Learning Low-Rank Latent Spaces with Simple Deterministic Autoencoder: Theoretical and Empirical Insights

Alokendu Mazumder[1]*, Tirthajit Baruah[1†], Bhartendu Kumar[2†], Rishab Sharma[3], Vishwajeet Pattanaik[1]
Punit Rathore[1]
[1]Indian Institute of Science, Bengaluru, [2]TCS Research, Bengaluru
[3]Dayananda Sagar College of Engineering, Bengaluru

Abstract

The autoencoder is an unsupervised learning paradigm that aims to create a compact latent representation of data by minimizing the reconstruction loss. However, it tends to overlook the fact that most data (images) are embedded in a lower-dimensional latent space, which is crucial for effective data representation. To address this limitation, we propose a novel approach called Low-Rank Autoencoder (LoRAE). In LoRAE, we incorporated a low-rank regularizer to adaptively learn a low-dimensional latent space while preserving the basic objective of an autoencoder. This helps embed the data in a lower-dimensional latent space while preserving important information. It is a simple autoencoder extension that learns low-rank latent space. Theoretically, we establish a tighter error bound for our model. Empirically, our model's superiority shines through various tasks such as image generation and downstream classification. Both theoretical and practical outcomes highlight the importance of acquiring low-dimensional embeddings.

1. Introduction

Learning effective representations remains a fundamental challenge in the field of artificial intelligence [1]. These representations, acquired through self-supervised or unsupervised learning, serve as valuable foundations for various downstream tasks like generation and classification. Among the methods employed for unsupervised representation learning, *autoencoders* (AEs) have gained popularity. AEs allow the extraction of meaningful features from data without the need for labelled examples. The process involves the transformation of data into a lower-dimensional space recognized as the latent space. Subsequently, this transformed data is reconstructed to match its original form, facilitating the acquisition of a meaningful and condensed representation known as a latent vector. To ensure the avoidance of trivial identity mapping, a key aspect is to restrict the information capacity within the autoencoder's internal representation. Over the course of time, various adaptations of autoencoders have been introduced to address this limitation. The Diabolo network [2], for instance, simplifies the approach by employing a low-dimensional representation. On the other hand, *variational autoencoders* (VAEs) [3] introduce controlled noise into latent vectors while constraining the distribution's variance. Denoising Autoencoders [4] are trained to intentionally generate substantial reconstruction errors by introducing random noise to their inputs. Meanwhile, Sparse Autoencoders [5] enforce a strict sparsity penalty on latent vectors. Quantized Autoencoders like *vector quantized* VAE (VQ-VAE) [6] discretize codes into distinct clusters, while Contrasting Autoencoders [7] minimize network function curvature beyond the manifold's boundaries. Notably, Low-rank Autoencoders, like *implicit rank minimizing autoencoder* (IR-MAE) [8], implicitly minimize the rank of the empirical covariance matrix of the latent space by leveraging the dynamics of *stochastic gradient descent* (SGD). In essence, the array of techniques for autoencoder refinement has expanded significantly, leading to a comprehensive toolkit of strategies for overcoming inherent drawbacks and enhancing the efficacy of these latent space learning models.

In this work, we let the network learn the best possible (low) rank/dimensionality of the latent space of an autoencoder by deploying a nuclear norm regularizer that promotes a low-rank solution. We call this model *Low-Rank Autoencoder* (LoRAE). This method consists of inserting a single linear layer between the encoder and decoder of a vanilla autoencoder. This layer is trained along the encoder and decoder networks. The primary objective of this methodology involves not only minimizing the conventional reconstruction loss but also minimizing the nuclear norm of the added linear layer. Due to the presence of nuclear norm minimization in the loss function, the network will now adjust to an effective low-dimensionality of

---

*corresponding author: alokendum@iisc.ac.in
†denotes equal contribution

latent space. The nuclear norm regularization of a matrix $\mathbf{A}$ is an l_1 regularization of the singular values of $\mathbf{A}$, and it, therefore, promotes a low-rank solution [9]. Similar to various regularization techniques, the additional linear layer remains inactive during inference. Consequently, the architecture of both the encoder and the decoder within the model remains consistent with the original design. In practical application, the linear layer (matrix) is treated as the last layer of the encoder during the inference phase.

