Efficient Distribution Matching of Representations via Noise-Injected Deep InfoMax

Ivan Butakov[1,2] Alexander Semenenko[1] Alexander Tolmachev[1,2] Andrey Gladkov[1]
Marina Munkhoeva[3] Alexey Frolov[1]

[1]Skolkovo Institute of Science and Technology [2]Moscow Institute of Physics and Technology [3]Artificial Intelligence Research Institute

Introduction

- **Representation Learning** extracts meaningful low-dimensional embeddings for AI tasks in vision, audio and NLP.
- **Applications**: multi-modal learning, statistical and topological analysis, data visualization, hypothesis testing.
- We focus on **Self-Supervised Learning (SSL)** to avoid relying on labeled data.
- **Deep InfoMax (DIM)** is an information-theoretic contrastive approach that maximizes useful information contained in the embeddings, offering strong performance.
- **Distribution Matching (DM)** enforces embeddings to follow a specific distribution. Crucial for Generative modeling, Statistical analysis, Disentanglement and Outlier detection.

Problem Setup

Let X be a random vector and f be an encoder (modeled by a neural network).

Aim: obtain a comprehensive low-dimensional representation $f(X)$ admitting a certain distribution (e.g., $\mathcal{N}(\mu, \Sigma)$).

Problem: distribution matching is usually performed via a full-fledged generative modelling:

- adversarial term (similar to GAN setup);
- post-hoc DM with flow models

These approaches are **expensive** and **require additional NNs** to be trained.

We achieve *the same result without introducing significant changes* to the SSRL pipeline.

Deep InfoMax

Mutual Information: $I(X;Y) = \mathrm{D}_{\mathrm{KL}}(\mathbb{P}_{X,Y} \parallel \mathbb{P}_X \otimes \mathbb{P}_Y)$ — an invariant measure of non-linear dependance between X and Y.

Information-Theoretic Approach. To obtain the most informative embeddings, enforce

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

Problem: In most cases (e.g. with X and $f(X)$ being continuous), $I(X; f(X)) = \infty$. Moreover, as X is high-dimensional, $I(X; \cdot)$ is hard to estimate.

Solution: Apply augmentations $X \to X'$, encode augmented data.

$$f(X) \longleftarrow X \longrightarrow X' \longrightarrow f(X'), \qquad I(f(X'); f(X)) \le I(X; f(X))$$

A similar augmentation-driven approach can facilitate **cheap Distribution Matching**. We propose adding independent noise Z to normalized representation $f(X)$.

Noise-injected chain $f(X) + Z \longleftarrow X \longrightarrow X' \longrightarrow f(X')$ leads to the new objective

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

If Z being independent of (X, X') in the following chain, then

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

Weak Invariance to Augmentations

Definition. An encoding mapping f is called *weakly invariant* under data augmentation $X \to X'$ if there exists a function g such that $f(X) = g(f(X)) = g(f(X'))$ a.s.

Lemma. Consider the following Markov chain of absolutely continuous random vectors:

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

with Z being independent of (X, X'). Let $\mathbb{P}(X = X' \mid X) \ge \alpha > 0$. Then, $I(f(X) + Z; f(X) \mid f(X')) = 0$ precisely when f is weakly invariant to $X \to X'$.

Therefore, *mutual information maximization* yields representations admitting a certain level of *invariance to the augmentations employed.*

Noise Injection Enables Distribution Matching

Theorem (Gaussian distribution matching). Assume $Z \sim \mathcal{N}(0, \sigma^2 \mathrm{I})$, $\mathbb{E}\,[f(X)_i]^2 = 1$ for all $i \in \overline{1,d}$, and some other mild constraints. Then,

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

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

Theorem (Uniform distribution matching). Let $Z \sim \mathrm{U}([-\varepsilon; \varepsilon]^d)$, $\operatorname{supp} f(X) \subseteq [0;1]^d$, and some other mild constraints hold. Then,

$$I(f(X'); f(X) + Z) \le d \log\left(1 + \frac{1}{2\varepsilon}\right),$$

with equality *iff* $1/\varepsilon \in \mathbb{N}$, f is weakly invariant, and $f(X) \sim \mathrm{U}(\{0, 2\varepsilon, 4\varepsilon, \dots, 1\})$.

