Self-supervised learning with rotation-invariant kernels

Léon Zheng, Gilles Puy, Elisa Riccietti, Patrick Pérez, Rémi Gribonval

Learning meaningful representations without labels

Learning meaningful representations without labels

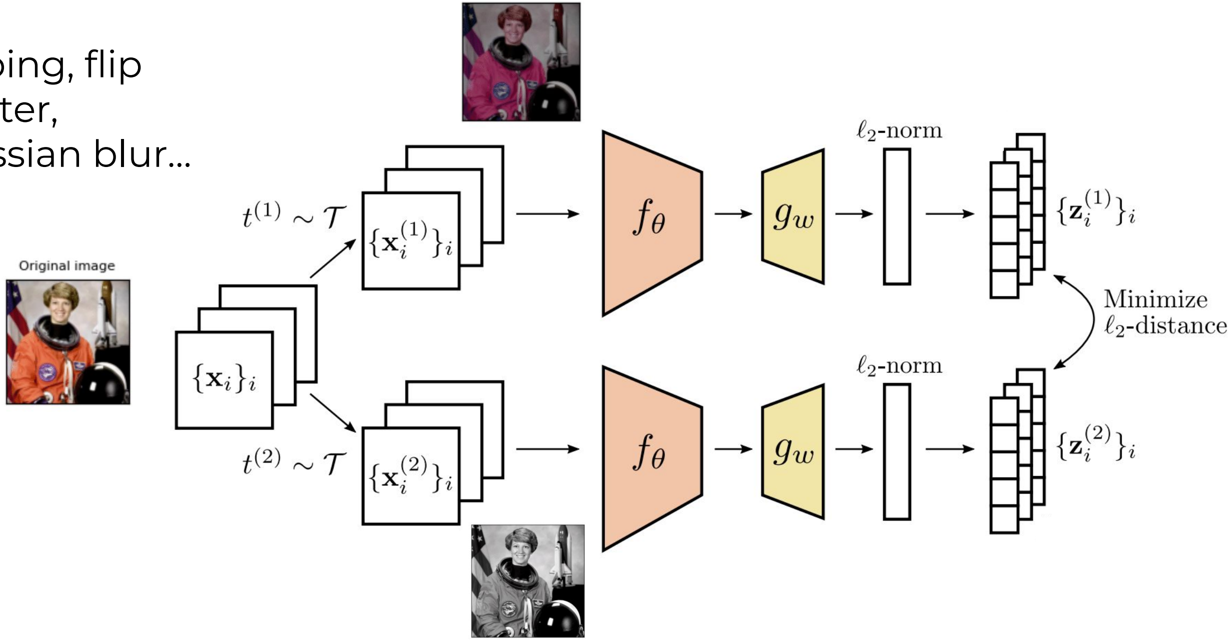
Approach: learn **invariance** to image transformations via a Siamese network

Random cropping, flip
image, color jitter,
grayscale, gaussian blur...

Learning meaningful representations without labels

Approach: learn **invariance** to image transformations via a Siamese network

Random cropping, flip image, color jitter, grayscale, gaussian blur...



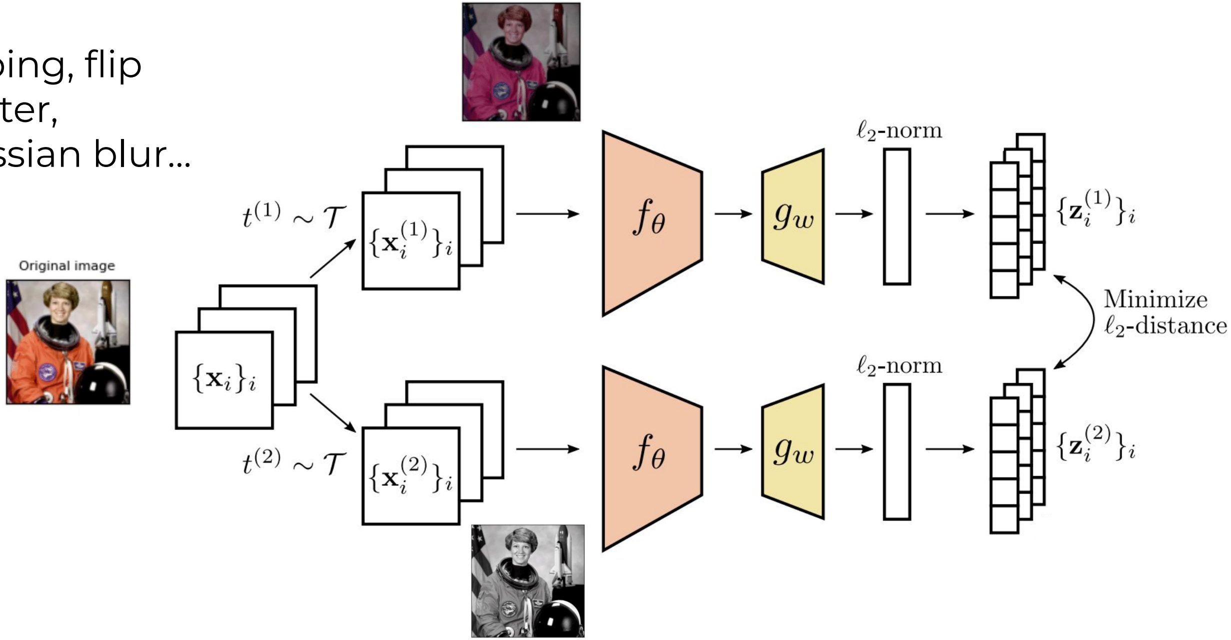
$$\mathbf{z} \in \mathbb{R}^q$$

$$\|\mathbf{z}\|_2 = 1$$

Learning meaningful representations without labels

Approach: learn **invariance** to image transformations via a Siamese network

Random cropping, flip image, color jitter, grayscale, gaussian blur...



