Efficient Distribution Matching of Representations via Noise-Injected Deep InfoMax

Ivan Butakov, Alexander Semenenko, Alexander Tolmachev, Andrey Gladkov, Marina Munkhoeva, Alexey Frolov

April 8, 2025

Center for Next Generation Wireless and IoT

Introduction

- **Representation Learning** involves extracting meaningful low-dimensional embeddings for AI tasks in vision, audio, and NLP. Such embeddings are especially useful for multi-modal learning, statistical and topological analysis, data visualization, and hypothesis testing.
- We focus on **self-supervised learning (SSL)** to eliminate the reliance on labeled data.
- **Contrastive learning**, a key SSL paradigm, encourages similar embeddings for augmented versions of the same data point.
- **Deep InfoMax (DIM)** is an information-theoretic contrastive approach that maximizes useful information contained in the embeddings, offering universality and strong performance.
- **Distribution Matching (DM)** refers to enforcing that embeddings follow a specific latent distribution.

Introduction

Why DM? Enforcing a specific latent distribution is crucial for:

- Generative modeling
- Statistical analysis
- Disentanglement
- Outlier detection

Our Contribution: We propose a simple, cost-effective DIM modification achieving exact DM via specific activation functions and noise injections—eliminating extra networks.

Key Information-Theoretic Quantities:

- $h(X) = -\mathbb{E} \log p(X) \quad (p \text{ is PDF of } X)$
- $I(X; Y) = h(X) - h(X \mid Y)$

Problem Setup
Let X be a high-dimensional random vector and f be an encoder (approximated by a neural network).

We aim to obtain low-dimensional representation $f(X)$.

Information-Theoretic Approach
One can maximize mutual information

$$I(X; f(X)) \to \max$$

to obtain the most informative embeddings.

Problem: In most cases, $I(X; f(X)) = +\infty$.

Deep InfoMax

1. Consider a random data augmentation: $X \to X'$ and a Markov chain $f(X) \to X \to X'$. Then by the data-processing inequality (DPI):

$$I(X'; f(X)) \leq I(X; f(X))$$

Deep InfoMax

1. Consider a random data augmentation: $X \to X'$ and a Markov chain $f(X) \to X \to X'$. Then by the data-processing inequality (DPI):

$$I(X'; f(X)) \leq I(X; f(X))$$

2. Next, apply f to X'. We now have the chain $f(X) \to X \to X' \to f(X')$ and, by DPI:

$$I(f(X'); f(X)) \leq I(X'; f(X)) \leq I(X; f(X))$$

Distribution Matching
We want to learn $f(X)$ that follow a given distribution, e.g., normal.

Cheap Modification of Deep InfoMax
We propose adding independent noise Z to the normalized representation of X. This produces the chain $f(X) + Z \longrightarrow X \longrightarrow X' \longrightarrow f(X')$ and leads to the objective

$$I(f(X'); f(X) + Z) \to \max$$

Note that

$$I(f(X'); f(X) + Z) \leq I(f(X'); f(X)) \leq I(X'; f(X)) \leq I(X; f(X))$$

Distribution Matching
We want to learn $f(X)$ that follow a given distribution, e.g., normal.

Cheap Modification of Deep InfoMax
We propose adding independent noise Z to the normalized representation of X. This produces the chain $f(X) + Z \longrightarrow X \longrightarrow X' \longrightarrow f(X')$ and leads to the objective

$$I(f(X'); f(X) + Z) \to \max$$

Note that

$$I(f(X'); f(X) + Z) \leq I(f(X'); f(X)) \leq I(X'; f(X)) \leq I(X; f(X))$$

Donsker-Varadhan bound

$$I(X; Y) = \sup_{T \colon \Omega \to \mathbb{R}} \left[\mathbb{E}_{\mathbb{P}_{X,Y}} T - \log \mathbb{E}_{\mathbb{P}_X \otimes \mathbb{P}_Y} \exp(T) \right],$$

Lemma 1

Consider the following Markov chain of absolutely continuous random vectors:

$$f(X) + Z \longrightarrow X \longrightarrow X' \longrightarrow f(X'),$$

with Z being independent of (X, X'). Then

$$I(f(X'); f(X) + Z) = h(f(X) + Z) - h(Z) - I(f(X) + Z; f(X) \mid f(X')).$$

Weak invariance

DM alone does not guarantee the learned representations be meaningful or useful for downstream tasks.

Definition 2

An encoder f is said to be a weakly invariant to data augmentation $X \to X'$ if there exists a function g such that $f(X) = g(f(X)) = g(f(X'))$ almost surely.

Lemma 3

Under the conditions of Lemma 1, let $\mathbb{P}(X = X' \mid X) \geq \alpha > 0$. Then, $I(f(X) + Z; f(X) \mid f(X')) = 0$ precisely when f is weakly invariant to $X \to X'$.

