Enhancing out-of-distribution detection through stochastic embeddings in self-supervised learning

Denis Janiak[1][0000−0003−1859−9093], Jakub Binkowski[1][0000−0001−7386−5150], Piotr Bielak[1][0000−0002−1487−2569], and Tomasz Kajdanowicz[1][0000−0002−8417−1012]

Wroclaw University of Science and Technology

Abstract. In recent years, self-supervised learning has played a pivotal role in advancing machine learning by allowing models to acquire meaningful representations from unlabeled data. An intriguing research avenue involves developing self-supervised models within an information-theoretic framework, e.g., feature decorrelation methods like Barlow Twins and VICReg, which can considered as particular implementations of the information bottleneck objective. However, many studies often deviate from the stochasticity assumptions inherent in the information-theoretic framework. Our research demonstrates that by adhering to these assumptions, specifically by employing stochastic embeddings in the form of a parametrized conditional density, we can not only achieve performance comparable to deterministic networks but also significantly improve the detection of out-of-distribution examples, surpassing even the performance of supervised detectors. With VICReg, specifically, we achieve an average AUROC of 0.858 for the stochastic unsupervised detector, compared to 0.796 for the supervised baseline. Remarkably, this improvement is achieved solely by leveraging information from the underlying embedding distribution.

Keywords: self-supervised learning · out-of-distribution detection · stochastic embeddings · uncertainty estimation · information theory

1 Introduction

Self-supervised learning (SSL) is an approach to learning representations of data without labels, often utilizing the data itself as a supervisory signal. In recent years, such methods have gained increasing popularity in computer vision and have demonstrated significant success in various downstream tasks [29]. The primary goal of SSL is to bring similar samples closer while pushing dissimilar samples further apart. This objective enhances the model's ability to discriminate between different data classes, contributing to its overall effectiveness.

One effective strategy for learning meaningful representations involves maximizing the similarity between various views of augmented images, thereby ensuring invariance to these augmentations [7]. However, this approach risks encountering trivial solutions (i.e., where all embeddings collapse into a single point).

ICCS Camera Ready Version 2024
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-63783-4_25

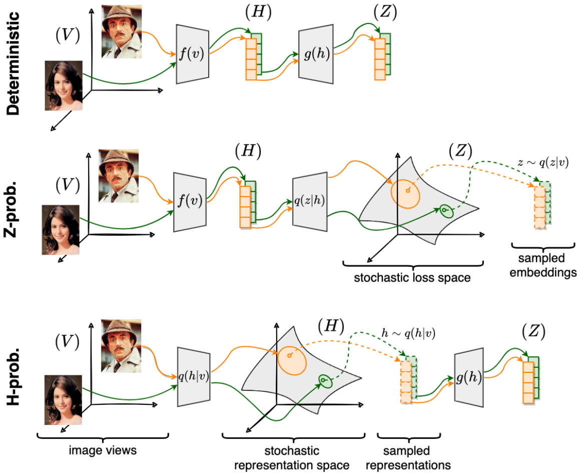

Fig. 1. This schematic illustrates the multi-view encoding process. It begins with image views (V), which are transformed via deterministic or stochastic pathways. The deterministic pathway processes the input through encoder function $f(\cdot)$, leading to deterministic representation space (H) and loss space (Z). The probabilistic pathways (H-prob. and Z-prob.) introduce stochasticity by sampling from associated spaces using stochastic encoders $q(h|v)$ and $q(z|v)$.

Methods such as Barlow Twins [40] and VICReg [4] employ feature decorrelation mechanisms to address this issue. Interestingly, they are closely linked to an information-theoretic framework as their objective can be derived using the information bottleneck principle [2,33]. Nevertheless, the information-theoretic framework typically assumes a source of stochasticity (noise) within the model. The aforementioned methods do not fulfil this stochastic condition, as they are simplified using deterministic networks and rely on a proxy objective. By aligning with the assumption of the information-theoretic framework (source of noise within the model), we could effectively benefit from stochasticity for tasks such as uncertainty estimation and out-of-distribution detection (OOD) [39]. Recent advancements in machine learning underscore the importance of quantifying these uncertainties and identifying distributionally shifted (OOD) samples, particularly in safety-critical applications such as medical imaging, active learning, and autonomous driving [11].

Our study directly introduces stochasticity into the feature decorrelation-based methods by parameterizing the conditional density of a model, i.e., predicting the parameters of the embeddings' distribution. In essence, our goal is twofold: first, to adhere to the stochastic assumptions required by the information-theoretic framework, and second, to harness this stochasticity to enhance our ability to detect OOD examples and handle uncertainty. We demonstrate that leveraging stochastic embeddings enables us to outperform other supervised and deterministic methods for OOD detection while also achieving comparable re-

DOI: 10.1007/978-3-031-63783-4_25

sults when evaluated across various downstream tasks (linear classification, semi-supervised learning, and transfer learning).

Our contributions are as follows:

1. We employed stochastic embeddings to feature decorrelation-based methods, adhering to the theoretical underpinnings of the information-theoretic framework.
2. We showcased the effectiveness of our approach in detecting OOD samples by leveraging the embedding distribution, outperforming traditional supervised detectors.
3. We explored novel strategies for exploiting stochastic embeddings to accurately identify OOD examples.
4. We conducted a comprehensive empirical evaluation to assess the impact of stochastic embeddings on downstream tasks.

The paper's remaining sections are organized as follows: Section 2 reviews related works in SSL, focusing on representation learning methods and addressing challenges like trivial solutions and lack of stochasticity. Section 3 details our approach for introducing stochasticity into feature decorrelation-based methods. In Section 4, we present experiments evaluating our approach's effectiveness in downstream tasks and OOD detection. Lastly, Section 5 offers conclusions and summarizes key findings.