We showcase the superiority of LoRAE in learning representations, surpassing the performance of conventional AE, VAE, IRMAE and several state-of-the-art deterministic autoencoders. This validation is carried out using the MNIST and CelebA datasets, employing diverse tasks such as generating samples from noise, interpolation, and classification. We additionally perform experiments to probe the influence of the regularizer on the rank of the latent space[1], achieved by varying its penalty parameter. Furthermore, we explore the impact of varying the dimension of the latent space while maintaining a constant penalty term.

The fundamental objective of this paper is to highlight the capacity of low-rank latent spaces acquired through the encoder to yield substantial data representations, thereby elevating both the generative potential and downstream performance. Moreover, we assert that the presence of a nuclear norm penalty in LoRAE offers several robust mathematical assurances. *It's important to note that our paper does not strive to propose a new generative model. Only one of the essences of this work lies in empirically confirming that a low-rank constraint on the latent space of a simple deterministic autoencoder promotes generative capabilities comparable to well-established generative models.*

This paper serves as a proclamation that the utilization of low-rank autoencoders with nuclear norm penalty can yield significant representation outcomes while simultaneously affording the opportunity for the establishment of robust mathematical guarantees.

We summarize our contribution as follows:

1. We introduce a novel framework to enrich the latent representation of autoencoders. This is achieved by incorporating a sparse/low-rank regularized projection layer, which dynamically reduces the dimensionality of the latent space.

2. We provide substantial mathematical underpinning for LoRAE through an analysis of distance prediction error bounds. This analysis sheds light on the reasons driving LoRAE's superior performance in downstream tasks such as classification. Additionally, we offer theoretical guarantees (under certain assumptions) on the convergence of our learning algorithm.

3. We showcased LoRAE's superior performance by comparing it with several baselines, like **(i)** a conventional deterministic AE, **(ii)** a VAE, and **(iii)** an IRMAE. We also conducted comparisons of LoRAE against several state-of-the-art deterministic autoencoders and various established *generative adversarial network* (GAN) and VAE based generative models across a variety of generative tasks.

2. Literature Survey

Training a fully linear network with SGD naturally results in a low-rank solution. This phenomenon can be interpreted as a form of implicit regularization, which has been extensively investigated across diverse learning tasks. Examples include deep matrix factorization [10, 11], convolutional neural networks [12], and logistic regression [13]. The implicit regularization offered by gradient descent is believed to be a pivotal element in enhancing the generalizability of deep neural networks. In the domain of deep matrix factorization, *Arora et al.* [10] extended this concept in the case of deep neural nets with solid theoretical and empirical results that a deep linear network can promote low-rank solutions. *Gunasekar et al.* [12] further extended this work towards *convolutional neural networks* (CNN) and proved that deep linear CNNs can derive low-rank solutions when optimized with gradient descent. Numerous other studies have concentrated on linear scenarios, aiming to empirically and theoretically investigate this phenomenon. The work by *Saxe et al.* [14] demonstrates theoretically that a simple two-layer linear regression model can attain a low-rank solution when optimized using continuous gradient descent. Later *Gidel et al.* [15] extended this same concept of low-rank solutions in linear regression problems for a discrete case of gradient descent.

While the previously mentioned works primarily aimed to comprehend how gradient descent contributes to generalization within established practices, our approach diverges by harnessing this phenomenon to design deep models capable of learning better latent representations of data. Autoencoders are simple yet powerful unsupervised deep models capable of learning latent representations of data. As most complex data (like images) are embedded in some low-dimensional subspace, it is necessary to limit their latent capacity. A significant category of these methods is founded on the concepts of variational autoencoders [3], including variations like β-VAE [16]. A notable limitation of these methods arises from their probabilistic/generative nature, often resulting in the generation of reconstructed images with reduced clarity (blurry images). Conversely, basic deterministic autoencoders encounter an issue characterized by the existence of *"holes"*[2] within their latent

---

[1]The rank of the latent space corresponds to the count of non-zero singular values of the empirical covariance matrix of the latent space.