Conclusion

- We propose a **novel and cheap** method to achieve **representations admitting a particular distribution**. No additional NNs are employed, only noise injection.
- Theoretical and empirical justifications are provided, with the latter including three normality tests and visual assessment. The results indicate **successful DM**.
- Moderate noise injection **does not affect performance** on downstream tasks.
- Our framework allows for additional theoretical analysis, e.g., for weak invariance.

Distributional Tests

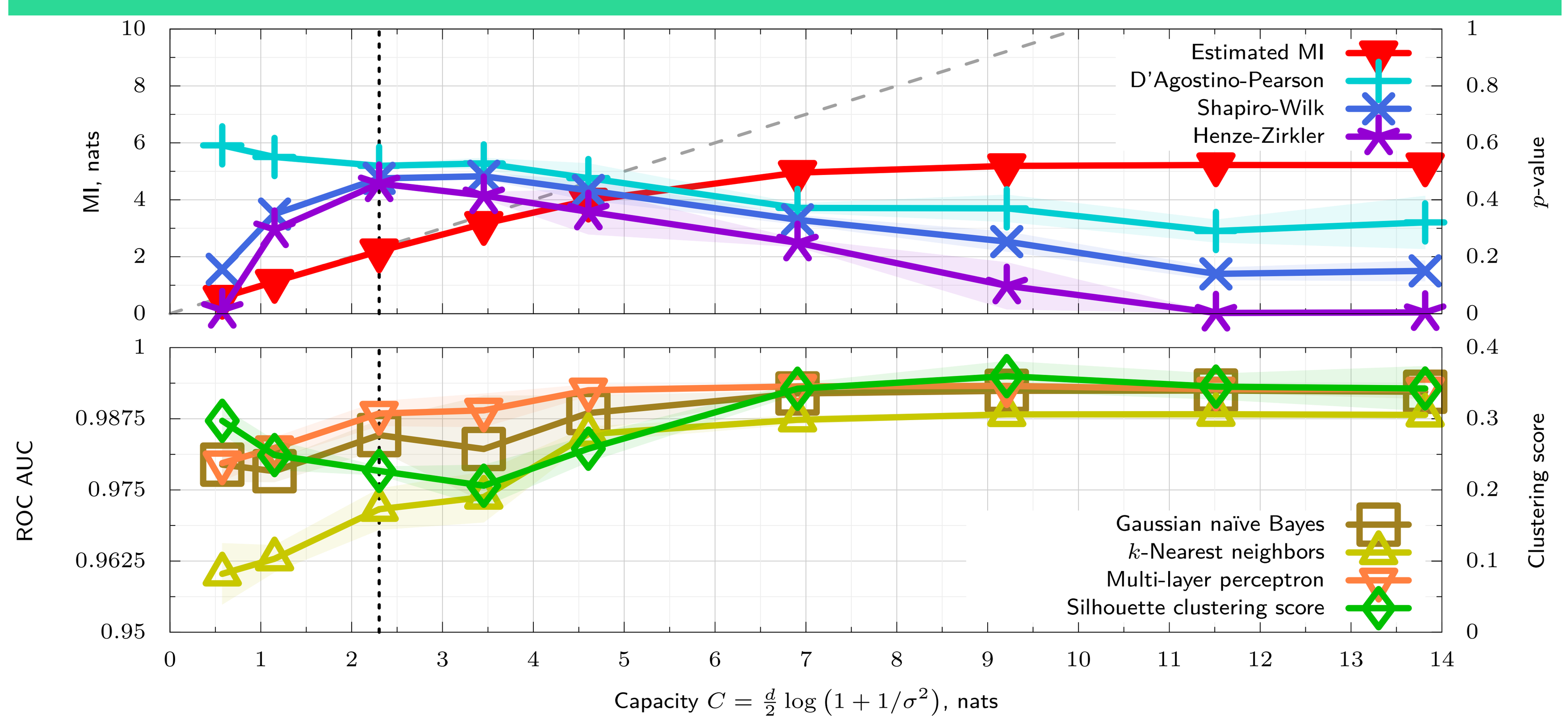


Figure 1. Results for MNIST in the Gaussian DM setup for $d = 2$ and varying capacity C. The vertical line denotes the minimal capacity required to preserve the information about the class labels in $f(X) + Z$. InfoNCE loss is used to approximate Donsker-Varadhan bound (red line). The dashed line represents the upper bound on MI. We report mean values and 99% asymptotic confidence intervals over 5 runs for each point.