2 Related works

The primary objective of **self-supervised learning** is to optimize a specific loss function tailored to capture meaningful patterns or relationships within unlabeled data. One approach, proven to be very successful in vision tasks, is contrastive learning [7], which aims to maximize the agreement between positive (similar) pairs of data samples while minimizing it for negative (dissimilar) pairs. More recent avenues are non-contrastive methods that adopt various mechanisms to prevent representation collapse, eliminating the need for negative samples [34]. These mechanisms could include architectural constraints [13], clustering-based objectives [6], or feature decorrelation [40,4]. These non-contrastive methods indirectly optimize the uniformity property of the representation, aiming to prevent representation collapse [36]. In this context, we focus on feature decorrelation-based techniques, particularly the Barlow Twins [40] and VICReg [4]. The Barlow Twins method involves computing the cross-correlation matrix from embeddings of augmented images. The objective is to minimize the off-diagonal elements, encouraging feature decorrelation while promoting data invariance by aiming to set the diagonal elements to one. VICReg introduces an additional term into the loss function, which controls the variance of each dimension within the embeddings. Consequently, this facilitates straightforward computation of the covariance of the embeddings and eliminates the necessity for normalization.

ICCS Camera Ready Version 2024
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-63783-4_25

OOD detection methods in machine learning are crucial for identifying instances during the testing phase that deviate semantically from the categories encountered in the training data, thereby preventing misclassification [11]. While supervised detectors [16,24] have demonstrated success, they rely on label information and a classifier to derive their scores. On the other hand, distance-based methods [23,31] utilize representations (features) for detecting OOD examples, yet many of them still necessitate computing class-conditional statistics from the training data. In fully unsupervised OOD detection, density-based and reconstruction-based methods leverage data density or reconstruction techniques, often incorporating generative models [39].

Our research focuses on **self-supervised** methods for **OOD detection**. Previous studies have explored various avenues, including combining self-supervised and discriminative objectives [16], employing hard data augmentations for sample separation [35], leveraging contrastively learned features [3,32], and integrating probabilistic modelling to estimate uncertainty [19,26]. For instance, [19] utilizes the von Mises Fischer distribution to model embeddings, with the concentration parameter serving as an uncertainty metric. Similarly, [26] investigates SimSiam [8] within the variational inference framework, employing a power spherical distribution to characterize the embedding distribution. Our study adopts a similar approach, leveraging embedding distribution characteristics to identify uncertain and OOD examples. However, we consider feature decorrelation-based methods within the information-theoretic framework, representing embeddings with a different distribution than the aforementioned studies and learning a different objective (Section 3.2).

The **information-theoretic perspective** offers crucial insights into the underlying mechanics of self-supervised learning. While models like InfoMax [17] prioritize capturing maximal data information, the Information Bottleneck (IB) principle emphasizes balancing informativeness and compression [2]. In this context, Barlow Twins exemplifies an IB aiming to maximize information between the image and representation while minimizing information about data augmentation, rendering the representation invariant to these augmentations. Another approach, the multi-view information bottleneck (MIB) framework [10], seeks to capture predictive information shared across different data views. This is achieved by maximizing mutual (shared) information between views in their embeddings while minimizing redundant information not shared between them. Shwartz-Ziv et al. [33] demonstrated that the VICReg objective can also be derived from an information-theoretic standpoint, exploiting a lower bound derived from the MIB framework.

3 Methodology

In this section, we begin by outlining the setup of the feature decorrelation-based SSL framework for deterministic (point estimate) embeddings. Following this, we extend this setup to incorporate stochasticity and introduce stochastic (probabilistic) embeddings with regularization. Lastly, we introduce methods

DOI: 10.1007/978-3-031-63783-4_25

to leverage the stochastic nature of the embeddings, offering stochastic OOD metrics. Figure 1 illustrates the workflow of deterministic and stochastic self-supervised learning variants.

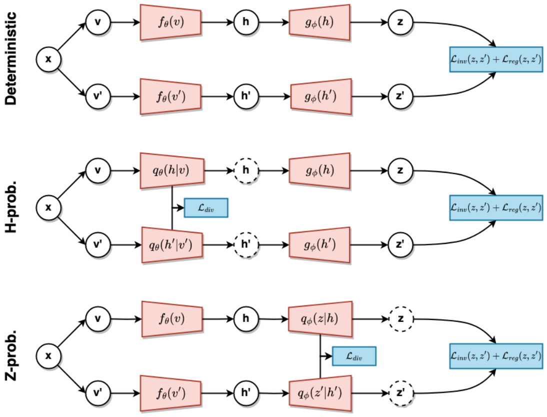

Fig. 2. Diagram of three approaches to SSL: (1) **Deterministic**, with deterministic mappings using f_θ and g_ϕ, (2) **H-prob**, introducing stochasticity with q_θ before deterministic projection g_ϕ, and (3) **Z-prob**, with deterministic mapping f_θ and stochastic projection q_ϕ. Losses $\mathcal{L}_{\text{inv}}(z,z')$ and $\mathcal{L}_{\text{reg}}(z,z')$ are computed using the embeddings. $\mathcal{L}_{\text{div}}$ loss is the stochastic regularization loss (see Section 3.3).

