Contrastive Learning Inverts the Data Generating Process

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

University of Tubingen, IMPRS-IS & EPFL

Dec 15, 2023

Presented by Qin Weng

Introduction

The paper demonstrates that feed-forward models trained with a contrastive loss from InfoNCE family can effectively invert the underlying generative model.

Significance:

- Establish theoretical connection between contrastive learning (CL), generative modeling, and nonlinear independent component analysis (ICA).
- Explain why contrastive learning with InfoNCE objectives, which is commonly used in self-supervised learning, can be effective in a wide rage of downstream practical tasks.
- Reveal the implicit assumptions under which contrastive learning works, and propose for methods improvement by avoid violating these assumptions.

Introduction

Framework

- Proved that training with InfoNCE inverts the data generating process under certain statistical assumptions.
- Conduct simulation experiments to empirically validate the theoretical findings, testing for identifiability of source signals whether the assumptions hold or be violated.
- Demonstrate that a contrastive loss derived from our theoretical framework can identify the ground-truth factors of complex, high-resolution images which mimics natural features.

Background

Contrastive learning (CL)

Despite the success of contrastive learning, the understanding of the learned representations remains limited.

CL motivation theories:

- InfoMAX principle: maximize the mutual information between different views
- Latent classes
- <u>**alignment** and **uniformity** properties of representations</u>

Background

Nonlinear Independent Components Analysis (ICA)

Demixing problem:

- Find the underlying source for observed data
- Given observed data, it aims to find a model f that equals the inversed generative model g^{-1}, which allows for the original sources to be recovered.

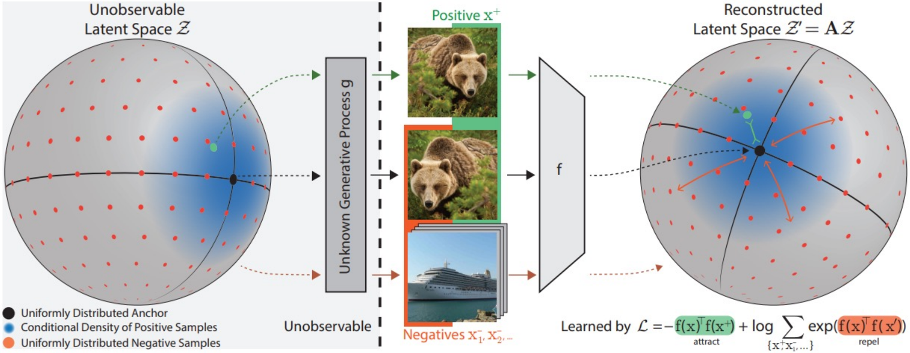
Theory (sphere)

Ground-truth source $z \in \mathbb{Z}$

Observations $\boldsymbol{x} = g(\boldsymbol{z})$, by generative model g

recovered source signals (representations) $\boldsymbol{z}' = f(\boldsymbol{x})$, by learned feature encoder f

h to map between true source signals $\boldsymbol{z}$ and estimated source signals $\boldsymbol{z}'$: $h = f \circ g$, $h(\boldsymbol{z}) = \boldsymbol{z}'$



Theory (sphere)

Latent source distribution assumptions: $(Z \in \mathbb{R}^n)$

Space: sphere

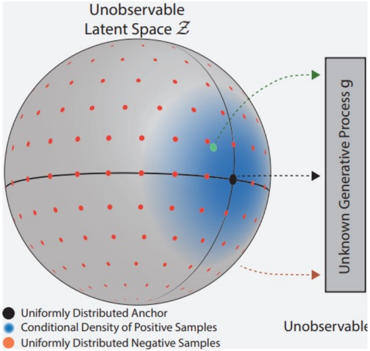
- normalize Z to hypersphere S^{N-1}

Source marginal distribution: **uniformity**

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

Source conditional distribution (for pairs): **alignment**

- $P(\cdot|\cdot)$: $p(\mathbf{z}|\tilde{\mathbf{z}}) = C_p^{-1} e^{\kappa \mathbf{z}^\top \tilde{\mathbf{z}}}$ with $C_p := \int e^{\kappa \mathbf{z}^\top \tilde{\mathbf{z}}} \, d\tilde{\mathbf{z}} = \text{const.}$,



