

Contrastive Learning Inverts the Data Generating Process

Roland S. Zimmermann, Yash Sharma, Steffen
Schneider,
Matthias Bethge, Wieland Brendel

Presenter: Yaorong Xiao





Content

1. Background overview
2. Theory and proof
3. Experiments results

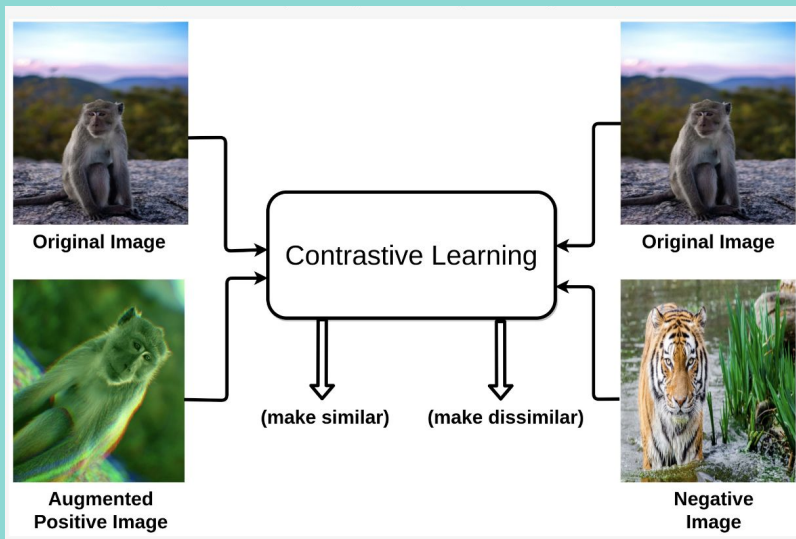


Contributions

1. InfoNCE inverts the data generating process if certain statistical assumptions on the data generating process hold.
2. Partially hold the assumption also works in some conditions.
3. They create the 3DIdent benchmark with natural environment features and show that a contrastive loss from their theory can identify its ground-truth factors.

Contrastive learning

1. Self-supervised representation learning technique
2. Goal: learn meaningful representations of data
3. Contrasting similar and dissimilar data pairs to encourage the model to map similar instances closer together in a lower-dimensional feature space while pushing dissimilar instances apart.





InfoNCE

$$I(x; c) = \sum_{x, c} p(x, c) \log \frac{p(x|c)}{p(x)}.$$

$$f_k(x_{t+k}, c_t) \propto \frac{p(x_{t+k}|c_t)}{p(x_{t+k})}$$

$$\mathcal{L}_N = -\mathbb{E}_X \left[\log \frac{f_k(x_{t+k}, c_t)}{\sum_{x_j \in X} f_k(x_j, c_t)} \right]$$

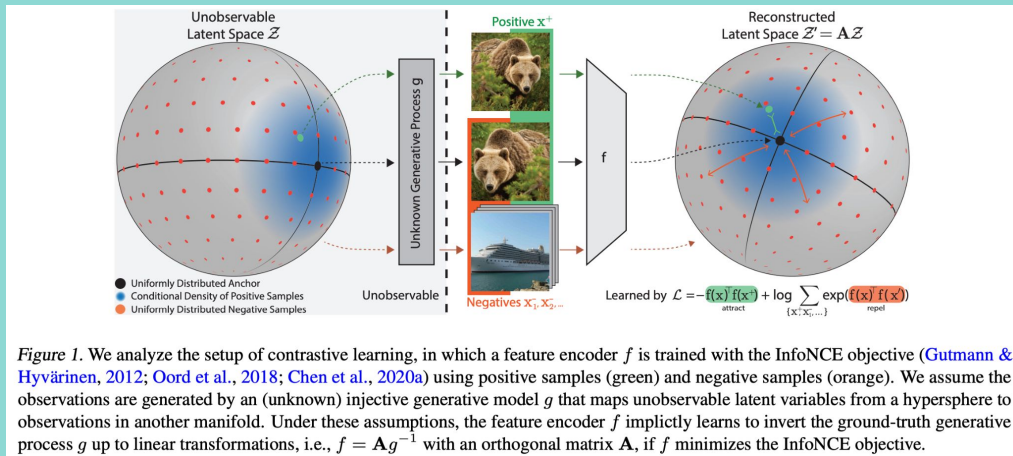
Reference: <https://arxiv.org/abs/1807.03748>



Why is so effectively to a large variety of downstream tasks?