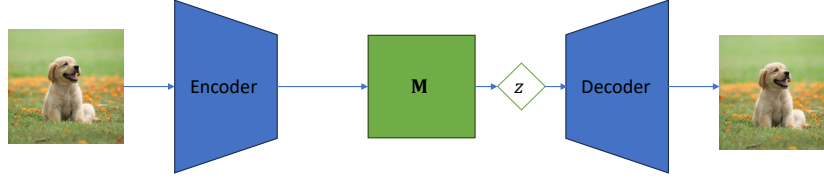[2]Discontinuous regions in the latent space of an autoencoder.

Figure 1. **Low-Rank Autoencoder (LoRAE)**: We employ a simple deterministic autoencoder coupled with a nuclear norm penalty in its loss function to facilitate the acquisition of a low-rank latent space. During the inference phase, the linear matrix **M** is condensed into the final layer of the encoder.

space, primarily arising from the absence of constraints on the distribution of its latent space. Several deterministic autoencoders are proposed to tackle this issue, namely *regularized autoencoders* (RAE) [17], *wasserstein autoecnoder* (WAE) [18] and *vector quantized VAE* (VQ-VAE) [6]. Recently, *Jing et al.* [8] proposed *implicit rank minimizing autoencoder* (IRMAE) which leverages the *"low-rank"* phenomena [10] in deep linear networks to learn a better latent representation. It uses a series of linear layers sandwiched between the encoder and decoder of a simple autoencoder. A sequence of linear layers maintains functional and expressive parity with a single linear layer. However, it's essential to note that implicit regularization doesn't manifest for unstructured datasets like random full-rank noise, as pointed out in [10]. This observation suggests that the occurrence of this phenomenon hinges on the underlying data structure. Moreover, it lacks vital mathematical assurances such as the convergence of its iterations and the underlying principles driving its effectiveness in downstream tasks.

3. Low-Rank Autoencoder (LoRAE)

In this section, we unveil our proposed architecture. Let $\mathbf{E} : \mathbb{R}^{m \times n \times c} \rightarrow \mathbb{R}^l$ and $\mathbf{D} : \mathbb{R}^l \rightarrow \mathbb{R}^{m \times n \times c}$ denote the encoder and decoder of a simple deterministic autoencoder respectively. Let $z \in \mathbb{R}^l$ denote a vector in its latent space, where l is the dimension of latent space. The latent space is modelled by $\mathbf{E}(x)$. Here $x \in \mathbb{R}^{m \times n \times c}$ be an image of size $m \times n$ with c number of channels. A simple (vanilla) deterministic autoencoder optimizes the L_2 reconstruction loss $\mathcal{L}_{vanilla} = \| x - \mathbf{D}(\mathbf{E}(x)) \|_2^2$ without any constrain over its latent space, hence promoting the presence of holes.

In LoRAE, we add an additional single linear layer between the encoder and decoder. Let $\mathbf{M} \in \mathbb{R}^{l \times l}$ denote a real matrix characterized by a linear layer. The diagram of LoRAE is shown in Figure 1. We explicitly regularize the matrix **M** with a nuclear norm penalty to encourage learning a low-rank latent space. We train the matrix **M** jointly with the encoder and decoder. Hence, the final loss of LoRAE can be written as:

$$\mathcal{L}(\mathbf{E}, \mathbf{D}, \mathbf{M}) = \| x - \mathbf{D}(\mathbf{M}(\mathbf{E}(x))) \|_2^2 + \lambda \| \mathbf{M} \|_* \quad (1)$$

where $\| \mathbf{M} \|_*$ is the nuclear norm[3] (also known as the trace norm) of matrix **M**. We will now minimize the loss given in Eq.(1) using ADAM [19] optimizer in batch form. Let $\mathcal{B}$ denote a mini-batch of training data and $|\mathcal{B}|$ be the number of training points in it that are randomly sampled from the training set. We can now write Eq.(1) in mini-batch form as follows:

$$\mathcal{L}_{batch}(\mathbf{E}, \mathbf{D}, \mathbf{M}) = \frac{1}{|\mathcal{B}|} \sum_{x \in \mathcal{B}} \| x - \mathbf{D}(\mathbf{M}(\mathbf{E}(x))) \|_2^2 + \lambda \| \mathbf{M} \|_*$$