3.1 Deterministic self-supervised learning

We sample an image x from a dataset $\mathcal{D}$ and create two views v and v' by applying transforms t and t' sampled from a distribution $\mathcal{T}$. These views are then fed into an encoder f_θ, parameterized by θ, to create representations $h = f_\theta(v)$ and $h' = f_\theta(v')$. Next, the representation vectors are passed through the projector g_ϕ, parameterized by ϕ, to obtain embeddings $z = g_\phi(h)$ and $z' = g_\phi(h')$. In a batch represented by $Z = [z_1, \ldots, z_n]$ and $Z' = [z'_1, \ldots, z'_n]$, each containing n embedding vectors of dimension d, corresponding to embedded image views, we define $z_{(i)} = [z_{1(i)}, \ldots, z_{n(i)}]$ as the i-th variable across all samples in the batch. The loss function $\mathcal{L}(Z, Z')$ is then applied to these embeddings Z and Z'.

In **Barlow Twins**, the loss function is computed using the cross-correlation matrix R on embeddings, which are mean-centred along the batch dimension:

$$R_{ij} = \text{corr}(z_{(i)}, z'_{(j)}) = \frac{\text{cov}(z_{(i)}, z'_{(j)})}{\sigma_{z_{(i)}} \cdot \sigma_{z'_{(j)}}}, \tag{1}$$

where $z_{(i)}$ and $z'_{(j)}$ represent the i-th and j-th component of embedding vectors across all samples in the batch, while $\sigma_{z_{(i)}}$ and $\sigma_{z'_{(j)}}$ denote the standard

DOI: 10.1007/978-3-031-63783-4_25

deviations of $z_{(i)}$ and $z'_{(j)}$ respectively. From the cross-correlation matrix, we compute: (1) invariance term $\mathcal{L}_{\text{inv}}$ that optimizes the diagonal elements to be close to 1, aiming to enforce invariance to data augmentations; and (2) regularization term $\mathcal{L}_{\text{reg}}$ that pushes the off-diagonal elements towards 0 to promote feature decorrelation and prevent collapse:

$$\mathcal{L}_{\text{inv}}(Z, Z') = \sum_i (1 - R_{(i,i)})^2, \quad \mathcal{L}_{\text{reg}}(Z, Z') = \lambda \sum_i \sum_{j \neq i} R^2_{(i,j)}$$

In contrast, **VICReg** computes the loss function with three terms. The invariance term is calculated using mean-squared error loss scaled by α coefficient and divided by number of samples in batch n:

$$\mathcal{L}_{\text{inv}}(Z, Z') = \frac{\alpha}{n} \sum_{b=1}^{n} \|z_b - z'_b\|_2^2. \tag{2}$$

The regularization term comprises two components - covariance and variance:

$$\mathcal{L}_{\text{cov}}(Z) = \frac{1}{d} \sum_i \sum_{j \neq i} C_{(i,j)}, \quad \mathcal{L}_{\text{var}}(Z) = \frac{1}{d} \sum_{i=1}^{d} \max(0, \gamma - \sigma_{z_{(i)}+\epsilon}), \tag{3}$$

where $C_{(i,j)}$ is the element of the covariance matrix, i.e., $C_{(i,j)} = \text{cov}(z_{(i)}, z_{(j)})$. The covariance term involves summing the squared off-diagonal coefficients of the covariance matrix. Meanwhile, the variance term is a hinge function that operates on the standard deviation of the embeddings across the batch dimension. Both regularization terms are calculated separately for Z and Z' using τ and ν scalars as loss coefficient:

$$\mathcal{L}_{\text{reg}}(Z, Z') = \tau[\mathcal{L}_{\text{var}}(Z) + \mathcal{L}_{\text{var}}(Z')] + \nu[\mathcal{L}_{\text{cov}}(Z) + \mathcal{L}_{\text{cov}}(Z')], \tag{4}$$

The final loss function in both Barlow Twins and VICReg is then formulated as the sum of invariance and regularization terms: $\mathcal{L}_{\text{SSL}} = \mathcal{L}_{\text{inv}} + \mathcal{L}_{\text{reg}}$.

3.2 Stochastic embeddings

Drawing inspiration from [10] and [33], we propose to reformulate our self-supervised objective as an information maximization problem and extend it to stochastic embeddings. We aim to maximize the mutual information between the views V and V' and their corresponding embeddings Z and Z', i.e., $I(Z; V')$ and $I(Z'; V)$, respectively. We utilize the lower bound from [33]:

$$\begin{aligned} I(Z; V') &= \mathcal{H}(Z) - \mathcal{H}(Z|V') \\ &\geq \mathcal{H}(Z) + \mathbb{E}_{v'}[\log q(z|v')] \\ &\geq \mathcal{H}(Z) + \mathbb{E}_{z|v}[\mathbb{E}_{z'|v'}[\log q(z|z')]] \end{aligned} \tag{5}$$

ICCS Camera Ready Version 2024
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-63783-4_25

where $\mathcal{H}(Z)$, is implicitly optimized by the regularization term $\mathcal{L}_{reg}$, while the expectations $\mathbb{E}_{z|v}[\mathbb{E}_{z'|v'}[\log q(z|z')]]$ (square log loss) are optimized by the invariance term $\mathcal{L}_{inv}$. To compute the expected loss, we evaluate these expectations over empirical data distribution. Specifically, we backpropagate through K Monte Carlo (MC) samples using the reparametrization trick [18]:

$$\mathbb{E}_{z|v}[\mathbb{E}_{z'|v'}[\log q(z|z')]] \simeq \frac{1}{nK}\sum_{i=1}^{n}\sum_{k=1}^{K}\log q(z_{ik}|z'_{ik}). \tag{6}$$

We introduce two variations of the model, which differ in terms of choice of the stochastic space and, therefore, the variational conditional density $q(z|v)$. Figure 2 depicts the workflow of probabilistic (stochastic) variations of self-supervised learning.