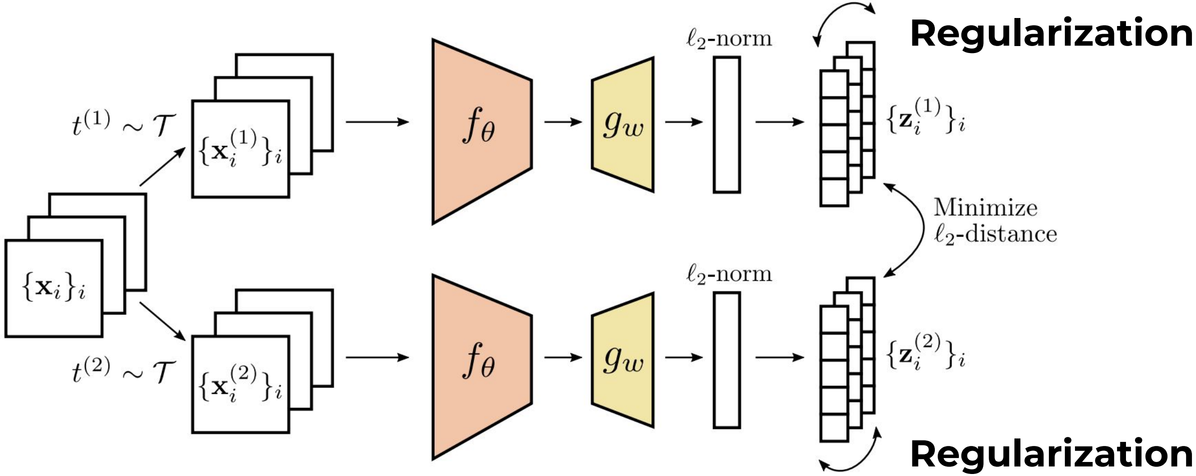
$\mathbf{z} \in \mathbb{R}^q$

$\|\mathbf{z}\|_2 = 1$

Avoid learning a low dimensional representation

Valeo

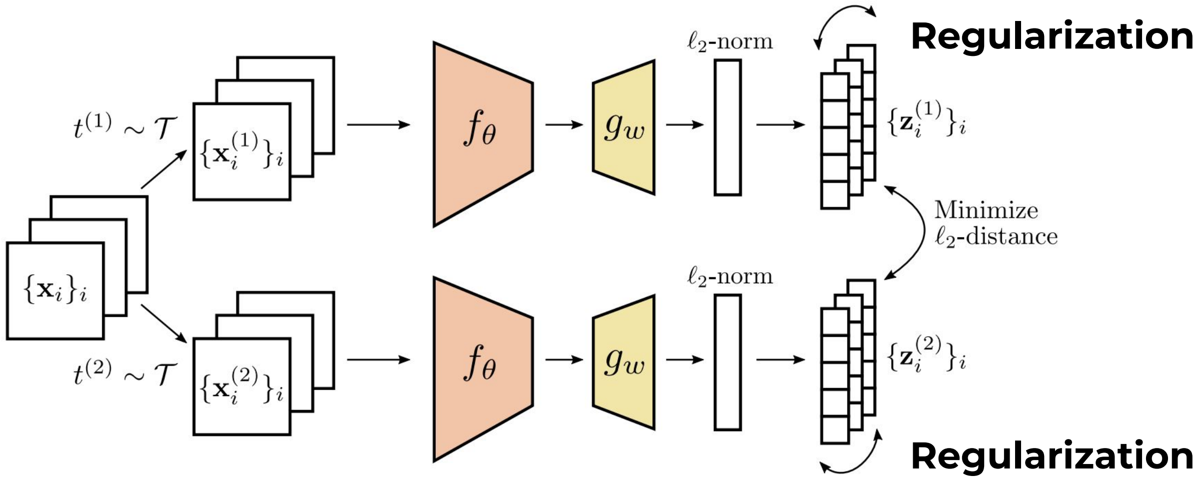
Regularizations of the embedding distribution



$$\mathbf{z} \in \mathbb{R}^q$$

$$\|\mathbf{z}\|_2 = 1$$

Regularizations of the embedding distribution

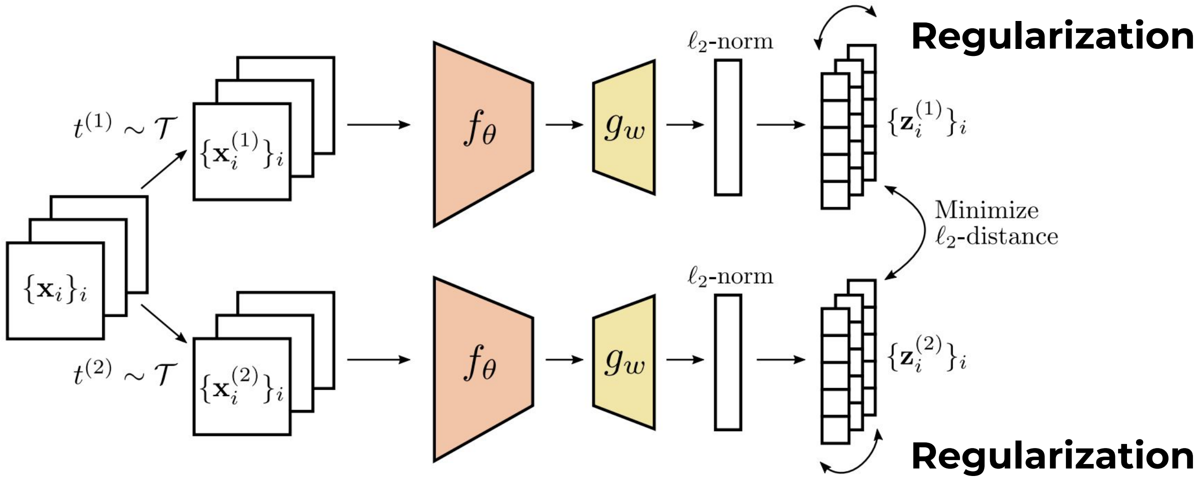


$$\mathbf{z} \in \mathbb{R}^q$$
$$\|\mathbf{z}\|_2 = 1$$

Existing methods differ in the way they impose this regularization:

- Sample-contrastive methods [Oord et al., 2018; Hjelm et al., 2019; Chen et al., 2020; He et al., 2020; Henaff, 2020]

Regularizations of the embedding distribution

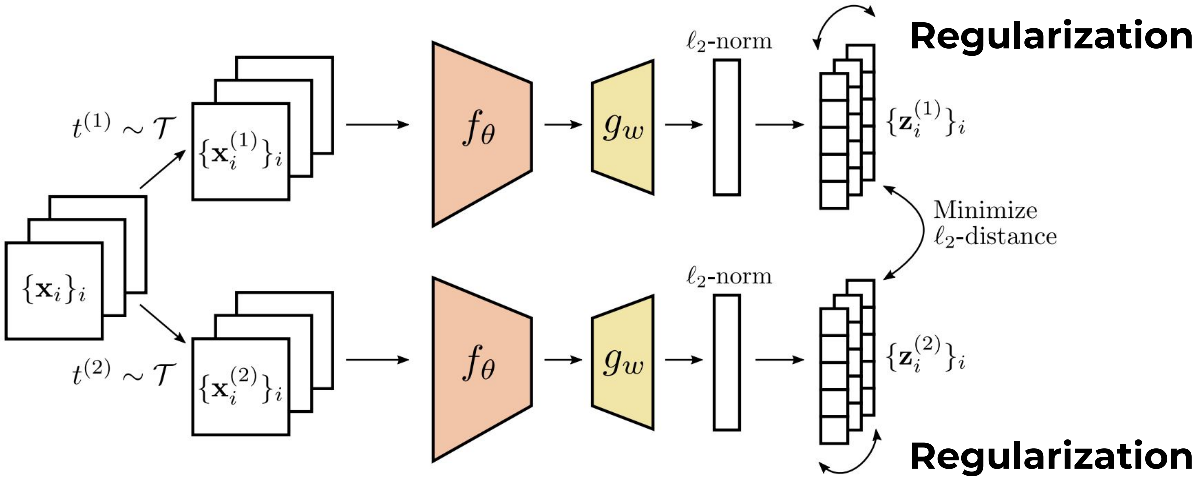


$$\mathbf{z} \in \mathbb{R}^q$$
$$\|\mathbf{z}\|_2 = 1$$

Existing methods differ in the way they impose this regularization:

- Sample-contrastive methods [Oord et al., 2018; Hjelm et al., 2019; Chen et al., 2020; He et al., 2020; Henaff, 2020]
- Distillation methods [Grill et al., 2020; Gidaris et al., 2020; 2021; Chen & He, 2021; Caron et al., 2021]

Regularizations of the embedding distribution

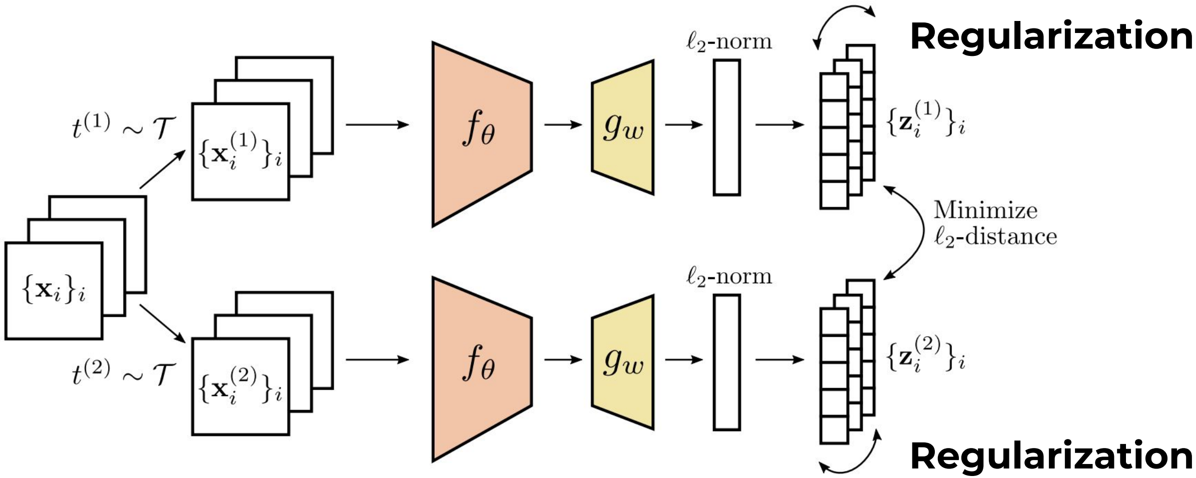


$$\mathbf{z} \in \mathbb{R}^q$$
$$\|\mathbf{z}\|_2 = 1$$

Existing methods differ in the way they impose this regularization:

- Sample-contrastive methods [Oord et al., 2018; Hjelm et al., 2019; Chen et al., 2020; He et al., 2020; Henaff, 2020]
- Distillation methods [Grill et al., 2020; Gidaris et al., 2020; 2021; Chen & He, 2021; Caron et al., 2021]
- Information-maximization methods [Zbontar et al., 2021; Ermolov et al., 2021; Bardes et al., 2022]

Regularizations of the embedding distribution



$$\mathbf{z} \in \mathbb{R}^q$$
$$\|\mathbf{z}\|_2 = 1$$

Existing methods differ in the way they impose this regularization:

- Sample-contrastive methods [Oord et al., 2018; Hjelm et al., 2019; Chen et al., 2020; He et al., 2020; Henaff, 2020]
- Distillation methods [Grill et al., 2020; Gidaris et al., 2020; 2021; Chen & He, 2021; Caron et al., 2021]
- Information-maximization methods [Zbontar et al., 2021; Ermolov et al., 2021; Bardes et al., 2022]

What is a good choice of regularization?