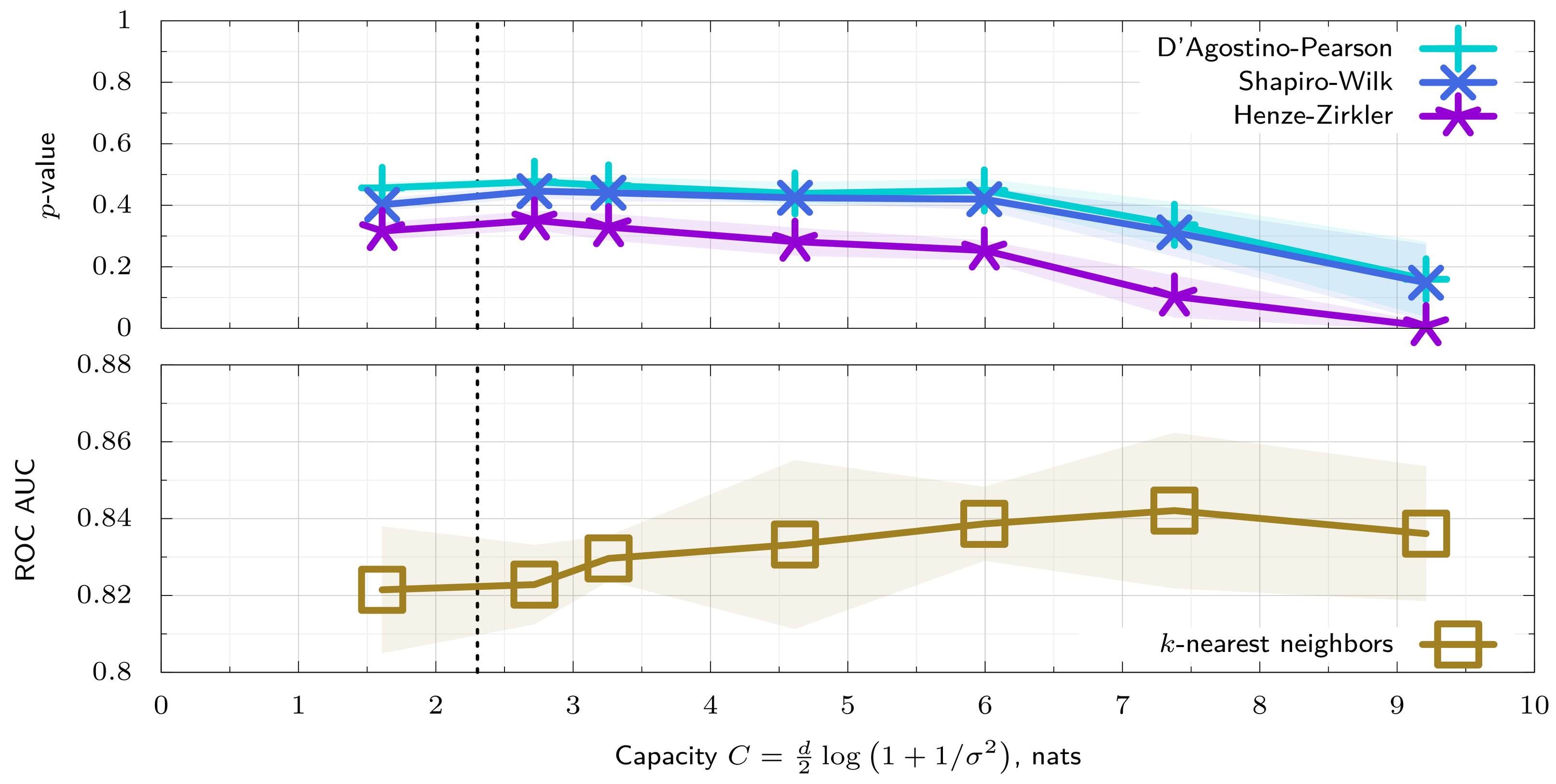
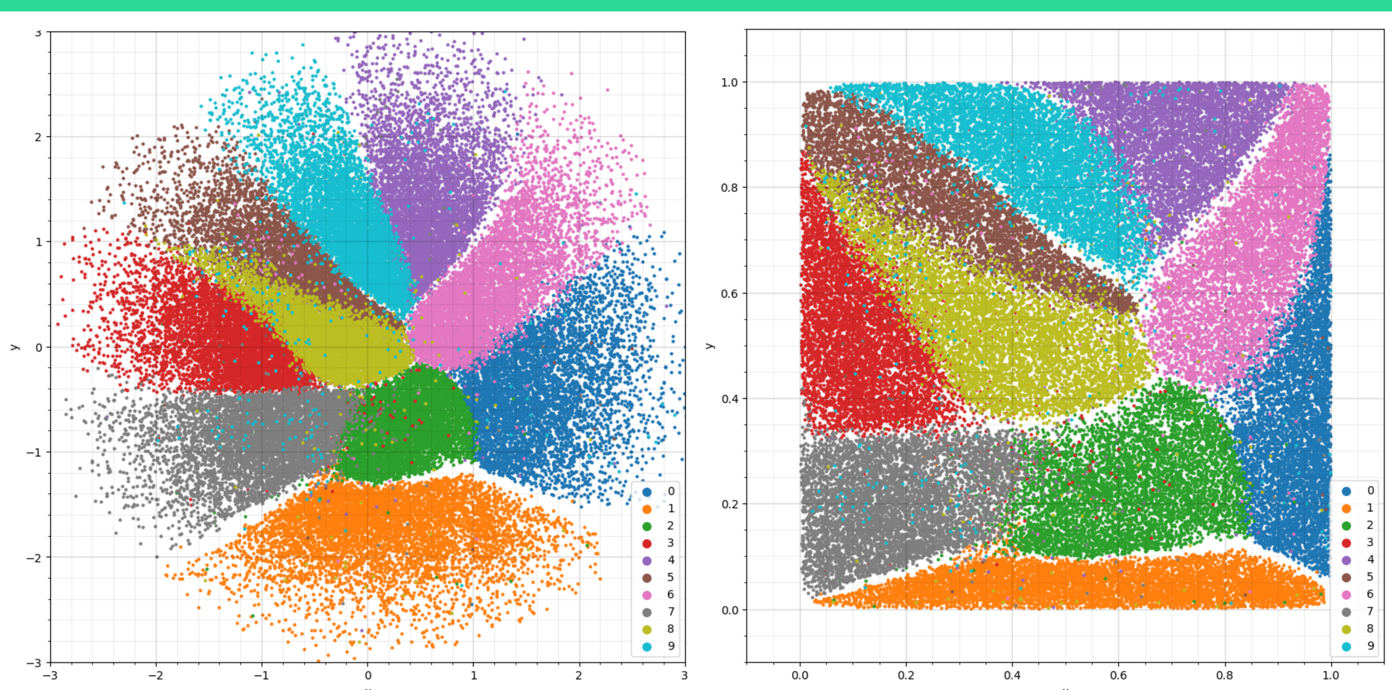