Assumption

1. \mathcal{Z} is the unit hypersphere $S^{(N-1)}$
2. The ground-truth marginal distribution of latents is uniform and is a von Mises-Fisher distribution



$$p(\mathbf{z}) = |\mathcal{Z}|^{-1}, \quad p(\mathbf{z}|\tilde{\mathbf{z}}) = C_p^{-1} e^{\kappa \mathbf{z}^\top \tilde{\mathbf{z}}} \quad \text{with} \quad (2)$$

$$C_p := \int e^{\kappa \mathbf{z}^\top \tilde{\mathbf{z}}} d\tilde{\mathbf{z}} = \text{const.}, \quad \mathbf{x} = g(\mathbf{z}), \quad \tilde{\mathbf{x}} = g(\tilde{\mathbf{z}}).$$



Proof

1. Contrastive learning is related to cross-entropy minimization
2. Contrastive learning identifies ground-truth factors on the hypersphere
3. Contrastive learning identifies ground-truth factors on convex bodies in \mathbb{R}^N

3.1 Contrastive learning is related to cross-entropy minimization

Theorem 1 ($\mathcal{L}_{\text{contr}}$ converges to the cross-entropy between latent distributions). *If the ground-truth marginal distribution p is uniform, then for fixed $\tau > 0$, as the number of negative samples $M \rightarrow \infty$, the (normalized) contrastive loss converges to*

$$\lim_{M \rightarrow \infty} \mathcal{L}_{\text{contr}}(f; \tau, M) - \log M + \log |\mathcal{Z}| = \mathbb{E}_{\mathbf{z} \sim p(\mathbf{z})} [H(p(\cdot | \mathbf{z}), q_h(\cdot | \mathbf{z}))] \quad (14)$$

$$\mathcal{L}_{\text{contr}}(f; \tau, M) := \mathbb{E}_{\substack{(\mathbf{x}, \tilde{\mathbf{x}}) \sim p_{\text{pos}} \\ \{\mathbf{x}_i^-\}_{i=1}^M \stackrel{\text{i.i.d.}}{\sim} p_{\text{data}}}} \left[-\log \frac{e^{f(\mathbf{x})^\top f(\tilde{\mathbf{x}})/\tau}}{e^{f(\mathbf{x})^\top f(\tilde{\mathbf{x}})/\tau} + \sum_{i=1}^M e^{f(\mathbf{x})^\top f(\mathbf{x}_i^-)/\tau}} \right]. \quad (1)$$

$$\lim_{M \rightarrow \infty} \mathcal{L}_{\text{contr}}(f; \tau, M) - \log M + \log |\mathcal{Z}| =$$

$$\mathbb{E}_{\mathbf{z} \sim p(\mathbf{z})} [H(p(\cdot|\mathbf{z}), q_h(\cdot|\mathbf{z}))]$$

$$h = f \circ g,$$

$$= -\frac{1}{\tau} \mathbb{E}_{\tilde{\mathbf{z}}, \mathbf{z} \sim p(\tilde{\mathbf{z}}, \mathbf{z})} [h(\tilde{\mathbf{z}})^\top h(\mathbf{z})] \\ + \mathbb{E}_{\mathbf{z} \sim p(\mathbf{z})} \left[\log \int_{\mathcal{Z}} e^{h(\tilde{\mathbf{z}})^\top h(\mathbf{z})/\tau} d\tilde{\mathbf{z}} \right].$$

$$q_h(\tilde{\mathbf{z}}|\mathbf{z}) = C_h(\mathbf{z})^{-1} e^{h(\tilde{\mathbf{z}})^\top h(\mathbf{z})/\tau} \\ \text{with } C_h(\mathbf{z}) := \int e^{h(\tilde{\mathbf{z}})^\top h(\mathbf{z})/\tau} d\tilde{\mathbf{z}},$$

$$p(\mathbf{z}) = |\mathcal{Z}|^{-1}$$

$$= -\frac{1}{\tau} \mathbb{E}_{\tilde{\mathbf{z}}, \mathbf{z} \sim p(\tilde{\mathbf{z}}, \mathbf{z})} [h(\tilde{\mathbf{z}})^\top h(\mathbf{z})] \quad (22)$$

$$+ \mathbb{E}_{\mathbf{z} \sim p(\mathbf{z})} \left[\log |\mathcal{Z}| \mathbb{E}_{\tilde{\mathbf{z}} \sim p(\tilde{\mathbf{z}})} \left[e^{h(\tilde{\mathbf{z}})^\top h(\mathbf{z})/\tau} \right] \right] \quad (23)$$