Unification of regularizers under a kernel loss

Kernel point of view: **MMD** between the **embedding distribution** and the **uniform distribution** on the hypersphere.

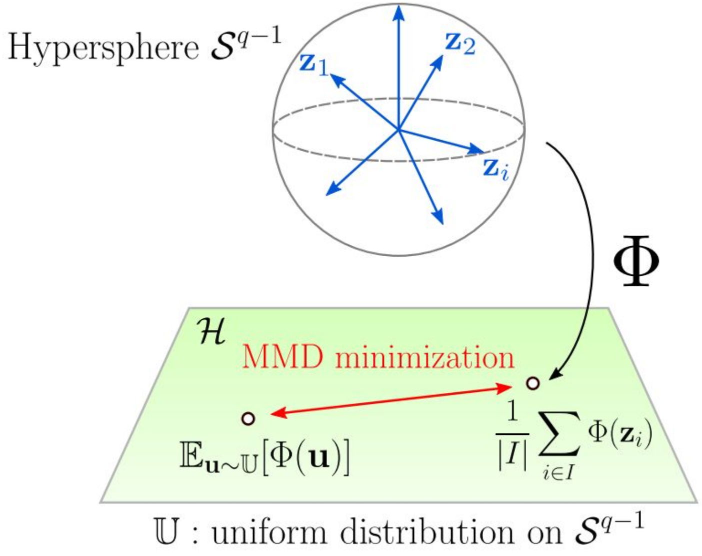
Valeo

Unification of regularizers under a kernel loss

<u>Kernel point of view</u>: **MMD** between the **embedding distribution** and the **uniform distribution** on the hypersphere.

Rotation-invariant kernel:

$$\mathcal{K}(\mathbf{u}, \mathbf{v}) = \varphi(\mathbf{u}^\intercal \mathbf{v}) = \langle \Phi(\mathbf{u}), \Phi(\mathbf{v}) \rangle_{\mathcal{H}}$$



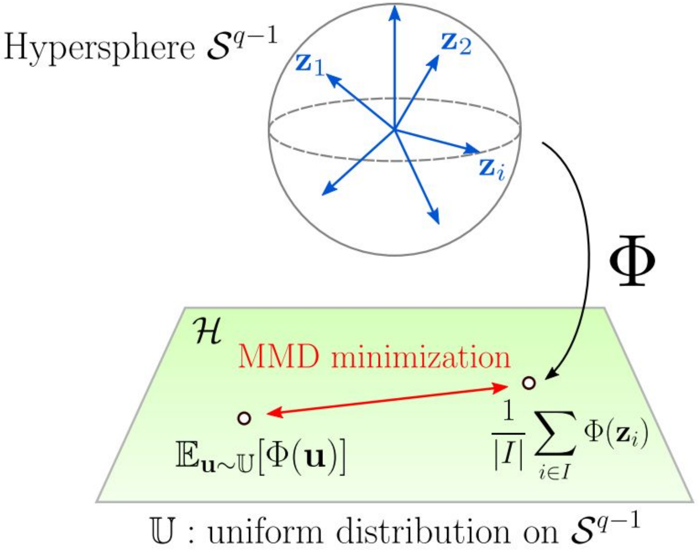
Valeo

Unification of regularizers under a kernel loss

Kernel point of view: **MMD** between the **embedding distribution** and the **uniform distribution** on the hypersphere.

Rotation-invariant kernel:

$$\mathcal{K}(\mathbf{u},\mathbf{v}) = \varphi(\mathbf{u}^\intercal \mathbf{v}) = \langle \Phi(\mathbf{u}), \Phi(\mathbf{v}) \rangle_{\mathcal{H}}$$



1) Unification

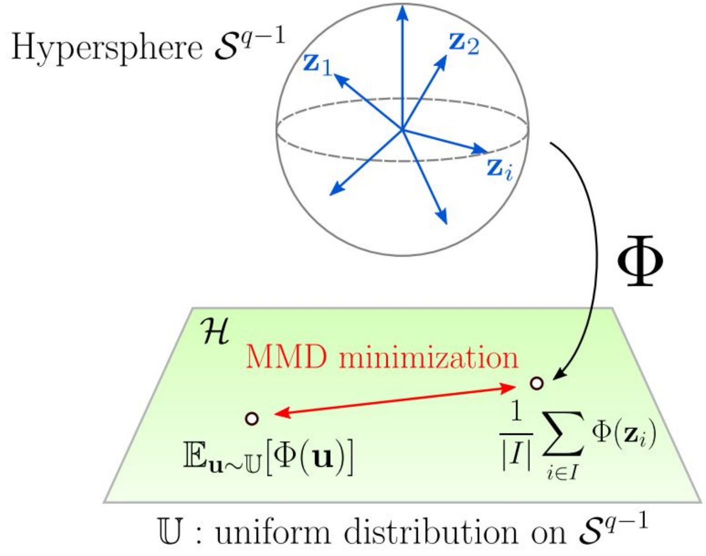
$\mathcal{K}(\mathbf{u},\mathbf{v})$	Method
$(\mathbf{u}\mathbf{v}^\top)^2$	Contrastive
$e^{-t\|\mathbf{u}-\mathbf{v}\|_2^2}$	Alignment & Uniformity on the Hypersphere
$C - \|\mathbf{u}-\mathbf{v}\|_2^{2s-q+1}$	PointContrast
$b_1\mathbf{u}\mathbf{v}^\top + b_2\frac{q(\mathbf{u}\mathbf{v}^\top)^2-1}{q-1}$	Analog to VICReg

Unification of regularizers under a kernel loss

Kernel point of view: **MMD** between the **embedding distribution** and the **uniform distribution** on the hypersphere.