Stochastic loss space (Z-prob.) In this model variant, we introduce the stochasticity into the loss space by parametrizing the projector conditional density $q_\phi(z|h)$, i.e., we make the projector stochastic (see Figure 2 at the bottom). Specifically, we employ a two-step process for encoding image views. Initially, we utilize a deterministic encoder f_θ to transform the image view v into a representation h, i.e., $h = f_\theta(v)$. Subsequently, we employ a stochastic projector $q_\phi(z|h)$ to sample latent variables z based on h. Our conditional density is defined as $q_{\phi,\theta}(z|v) = q_\phi(z|f_\theta(v))$, and the sampling process is represented as $z \sim \mathcal{N}\left(z|\mu_\phi(h), \sigma_\phi^2(h)I\right)$, where $\mu_\phi(h)$ and $\sigma_\phi^2(h)$ denote the mean and variance functions determined by the stochastic projector, respectively. The same procedure is applied to the second image view v' to generate the representation h' and the corresponding embedding z', utilizing identical encoder and projector parameters denoted by θ and ϕ.

Stochastic representation space (H-prob.) In this model variant, we shift the stochasticity from the loss space Z to the representation space H (see Figure 2 in the middle). We define the conditional density as $q_{\phi,\theta}(z|v) = g_\phi(q_\theta(h|v))$ and sampling process as $h \sim \mathcal{N}(h|\mu_\theta(v), \sigma_\theta^2(v)I)$. Then, we obtain the embedding z by mapping the representation h with a projector, g_ϕ. We apply the same procedure for the second image view v' to produce the representation h' and the embedding z', utilizing the same encoder and projector parameters θ and ϕ. In particular, to utilize the bound from Eq. 5, we must also account for the presence of h. We decompose the joint distribution as $q(v, h, z) = q(z|h)q(h|v)q(v)$, where z depends on h. The computation of $q(z|v)$ requires the marginalization of h, i.e. $q(z|v) = \int dh\, q(z, h|v) = \int dh\, q(z|h)q(h|v)$. Consequently, we can take expectations with respect to $q(h|v)$ and lower bound term from Eq. 5 using Jensen's inequality:

$$\mathbb{E}_{v'}[\log q(z|v')] \geq \mathbb{E}_{h'|v'}[\log q(z|h')]. \tag{7}$$

By taking the expectation over both Z and Z', we will obtain the final objective:

$$\mathbb{E}_{z|h,h|v}[\mathbb{E}_{z'|h',h'|v'}[\log q(z|z')]], \tag{8}$$

ICCS Camera Ready Version 2024
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-63783-4_25

which involves an additional expectation step that we take using MC estimation.

3.3 Stochastic regularization

Moving from point estimates to stochastic embeddings, we introduce an additional layer of uncertainty, which helps capture the inherent ambiguity and variability in the data. However, it also raises the challenge of regularizing this stochasticity to prevent trivial solutions and obtain reliable uncertainty estimates. To address this issue, we follow [2] and formulate an additional regularization term to the loss function in the form of a KL divergence between the stochastic embeddings $q(\cdot|v)$ and $q(\cdot|v')$, and a predefined prior $\hat{q}(\cdot)$, typically $\mathcal{N}(0,1)$:

$$\mathcal{L}_{\text{div}}(\cdot,\cdot) = \frac{\beta}{2}[\text{KL}(q(\cdot|v)||\hat{q}(\cdot)) + \text{KL}(q(\cdot|v')||\hat{q}(\cdot))], \tag{9}$$

where $(\cdot,\cdot)$ is either (h,h') or (z,z'). This regularization, controlled by β parameter, acts as a bottleneck, constraining the capacity of our stochastic embeddings, which proved to be effective in previous work [1,2] in terms of improving robustness and disentanglement.

Consequently, the overall objective for our stochastic self-supervised learning framework is defined as $\mathcal{L}_{\text{Stochastic-SSL}} = \mathcal{L}_{\text{inv}} + \mathcal{L}_{\text{reg}} + \mathcal{L}_{\text{div}}$.

3.4 Stochastic OOD detectors

We aim to utilize the stochastic nature of our embeddings to improve their ability to distinguish between in-distribution and OOD samples. To achieve this, we introduce **new stochastic scoring methods** that exploit the inherent variability in the embeddings. Notably, these methods do not require any training labels and do not depend on the information of OOD data. We propose the following scoring methods:

- **Log-prob prior (LogP)**: This score is derived by assessing the prior density $\hat{q}$ at the test input x^*, denoted as $\hat{q}(x^*)$. As we use KL regularization, the OOD examples may not be pushed towards the prior, and the log-prob for these examples could be higher.
- **Sigma mean (Sigma)**: This score computes the mean sigma value of the embedding, expressed as $\frac{1}{d}\sum_{i=1}^{d}\sigma(x^*)_{(i)}$, where i denotes the index of the embedding vector with dimension d, and σ is the variance predictor of the conditional density model. Higher sigma values indicate greater uncertainty in the embedding distribution, potentially signalling an unfamiliar example.
- **KLD-Kth-nearest (KLD-KN)**: In this approach, based on the test input x^*, we find its distance to the K-th nearest example x_k from the training samples $\mathcal{D}$, measured by the KL divergence, i.e., $\text{KL}(q(\cdot|x_k)||q(\cdot|x^*))$. This method operates under the assumption that OOD examples, which the model has not encountered, may appear in a ”hole” on a manifold. Thus, examining nearby examples can serve as a meaningful scoring metric.