$$= -\frac{1}{\tau} \mathbb{E}_{\tilde{\mathbf{z}}, \mathbf{z} \sim p(\tilde{\mathbf{z}}, \mathbf{z})} [h(\tilde{\mathbf{z}})^\top h(\mathbf{z})] \quad (24)$$

$$+ \mathbb{E}_{\mathbf{z} \sim p(\mathbf{z})} \left[\log \mathbb{E}_{\tilde{\mathbf{z}} \sim p(\tilde{\mathbf{z}})} \left[e^{h(\tilde{\mathbf{z}})^\top h(\mathbf{z})/\tau} \right] \right] + \log |\mathcal{Z}|.$$

$$\begin{aligned}
& - \frac{1}{\tau} \mathbb{E}_{\tilde{\mathbf{z}}, \mathbf{z} \sim p(\tilde{\mathbf{z}}, \mathbf{z})} [h(\tilde{\mathbf{z}})^\top h(\mathbf{z})] \\
& + \mathbb{E}_{\mathbf{z} \sim p(\mathbf{z})} \left[\log \mathbb{E}_{\tilde{\mathbf{z}} \sim p(\tilde{\mathbf{z}})} \left[e^{h(\tilde{\mathbf{z}})^\top h(\mathbf{z}) / \tau} \right] \right] + \log |\mathcal{Z}|.
\end{aligned} \tag{24}$$

$$h = f \circ g,$$

$$\begin{aligned}
& = - \frac{1}{\tau} \mathbb{E}_{\tilde{\mathbf{z}}, \mathbf{z} \sim p(\tilde{\mathbf{z}}, \mathbf{z})} [(f \circ g)(\tilde{\mathbf{z}})^\top (f \circ g)(\mathbf{z})] \\
& \quad + \mathbb{E}_{\mathbf{z} \sim p(\mathbf{z})} \left[\log \mathbb{E}_{\tilde{\mathbf{z}} \sim p(\tilde{\mathbf{z}})} \left[e^{(f \circ g)(\tilde{\mathbf{z}})^\top (f \circ g)(\mathbf{z}) / \tau} \right] \right] \\
& \quad + \log |\mathcal{Z}|,
\end{aligned}$$



$$\begin{aligned}
\mathcal{L}_{\text{align}}(f; \tau) &:= - \frac{1}{\tau} \mathbb{E}_{(\tilde{\mathbf{z}}, \mathbf{z}) \sim p(\tilde{\mathbf{z}}, \mathbf{z})} [(f \circ g)(\tilde{\mathbf{z}})^\top (f \circ g)(\mathbf{z})] \\
\mathcal{L}_{\text{uni}}(f; \tau) &:= \mathbb{E}_{\mathbf{z} \sim p(\mathbf{z})} \left[\log \mathbb{E}_{\tilde{\mathbf{z}} \sim p(\tilde{\mathbf{z}})} \left[e^{(f \circ g)(\tilde{\mathbf{z}})^\top (f \circ g)(\mathbf{z}) / \tau} \right] \right].
\end{aligned}$$

$$= \mathcal{L}_{\text{align}}(f; \tau) + \mathcal{L}_{\text{uni}}(f; \tau) + \log |\mathcal{Z}|,$$

$$= \lim_{M \rightarrow \infty} \mathcal{L}_{\text{contr}}(f; \tau, M) - \log M + \log |\mathcal{Z}|.$$

Proposition A (Asymptotics of $\mathcal{L}_{\text{contr}}$, Wang & Isola, 2020).
For fixed $\tau > 0$, as the number of negative samples $M \rightarrow \infty$,
the (normalized) contrastive loss converges to

$$\begin{aligned}
& \lim_{M \rightarrow \infty} \mathcal{L}_{\text{contr}}(f; \tau, M) - \log M \\
& = \mathcal{L}_{\text{align}}(f; \tau) + \mathcal{L}_{\text{uni}}(f; \tau),
\end{aligned} \tag{12}$$