Rotation-invariant kernel:

$$\mathcal{K}(\mathbf{u},\mathbf{v}) = \varphi(\mathbf{u}^\intercal\mathbf{v}) = \langle \Phi(\mathbf{u}), \Phi(\mathbf{v}) \rangle_{\mathcal{H}}$$



1) Unification

$\mathcal{K}(\mathbf{u},\mathbf{v})$	Method
$(\mathbf{u}\mathbf{v}^\top)^2$	Contrastive
$e^{-t\|\mathbf{u}-\mathbf{v}\|_2^2}$	Alignment & Uniformity on the Hypersphere
$C - \|\mathbf{u}-\mathbf{v}\|_2^{2s-q+1}$	PointContrast
$b_1\mathbf{u}\mathbf{v}^\top + b_2\frac{q(\mathbf{u}\mathbf{v}^\top)^2-1}{q-1}$	Analog to VICReg

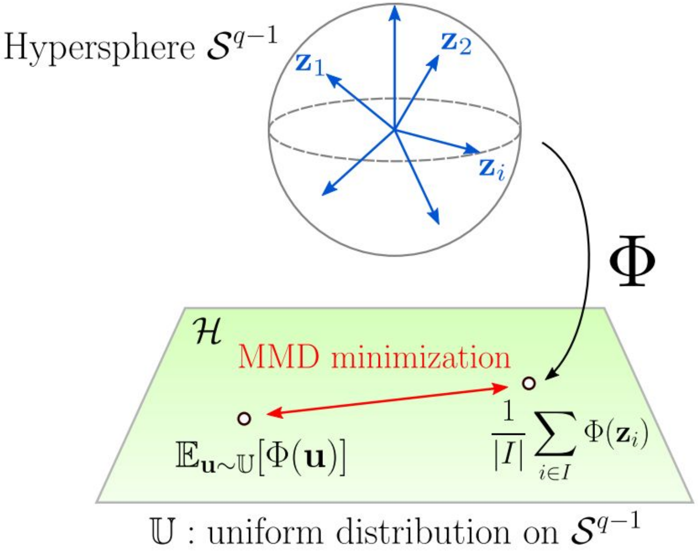
2) Good kernel choices?

Unification of regularizers under a kernel loss

<u>Kernel point of view</u>: **MMD** between the **embedding distribution** and the **uniform distribution** on the hypersphere.

Rotation-invariant kernel:

$$\mathcal{K}(\mathbf{u},\mathbf{v}) = \varphi(\mathbf{u}^\intercal \mathbf{v}) = \langle \Phi(\mathbf{u}), \Phi(\mathbf{v}) \rangle_{\mathcal{H}}$$



1) Unification

$\mathcal{K}(\mathbf{u},\mathbf{v})$	Method
$(\mathbf{u}\mathbf{v}^\top)^2$	Contrastive
$e^{-t\|\mathbf{u}-\mathbf{v}\|_2^2}$	Alignment & Uniformity on the Hypersphere
$C - \|\mathbf{u}-\mathbf{v}\|_2^{2s-q+1}$	PointContrast
$b_1\mathbf{u}\mathbf{v}^\top + b_2\frac{q(\mathbf{u}\mathbf{v}^\top)^2-1}{q-1}$	Analog to VICReg

2) Good kernel choices?

3) Identifying a new competitive kernel

Rotation-invariant kernel (a.k.a. dot-product kernel)

Theorem: $\mathcal{K}(\mathbf{u}, \mathbf{v}) = \varphi(\mathbf{u}^\intercal \mathbf{v})$ is positive definite iff φ admits an expansion

[Schoenberg, 1942]

$$\varphi(t) = \sum_{\ell=0}^{+\infty} b_\ell P_\ell(q;t), \quad with \quad b_\ell \geq 0$$

Legendre polynomial of **degree** ℓ in dimension q

Rotation-invariant kernel (a.k.a. dot-product kernel)

Theorem: $\mathcal{K}(\mathbf{u},\mathbf{v}) = \varphi(\mathbf{u}^\intercal \mathbf{v})$ is positive definite iff φ admits an expansion

[Schoenberg, 1942]

$$\varphi(t) = \sum_{\ell=0}^{+\infty} b_\ell P_\ell(q;t), \quad \textit{with} \quad b_\ell \geq 0$$

Legendre polynomial of **degree** ℓ in dimension q

$$\mathrm{MMD}(\mathbb{Q},\mathbb{U})^2 := \left\| \int_{\mathcal{S}^{q-1}} \mathcal{K}(\cdot,\mathbf{z})d\mathbb{Q}(\mathbf{z}) - \int_{\mathcal{S}^{q-1}} \mathcal{K}(\cdot,\mathbf{u})d\mathbb{U}(\mathbf{u}) \right\|_{\mathcal{H}}^2$$

Uniform distribution on hypersphere $\mathcal{S}^{q-1}$

Valeo

Rotation-invariant kernel (a.k.a. dot-product kernel)

Theorem: $\mathcal{K}(\mathbf{u},\mathbf{v}) = \varphi(\mathbf{u}^\intercal \mathbf{v})$ is positive definite iff φ admits an expansion

[Schoenberg, 1942]

$$\varphi(t) = \sum_{\ell=0}^{+\infty} b_\ell P_\ell(q;t), \quad \textit{with} \quad b_\ell \geq 0$$