$$(2)$$

Let $\theta \in \{\mathbf{E}, \mathbf{D}, \mathbf{M}\}$, β_1 and $\beta_2 \in [0, 1)$ and $\epsilon > 0$. Here, $\mathbf{m}_{\theta,t} = \beta_1 \mathbf{m}_{\theta,t-1} + (1 - \beta_1) \nabla_\theta \mathcal{L}_{batch}$, $\mathbf{v}_{\theta,t} = \beta_2 \mathbf{v}_{\theta,t-1} + (1 - \beta_2)(\nabla_\theta \mathcal{L}_{batch})^2$ and $V_{\theta,t} = diag(\mathbf{v}_{\theta,t})$. Now, we can solve for Eq.(2) using Algorithm 1.

Algorithm 1: Minimizing Eq. (2) using ADAM

Input: Training data $\{x\}_{i=1}^N$, Batch size $|\mathcal{B}| \in \mathbb{Z}^+$,
          Learning rate: $\alpha \in (0, 1]$

1 **Initialization:**
2 **for** t *from 1 to* T*:* **do**
3   (i) Randomly sample $|\mathcal{B}|$ number of data points from training set.
4   (ii) Compute gradient of $\mathcal{L}_{batch}$ with respect to **E**, **D** and **M**.
5   (iii) Update the parameters using the ADAM update rule:
6     $\mathbf{E}_{t+1} = \mathbf{E}_t - \alpha(V_{\mathbf{E},t}^{1/2} + diag(\epsilon\mathbb{I}))^{-1}\mathbf{m}_{\mathbf{E},t}$
7     $\mathbf{D}_{t+1} = \mathbf{D}_t - \alpha(V_{\mathbf{D},t}^{1/2} + diag(\epsilon\mathbb{I}))^{-1}\mathbf{m}_{\mathbf{D},t}$
8     $\mathbf{M}_{t+1} = \mathbf{M}_t - \alpha(V_{\mathbf{M},t}^{1/2} + diag(\epsilon\mathbb{I}))^{-1}\mathbf{m}_{\mathbf{M},t}$
9 **End:**
10 **Output:** The learned $\mathbf{E}^*$, $\mathbf{D}^*$ and $\mathbf{M}^*$

During the training phase, the use of the nuclear norm penalty prompts the latent variables to occupy an even lower-dimensional subspace. As a result, this process reduces the rank of the empirical covariance matrix of the la-

---

[3]$\| \mathbf{M} \|_* = \sum_{i=1}^l \sigma_i(\mathbf{M}) = trace(\sqrt{\mathbf{M}^T\mathbf{M}})$, here $\sigma_i's$ denote singular values of matrix **M**.

tent space. To enhance the impact of this regularization, one can amplify its effect by adjusting the penalty term λ to a higher value.

During inference time, the linear layer is *"collapsed"* into the encoder. Hence, we can directly use this new encoder for generative tasks. We can also use the encoder only for downstream tasks.

4. Experiments

In this section, we conduct an empirical assessment of LoRAE in comparison to baselines (AE, VAE, IRMAE) and state-of-the-art models (WAE, RAE). In Section 4.1, we highlight that LoRAE occupies a relatively smaller latent space compared to a standard AE and achieves smooth dimensionality reduction, unlike IRMAE. Moving to Section 4.2, we provide empirical evidence of LoRAE's ability to generate images of superior quality compared to the fundamental vanilla AE. This superiority is attributed to its low-rank latent space. Additionally, the model demonstrates enhanced quantitative performance when compared against simple AE, VAE, and IRMAE. Next, in Section 4.3, we leverage the encoder component of our trained model to effectively classify images within the MNIST dataset. Lastly, we study the effect of two crucial hyper-parameters of our model in Section 4.4.

4.1. Dimensionality Reduction

In Figure 2, we present the dimensionality reduction achieved by LoRAE and IRMAE in the latent space. LoRAE's gradual and smooth decay in the singular value plot stands in contrast to IRMAE's sharp decline. This discrepancy contributes to IRMAE occupying an even smaller-dimensional latent space than LoRAE. This distinction arises because nuclear norm minimization penalizes larger singular values more. Conversely, in the IRMAE plot, a rapid transition toward zero is observed compared to ours, potentially stemming from the fact that gradient descent dynamics penalize smaller singular values more than larger ones, driving them toward zero. LoRAE employs a considerably smaller latent space compared to a basic deterministic AE.