ICCS Camera Ready Version 2024
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-63783-4_25

- **Euclidean-Kth-nearest (Euclid-KN)**: This score serves as the deterministic counterpart to the KLD-KN score. We utilize the Euclidean distance instead of KL divergence, i.e., $||x_k - x^*||_2$. We implement this score to assess whether the KLD-KN score effectively exploits the embedding variance.

4 Experiments

We pretrain our model in a self-supervised manner (without labels), adopting the same image augmentations and closely adhering to the original works in determining the loss coefficients [40,4]. Due to computational constraints, we opt for the smaller ResNet-18 [14] architecture as our backbone encoder and a smaller non-linear projection head (3-layer MLP, each of 1024 dimensions). We train the model using the AdamW [25] optimizer with a batch size of 256. We make our code publicly available at `https://github.com/graphml-lab-pwr/stochastic-embeddings-ssl`.

4.1 Downstream tasks evaluation

Setup. In these tasks, the model is pretrained once for 100 epochs on the ImageNet dataset [30]. Our experiments and previous works showed that Barlow Twins and VICReg exhibit low sensitivity to model intialization [4]. Moreover, we utilize the $\mathcal{N}(0,1)$ prior $\hat{q}(\cdot)$ and 12 MC samples. Next, we employ three downstream tasks: linear classification, semi-supervised and transfer learning [12]. For linear classification (`Linear`), we train a linear classifier (single linear layer) on the frozen representations from our pretrained backbone encoder and corresponding image labels. A similar process is repeated for `Transfer learning` task, where we employ the SUN397 [37] and the Flowers-102 [28] datasets. In the `Semi-supervised` learning task, we fine-tune both the backbone encoder and the linear classifier. We utilize subsets of the ImageNet dataset corresponding to 1% and 10% of the labels [8].

Table 1. Comparison of top-1 accuracy (Acc@1) and expected calibration error (ECE) for ImageNet tasks. Both Barlow Twins and VICReg exhibit low variance, allowing for a single-run performance comparison.

		`Linear`		`Semi-supervised`				`Transfer learning`			
				1%		10%		SUN397		flowers-102	
		Acc@1(↑)	ECE(↓)	Acc@1(↑)	ECE(↓)	Acc@1(↑)	ECE(↓)	Acc@1(↑)	ECE(↓)	Acc@1(↑)	ECE(↓)
VICReg	Deterministic	0.490	0.011	0.315	0.044	0.509	0.055	0.477	0.112	0.649	0.445
	H-prob.	0.451	0.013	0.313	0.220	0.498	0.114	0.460	0.038	0.622	0.315
	Z-prob.	0.484	0.012	0.310	0.042	0.507	0.054	0.478	0.112	0.629	0.435
Barlow	Deterministic	0.495	0.008	0.316	0.040	0.518	0.138	0.482	0.115	0.645	0.428
	H-prob.	0.451	0.010	0.309	0.217	0.495	0.111	0.457	0.023	0.637	0.302
	Z-prob.	0.489	0.010	0.313	0.039	0.506	0.054	0.481	0.111	0.645	0.447

Results. We present the results in Table 1. Our observations show that stochastic embeddings, particularly those in the loss space (Z-prob.), exhibit competitive performance compared to deterministic embeddings. For linear classification and transfer learning tasks, Z-prob. embeddings tend to outperform stochastic embeddings in the representation space (H-prob.) across both the Barlow Twins and VICReg methods in terms of accuracy. However, the difference is lower in the semi-supervised task, especially for 1% of available labels. Noticeably, in transfer learning, H-prob embeddings maintain lower ECE scores for both datasets, indicating better calibration, but they tend to exhibit higher ECE for semi-supervised. The performance variation across different datasets underscores the importance of dataset characteristics in model evaluation. We hypothesize that the superior performance of Z-prob. embeddings compared to H-prob. is attributable to the representation bottleneck created by the H-prob. model, which leads to an early representation compression, potentially removing data invariance at the representation level. As demonstrated in previous studies [5], such premature compression can negatively impact the performance of SSL models.

4.2 OOD detection

Setup. We utilize the same setup for the backbone as in the ablation study and select the best-performing model (see Section 4.3). Next, we investigate the OOD capabilities of our methods, following a similar evaluation procedure to [38] and report the results with the commonly used AUROC metric. We consider the original test set of CIFAR-10 as IN data and assess its ability to distinguish between other `Near` (MNIST [22], SVHN [27], Places365 [41], Textures [9]) and `Far` (CIFAR-100 [20], TinyImageNet [TIN] [21]) OOD datasets. We evaluate our proposed stochastic detectors (`LogP`, `Sigma`, `KLD-KN`) against commonly used methods in OOD detection problems. Specifically, we compare them with classification-based methods such as MaxSoftmax probability (`MSP`) [15] and `ODIN` [24], as well as distance-based methods like `Gram` matrices [31] and Mahalanobis distance (`MDS`) [23]. Contrary to our detectors, these methods require label information from the training data: MaxSoftmax and ODIN rely on a trained classifier, while Gram and Mahalanobis necessitate the computation of class-conditional statistics. Moreover, we provide a comparison between stochastic `KLD-KN` and its deterministic counterpart `Euclid-KN` (we select K based on the hyperparameter search [38]). Finally, we evaluate our methods against `SSD` [32], a framework for unsupervised OOD detection in self-supervised learning, which leverages Mahalanobis distance on k-means detected clusters.

