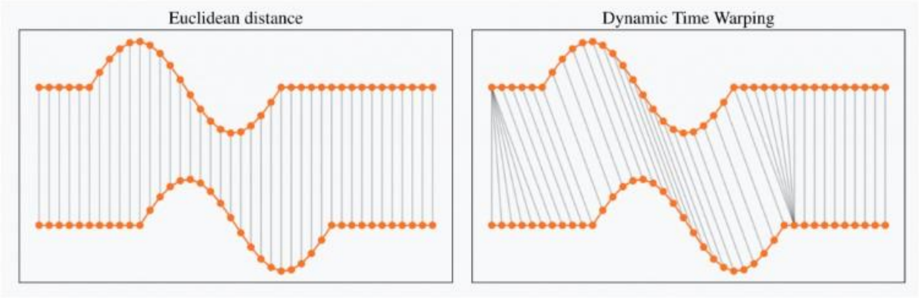
- **CATS addresses the limitations of traditional CL with temporal similarity and negative data augmentation.**
- **Empirical evaluations demonstrate performance in AD tasks on different datasets while being robust to data contamination.**
- **Some limitations remain:**
  - **Increased training time due to the SoftDTW time complexity $O(N^2)$**
  - **Triplet loss in TCL hinders the efficiency of temporal modeling due to the use of 1 negative.**

Orange
Innovation

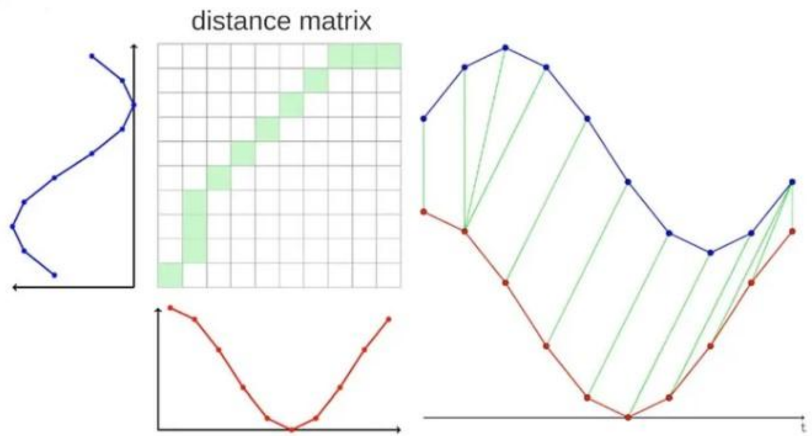
Loria
Laboratoire lorrain de recherche en informatique et ses applications

orange™

Dynamic Time Warping



Dynamic Time Warping (source : https://rtavenar.github.io/blog/dtw.html)

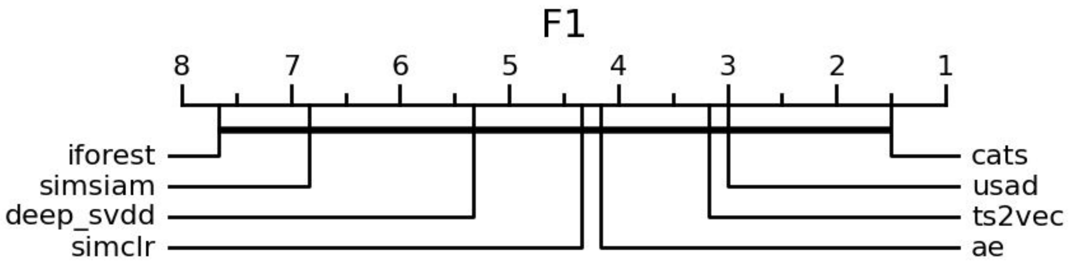


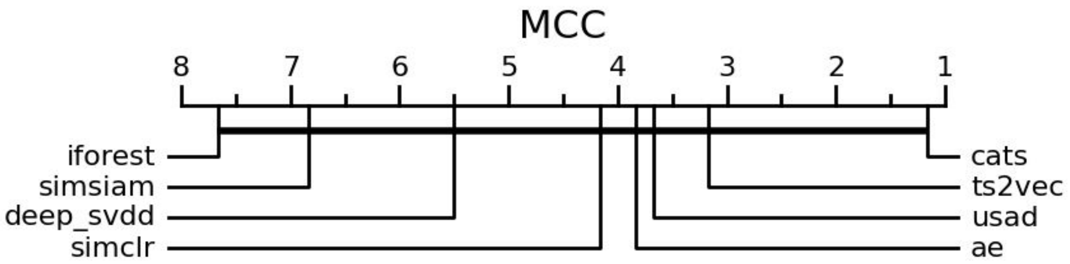
$$DTW_q(x, x') = \min_{\pi \in \mathcal{A}(x,x')} \left(\sum_{(i,j) \in \pi} d(x_i, x'_j)^q \right)^{\frac{1}{q}}$$

