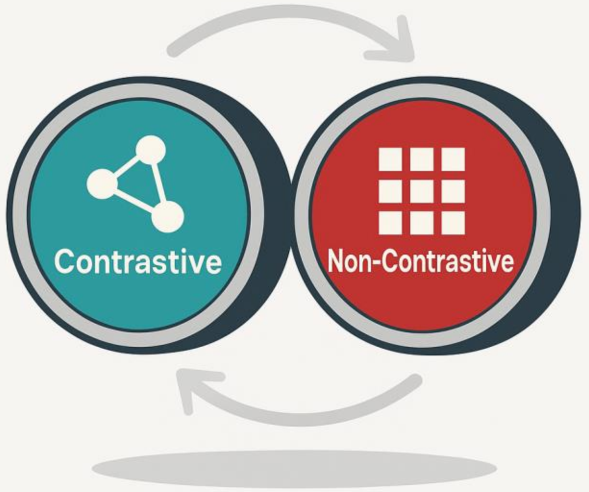
Matteo D'Agostino, EPFL, Lausanne, Switzerland

ON THE DUALITY BETWEEN CONTRASTIVE AND NON-CONTRASTIVE SSL



Published as a conference paper @ ICLR 2023

Garrido, Chen, Bardes, Najman, LeCun

Roadmap

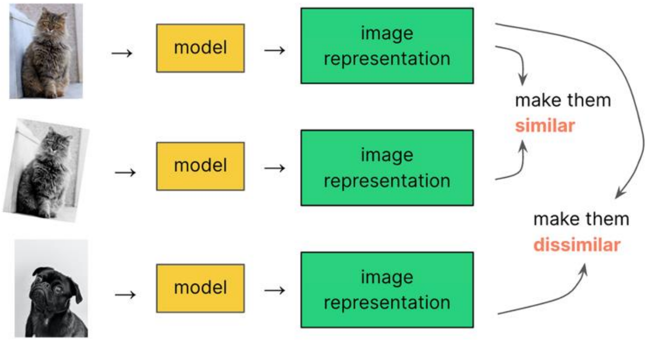
1) Introduction
2) The point of the paper
3) Taxonomy and Framework definition
4) Theoretical equivalence between the two approaches
5) Why do they (apparently) behave differently in practice?
6) Hyperparams Tuning is all you need! (...to unify the SOTA)
7) Conclusion and Takeaways

Introduction

Self-Supervised Learning (SSL)

Contrastive methods
- explicitly push away
- use negative examples
- **SimCLR, DCL**

sample-contrastive methods



Non-Contrastive methods
- no need for negative pairs
- regularization of covariance
- **VICReg, Barlow Twins**

dimension-contrastive methods

$$\mathcal{L}_{\mathrm{DCL}} = \sum_{i=1}^{N} -\log\left(\frac{e^{\mathcal{K}_{\cdot,i}^T \mathcal{K}'_{\cdot,i}/\tau}}{\sum_{j\neq i} e^{\mathcal{K}_{\cdot,i}^T \mathcal{K}_{\cdot,j}/\tau}}\right) = \sum_{i=1}^{N} -\frac{\mathcal{K}_{\cdot,i}^T \mathcal{K}'_{\cdot,i}}{\tau} + \log\left(\sum_{j\neq i} e^{\mathcal{K}_{\cdot,i}^T \mathcal{K}_{\cdot,j}/\tau}\right)$$

$$\mathcal{L}_{\mathrm{SimCLR}} = \sum_{i=1}^{N} -\log\left(\frac{e^{\mathcal{K}_{\cdot,i}^T \mathcal{K}'_{\cdot,i}/\tau}}{e^{\mathcal{K}_{\cdot,i}^T \mathcal{K}'_{\cdot,i}/\tau} + \sum_{j\neq i} e^{\mathcal{K}_{\cdot,i}^T \mathcal{K}_{\cdot,j}/\tau}}\right)$$