Results. Table 2 presents the results from our experiment on stochastic embeddings. As observed, leveraging the intrinsic properties of stochastic embeddings, such as their variance (`Sigma`) or latent space manifold (`KLD-KN`), can be highly effective as an OOD detector. For both VICReg and Barlow Twins, the `KLD-KN` detection score surpasses the performance of supervised detectors (`MSP`, `ODIN`) and significantly exceeds that of the `Euclidean-KN`, its deterministic counterpart. Furthermore, simple `Sigma` provides better scoring than the

ICCS Camera Ready Version 2024
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-63783-4_25

distances-based methods such as Mahalanobis, Gram and SSD for both VICReg and Barlow Twins while reaching the performance of classification-based detectors for VICReg, thereby demonstrating its ability to take into account the stochasticity of the embeddings.

Table 2. The AUROC performance of OOD detection methods. * denotes supervised detectors, while † denotes unsupervised detectors requiring fitting labels.

		Near OOD				Far OOD		
	Detector	MNIST	SVHN	Places365	Texture	CIFAR-100	TIN	Avg.
VICReg	MSP*	0.837	0.790	0.780	0.769	0.799	0.800	0.796
	ODIN*	0.862	0.668	0.806	0.778	**0.818**	**0.819**	0.792
	Gram†	0.946	0.927	0.599	0.727	0.545	0.583	0.721
	MDS†	0.935	0.928	0.577	0.851	0.506	0.546	0.724
	SSD	0.935	0.928	0.565	0.850	0.514	0.538	0.722
	LogP	0.926	0.742	0.651	0.550	0.658	0.598	0.688
	Sigma	0.890	0.858	0.815	0.645	0.730	0.769	0.784
	KLD-KN	**0.954**	**0.972**	**0.821**	**0.874**	0.742	0.783	**0.858**
	Euclid-KN	0.936	0.907	0.776	0.805	0.706	0.731	0.810
Barlow Twins	MSP*	0.888	0.812	0.768	0.783	0.787	0.790	0.805
	ODIN*	0.929	0.672	0.800	0.821	**0.809**	0.815	0.808
	Gram†	0.956	**0.960**	0.594	0.771	0.551	0.598	0.738
	MDS†	0.983	0.935	0.589	0.817	0.508	0.537	0.728
	SSD	0.978	0.946	0.544	**0.856**	0.504	0.530	0.726
	LogP	0.830	0.749	0.628	0.549	0.648	0.572	0.663
	Sigma	0.858	0.913	0.727	0.791	0.570	0.611	0.745
	KLD-KN	0.906	0.910	**0.840**	0.759	0.705	0.759	**0.813**
	Euclid-KN	**0.992**	0.954	0.721	0.825	0.668	0.688	0.808

Table 3 compares the performance of probabilistic and deterministic embeddings as an average over all datasets. We can see that classification-based detectors work best for deterministic embeddings, as their performance is often correlated with downstream performance. However, other distance- and feature-based detectors have higher AUROC for the stochastic embeddings, meaning we have a latent space more suited for detecting examples outside of IN distribution.

4.3 Ablation study

Setup. The model is pretrained three times with different seeds for 200 epochs each time on the CIFAR-10 dataset [20] to evaluate different model hyperparameters. In particular, we assess the impact of various priors, β scales, and the number of MC samples. We compare the standard normal prior, $\mathcal{N}(0, 1)$,

Table 3. Comparison of AUROC performance (averaged over all datasets) for deterministic and stochastic embeddings in OOD detection.

		MSP	ODIN	Gram	MDS	SSD	LogP	Sigma	KLD-KN	Euclid-KN
VICReg	Deterministic	**0.806**	**0.805**	0.694	0.695	0.695	0.000	0.000	0.000	0.769
	H-prob.	0.792	0.790	**0.721**	**0.723**	0.720	**0.688**	0.736	**0.858**	**0.810**
	Z-prob.	0.796	0.792	0.704	**0.724**	**0.722**	0.683	**0.784**	0.785	0.796
Barlow	Deterministic	**0.808**	**0.812**	**0.739**	0.693	0.694	0.000	0.000	0.000	0.774
	H-prob.	0.805	0.808	**0.738**	0.687	0.688	0.663	**0.745**	**0.813**	0.762
	Z-prob.	0.803	0.802	0.719	**0.728**	**0.726**	0.659	0.687	0.774	**0.808**

with a Mixture of Gaussians (MoG),[1] aiming to assess the impact of employing a more expressive distribution for modelling stochastic embeddings. Additionally, we explore how the β scale influences the bottleneck and, consequently, the capacity of the embeddings.[2] Finally, we explore the advantages of utilizing multiple MC samples to estimate the expectation from Equation 6. Like downstream task evaluations, we train a linear classifier on the fixed representation from our pretrained backbone encoder.

Table 4. Comparison of top-1 accuracy (Acc@1) and expected calibration error (ECE) for the ablation study.