Theory (sphere)

InfoNCE loss

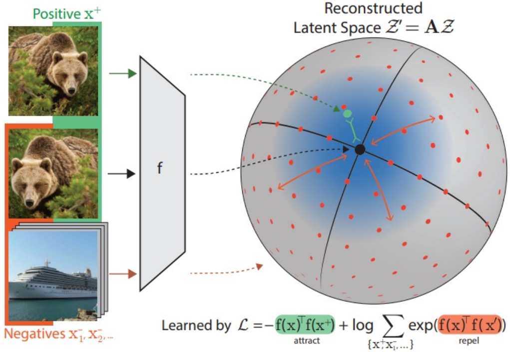
M: fixed number of negative samples

p_{data}: distribution of all observations

p_{pos}: distribution of positive pairs

$$\mathcal{L}_{\text{contr}}(f;\tau,M) := \tag{1}$$

$$\mathop{\mathbb{E}}_{\substack{(\mathbf{x},\tilde{\mathbf{x}})\sim p_{\text{pos}} \\ \{\mathbf{x}_i^-\}_{i=1}^M \overset{\text{i.i.d.}}{\sim} p_{\text{data}}}} \left[-\log \frac{e^{f(\mathbf{x})^\mathsf{T} f(\tilde{\mathbf{x}})/\tau}}{e^{f(\mathbf{x})^\mathsf{T} f(\tilde{\mathbf{x}})/\tau} + \sum_{i=1}^{M} e^{f(\mathbf{x}_i^-)^\mathsf{T} f(\tilde{\mathbf{x}})/\tau}} \right].$$



Theory (sphere)

Step 1: interpret CL loss as cross-entropy for source conditional distribution

Theorem (Wang & Isola, 2020):
Contrastive learning loss converges to cross-entropy between latent distributions.

Given uniform marginal distribution:

$$\lim_{M\to\infty} \mathcal{L}_{\mathsf{contr}}(f;\tau,M) - \log M + \log|\mathcal{Z}| = \underset{\mathbf{z}\sim p(\mathbf{z})}{\mathbb{E}}\left[H(p(\cdot|\mathbf{z}), q_h(\cdot|\mathbf{z}))\right]$$

- True: $p(\mathbf{z}|\tilde{\mathbf{z}}) = C_p^{-1} e^{\kappa \mathbf{z}^\top \tilde{\mathbf{z}}}$ with $C_p := \int e^{\kappa \mathbf{z}^\top \tilde{\mathbf{z}}} \, \mathrm{d}\tilde{\mathbf{z}}$
- Estimated: $q_h(\tilde{\mathbf{z}}|\mathbf{z}) = C_h(\mathbf{z})^{-1} e^{h(\tilde{\mathbf{z}})^\top h(\mathbf{z})/\tau}$ *with* $C_h(\mathbf{z}) := \int e^{h(\tilde{\mathbf{z}})^\top h(\mathbf{z})/\tau} \, \mathrm{d}\tilde{\mathbf{z}},$

Theory (sphere)

Step 2: minimizer h^* preserves the dot product (distance)

- True: $p(\mathbf{z}|\tilde{\mathbf{z}}) = C_p^{-1} e^{\kappa \mathbf{z}^\top \tilde{\mathbf{z}}}$ with $C_p := \int e^{\kappa \mathbf{z}^\top \tilde{\mathbf{z}}} \, \mathrm{d}\tilde{\mathbf{z}}$
- Estimated: $q_{\mathrm{h}}(\tilde{\mathbf{z}}|\mathbf{z}) = C_h(\mathbf{z})^{-1} e^{h(\tilde{\mathbf{z}})^\top h(\mathbf{z})/\tau}$ *with* $C_h(\mathbf{z}) := \int e^{h(\tilde{\mathbf{z}})^\top h(\mathbf{z})/\tau} \, \mathrm{d}\tilde{\mathbf{z}},$