$$h = f \circ g,$$

Proposition 1 (Minimizers of the cross-entropy maintain the dot product). *Let $\mathcal{Z} = \mathbb{S}^{N-1}$, $\tau > 0$ and consider the ground-truth conditional distribution of the form $p(\tilde{\mathbf{z}}|\mathbf{z}) = C_p^{-1} \exp(\kappa \tilde{\mathbf{z}}^\top \mathbf{z})$. Let h map onto a hypersphere with radius $\sqrt{\tau \kappa}$.⁴ Consider the conditional distribution q_h parameterized by the model, as defined above in Theorem 1, where the hypothesis class for h is assumed to be sufficiently flexible such that $p(\tilde{\mathbf{z}}|\mathbf{z})$ and $q_h(\tilde{\mathbf{z}}|\mathbf{z})$ can match. If h is a minimizer of the cross-entropy $\mathbb{E}_{p(\tilde{\mathbf{z}}|\mathbf{z})}[-\log q_h(\tilde{\mathbf{z}}|\mathbf{z})]$, then $p(\tilde{\mathbf{z}}|\mathbf{z}) = q_h(\tilde{\mathbf{z}}|\mathbf{z})$ and $\forall \mathbf{z}, \tilde{\mathbf{z}} : \kappa \mathbf{z}^\top \tilde{\mathbf{z}} = h(\mathbf{z})^\top h(\tilde{\mathbf{z}})$.*

⁴Note that in practice this can be implemented as a learnable rescaling operation of the network f .

Proof. By assumption, $q_h(\tilde{\mathbf{z}}|\mathbf{z})$ is powerful enough to match $p(\tilde{\mathbf{z}}|\mathbf{z})$ for the correct choice of h — in particular, for $h(\mathbf{z}) = \sqrt{\tau \kappa} \mathbf{z}$.

$$p(\tilde{\mathbf{z}}|\mathbf{z}) = q_h(\tilde{\mathbf{z}}|\mathbf{z}).$$

$$p(\mathbf{z}|\mathbf{z}) = q_h(\mathbf{z}|\mathbf{z})$$

$$C_p^{-1} e^{\kappa \mathbf{z}^\top \mathbf{z}} = C_h(\mathbf{z})^{-1} e^{h(\mathbf{z})^\top h(\mathbf{z})/\tau}$$

$$C_p^{-1} e^\kappa = C_h(\mathbf{z})^{-1} e^\kappa$$

$$C_p = C_h.$$

$$e^{\kappa \mathbf{z}^\top \tilde{\mathbf{z}}} = e^{h(\mathbf{z})^\top h(\tilde{\mathbf{z}})} \Leftrightarrow \kappa \mathbf{z}^\top \tilde{\mathbf{z}} = h(\mathbf{z})^\top h(\tilde{\mathbf{z}}).$$

3.2 Contrastive learning identifies ground-truth factors on the hypersphere

Proposition 2 (Extension of the Mazur-Ulam theorem to hyperspheres and the dot product). *Let $\mathcal{Z} = \mathbb{S}^{N-1}$ and $\mathcal{Z}' = \mathbb{S}_r^{N-1}$ be the hyperspheres with radius 1 and $r > 0$, respectively. If $h : \mathbb{R}^N \rightarrow \mathcal{Z}'$ is differentiable in the vicinity of \mathcal{Z} and its restriction to \mathcal{Z} maintains the dot product up to a constant factor, i.e., $\forall \mathbf{z}, \tilde{\mathbf{z}} \in \mathcal{Z} : r^2 \mathbf{z}^\top \tilde{\mathbf{z}} = h(\mathbf{z})^\top h(\tilde{\mathbf{z}})$, then h is an orthogonal linear transformation scaled by r for all $\mathbf{z} \in \mathcal{Z}$.*

Proof. First, we begin with the case $r = 1$. As h maintains the dot product we have:

$$\forall \mathbf{z}, \tilde{\mathbf{z}} \in \mathcal{Z} : \mathbf{z}^\top \tilde{\mathbf{z}} = h(\mathbf{z})^\top h(\tilde{\mathbf{z}}). \quad (37)$$

We consider the partial derivative w.r.t. \mathbf{z} and obtain:

$$\forall \mathbf{z}, \tilde{\mathbf{z}} \in \mathcal{Z} : \tilde{\mathbf{z}} = \mathbf{J}_h^\top(\mathbf{z}) h(\tilde{\mathbf{z}}). \quad (38)$$


Taking the partial derivative w.r.t. $\tilde{\mathbf{z}}$ yields

$$\forall \mathbf{z}, \tilde{\mathbf{z}} \in \mathcal{Z} : \mathbf{I} = \mathbf{J}_h^\top(\mathbf{z}) \mathbf{J}_h(\tilde{\mathbf{z}}). \quad (39)$$