Legendre polynomial of **degree** ℓ in dimension q

$$\mathrm{MMD}(\mathbb{Q},\mathbb{U})^2 := \left\| \int_{\mathcal{S}^{q-1}} \mathcal{K}(\cdot,\mathbf{z})d\mathbb{Q}(\mathbf{z}) - \int_{\mathcal{S}^{q-1}} \mathcal{K}(\cdot,\mathbf{u})d\mathbb{U}(\mathbf{u}) \right\|_{\mathcal{H}}^2 = \mathbb{E}_{\mathbf{z},\mathbf{z}'\sim\mathbb{Q}}\left[\tilde{\mathcal{K}}(\mathbf{z},\mathbf{z}')\right]$$

Uniform distribution on hypersphere $\mathcal{S}^{q-1}$

where $\tilde{\mathcal{K}}(\cdot,\cdot) = \mathcal{K}(\cdot,\cdot) - b_0$

Rotation-invariant kernel (a.k.a. dot-product kernel)

Theorem: $\mathcal{K}(\mathbf{u},\mathbf{v}) = \varphi(\mathbf{u}^\intercal \mathbf{v})$ is positive definite iff φ admits an expansion

[Schoenberg, 1942]

$$\varphi(t) = \sum_{\ell=0}^{+\infty} b_\ell P_\ell(q;t), \quad \textit{with} \quad b_\ell \geq 0$$

Legendre polynomial of **degree** ℓ in dimension q

$$\mathrm{MMD}(\mathbb{Q},\mathbb{U})^2 := \left\| \int_{\mathcal{S}^{q-1}} \mathcal{K}(\cdot,\mathbf{z})d\mathbb{Q}(\mathbf{z}) - \int_{\mathcal{S}^{q-1}} \mathcal{K}(\cdot,\mathbf{u})d\mathbb{U}(\mathbf{u}) \right\|_{\mathcal{H}}^2 = \mathbb{E}_{\mathbf{z},\mathbf{z}'\sim\mathbb{Q}}\left[\tilde{\mathcal{K}}(\mathbf{z},\mathbf{z}')\right]$$

Uniform distribution on hypersphere $\mathcal{S}^{q-1}$

where $\tilde{\mathcal{K}}(\cdot,\cdot) = \mathcal{K}(\cdot,\cdot) - b_0$

Proposed generic regularization loss:

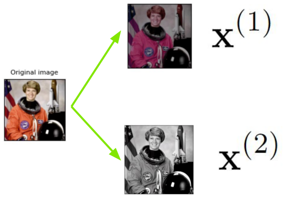
$$\mathcal{L}_{reg}(\{\mathbf{z}_i\}_{i=1}^n) := \widehat{\mathrm{MMD}}^2(\mathbb{Q},\mathbb{U}) = \frac{1}{n^2}\sum_{i=1}^n \sum_{i'=1}^n \tilde{\mathcal{K}}(\mathbf{z}_i,\mathbf{z}_{i'})$$

→ interpretation as an energy functional

Former methods correspond to different kernels

Invariance — Kernel regularization

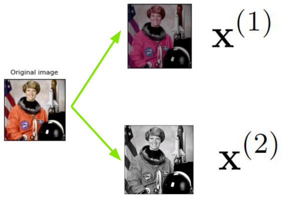
$$\frac{1}{n}\sum_{i=1}^{n}\|\mathbf{z}_i^{(1)}-\mathbf{z}_i^{(2)}\|_2^2+\frac{\lambda}{2}\left(\mathcal{L}_{reg}(\{\mathbf{z}_i^{(1)}\}_{i=1}^{n})+\mathcal{L}_{reg}(\{\mathbf{z}_i^{(2)}\}_{i=1}^{n})\right)$$



Former methods correspond to different kernels

Invariance — Kernel regularization

$$\frac{1}{n}\sum_{i=1}^{n}\|\mathbf{z}_i^{(1)}-\mathbf{z}_i^{(2)}\|_2^2+\frac{\lambda}{2}\Big(\mathcal{L}_{reg}(\{\mathbf{z}_i^{(1)}\}_{i=1}^{n})+\mathcal{L}_{reg}(\{\mathbf{z}_i^{(2)}\}_{i=1}^{n})\Big)$$



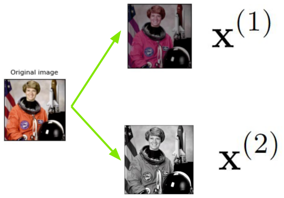
Sample-contrastive loss
[Garrido et al., 2023]:

$$\frac{1}{n^2}\sum_{i=1}^{n}\sum_{i'=1}^{n}(\mathbf{z}_i^\top \mathbf{z}_{i'})^2 \longrightarrow \text{quadratic kernel}$$

Former methods correspond to different kernels

Invariance | Kernel regularization

$$\frac{1}{n}\sum_{i=1}^{n}\|\mathbf{z}_i^{(1)}-\mathbf{z}_i^{(2)}\|_2^2+\frac{\lambda}{2}\left(\mathcal{L}_{reg}(\{\mathbf{z}_i^{(1)}\}_{i=1}^{n})+\mathcal{L}_{reg}(\{\mathbf{z}_i^{(2)}\}_{i=1}^{n})\right)$$



Sample-contrastive loss
[Garrido et al., 2023]:

$$\frac{1}{n^2}\sum_{i=1}^{n}\sum_{i'=1}^{n}(\mathbf{z}_i^\top\mathbf{z}_{i'})^2 \longrightarrow \text{quadratic kernel}$$

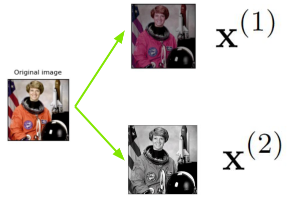
Alignment & Uniformity on the Hypersphere
[Wang & Isola, 2020]:

$$\frac{1}{n^2}\sum_{i=1}^{n}\sum_{i'=1}^{n}e^{-t\|\mathbf{z}_i-\mathbf{z}_{i'}\|_2^2} \longrightarrow \text{RBF kernel}$$