Assume h (thus f) is sufficiently flexible that estimation can match with truth

$$p(\tilde{\mathbf{z}}|\mathbf{z}) = q_{\mathrm{h}}(\tilde{\mathbf{z}}|\mathbf{z})$$

For minimizer h^* of the cross-entropy loss:

$$\forall \mathbf{z}, \tilde{\mathbf{z}} : \kappa \mathbf{z}^\top \tilde{\mathbf{z}} = h(\mathbf{z})^\top h(\tilde{\mathbf{z}})$$

Theory (sphere)

Step 3: leverage distance preservation to show generative model g has been inverted

$$\forall \mathbf{z}, \tilde{\mathbf{z}} : \kappa \mathbf{z}^\top \tilde{\mathbf{z}} = h(\mathbf{z})^\top h(\tilde{\mathbf{z}})$$

h^* is an orthogonal linear transformation

h^* (and thus f^*) solves demixing problem up to orthogonal linear transformations

(i.e., h^* recovers latent source space Z in the representation space Z', except for permutation, rotation, and sign flips)

Theory (convex body)

***Similar results for general convex bodies with general similarity measures**

Re-define latent source distribution assumptions: ($Z \in \mathbb{R}^n$)

- Space: convex body
- Source marginal distribution: uniform
- Source conditional distribution (for pairs):

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

δ is a general similarity metric induced by a norm

Re-define InfoNCE loss:

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

Theory (convex body)

$$q_{\mathrm{h}}(\tilde{\mathbf{z}}|\mathbf{z}) = C_q^{-1}(\mathbf{z})e^{-\delta(h(\tilde{\mathbf{z}}),h(\mathbf{z}))/\tau} \quad with \quad C_q(\mathbf{z}) := \int e^{-\delta(h(\tilde{\mathbf{z}}),h(\mathbf{z}))/\tau} \, \mathrm{d}\tilde{\mathbf{z}},$$

Let minimizer h^* of the cross-entropy loss to match $p(\tilde{\mathbf{z}}|\mathbf{z}) = q_{\mathrm{h}}(\tilde{\mathbf{z}}|\mathbf{z})$

h^* (and thus f^*) solves demixing problem up to affine transformations

Under the mild restriction that the source conditional distribution is based on an L^p similarity measure for p>2

h^* solves demixing problem up to generalized permutations

(i.e., recovers the latent sources except for permutation, sign flips and rescaling)

Experiments (validation)

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

Assumption violations do not lead to performance drop in affine identifiability

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

Assumption violations lead to performance drop in permutation identifiability

Experiments (validation)

Violation of uniform marginal assumptions influence the identifiability:

Performance drop drastically once the marginal distribution is more concentrated than the conditional distribution of positive pairs.
In such scenarios, positive pairs are indistinguishable from negative pairs.

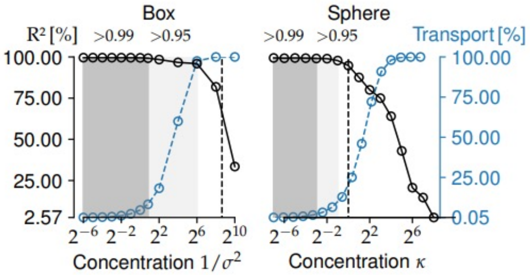
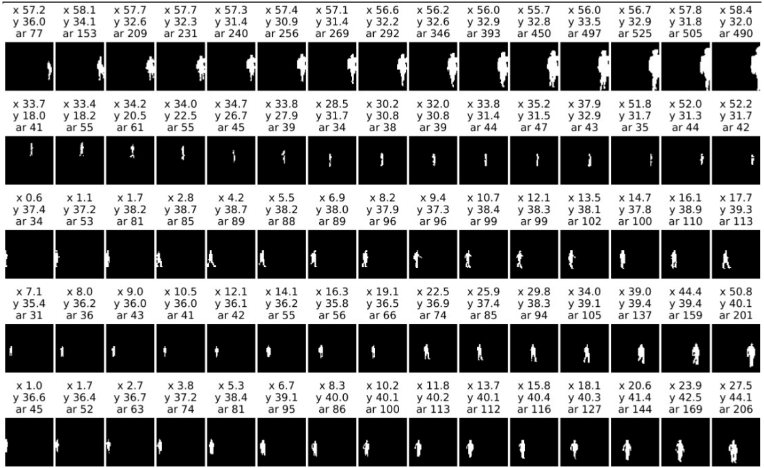