Theorem 4 (Gaussian distribution matching)

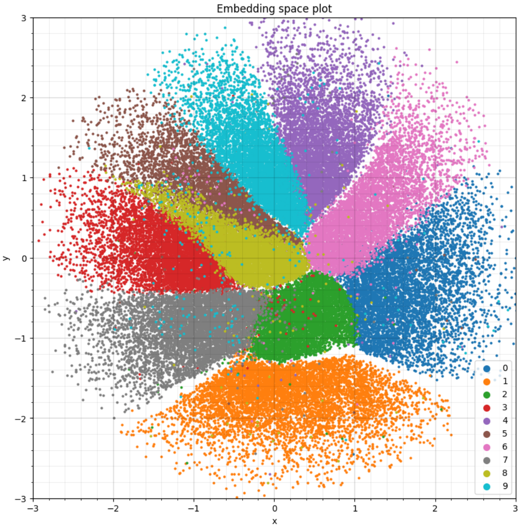
Let the conditions of Lemma 3 be satisfied. Assume $Z \sim \mathcal{N}(0, \sigma^2 \mathrm{I})$, $\mathbb{E}\, f(X) = 0$ *and* $\mathrm{Var}\,(f(X)_i) = 1$ *for all* $i \in d$*. Then, the mutual information* $I(f(X'); f(X) + Z)$ *can be upper bounded as follows*

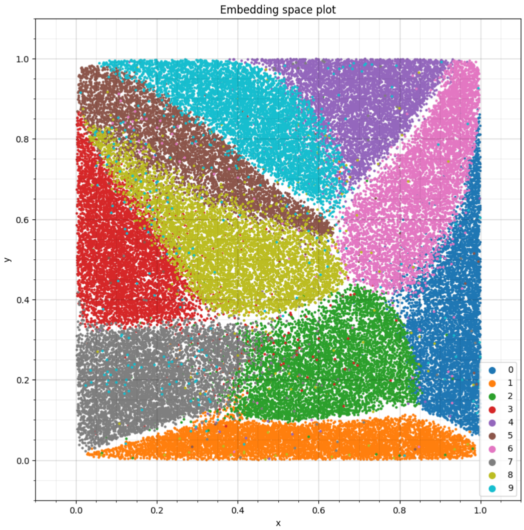
$$I(f(X'); f(X) + Z) \leq \frac{d}{2} \log \left(1 + \frac{1}{\sigma^2} \right), \tag{1}$$

with the equality holding exactly when f *is weakly invariant and* $f(X) \sim \mathcal{N}(0, \mathrm{I})$*. Moreover,*

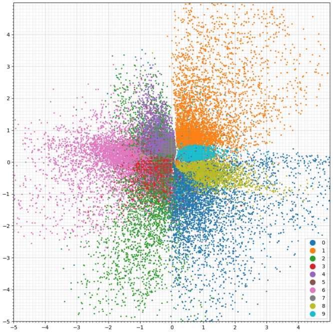
$$\mathrm{D}_{\mathrm{KL}}\,(f(X) \,\|\, \mathcal{N}(0, \mathrm{I})) \leq I(Z; f(X) + Z) - I(f(X'); f(X) + Z) - d \log \sigma.$$

Two-dimensional embeddings for MNIST dataset

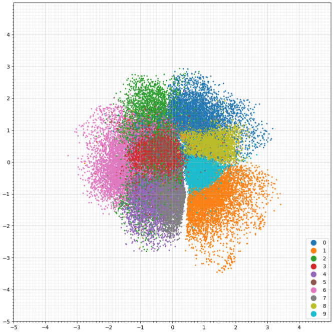




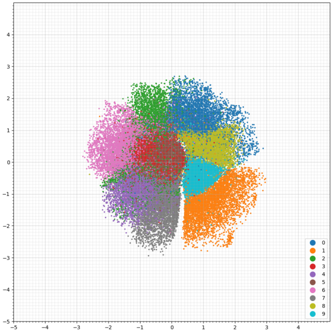
Two-dimensional embeddings for CIFAR10 dataset



(a) No noise injection



(b) Gaussian, $\sigma = 0.05$



(c) Gaussian, $\sigma = 0.1$

Theorem 5 (Dual form of Gaussian distribution matching)
Under the conditions of Theorem 4,

$$I(f(X'); f(X) + Z) \geq \mathbb{E}_{\mathbb{P}^+}\left[T^*_{\mathcal{N}(0,\sigma^2\mathrm{I})}\right] - \log \mathbb{E}_{\mathbb{P}^-}\left[\exp\left(T^*_{\mathcal{N}(0,\sigma^2\mathrm{I})}\right)\right],$$

$$T^*_{\mathcal{N}(0,\sigma^2\mathrm{I})}(x, y) = \frac{\|y\|^2}{2(1+\sigma^2)} - \frac{\|y-x\|^2}{2\sigma^2} = \frac{1}{\sigma^2}\left(\langle x, y\rangle - \frac{\|x\|^2 + \|y\|^2/(1+\sigma^2)}{2}\right),$$

with the equality holding precisely when f is weakly invariant and $f(X) \sim \mathcal{N}(0, \mathrm{I})$.

Thank you for your attention!