			Prior (# of MC samples)							
			Normal(1)		Normal(12)		MoG(1)		MoG(12)	
	Embeddings	Beta	Acc@1(↑)	ECE(↓)	Acc@1(↑)	ECE(↓)	Acc@1(↑)	ECE(↓)	Acc@1(↑)	ECE(↓)
VICReg	Z-prob.	1e-3	0.824	0.022	0.826	0.021	0.793	0.021	0.793	0.022
		1e-4	0.834	0.020	0.831	0.021	0.830	0.019	0.832	0.020
		1e-5	0.834	0.018	0.831	0.022	0.836	0.021	0.830	0.021
	H-prob.	1e-3	0.804	0.018	0.817	0.025	0.802	0.010	0.810	0.012
		1e-4	0.823	0.010	0.828	0.011	0.826	0.009	0.825	0.010
		1e-5	0.826	0.009	0.829	0.011	0.824	0.011	0.824	0.011
Barlow Twins	Z-prob.	1e-1	0.821	0.022	0.819	0.026	0.748	0.021	0.746	0.025
		1e-2	0.827	0.031	0.827	0.033	0.817	0.020	0.821	0.019
		1e-3	0.823	0.031	0.826	0.025	0.823	0.020	0.824	0.019
	H-prob.	1e-2	0.790	0.014	0.805	0.015	0.788	0.011	0.801	0.010
		1e-3	0.799	0.010	0.809	0.011	0.782	0.031	0.804	0.012
		1e-4	0.801	0.011	0.803	0.009	0.796	0.013	0.800	0.010

[1] The MoG prior has the following form: $\frac{1}{M}\sum_{m=1}^{M}\mathcal{N}(\mu_m, \mathrm{diag}(\sigma_m^2))$, where M denotes the number of mixtures, while μ_m and σ_m denote trainable parameters of a specific Gaussian in the mixture model.

[2] The variability in the loss function's magnitude and method-specific sensitivities necessitated the selection of distinct beta (β) scale hyperparameters for each approach, as documented in Table 4.

ICCS Camera Ready Version 2024
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-63783-4_25

Results. We report the results for the classification task in Table 4. Contrary to our initial expectations, the influence of MoG on performance appears insignificant, often leading to a deterioration in the model's efficacy. We have observed that smaller values of β tend to yield superior model performance, whereas higher values may degrade efficacy. However, excessively reducing β results in a corresponding reduction in the variance of the embeddings, rendering them more deterministic. In the case of H-prob. embeddings, increasing the number of MC samples enhances model performance, particularly with higher values of β. This suggests that employing more MC samples provides a more accurate and less biased estimation of expectations. Conversely, we found that the number of MC samples has a less significant effect on the performance of Z-prob. embeddings. While Z-prob. embeddings generally outperform H-prob. embeddings in terms of accuracy, the H-prob. embeddings offer better calibration, measured through the ECE.

5 Conclusions

In our study, we make significant strides in advancing the field by integrating stochastic assumptions directly into the information-theoretic-based self-supervised methods. Specifically, we introduce stochastic embeddings within feature decorrelation-based methods, demonstrating their potential to achieve performance competitive with the fully deterministic networks. Additionally, we delve into innovative strategies for effectively leveraging stochastic embeddings to identify OOD examples accurately. Our findings reveal that our methods exhibit robust OOD sample detection capabilities, surpassing traditional supervised detectors' performance. Moreover, we provide a comprehensive empirical evaluation, elucidating the impact of various hyperparameters on the training process. This showcases the potential of our approach and suggests avenues for future research in self-supervised learning optimization.

References

1. Achille, A., Soatto, S.: Emergence of invariance and disentanglement in deep representations. The Journal of Machine Learning Research **19**(1), 1947–1980 (2018)
2. Alemi, A.A., Fischer, I., Dillon, J.V., Murphy, K.: Deep Variational Information Bottleneck (Oct 2019), arXiv:1612.00410 [cs, math]
3. Ardeshir, S., Azizan, N.: Uncertainty in Contrastive Learning: On the Predictability of Downstream Performance (Jul 2022), arXiv:2207.09336 [cs, eess, stat]
4. Bardes, A., Ponce, J., LeCun, Y.: VICReg: Variance-Invariance-Covariance Regularization for Self-Supervised Learning (Jan 2022), arXiv:2105.04906 [cs]
5. Bordes, F., Balestriero, R., Garrido, Q., Bardes, A., Vincent, P.: Guillotine regularization: Why removing layers is needed to improve generalization in self-supervised learning. arXiv preprint arXiv:2206.13378 (2022)
6. Caron, M., Misra, I., Mairal, J., Goyal, P., Bojanowski, P., Joulin, A.: Unsupervised learning of visual features by contrasting cluster assignments. Advances in neural information processing systems **33**, 9912–9924 (2020)

ICCS Camera Ready Version 2024
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-63783-4_25