In Section 4.4.1, we showcased that there exists a *sweet* point of latent space rank which signifies a balance wherein the model's generative capacity is empirically maximized. Hence, latent spaces with very low ranks/low dimensions are also unfavourable as they lead to a deterioration in generative performance.

4.2. Generative Tasks

The effectiveness and quality of a learned latent space can be assessed by generating images from it and examining the smoothness during transitions from one point to another within it. When the generated images demonstrate a
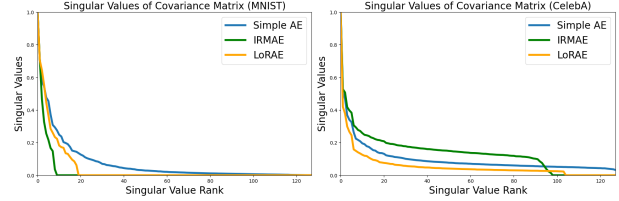

Figure 2. **Singular value plot of the empirical covariance matrix of the latent space**: A comparison between a basic AE, IRMAE and LoRAE on MNIST and CelebA datasets. The empirical covariance matrix is computed from the test set of both datasets.
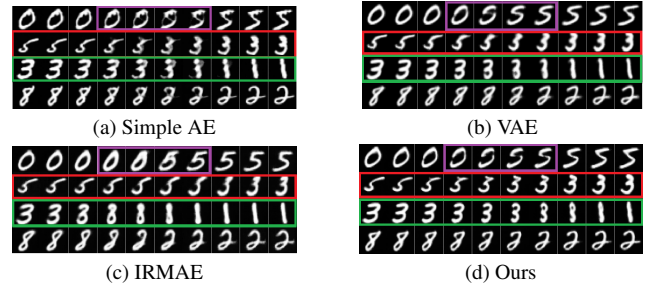

(a) Simple AE   (b) VAE

(c) IRMAE   (d) Ours

Figure 3. **Interpolation**: Employing linear interpolation among data points on the MNIST dataset, we observe three instances in a left-to-right top-to-bottom sequence: (a) Basic AE, (b) VAE, (c) IRMAE and (d) LoRAE.

high level of quality, it indicates the latent space's successful learning. We trained our model using the MNIST and CelebA datasets and conducted a comparison with a basic deterministic AE, IRMAE as well as a VAE. We set the latent space dimensions to 128 for MNIST and CelebA, training for 50 and 100 epochs respectively (Additional hyperparameter details are available in our *supplementary material*). In this section, we conducted two experiments: **(i)** linearly interpolating between two data points, and **(ii)** generating images from the latent space using *Gaussian Mixture Model* (GMM) and *Multivariate Gaussian* (MVG) noise fitting. Additionally, we quantitatively evaluated our model's generative capability using the *Fréchet Inception Distance* (FID) [20, 21] score. Across the tasks mentioned above, we compare our model qualitatively (see Figure 3, Table 1 and Table 2) and quantitatively (see Table 3). with a simple AE, VAE and IRMAE. All models reported in Table 3 shared the exact same architecture.

4.2.1 Interpolation Between Datapoints

We conduct linear interpolation on the latent variables of two randomly selected images from the test set. The results depicted in Figure 3 show that LoRAE significantly outperforms the conventional AE. Its generated quality aligns with that of a VAE and surpasses IRMAE. The progression

Table 1. **Generation from GMM**: MNIST and CelebA images generated from GMM with 4 and 10 clusters respectively. We show the images generated by a simple autoencoder, VAE, IRMAE and LoRAE (from left to right).