Former methods correspond to different kernels

Invariance — Kernel regularization

$$\frac{1}{n}\sum_{i=1}^{n}\|\mathbf{z}_i^{(1)}-\mathbf{z}_i^{(2)}\|_2^2+\frac{\lambda}{2}\left(\mathcal{L}_{reg}(\{\mathbf{z}_i^{(1)}\}_{i=1}^{n})+\mathcal{L}_{reg}(\{\mathbf{z}_i^{(2)}\}_{i=1}^{n})\right)$$



Sample-contrastive loss [Garrido et al., 2023]: $\frac{1}{n^2}\sum_{i=1}^{n}\sum_{i'=1}^{n}(\mathbf{z}_i^\top \mathbf{z}_{i'})^2$ → quadratic kernel

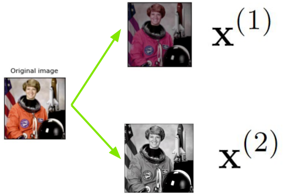
Alignment & Uniformity on the Hypersphere [Wang & Isola, 2020]: $\frac{1}{n^2}\sum_{i=1}^{n}\sum_{i'=1}^{n}e^{-t\|\mathbf{z}_i-\mathbf{z}_{i'}\|_2^2}$ → RBF kernel

Information-maximization method [Bardes et al., 2023]: $\frac{1}{n^2}\sum_{i=1}^{n}\sum_{i'=1}^{n}\left(b_1\mathbf{z}_i^\top \mathbf{z}_{i'}+b_2\frac{q(\mathbf{z}_i^\top \mathbf{z}_{i'})^2-1}{q-1}\right)$ → combination of linear and quadratic

Former methods correspond to different kernels

Invariance — Kernel regularization

$$\frac{1}{n}\sum_{i=1}^{n}\|\mathbf{z}_i^{(1)}-\mathbf{z}_i^{(2)}\|_2^2+\frac{\lambda}{2}\left(\mathcal{L}_{reg}(\{\mathbf{z}_i^{(1)}\}_{i=1}^{n})+\mathcal{L}_{reg}(\{\mathbf{z}_i^{(2)}\}_{i=1}^{n})\right)$$



Sample-contrastive loss [Garrido et al., 2023]: $\frac{1}{n^2}\sum_{i=1}^{n}\sum_{i'=1}^{n}(\mathbf{z}_i^\top\mathbf{z}_{i'})^2$ → quadratic kernel

Alignment & Uniformity on the Hypersphere [Wang & Isola, 2020]: $\frac{1}{n^2}\sum_{i=1}^{n}\sum_{i'=1}^{n}e^{-t\|\mathbf{z}_i-\mathbf{z}_{i'}\|_2^2}$ → RBF kernel

Information-maximization method [Bardes et al., 2023]: $\frac{1}{n^2}\sum_{i=1}^{n}\sum_{i'=1}^{n}\left(b_1\mathbf{z}_i^\top\mathbf{z}_{i'}+b_2\frac{q(\mathbf{z}_i^\top\mathbf{z}_{i'})^2-1}{q-1}\right)$ → combination of linear and quadratic

For this kernel: $\mathrm{MMD}(\mathbb{Q},\mathbb{U})=0 \implies \mathbb{E}_{\mathbf{z}\sim\mathbb{Q}}\left[(\mathbf{z}-\mathbb{E}(\mathbf{z}))(\mathbf{z}-\mathbb{E}(\mathbf{z}))^\top\right]=\frac{1}{q}\mathbf{I}_q$

Same goal as the regularizer in VICReg

[Bardes et al., 2023]

What is a good kernel choice?

Legendre expansion:

$$\varphi(t) = \sum_{\ell=0}^{+\infty} b_\ell P_\ell(q; t)$$

What is a good kernel choice?

Legendre expansion:

$$\varphi(t) = \sum_{\ell=0}^{+\infty} b_\ell P_\ell(q; t)$$

Truncated Legendre kernel: $$\mathcal{K}(\mathbf{u}, \mathbf{v}) = \sum_{\ell=1}^{L} b_\ell P_\ell(q; \mathbf{u}^\top \mathbf{v})$$

What is a good kernel choice?

Legendre expansion:

$$\varphi(t) = \sum_{\ell=0}^{+\infty} b_\ell P_\ell(q; t)$$

Truncated Legendre kernel: $\mathcal{K}(\mathbf{u}, \mathbf{v}) = \sum_{\ell=1}^{L} b_\ell P_\ell(q; \mathbf{u}^\top \mathbf{v})$

ResNet-18 on a subset of 20% of ImageNet-1k. Evaluation by linear probing.

	SimCLR†	AUH†	VICReg†	SFRIK (ours)		
				$L=1$	$L=2$	$L=3$
$q=1024$	45.2	45.3	40.6	-	45.2	-
$q=2048$	45.8	45.9	44.0	-	45.9	-
$q=4096$	46.0	46.7	44.9	-	46.9	-
$q=8192$	46.1	46.8	46.0	27.7	47.0	**47.5**

Conclusion: the first three orders are the most important.

Competitive results on ImageNet-1k

Pretraining with ResNet-50 during 200 epochs.