7. Chen, T., Kornblith, S., Norouzi, M., Hinton, G.: A simple framework for contrastive learning of visual representations. In: International conference on machine learning. pp. 1597–1607. PMLR (2020)
8. Chen, X., He, K.: Exploring simple siamese representation learning. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 15750–15758 (2021)
9. Cimpoi, M., Maji, S., Kokkinos, I., Mohamed, S., Vedaldi, A.: Describing textures in the wild. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 3606–3613 (2014)
10. Federici, M., Dutta, A., Forré, P., Kushman, N., Akata, Z.: Learning Robust Representations via Multi-View Information Bottleneck (Feb 2020), arXiv:2002.07017 [cs, stat]
11. Gawlikowski, J., Tassi, C.R.N., Ali, M., Lee, J., Humt, M., Feng, J., Kruspe, A.M., Triebel, R., Jung, P., Roscher, R., Shahzad, M., Yang, W., Bamler, R., Zhu, X.: A survey of uncertainty in deep neural networks. Artificial Intelligence Review **56**, 1513–1589 (2021)
12. Goyal, P., Mahajan, D., Gupta, A., Misra, I.: Scaling and benchmarking self-supervised visual representation learning. In: Proceedings of the ieee/cvf International Conference on computer vision. pp. 6391–6400 (2019)
13. Grill, J.B., Strub, F., Altché, F., Tallec, C., Richemond, P., Buchatskaya, E., Doersch, C., Avila Pires, B., Guo, Z., Gheshlaghi Azar, M., et al.: Bootstrap your own latent-a new approach to self-supervised learning. Advances in neural information processing systems **33**, 21271–21284 (2020)
14. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016)
15. Hendrycks, D., Gimpel, K.: A Baseline for Detecting Misclassified and Out-of-Distribution Examples in Neural Networks (Oct 2018), `http://arxiv.org/abs/1610.02136`, arXiv:1610.02136 [cs]
16. Hendrycks, D., Mazeika, M., Kadavath, S., Song, D.: Using self-supervised learning can improve model robustness and uncertainty. Advances in neural information processing systems **32** (2019)
17. Hjelm, R.D., Fedorov, A., Lavoie-Marchildon, S., Grewal, K., Bachman, P., Trischler, A., Bengio, Y.: Learning deep representations by mutual information estimation and maximization (Feb 2019), arXiv:1808.06670 [cs, stat]
18. Kingma, D.P., Welling, M.: Auto-encoding variational bayes. CoRR **abs/1312.6114** (2013)
19. Kirchhof, M., Kasneci, E., Oh, S.J.: Probabilistic contrastive learning recovers the correct aleatoric uncertainty of ambiguous inputs. In: Proceedings of the 40th International Conference on Machine Learning. JMLR.org (2023)
20. Krizhevsky, A.: Learning Multiple Layers of Features from Tiny Images (2009)
21. Le, Y., Yang, X.S.: Tiny ImageNet Visual Recognition Challenge (2015)
22. LeCun, Y., Cortes, C., Burges, C.: The MNIST database of handwritten digits (1998)
23. Lee, K., Lee, K., Lee, H., Shin, J.: A simple unified framework for detecting out-of-distribution samples and adversarial attacks. Advances in neural information processing systems **31** (2018)
24. Liang, S., Li, Y., Srikant, R.: Enhancing The Reliability of Out-of-distribution Image Detection in Neural Networks (Aug 2017), arXiv:1706.02690 [cs, stat]
25. Loshchilov, I., Hutter, F.: Decoupled Weight Decay Regularization (Jan 2019), arXiv:1711.05101 [cs, math]

26. Nakamura, H., Okada, M., Taniguchi, T.: Representation uncertainty in self-supervised learning as variational inference. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 16484–16493 (2023)
27. Netzer, Y., Wang, T., Coates, A., Bissacco, A., Wu, B., Ng, A.Y.: Reading digits in natural images with unsupervised feature learning. In: NIPS Workshop on Deep Learning and Unsupervised Feature Learning (2011)
28. Nilsback, M.E., Zisserman, A.: Automated Flower Classification over a Large Number of Classes. In: 2008 Sixth Indian Conference on Computer Vision, Graphics & Image Processing. pp. 722–729 (Dec 2008)
29. Ozbulak, U., Lee, H.J., Boga, B., Anzaku, E.T., Park, H., Van Messem, A., De Neve, W., Vankerschaver, J.: Know Your Self-supervised Learning: A Survey on Image-based Generative and Discriminative Training (May 2023), arXiv:2305.13689 [cs]
30. Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A.C., Fei-Fei, L.: ImageNet Large Scale Visual Recognition Challenge (Jan 2015), arXiv:1409.0575 [cs]
31. Sastry, C.S., Oore, S.: Detecting Out-of-Distribution Examples with Gram Matrices. In: Proceedings of the 37th International Conference on Machine Learning. pp. 8491–8501. PMLR (Nov 2020)
32. Sehwag, V., Chiang, M., Mittal, P.: Ssd: A unified framework for self-supervised outlier detection. arXiv preprint arXiv:2103.12051 (2021)
33. Shwartz-Ziv, R., Balestriero, R., Kawaguchi, K., Rudner, T.G.J., LeCun, Y.: An Information-Theoretic Perspective on Variance-Invariance-Covariance Regularization (Mar 2023), arXiv:2303.00633 [cs, math]
34. Shwartz-Ziv, R., LeCun, Y.: To Compress or Not to Compress – Self-Supervised Learning and Information Theory: A Review (2023), arXiv:2304.09355 [cs, math]
35. Tack, J., Mo, S., Jeong, J., Shin, J.: Csi: Novelty detection via contrastive learning on distributionally shifted instances. Advances in neural information processing systems **33**, 11839–11852 (2020)
36. Wang, T., Isola, P.: Understanding contrastive representation learning through alignment and uniformity on the hypersphere. In: International Conference on Machine Learning. pp. 9929–9939. PMLR (2020)
37. Xiao, J., Hays, J., Ehinger, K.A., Oliva, A., Torralba, A.: SUN database: Large-scale scene recognition from abbey to zoo. In: Computer Society Conference on Computer Vision and Pattern Recognition. pp. 3485–3492. IEEE (Jun 2010)
38. Yang, J., Wang, P., Zou, D., Zhou, Z., Ding, K., Peng, W., Wang, H., Chen, G., Li, B., Sun, Y., et al.: Openood: Benchmarking generalized out-of-distribution detection. Advances in Neural Information Processing Systems **35**, 32598–32611 (2022)
39. Yang, J., Zhou, K., Li, Y., Liu, Z.: Generalized Out-of-Distribution Detection: A Survey (Jan 2024), arXiv:2110.11334 [cs]
40. Zbontar, J., Jing, L., Misra, I., LeCun, Y., Deny, S.: Barlow twins: Self-supervised learning via redundancy reduction. In: International Conference on Machine Learning. pp. 12310–12320. PMLR (2021)
41. Zhou, B., Lapedriza, A., Khosla, A., Oliva, A., Torralba, A.: Places: A 10 million image database for scene recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence (2017)