We can now conclude

$$\forall \mathbf{z}, \tilde{\mathbf{z}} \in \mathcal{Z} : \mathbf{J}_h(\tilde{\mathbf{z}})^{-1} = \mathbf{J}_h^\top(\mathbf{z}). \quad (40)$$

which implies a constant Jacobian matrix $\mathbf{J}_h(\mathbf{z}) = \mathbf{J}_h$ as the identity holds on all points in \mathcal{Z} , and further that the Jacobian \mathbf{J}_h is orthogonal. Hence, $\forall \mathbf{z} \in \mathcal{Z} : h(\mathbf{z}) = \mathbf{J}_h \mathbf{z}$ is an orthogonal linear transformation.



Theorem 2. *Let $\mathcal{Z} = \mathbb{S}^{N-1}$, the ground-truth marginal be uniform, and the conditional a vMF distribution (cf. Eq. 2). Let the restriction of the mixing function g to \mathcal{Z} be injective and h be differentiable in a vicinity of \mathcal{Z} . If the assumed form of q_h , as defined above, matches that of p , and if f is differentiable and minimizes the CL loss as defined in Eq. (1), then for fixed $\tau > 0$ and $M \rightarrow \infty$, $h = f \circ g$ is linear, i.e., f recovers the latent sources up to an orthogonal linear transformation and a constant scaling factor.*

3.3 Contrastive learning identifies ground-truth factors on convex bodies in \mathbb{R}^N



1. From hypersphere to a convex body, such as hyperrectangle

2. Three steps of proof:
 - a. Extend based on ‘alignment and uniformity’ paper.
 - b. Derive minimizers are isometries of the latent space.
 - c. Show that minimizers must be affine transformations.

$$p(\mathbf{z}) = |\mathcal{Z}|^{-1}, \quad p(\mathbf{z}|\tilde{\mathbf{z}}) = C_p^{-1} e^{-\delta(\mathbf{z}, \tilde{\mathbf{z}})} \quad \text{with} \\ C_p(\mathbf{z}) := \int e^{-\delta(\mathbf{z}, \tilde{\mathbf{z}})} d\tilde{\mathbf{z}}, \quad \mathbf{x} = g(\mathbf{z}), \quad \tilde{\mathbf{x}} = g(\tilde{\mathbf{z}}),$$

δ is similarity measure.

$$\mathcal{L}_{\delta\text{-contr}}(f; \tau, M) \quad := \quad (6) \\ \mathbb{E}_{\substack{(\mathbf{x}, \tilde{\mathbf{x}}) \sim p_{\text{pos}} \\ \{\mathbf{x}_i^-\}_{i=1}^M \stackrel{\text{i.i.d.}}{\sim} p_{\text{data}}}} \left[-\log \frac{e^{-\delta(f(\mathbf{x}), f(\tilde{\mathbf{x}}))/\tau}}{e^{-\delta(f(\mathbf{x}), f(\tilde{\mathbf{x}}))/\tau} + \sum_{i=1}^M e^{-\delta(f(\mathbf{x}), f(\mathbf{x}_i^-))/\tau}} \right].$$

Theorem 3. Let δ be a semi-metric and $\tau, \lambda > 0$ and let the ground-truth marginal distribution p be uniform. Consider a ground-truth conditional distribution $p(\tilde{\mathbf{z}}|\mathbf{z}) = C_p^{-1}(\mathbf{z}) \exp(-\lambda\delta(\tilde{\mathbf{z}}, \mathbf{z}))$ and the model conditional distribution

$$q_h(\tilde{\mathbf{z}}|\mathbf{z}) = C_h^{-1}(\mathbf{z}) e^{-\delta(h(\tilde{\mathbf{z}}), h(\mathbf{z}))/\tau}$$

with $C_h(\mathbf{z}) := \int_{\mathcal{Z}} e^{-\delta(h(\tilde{\mathbf{z}}), h(\mathbf{z}))/\tau} d\tilde{\mathbf{z}}.$ (48)

Then the cross-entropy between p and q_h is given by

$$\lim_{M \rightarrow \infty} \mathcal{L}_{\delta\text{-contr}}(f; \tau, M) - \log M + \log |\mathcal{Z}| = \mathbb{E}_{\mathbf{z} \sim p(\mathbf{z})} [H(p(\cdot|\mathbf{z}), q_h(\cdot|\mathbf{z}))],$$
 (49)

which can be implemented by sampling data from the accessible distributions.