"Two sides of the same coin"
- using the Frobenius norm
- assumption of normalization

$$\mathcal{L}_{BT} = \sum_{j=1}^{M} \left(1 - (\mathcal{K}\mathcal{K}'^T)_{j,j}\right)^2 + \lambda \sum_{i,j,i\neq j}^{M} (\mathcal{K}\mathcal{K}'^T)_{j,i}^2$$

$$\mathcal{L}_{VICReg} = \lambda \sum_{i=1}^{N} \|\mathcal{K}_{\cdot,i} - \mathcal{K}'_{\cdot,i}\|_2^2 + \mu\left(v(\mathcal{K}) + v(\mathcal{K}')\right) + \nu\left(c(\mathcal{K}) + c(\mathcal{K}')\right)$$

$$c(\mathcal{K}) = \sum_{i\neq j} \mathrm{Cov}(\mathcal{K})_{i,j}^2 = \|\mathcal{K}\mathcal{K}^T - \mathrm{diag}(\mathcal{K}\mathcal{K}^T)\|_F^2$$

EPFL

The point of the paper

- Unifying the SOTA through the equivalence proof
- Providing a methods interpolation between the two paradigms
- Strengths and weaknesses of empirical implementations
- Clearing up misconceptions in the field

Taxonomy and Framework Definition

Dataset: $\mathcal{D} = \{ d_i \in \mathbb{R}^{c \times h \times w} \}_{i=1}^{N}$

Two augmented views: $x_i,\ x_i'$

Encoders: $f_\theta,\ f_{\theta'} \Rightarrow f_\theta(x_i),\ f_{\theta'}(x_i')$

Projectors: $p_\theta,\ p_{\theta'} \Rightarrow p_\theta(f_\theta(x_i)),\ p_{\theta'}(f_{\theta'}(x_i'))$

Embedding matrices: $\mathcal{K}, \mathcal{K}' \in \mathbb{R}^{M \times N},\ \mathcal{K}_{\cdot,i} = p_\theta(f_\theta(x_i)),\ \mathcal{K}'_{\cdot,i} = p_{\theta'}(f_{\theta'}(x_i'))$

Positive pair: $(x_i,\ x_i')$

Negative pairs (full): $\{\forall j \neq i,\ (x_i, x_j)\} \cup \{\forall j \neq i,\ (x_i, x_j')\}$

Negative pairs (simplified): $\{\forall j \neq i,\ (x_i, x_j)\}$

Definition: $\mathrm{diag}(A)_{ij} = A_{ii}$ if $i = j$, else 0

$$\|A\|_F = \sqrt{\sum_i^m \sum_j^n |a_{ij}|^2} = \sqrt{\mathrm{trace}(A^* A)} = \sqrt{\sum_{i=1}^{\min\{m,n\}} \sigma_i^2(A)},$$

$$L_{SSL} = L_{inv} + L_{reg}$$

Invariance criterion, cos similarity or MSE

Gram

$$L_c = \|\mathcal{K}^T\mathcal{K} - \mathrm{diag}(\mathcal{K}^T\mathcal{K})\|_F^2$$

Covariance

$$L_{nc} = \|\mathcal{K}\mathcal{K}^T - \mathrm{diag}(\mathcal{K}\mathcal{K}^T)\|_F^2$$

This is the source of "difference"

Adpating the methods to the framework

VICReg is Dimension-Contrastive

$$\mathcal{L}_{VICReg} = \lambda \sum_{i=1}^{N} \|\mathcal{K}_{\cdot,i} - \mathcal{K}'_{\cdot,i}\|_2^2 + \mu\left(v(\mathcal{K}) + v(\mathcal{K}')\right) + \nu\left(c(\mathcal{K}) + c(\mathcal{K}')\right)$$

$$c(\mathcal{K}) = \sum_{i\neq j} \mathrm{Cov}(\mathcal{K})_{i,j}^2 = \|\mathcal{K}\mathcal{K}^T - \mathrm{diag}(\mathcal{K}\mathcal{K}^T)\|_F^2 = L_{nc}$$