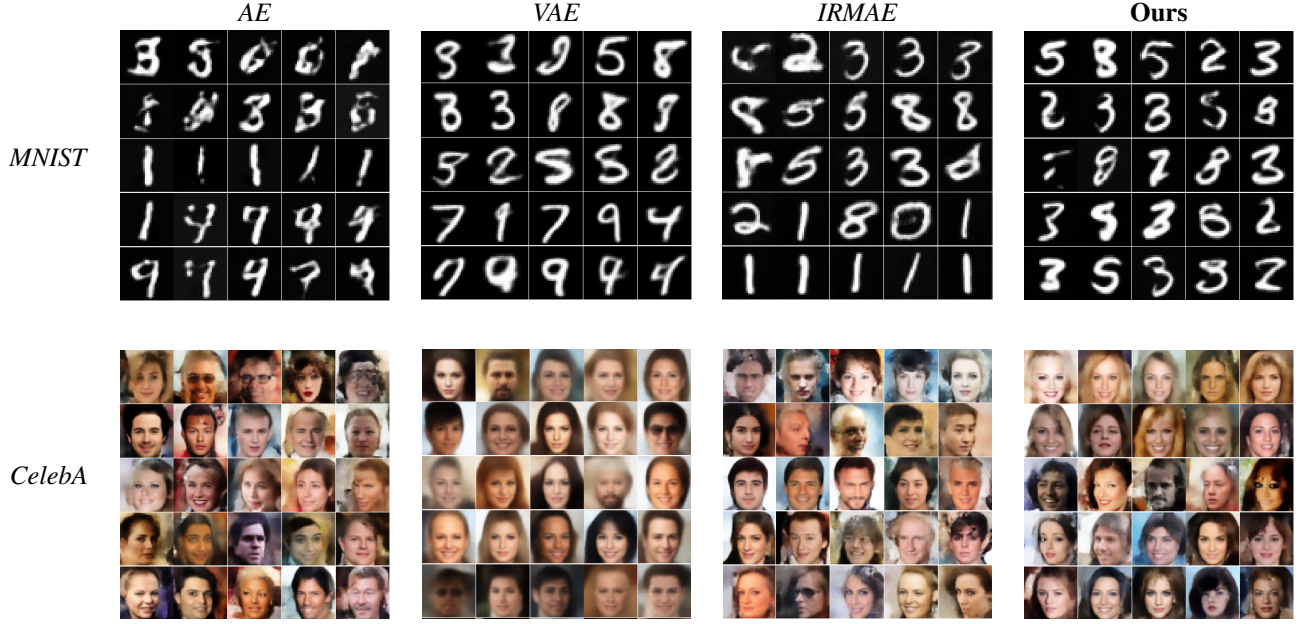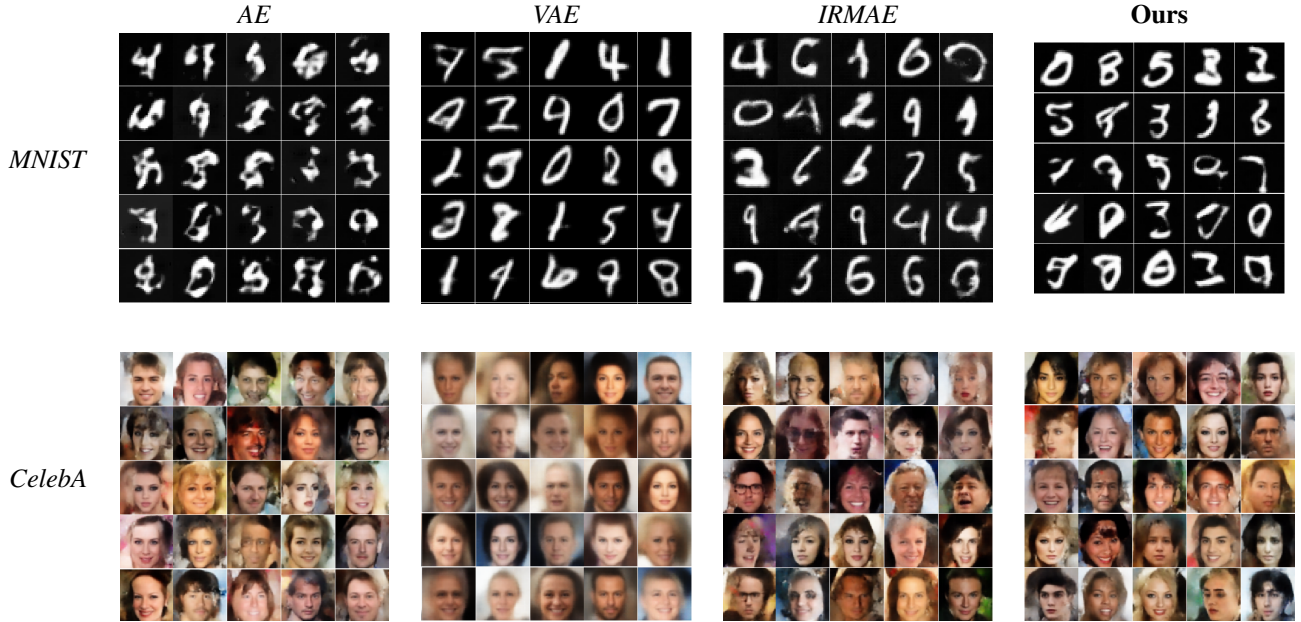|	*AE*	*VAE*	*IRMAE*	**Ours**
MNIST				
CelebA				