Proposition 4 (Minimizers of the cross-entropy are isometries). Let δ be a semi-metric. Consider the conditional distributions of the form $p(\tilde{\mathbf{z}}|\mathbf{z}) = C_p^{-1}(\mathbf{z}) \exp(-\delta(\tilde{\mathbf{z}}, \mathbf{z})/\lambda)$ and

$$q_h(\tilde{\mathbf{z}}|\mathbf{z}) = C_h^{-1}(\mathbf{z}) e^{-\delta(h(\tilde{\mathbf{z}}), h(\mathbf{z}))/\tau}$$

with $C_h(\mathbf{z}) := \int_{\mathcal{Z}} e^{-\delta(h(\tilde{\mathbf{z}}), h(\mathbf{z}))/\tau} d\tilde{\mathbf{z}},$ (64)

where the hypothesis class for h is assumed to be sufficiently flexible such that $p(\tilde{\mathbf{z}}|\mathbf{z})$ and $q_h(\tilde{\mathbf{z}}|\mathbf{z})$ can match for any point \mathbf{z} . If h is a minimizer of the cross-entropy $\mathcal{L}_{\text{CE}} = \mathbb{E}_{p(\tilde{\mathbf{z}}|\mathbf{z})} [-\log q_h(\tilde{\mathbf{z}}|\mathbf{z})]$, then h is an isometry, i.e., $\forall \mathbf{z}, \tilde{\mathbf{z}} \in \mathcal{Z} : \lambda\tau\delta(\mathbf{z}, \tilde{\mathbf{z}}) = \delta(h(\mathbf{z}), h(\tilde{\mathbf{z}})).$

Theorem 4. Let $\mathcal{Z} = \mathcal{Z}'$ be a convex body in \mathbb{R}^N . Let the mixing function g be differentiable and invertible. If the assumed form of q_h as defined in Eq. (42) matches that of p , and if f is differentiable and minimizes the cross-entropy between p and q_h , then we find that $h = f \circ g$ is affine, i.e., we recover the latent sources up to affine transformations.

Theorem 5. Let \mathcal{Z} be a convex body in \mathbb{R}^N , $h = f \circ g : \mathcal{Z} \rightarrow \mathcal{Z}$, and δ be a metric or a semi-metric (cf. Lemma 1 in Appx. A.2.4), induced by a norm. Further, let the ground-truth marginal distribution be uniform and the conditional distribution be as Eq. (5). Let the mixing function g be differentiable and injective. If the assumed form of q_h matches that of p , i.e.,

$$q_h(\tilde{\mathbf{z}}|\mathbf{z}) = C_q^{-1}(\mathbf{z}) e^{-\delta(h(\tilde{\mathbf{z}}), h(\mathbf{z}))/\tau}$$

with $C_q(\mathbf{z}) := \int e^{-\delta(h(\tilde{\mathbf{z}}), h(\mathbf{z}))/\tau} d\tilde{\mathbf{z}},$ (7)

and if f is differentiable and minimizes the $\mathcal{L}_{\delta\text{-contr}}$ objective in Eq. (6) for $M \rightarrow \infty$, we find that $h = f \circ g$ is invertible and affine, i.e., we recover the latent sources up to affine transformations.

Theorem 6. *Let \mathcal{Z} be a convex body in \mathbb{R}^N , $h : \mathcal{Z} \rightarrow \mathcal{Z}$, and δ be an L^α metric or semi-metric (cf. Lemma 1 in Appx. A.2.4) for $\alpha \geq 1, \alpha \neq 2$. Further, let the ground-truth marginal distribution be uniform and the conditional distribution be as Eq. (5), and let the mixing function g be differentiable and invertible. If the assumed form of $q_h(\cdot|\mathbf{z})$ matches that of $p(\cdot|\mathbf{z})$, i.e., both use the same metric δ up to a constant scaling factor, and if f is differentiable and minimizes the $\mathcal{L}_{\delta\text{-contr}}$ objective in Eq. (6) for $M \rightarrow \infty$, we find that $h = f \circ g$ is a composition of input independent permutations, sign flips and rescaling.*



Experiments

Metrics:

coefficient of determination (R^2)

mean correlation coefficient (MCC).