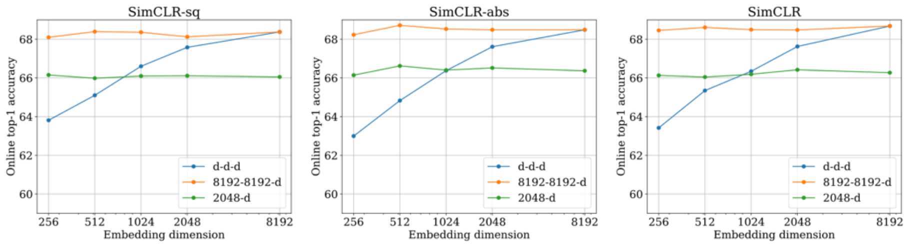
SimCLR is Sample-Contrastive ?

$$\mathcal{L}_{\text{SimCLR}} = \sum_{i=1}^{N} -\log\left(\frac{e^{\mathcal{K}_{\cdot,i}^T\mathcal{K}'_{\cdot,i}/\tau}}{e^{\mathcal{K}_{\cdot,i}^T\mathcal{K}'_{\cdot,i}/\tau} + \sum_{j\neq i} e^{\mathcal{K}_{\cdot,i}^T\mathcal{K}_{\cdot,j}/\tau}}\right)$$

- Cannot be easily linked to L_c
- Relies on cosine sim, not squared errors

Proposition 3.1. Considering an infinite amount of available negative samples, SimCLR and DCL's criteria lead to embeddings where for negative pairs $(x, x^-) \in \mathbb{R}^M$ we have

$$\mathbb{E}\left[x^T x^-\right] = 0 \quad \text{and} \quad \mathrm{Var}\left[x^T x^-\right] = \frac{1}{M}. \tag{2}$$



Equivalence between the two approaches

Theorem

$$L_{nc} + \sum_{j=1}^{M} \|\mathcal{K}_{j,\cdot}\|_2^4 = L_c + \sum_{i=1}^{N} \|\mathcal{K}_{\cdot,i}\|_2^4$$

We have

$$\begin{aligned}
L_{nc} &= \|\mathcal{K}\mathcal{K}^T - \text{diag}(\mathcal{K}\mathcal{K}^T)\|_F^2 && (40)\\
&= tr\left[(\mathcal{K}\mathcal{K}^T - \text{diag}(\mathcal{K}\mathcal{K}^T))^T(\mathcal{K}\mathcal{K}^T - \text{diag}(\mathcal{K}\mathcal{K}^T))\right] && (41)\\
&= tr(\mathcal{K}\mathcal{K}^T\mathcal{K}\mathcal{K}^T) - 2tr(\mathcal{K}\mathcal{K}^T\,\text{diag}(\mathcal{K}\mathcal{K}^T)) + tr(\text{diag}(\mathcal{K}\mathcal{K}^T)\,\text{diag}(\mathcal{K}\mathcal{K}^T)) && (42)\\
&= tr(\mathcal{K}\mathcal{K}^T\mathcal{K}\mathcal{K}^T) - tr(\mathcal{K}\mathcal{K}^T\,\text{diag}(\mathcal{K}\mathcal{K}^T)) && (43)\\
&= tr(\mathcal{K}^T\mathcal{K}\mathcal{K}^T\mathcal{K}) - tr(\mathcal{K}\mathcal{K}^T\,\text{diag}(\mathcal{K}\mathcal{K}^T)). && (44)
\end{aligned}$$

Similarly for L_c, we obtain

$$\begin{aligned}
L_c &= \|\mathcal{K}^T\mathcal{K} - \text{diag}(\mathcal{K}^T\mathcal{K})\|_F^2 && (45)\\
&= tr(\mathcal{K}^T\mathcal{K}\mathcal{K}^T\mathcal{K}) - tr(\mathcal{K}^T\mathcal{K}\,\text{diag}(\mathcal{K}^T\mathcal{K})). && (46)
\end{aligned}$$