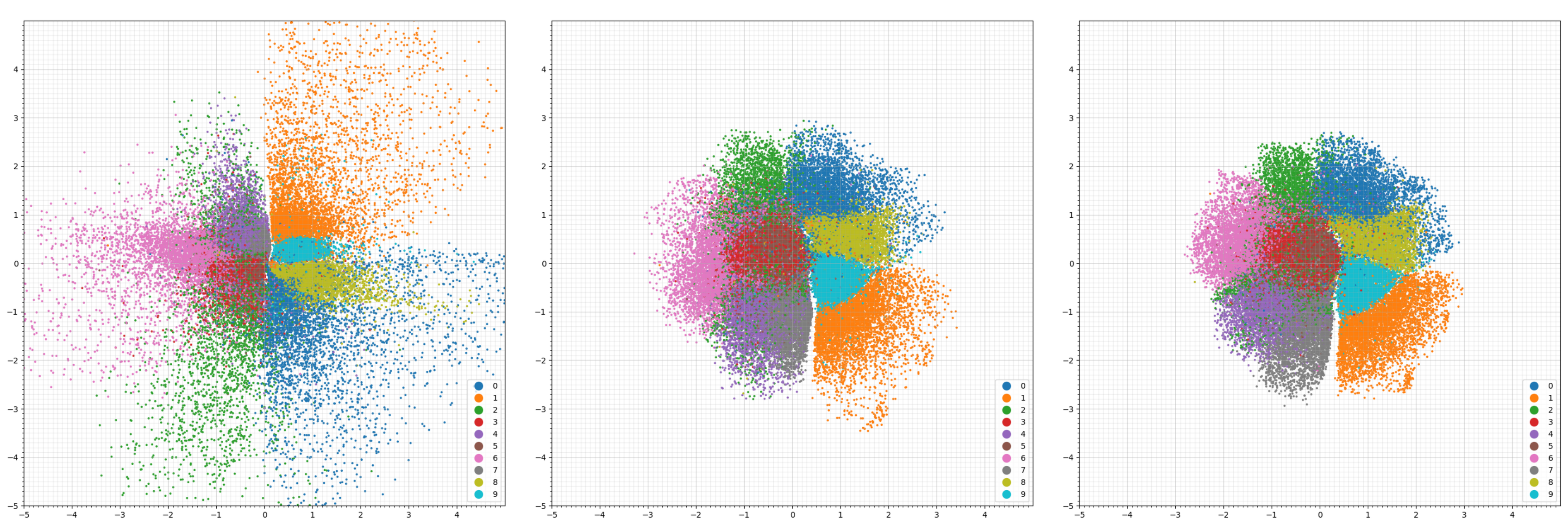


