Integrating Prior Knowledge in Contrastive Learning with Kernel

B. Dufumier[1,2], C.A. Barbano[2,3], R. Louiset[1], E. Duchesnay[1], P. Gori[2]
[1]NeuroSpin, CEA Saclay, Université Paris-Saclay
[2]LTCI, Télécom Paris, IPParis
[3]University of Turin

ICML International Conference On Machine Learning

1. Motivation

- Contrastive visual representation learning (CL) highly depends on data augmentation
- Data augmentation is domain-specific and it requires a priori hypothesis about the invariants of the decision function
- Can we learn visual representations from other prior knowledge, such as generative model's representation or auxiliary attributes ?

  → We propose a new contrastive loss, integrating prior knowledge through a kernel function
  → We derive theoretical guarantees on the downstream classification task
  → We outperform previous unsupervised approaches using generative models as prior

2. NPC problem solving

Problem setup. The general problem in contrastive learning is to learn a data representation using an encoder $f : \mathcal{X} \to \mathbb{S}^{d-1}$ that is pre-trained with a set of n original samples $(\bar{x}_i)_{i\in[1..n]} \in \bar{\mathcal{X}}$ and their augmented views $x_i \sim \mathcal{A}(\cdot|\bar{x}_i)$.

Negative-Positive Coupling (NPC) issue in CL. InfoNCE loss asympotically imposes 1) alignment between positives and 2) uniformity between negatives+positives→ by repelling and attracting positives, InfoNCE cannot achieve both perfect alignment and uniformity.

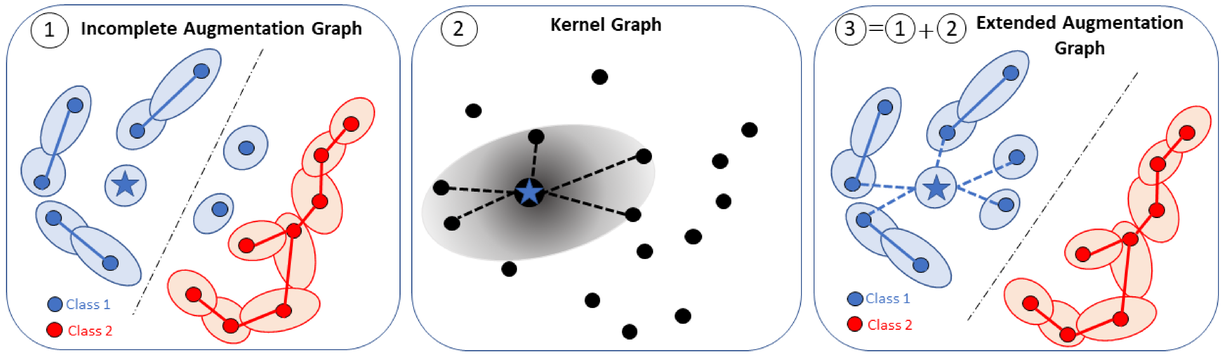
Decoupled Uniformity (DU) Loss. We propose to solve the NPC issue in CL by optimizing a loss relying only on centroids $\mu_{\bar{x}} = \mathbb{E}_{x\sim\mathcal{A}(\cdot|\bar{x})} f(x)$, called Decoupled Uniformity (DU) loss:

$$\mathcal{L}^{de}_{unif}(f) = \log \mathbb{E}_{p(\bar{x})p(\bar{x}')} e^{-||\mu_{\bar{x}} - \mu_{\bar{x}'}||^2}$$

Table 1: Linear evaluation accuracy of DU w/o prior

Dataset	Network	$\mathcal{L}_{InfoNCE}$	$\mathcal{L}_{DC}$	$\mathcal{L}^{de}_{unif}$
CIFAR-10	ResNet18	$82.18_{\pm 0.30}$	$84.87_{\pm 0.27}$	**85.05**$_{\pm 0.37}$
CIFAR-100	ResNet18	$55.11_{\pm 0.20}$	$58.27_{\pm 0.34}$	**58.41**$_{\pm 0.05}$
ImageNet100	ResNet50	68.76	73.98	**77.18**

3. Contrastive learning with prior knowledge



Step 1. Build an augmentation graph $G_{\mathcal{A}}$ for $\mathcal{A}$:
- Nodes $\mathcal{V}_{\mathcal{A}} = \bar{\mathcal{X}}$
- Edges connect $(\bar{x}, \bar{x}')$ if they can be mapped to the same augmented sample, *i.e.* $\text{supp}\mathcal{A}(\cdot|\bar{x}) \cap \text{supp}\mathcal{A}(\cdot|\bar{x}') \neq \emptyset$

Step 2. Build a kernel graph G_K for a given kernel K:
- Nodes $\mathcal{V}_K = \bar{\mathcal{X}}$
- Edges connect two images $(\bar{x}, \bar{x}')$ if they are close in the kernel space, *i.e.* $d_K(\bar{x}, \bar{x}') \leq 2\epsilon$ for K with constant norm.