Since $\left(\mathcal{K}^T\mathcal{K}\right)_{i,i} = \|\mathcal{K}_{\cdot,i}\|_2^2$ we deduce that $tr(\mathcal{K}^T\mathcal{K}\,\text{diag}(\mathcal{K}^T\mathcal{K})) = \sum_{i=1}^{N}\|K_{\cdot,i}\|_2^4$. Similarly, we obtain that $tr(\mathcal{K}\mathcal{K}^T\,\text{diag}(\mathcal{K}\mathcal{K}^T)) = \sum_{j=1}^{M}\|K_{j,\cdot}\|_2^4$.
Plugging this back in, we finally deduce that

$$L_{nc} = L_c + \sum_{i=1}^{N}\|\mathcal{K}_{\cdot,i}\|_2^4 - \sum_{j=1}^{M}\|\mathcal{K}_{j,\cdot}\|_2^4,$$

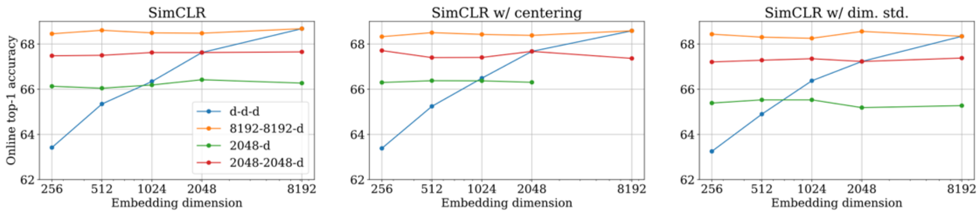
In the **specific case** of K as double stochastic matrix, then:

$$L_{nc} = L_c + N - M$$

Criteria equivalent from an optimization point of view

Normalization is then the culprit?
- Normalising embeddings vs dimensions
- Criteria not far apart in practical scenarios

Importance of optimization and implementation process



Interpolating between methods

$$\text{VICReg} \xrightarrow{\text{LogSumExp}} \text{VICReg-exp} \xrightarrow{\text{Contrastive}} \text{VICReg-ctr} \xrightarrow{\text{Neg. pair sampling}} \text{SimCLR}$$

$$\mathcal{L}_{VICReg-exp} = \lambda \sum_{i=1}^{N} \|\mathcal{K}_{\cdot,i} - \mathcal{K}'_{\cdot,i}\|_2^2 + \mu\left(v(\mathcal{K}) + v(\mathcal{K}')\right) + \nu\left(c_{exp}(\mathcal{K}) + c_{exp}(\mathcal{K}')\right)$$

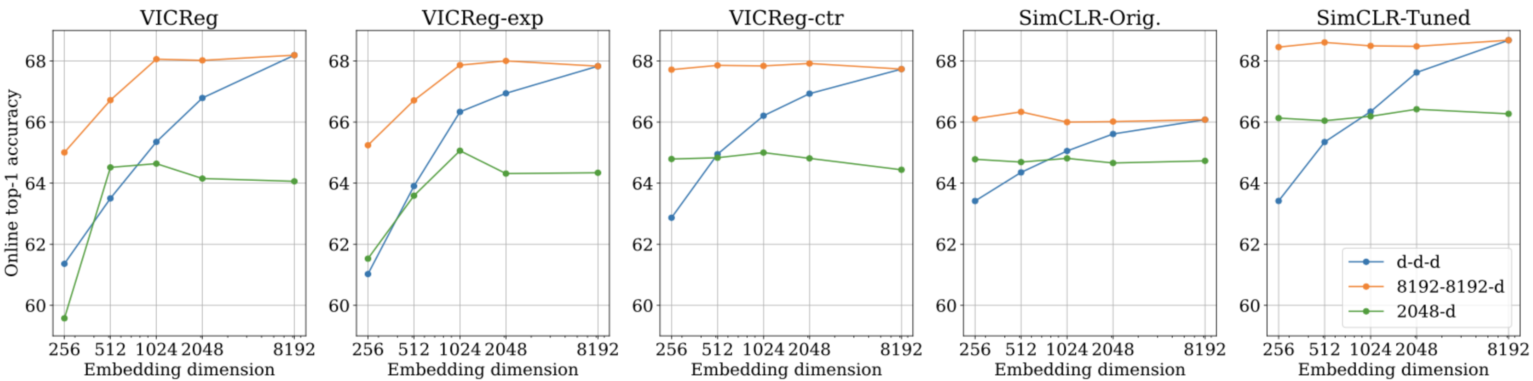
$$c_{exp}(\mathcal{K}) = \frac{1}{d}\sum_{i} \log\left(\sum_{j\neq i} e^{C(\mathcal{K})_{i,j}/\tau}\right)$$