Reference: [Unsupervised feature extraction by time-contrastive learning and nonlinear ICA](#)

Table 1. Identifiability up to affine transformations. Mean \pm standard deviation over 5 random seeds. Note that only the first row corresponds to a setting that matches (\checkmark) our theoretical assumptions, while the others show results for violated assumptions (\times ; see column M). Note that the identity score only depends on the ground-truth space and the marginal distribution defined for the generative process, while the supervised score additionally depends on the space assumed by the model.

Generative process g			Model f		M	R^2 Score [%]		
Space	$p(\cdot)$	$p(\cdot \cdot)$	Space	$q_h(\cdot \cdot)$		Identity	Supervised	Unsupervised
Sphere	Uniform	vMF($\kappa=1$)	Sphere	vMF($\kappa=1$)	\checkmark	66.98 ± 2.79	99.71 ± 0.05	99.42 ± 0.05
Sphere	Uniform	vMF($\kappa=10$)	Sphere	vMF($\kappa=1$)	\times	— —	— —	99.86 ± 0.01
Sphere	Uniform	Laplace($\lambda=0.05$)	Sphere	vMF($\kappa=1$)	\times	— —	— —	99.91 ± 0.01
Sphere	Uniform	Normal($\sigma=0.05$)	Sphere	vMF($\kappa=1$)	\times	— —	— —	99.86 ± 0.00
Box	Uniform	Normal($\sigma=0.05$)	Unbounded	Normal	\times	67.93 ± 7.40	99.78 ± 0.06	99.60 ± 0.02
Box	Uniform	Laplace($\lambda=0.05$)	Unbounded	Normal	\times	— —	— —	99.64 ± 0.02
Box	Uniform	Laplace($\lambda=0.05$)	Unbounded	GenNorm($\beta=3$)	\times	— —	— —	99.70 ± 0.02
Box	Uniform	Normal($\sigma=0.05$)	Unbounded	GenNorm($\beta=3$)	\times	— —	— —	99.69 ± 0.02
Sphere	Normal($\sigma=1$)	Laplace($\lambda=0.05$)	Sphere	vMF($\kappa=1$)	\times	63.37 ± 2.41	99.70 ± 0.07	99.02 ± 0.01
Sphere	Normal($\sigma=1$)	Normal($\sigma=0.05$)	Sphere	vMF($\kappa=1$)	\times	— —	— —	99.02 ± 0.02
Unbounded	Laplace($\lambda=1$)	Normal($\sigma=1$)	Unbounded	Normal	\times	62.49 ± 1.65	99.65 ± 0.04	98.13 ± 0.14
Unbounded	Normal($\sigma=1$)	Normal($\sigma=1$)	Unbounded	Normal	\times	63.57 ± 2.30	99.61 ± 0.17	98.76 ± 0.03

Table 2. Identifiability up to generalized permutations, averaged over 5 runs. Note that while Theorem 6 requires the model latent space to be a convex body and $p(\cdot|\cdot) = q_h(\cdot|\cdot)$, we find that empirically either is sufficient. The results are grouped in four blocks corresponding to different types and degrees of violation of assumptions of our theory showing identifiability up to permutations: (1) no violation, violation of the assumptions on either the (2) space or (3) the conditional distribution, or (4) both.

Space	Generative process g		Space	Model f $q_h(\cdot \cdot)$	M.	Identity	MCC Score [%]	
	$p(\cdot)$	$p(\cdot \cdot)$					Supervised	Unsupervised
Box	Uniform	Laplace($\lambda=0.05$)	Box	Laplace	✓	46.55 ± 1.34	99.93 ± 0.03	98.62 ± 0.05
Box	Uniform	GenNorm($\beta=3; \lambda=0.05$)	Box	GenNorm($\beta=3$)	✓	— —	— —	99.90 ± 0.06
Box	Uniform	Normal($\sigma=0.05$)	Box	Normal	✗	— —	— —	99.77 ± 0.01
Box	Uniform	Laplace($\lambda=0.05$)	Box	Normal	✗	— —	— —	99.76 ± 0.02
Box	Uniform	GenNorm($\beta=3; \lambda=0.05$)	Box	Laplace	✗	— —	— —	98.80 ± 0.02
Box	Uniform	Laplace($\lambda=0.05$)	Unbounded	Laplace	✗	— —	99.97 ± 0.03	98.57 ± 0.02
Box	Uniform	GenNorm($\beta=3; \lambda=0.05$)	Unbounded	GenNorm($\beta=3$)	✗	— —	— —	99.85 ± 0.01
Box	Uniform	Normal($\sigma=0.05$)	Unbounded	Normal	✗	— —	— —	58.26 ± 3.00
Box	Uniform	Laplace($\lambda=0.05$)	Unbounded	Normal	✗	— —	— —	59.67 ± 2.33
Box	Uniform	Normal($\sigma=0.05$)	Unbounded	GenNorm($\beta=3$)	✗	— —	— —	43.80 ± 2.15