Figure 2. Results for CIFAR10 dataset in the Gaussian DM setup. The designations are described above.

2D Embeddings Visualization



(a) Gaussian distribution, $\sigma = 0.1$, MNIST (b) Uniform distribution, $\varepsilon = 0.05$, MNIST



(a) No noise injection, CIFAR10 (b) Gaussian, $\sigma = 0.05$, CIFAR10 (c) Gaussian, $\sigma = 0.1$, CIFAR10

Noise Injection Preserves Performance

Table 1. Linear probing accuracy (in %) on CIFAR-10/100 under noise injection (800 pretrain epochs).

	CIFAR-10		CIFAR-100	
	top-1	top-5	top-1	top-5
SimCLR	90.83	99.76	65.64	89.91
SimCLR $\sigma = 0.1$	90.96	99.72	67.03	90.49
SimCLR $\sigma = 0.3$	91.56	99.77	65.72	89.76
SimCLR $\sigma = 0.5$	90.51	99.74	65.58	89.56

	CIFAR-10		CIFAR-100	
	top-1	top-5	top-1	top-5
VICReg	90.63	99.67	65.71	88.96
VICReg $\sigma = 0.1$	91.09	99.68	68.92	90.50
VICReg $\sigma = 0.3$	90.75	99.61	67.31	89.89
VICReg $\sigma = 0.5$	91.02	99.75	66.52	89.66

Table 2. Linear probing accuracy (in %) on ImageNet under noise injection (VICReg, 100 pretrain epochs).

	ImageNet-100		ImageNet-1k	
σ	top-1	top-5	top-1	top-5
0	72.18 ± 0.40	92.02 ± 0.12	67.41 ± 0.17	87.43 ± 0.08
0.05	72.27 ± 0.38	91.99 ± 0.18	67.29 ± 0.20	87.47 ± 0.06
0.1	72.07 ± 0.27	91.65 ± 0.13	67.30 ± 0.13	87.43 ± 0.02
0.2	71.68 ± 0.50	91.61 ± 0.24	67.19 ± 0.12	87.32 ± 0.09

The Thirteenth International Conference on Learning Representations 2025, Singapore; www.iclr.cc

butakov.id@phystech.edu, tolmachev.ad@phystech.edu, al.frolov@skoltech.ru