Step 3. Estimate each centroid $\mu_{\bar{x}}$ on $G = G_K \cup G_{\mathcal{A}}$:

$$\hat{\mu}_{\bar{x}} = \sum_{i=1}^{n} \alpha_i(\bar{x}) f(x_i) \xrightarrow{\ell_2} \mu_{\bar{x}}$$

where $\alpha_i(\bar{x}) = \sum_{j=1}^{n} [(K_n + n\lambda \mathbf{I}_n)^{-1}]_{ij} K(\bar{x}_j, \bar{x})$ and $K_n = [K(\bar{x}_i, \bar{x}_j)]_{i,j\in[1..n]}$.

Theorem 1 *(Tight bounds on the supervised risk) We assume that G is class-wise connected and K is expressive enough. Let $(x_i, \bar{x}_i)_{i\in[1..n]} \overset{iid}{\sim} \mathcal{A}(x, \bar{x})$. For any ϵ'-weak aligned encoder f:*

$$\hat{\mathcal{L}}^{de}_{unif}(f) - O\left(n^{-1/4}\right) \leq \mathcal{L}_{sup}(f) \leq \hat{\mathcal{L}}^{de}_{unif}(f) + 4D(2\epsilon' + \epsilon) + O\left(n^{-1/4}\right)$$

where D is the maximal diameter of all class-connected sub-graphs and $\lambda_{min}(K_n) > 0$ the minimal eign(K_n).

4. Generative models improves CL representation

- We use generative model's representation $z(\bar{x})$ to set the kernel $K(\bar{x}_i, \bar{x}_j) = K_{rbf}(z(\bar{x}_i), z(\bar{x}_j))$

Table 2: BigBiGAN improves CL

Model	ImageNet100
SimCLR	68.76
BYOL	72.26
CMC	73.58
DCL	74.6
AlignUnif	76.3
DC	73.98
SwAV (w/o m-c)	73.5
BigBiGAN	72.0
DU (ours)	77.18
K_{GAN} DU (ours)	**78.02**
Supervised	$82.1_{\pm 0.59}$

Table 3: Can we remove data augmentation from CL?

Model	CIFAR-10			CIFAR-100		
	All	w/o Color	w/o Color+Crop	*All*	w/o Color	w/o Color+Crop
SimCLR	83.06	65.00	24.47	55.11	37.63	6.62
BYOL	84.71	81.45	50.17	53.15	49.59	27.9
Barlow Twins	81.61	53.97	47.52	52.27	28.52	24.17
VAE	41.37	41.37	41.37	14.34	14.34	14.34
DCGAN	66.71	66.71	66.71	26.17	26.17	26.17
K_{GAN} DU (ours)	**85.85**	**82.0**	**69.19**	**58.42**	**54.17**	**35.98**

- Small batch size $n = 256$ and $f = ResNet$ in all experiments
- No data aug. used for generative models training
- Competitive results even when removing color distorsion

5. Weak attributes help

Table 4: If weak attributes are accessible (e.g birds color or size for CUB200), they can be leveraged as prior in our framework to improve the representation.

Model	CUB	ImageNet100	UT-Zappos
SimCLR	17.48	65.30	84.08
BYOL	16.82	72.20	85.48
CosKernel CCLK	15.61	74.34	83.23
RBFKernel CCLK	30.49	77.24	84.65
CosKernel DU (ours)	27.77	**79.02**	**85.56**
RBFKernel DU (ours)	**32.87**	76.34	84.78

- Weak attributes $z(\bar{x})$ plugged in RBF or Cosine Kernel K:
  → Bird's attributes for CUB200
  → CLIP image's encoder for ImageNet100
  → Brand sub-categories for UT-Zappos

6. CL for Medical Imaging

Table 5: ROC-AUC for classifying 5 pathologies on CheXpert. GloRIA's representation $z(\bar{x})$ plugged in RBF kernel as prior K_{Gl}.

Model	At.	Cardio.	Consol.	Edema	Eff.
SimCLR	82.42	77.62	90.52	89.08	86.83
BYOL	83.04	81.54	90.98	90.18	85.99
MoCo-CXR	75.8	73.7	77.1	86.7	85.0
GLoRIA	86.70	**86.39**	90.41	90.58	91.82
CCLK	86.31	83.67	92.45	91.59	91.23
K_{Gl} DU (ours)	**86.92**	85.88	**93.03**	**92.39**	**91.93**
Supervised	81.6	79.7	90.5	86.8	89.9

Conclusion & Future Work

[Github] [Paper]

- Theoretically-grounded CL loss integrating prior knowledge and solving NPC problem

Next steps ?

- Integrate language models in CL
- Trainable kernel for centroid's estimation
- Geometrical analysis of centroids's distribution