

f-Mutual Information Contrastive Learning

Guojun Zhang*, Yiwei Lu*, Sun Sun, Hongyu Guo, Yaoliang Yu



Our Contribution:

- (1) we propose a novel framework for contrastive learning with a general f -divergence family
- (2) we provide an optimal design for the similarity function with Gaussian kernels
- (3) Experimentally, our objectives consistently outperform InfoNCE loss

1. Contrastive Learning: learning an informative representation $g(\cdot)$ by encouraging the contrastiveness between similar (positive: different views of the same image) and dissimilar (negative: different images) sample pairs.

To learn a good representation: (1) positive pairs should be close to each other in the feature space, (2) negative pairs should be far away from each other in the feature space.

2. InfoNCE loss: the most popular contrastive learning loss

$$\mathcal{L} = \mathbb{E}_{(x,y) \sim p_{\text{pos}}} [k(g(x), g(y))] - \mathbb{E}_{(x,y) \sim p_{\text{data}} \otimes p_{\text{data}}} \log \left(\sum_{i=1}^N \exp(k(g(x), g(y_i))) \right)$$

Similarity score between positive pairs Similarity score between negative pairs

InfoNCE can be seen as a lower bound on **mutual information**: $I(X, Y) \geq \log(K) + \mathcal{L}$,

where $I(X; Y) := D_{KL}(p(x, y) \| p(x)p(y))$.

We aim at generalizing the KL divergence to f -divergence.

3. f-Mutual Information and f-MICL objective

Definition 1 (f-mutual information, Csiszár 1967). Consider a pair of random variables (X, Y) with density function $p(x, y)$. The f -mutual information I_f between X and Y is defined as

$$I_f(X; Y) := D_f(p(x, y) \| p(x)p(y)) = \int f \left(\frac{p(x, y)}{p(x)p(y)} \right) p(x)p(y) \cdot d\lambda(x, y), \quad (1)$$

where $f: \mathbb{R}_+ \rightarrow \mathbb{R}$ is (closed) convex with $f(1) = 0$, and recall that $p(x)$ and $p(y)$ are the marginal densities of $p(x, y)$, whereas λ is a dominating measure (e.g. Lebesgue).

(1) Estimating $I_f(X; Y)$ directly is generally challenging so we consider the dual problem instead:

$$I_f(X; Y) \geq \sup_{T \in \mathcal{T}} i_f(X; Y) := \mathbb{E}_{(x,y) \sim p_{\text{pos}}} [T(x, y)] - \mathbb{E}_{(x,y) \sim p_{\text{data}} \otimes p_{\text{data}}} [f^*(T(x, y))],$$

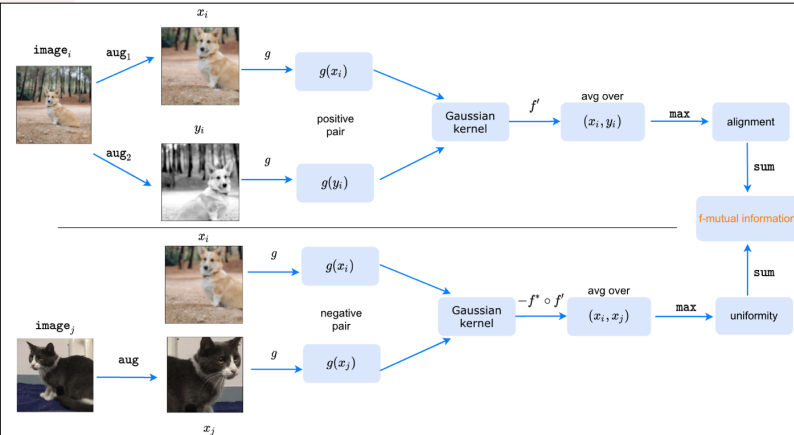
where $f^*(t) := \sup_{x \in \mathbb{R}_+} (xt - f(x))$ is the (monotone) Fenchel conjugate of f , and is always **monotonically increasing**.

(2) f -MICL objective: In contrastive learning:

$$T(x, y) = k(g(x), g(y))$$

Thus our f -MICL objective:

$$\sup_{T \in \mathcal{T}} i_f(X; Y) := \mathbb{E}_{(x,y) \sim p_{\text{pos}}} [k(g(x), g(y))] - \mathbb{E}_{(x,y) \sim p_{\text{data}} \otimes p_{\text{data}}} [f^*(k(g(x), g(y)))]$$



* g : feature embedding; f' : the derivative; f^* : the Fenchel conjugate.

4. Optimal similarity function

We assume that:

$$p_g(g(x), g(y)) \propto G_\sigma(\|g(x) - g(y)\|^2) := \mu \exp\left(-\frac{\|g(x) - g(y)\|^2}{2\sigma^2}\right)$$

And derive: $k^*(g(x), g(y)) = f'(CG_\sigma(\|g(x) - g(y)\|^2))$

Our complete objective is :

$$\mathbb{E}_{(x,y) \sim p_{\text{pos}}} [f' \circ G_\sigma(\|g(x) - g(y)\|^2)] - \mathbb{E}_{(x,y) \sim p_{\text{data}} \otimes p_{\text{data}}} [f^* \circ f' \circ G_\sigma(\|g(x) - g(y)\|^2)]$$

5. Experiments

(1) Overall comparison between SOTA and our f -MICL

Table 1: Test classification accuracy (%) on various datasets with linear evaluation.

Dataset	Baselines			f-MICL					
	SimCLR	Uniformity	RPC	KL	JS	Pearson	SH	Tsallis	VLC
CIFAR-10	89.71	90.41	90.39	90.61	89.66	89.35	89.52	89.15	89.13
CIFAR-100	62.75	62.51	62.66	63.00	63.11	61.69	61.47	60.55	61.19
STL-10	82.97	84.44	82.41	85.33	85.94	82.64	82.80	84.79	83.27
TinyImageNet	30.54	41.10	34.93	39.16	42.88	38.42	40.87	32.95	38.61
ImageNet	57.66	59.12	56.11	58.91	61.11	55.33	52.37	53.11	54.26

(2) Gaussian kernel

Similarity	KL	JS	Pearson	SH	Tsallis	VLC
Cosine	88.95	87.06	87.79	87.06	88.55	10.00
Gaussian	89.13	88.94	89.41	88.24	89.26	89.04

(3) Uniformity and Alignment