We add the **LogSumExp**, one of the most evident difference, useful for the repulsive force (inspired by the **InfoNCE** criterion)

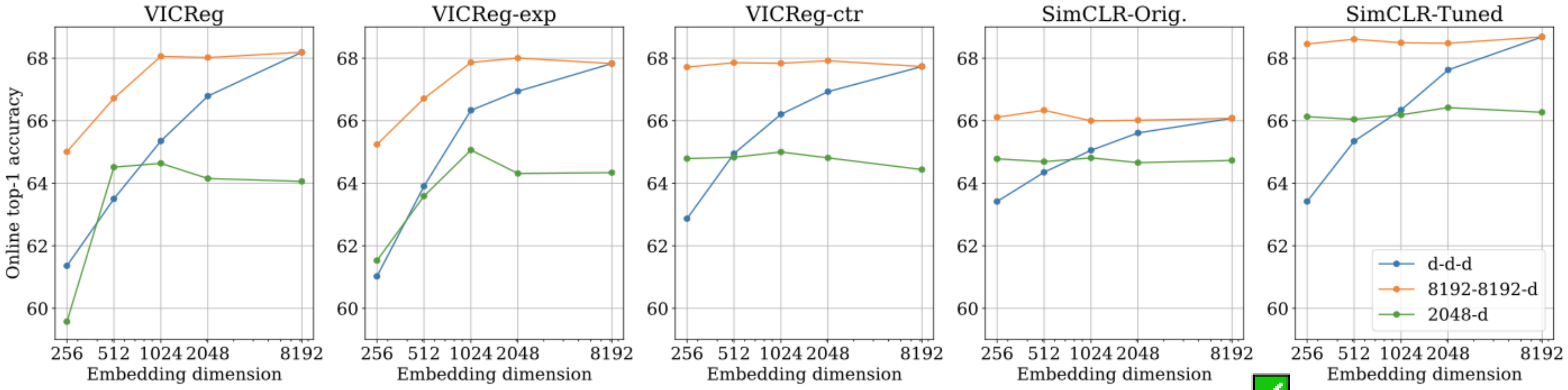
$$\mathcal{L}_{VICReg-ctr} = \lambda \sum_{i=1}^{N} \|\mathcal{K}_{\cdot,i} - \mathcal{K}'_{\cdot,i}\|_2^2 + \mu\left(v(\mathcal{K}^T) + v(\mathcal{K}'^T)\right) + \nu\left(c_{exp}(\mathcal{K}^T) + c_{exp}(\mathcal{K}'^T)\right)$$

We transpose the embedding matrix before applying **Var/Cov** regularization: the first for norm of embeddings, the second for penalizing the Gram matrix

- 3 layer for projector
- ReLu + Batch Norm
- Linear classifier as readout



Hyperparams Tuning & Misconceptions



slight differences on low embeddings dim and low proj capacity only

known performance of SimCLR is suboptimal

- projector’s size plays a role
- hyperparams tuning is required
- non-contrastive methods are not dimension inefficient
- large embedding size not a deciding factor

Matteo D’Agostino, EPFL,
Lausanne, Switzerland