$$\text{soft-}DTW^{\gamma}(x, x') = \min_{\pi \in \mathcal{A}(x,x')}{}^{\gamma} \sum_{(i,j) \in \pi} d(x_i, x'_j)^2$$

$$\min{}^{\gamma}(a_1, \ldots, a_n) = -\gamma \log \sum_i e^{-a_i/\gamma}$$

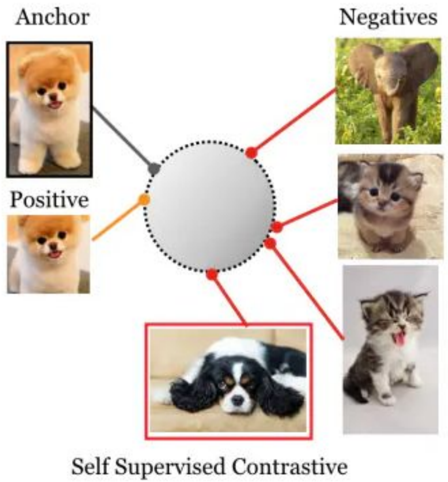
Experimental results: Performance





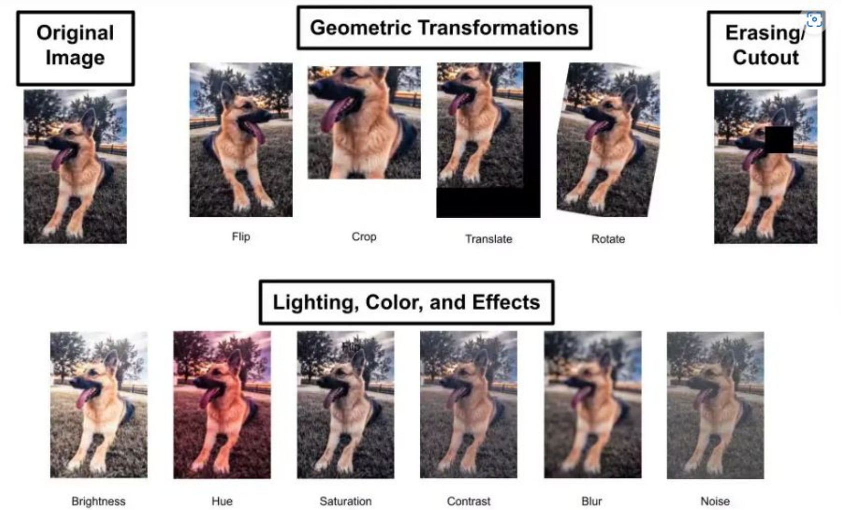
2. Background on Contrastive Learning

- Contrastive Learning (CL) consist in learning representation without label information while ensuring that semantically-similar samples are closed (positives) and far from others (negatives).



2-1. Data augmentation

- The key ingredients to the success of CL are data augmentation and loss functions.
  - Data augmentation generate different views of a sample and then help learn representations by maximizing the similarity of views from the same samples and minimize those of others.

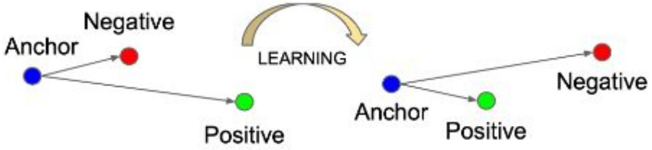


2-2. Loss functions

Some popular CL loss functions are:

- **Triplet loss:**

$$\mathcal{L}(A, P, N) = \max(\| f(A) - f(P)\|_2 - \| f(A) - f(N)\|_2 + \alpha, 0)$$



- **N-pair loss: extension of triplet to many negative samples**

- **NT-Xent loss (proposed in SimCLR): extension of N-pair loss with a temperature parameter to scale cosine similarity**

$$\mathcal{L}(\mathbf{z_i}, \mathbf{z_j}) = -\log \frac{\exp(\mathbf{z_i}\mathbf{z_j}/\tau)}{\sum_{k=1}^{2N} \mathbb{1}_{k \neq i} \exp(\mathbf{z_i}\mathbf{z_k}/\tau)}$$

Self-supervised NT-Xent loss

2-3. Contributions

- Negative sampling is important for CL to avoid collapse issues.
- CL losses do not handle temporal dependencies
- Contributions for time-series anomaly detection:
  - Use negative data augmentation techniques for time-series to be considered as anomalies (anomaly injection)
  - Consider temporal dependencies with a novel (Dynamic Time Warping) DTW-based temporal loss