Method	Epochs	Linear classification					Semi-supervised			
		IN100%		Places205		VOC07	1% labels		10% labels	
		Top-1	Top-5	Top-1	Top-5	mAP	Top-1	Top-5	Top-1	Top-5
SimCLR* (Chen et al., 2020a)	200	68.3	-	-	-	-	-	-	-	-
SwAV* (Caron et al., 2020) (no multi-crop)	200	69.1	-	-	-	-	-	-	-	-
SimSiam (Chen & He, 2021)	200	70.0	-	-	-	-	-	-	-	-
VICReg† (Bardes et al., 2022) ($q = 8192$)	200	70.0	89.3	54.1	83.4	84.9	**49.4**	**75.1**	65.9	87.2
SFRIK ($L = 2, q = 8192$)	200	70.1	89.3	53.8	83.0	85.1	46.6	73.3	65.7	87.3
SFRIK ($L = 3, q = 8192$)	200	70.2	89.6	**54.5**	**83.9**	84.6	46.9	73.6	**66.0**	**87.7**
SFRIK ($L = 2, q = 16384$)	200	**70.3**	89.6	54.3	83.4	**85.2**	46.0	73.0	65.3	87.2
SFRIK ($L = 2, q = 32768$)	200	**70.3**	89.6	54.1	83.0	85.0	46.1	73.0	65.4	87.3
SFRIK ($L = 3, q = 32768$)	200	**70.3**	**89.7**	54.4	83.2	85.1	46.6	73.0	65.8	87.5

(ours)

Competitive results on ImageNet-1k

Pretraining with ResNet-50 during 200 epochs.

Method	Epochs	Linear classification					Semi-supervised			
		IN100%		Places205		VOC07	1% labels		10% labels	
		Top-1	Top-5	Top-1	Top-5	mAP	Top-1	Top-5	Top-1	Top-5
SimCLR* (Chen et al., 2020a)	200	68.3	-	-	-	-	-	-	-	-
SwAV* (Caron et al., 2020) (no multi-crop)	200	69.1	-	-	-	-	-	-	-	-
SimSiam (Chen & He, 2021)	200	70.0	-	-	-	-	-	-	-	-
VICReg† (Bardes et al., 2022) ($q = 8192$)	200	70.0	89.3	54.1	83.4	84.9	**49.4**	**75.1**	65.9	87.2
(ours) SFRIK ($L = 2$, $q = 8192$)	200	70.1	89.3	53.8	83.0	85.1	46.6	73.3	65.7	87.3
(ours) SFRIK ($L = 3$, $q = 8192$)	200	70.2	89.6	**54.5**	**83.9**	84.6	46.9	73.6	**66.0**	**87.7**
(ours) SFRIK ($L = 2$, $q = 16384$)	200	**70.3**	89.6	54.3	83.4	**85.2**	46.0	73.0	65.3	87.2
(ours) SFRIK ($L = 2$, $q = 32768$)	200	**70.3**	89.6	54.1	83.0	85.0	46.1	73.0	65.4	87.3
(ours) SFRIK ($L = 3$, $q = 32768$)	200	**70.3**	**89.7**	54.4	83.2	85.1	46.6	73.0	65.8	87.5

Kernel trick during pretraining:

- 19% faster
- 8% less memory per GPU

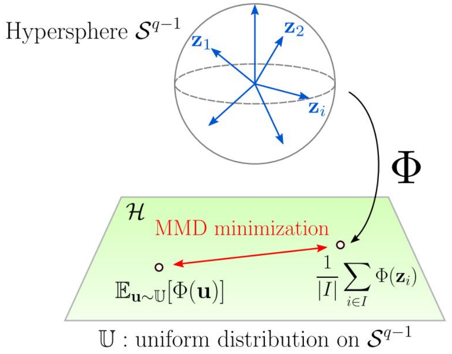
compared to VICReg (q=16384, batch size=2048).

Memory per GPU at q=32768

Batch size	VICReg	SFRIK	(ratio)
256	22.5GB	10.3GB	(2.2)
512	25.4GB	13.1GB	(1.9)
1024	31.1GB	18.8GB	(1.7)

Valeo

Conclusion

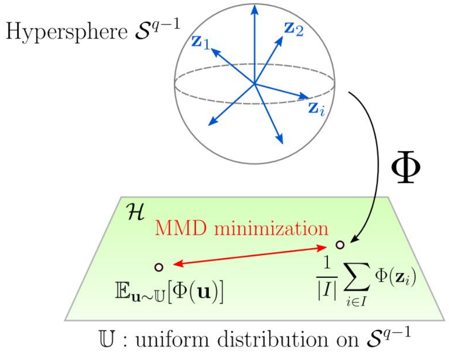


Embedding distribution

Uniform distribution on the hypersphere

Many existing regularizers turn out to minimize $\mathrm{MMD}(\mathbb{Q}, \mathbb{U})$ for different rotation-invariant kernels.

Conclusion



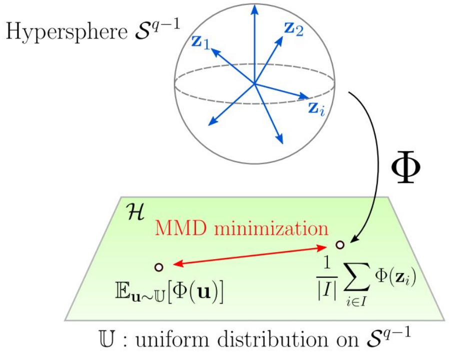
Embedding distribution

Uniform distribution on the hypersphere

Many existing regularizers turn out to minimize $\mathrm{MMD}(\mathbb{Q}, \mathbb{U})$ for different rotation-invariant kernels.

A truncated Legendre kernel is competitive.

Conclusion



Embedding distribution

Uniform distribution on the hypersphere

Many existing regularizers turn out to minimize $\mathrm{MMD}(\mathbb{Q}, \mathbb{U})$ for different rotation-invariant kernels.

A truncated Legendre kernel is competitive.

Perspectives: leverage the kernel framework for better self-supervision methods.

Thank you!

Self-supervised learning with rotation-invariant kernels

Léon Zheng · Gilles Puy · Elisa Riccietti · Patrick Perez · Rémi Gribonval

Poster: MH1-2-3-4 #166