Experiments

Dataset : 3DIDENT

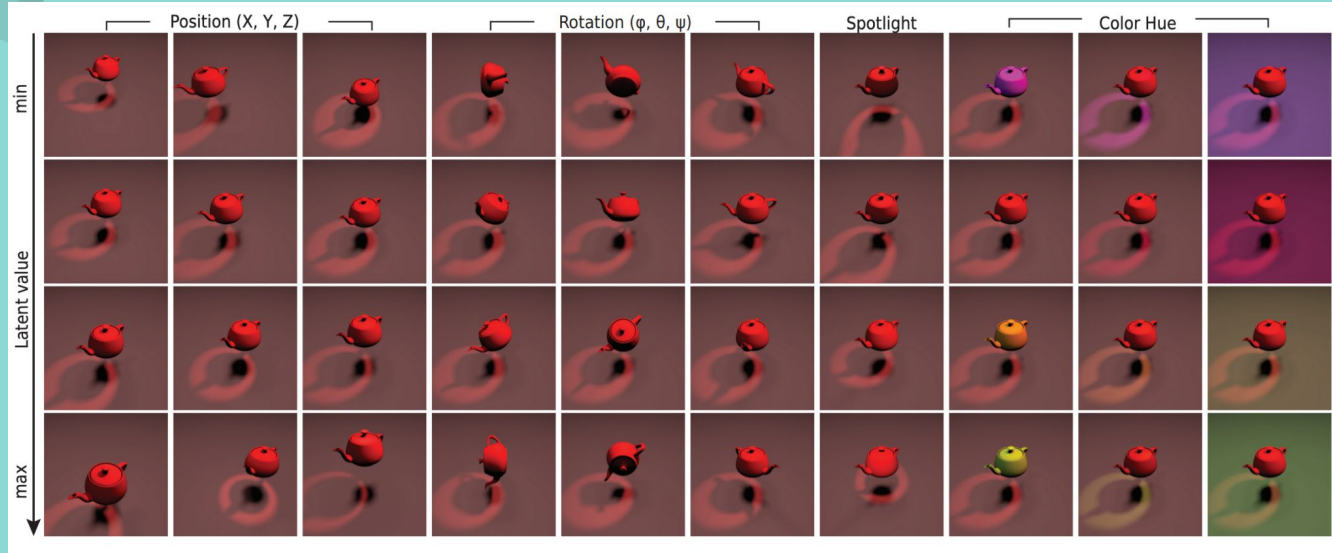


Table 4. Identifiability up to affine transformations on the test set of 3DIdent. Mean \pm standard deviation over 3 random seeds. As earlier, only the first row corresponds to a setting that matches the theoretical assumptions for linear identifiability; the others show distinct violations. Supervised training with unbounded space achieves scores of $R^2 = (98.67 \pm 0.03)\%$ and $\text{MCC} = (99.33 \pm 0.01)\%$. The last row refers to using the image augmentations suggested by [Chen et al. \(2020a\)](#) to generate positive image pairs. For performance on the training set, see Appx. Table 5.

Dataset $p(\cdot \cdot)$	Space	Model f $q_h(\cdot \cdot)$	M.	Identity [%] R^2	Unsupervised [%]	
					R^2	MCC
Normal	Box	Normal	✓	5.25 ± 1.20	96.73 ± 0.10	98.31 ± 0.04
Normal	Unbounded	Normal	✗	— —	96.43 ± 0.03	54.94 ± 0.02
Laplace	Box	Normal	✗	— —	96.87 ± 0.08	98.38 ± 0.03
Normal	Sphere	vMF	✗	— —	65.74 ± 0.01	42.44 ± 3.27
Augm.	Sphere	vMF	✗	— —	45.51 ± 1.43	46.34 ± 1.59



Conclusion

1. InfoNCE objectives reveal underlying data generative factors in self-supervised learning. Effective even when theoretical assumptions are not precisely met.
2. Representations learned through contrastive learning implicitly reverse data generation processes. Explains their utility in downstream tasks.

Future works:

1. Adapting the framework to different marginal distributions beyond uniform.



THANK YOU