Figure 2. Varying degrees of violation of the uniformity assumption for the marginal distribution. The figure shows the R^2 score measuring identifiability up to linear transformations (black) as well as the difference between the used marginal and assumed uniform distribution in terms of probability mass (blue) as a function of the marginal's concentration. The black dotted line indicates the concentration of the used conditional distribution.

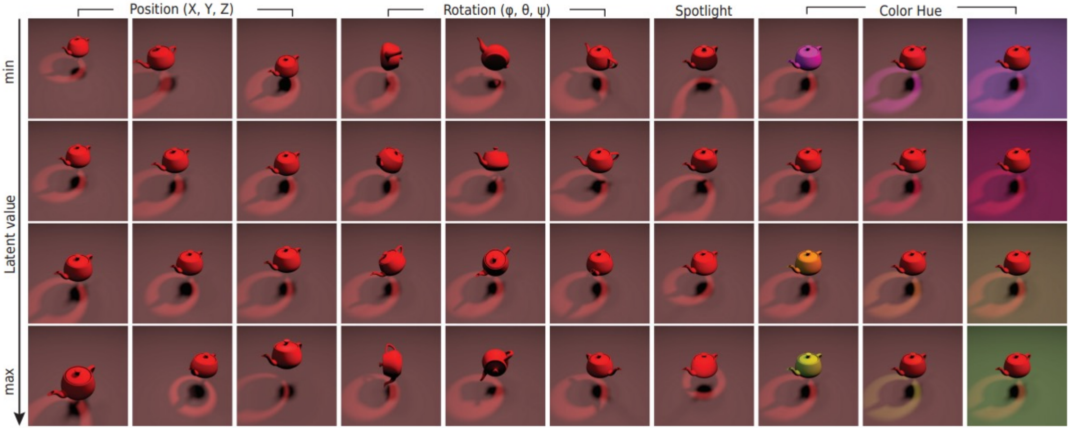
Experiments (KITTI Masks)

Table 3. **KITTI Masks**. Mean ± standard deviation over 10 random seeds. $\overline{\Delta t}$ indicates the average temporal distance of frames used.

	Model	Model Space	MCC [%]
$\overline{\Delta t} = 0.05s$	SlowVAE	Unbounded	66.1 ± 4.5
	Laplace	Unbounded	77.1 ± 1.0
	Laplace	Box	74.1 ± 4.4
	Normal	Unbounded	58.3 ± 5.4
	Normal	Box	59.9 ± 5.5
$\overline{\Delta t} = 0.15s$	SlowVAE	Unbounded	79.6 ± 5.8
	Laplace	Unbounded	79.4 ± 1.9
	Laplace	Box	80.9 ± 3.8
	Normal	Unbounded	60.2 ± 8.7
	Normal	Box	68.4 ± 6.7



Experiments (3DIdent)



The dataset contains 250 000 observation-latent pairs where the latents are uniformly sampled from the hyperrectangle Z.

Experiments (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 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	Model f			Identity [%]	Unsupervised [%]	
$p(\cdot\|\cdot)$	Space	$q_h(\cdot\|\cdot)$	M.	R^2	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

- InfoNCE objectives can uncover the true generative factors of data variability.
- Weak statistical assumptions are enough to identify these factors, even if not perfectly aligned with theory.
- Learned representations can approximate the data's generative process, beneficial for downstream tasks.

Recommendation

Is it worth reading? Yes

- Contributes to the theoretical foundations for a number of advancing self-supervised learning algorithms.
- Suggests ways to construct more effective contrastive learning.
- The research framework is a good example for theoretical study

Future work

Potential for extending the framework beyond the uniform implicitly encoded in InfoNCE.