Table 2. **Generation from MVG**: MNIST and CelebA images generated from MVG noise. We show the images generated by a simple autoencoder, VAE, IRMAE and LoRAE (from left to right).

	AE	*VAE*	*IRMAE*	**Ours**
MNIST				
CelebA				

from digit 5 to 3 showcases a high degree of smoothness in LoRAE, setting it apart from both VAE and IRMAE (highlighted by the red box). VAE's probabilistic nature introduces a slightly blurred transition for the same sequence. Conversely, transitioning from 3 to 1, IRMAE generates a digit resembling 8 (marked by the green box). Likewise, in the transition from 0 to 5, IRMAE produces a somewhat blurry transformation (indicated by the violet box in Figure 3 (c)). Remarkably, LoRAE achieves seamless transitions between individual digits, thereby effectively learning a better and refined latent space.

4.2.2 Generating Images from Noise

Deterministic autoencoders lack generative capability. Here, to demonstrate the generative capability of LoRAE, we fit a Gaussian noise. We demonstrate the capability of LoRAE to generate high-quality images from Gaussian noise. In particular, we accomplish this by fitting **(i)** a *Gaussian Mixture Model* (GMM) and **(ii)** a *Multivariate Gaussian (MVG)* (which we denote as $\mathcal{N}$) distribution to its latent space. After fitting the noise distribution, we proceed to sample from it and feed these samples through the decoder to generate images. We take the number of clusters 4 and 10 for MNIST and CelebA respectively in the case of GMM to avoid overfitting. We quantitatively evaluate the generating capacity of each model using the FID score.

Analyzing the outcomes presented in Table 1 and Table 2, a clear trend emerges: our model excels in generating images with higher visual quality when compared to the simple AE and IRMAE models. Furthermore, its image generation quality stands on par with that of the VAE. Differing from VAE, which often produces images with blurred backgrounds due to its probabilistic nature, our model, being deterministic, avoids generating images with such blurriness. This distinction results in clearer and more defined images produced by LoRAE. These findings are supported by the data showcased in Table 3 where our model achieves the best FID score among the three, highlighting the pivotal role of low-rank latent space achieved through explicit regularization using nuclear norm penalty. This strategic enhancement of generative capabilities effectively surpasses even the performance of a conventional VAE.

Table 3. FID scores obtained between the generated samples from MVG and GMM noise and real test samples for all models.

	MNIST		CelebA	
	$\mathcal{N}$	GMM	$\mathcal{N}$	GMM
AE	103.08	68.97	68.13	59.43
VAE	21.01	18.86	61.87	53.63
IRMAE	26.58	22.31	58.98	48.56
LoRAE	**19.50**	**11.09**	56.29	**45.43**

4.2.3 Comparing with other SOTA Models

We perform a comparative analysis of LoRAE against state-of-the-art deterministic autoencoders such as *wasserstein's autoencoder* (WAE) [18] and *regularized autoencoders* (RAE) [17], in terms of FID scores (we retrained these models in our system using the exact parameters from their original paper.). Based on the data presented in Table 4, it can be observed that LoRAE attains the top rank in all cases, except for the MVG ($\mathcal{N}$) case of CelebA, where it shares the second position with RAE.

It is evident that LoRAE exhibits enhanced performance

on the MNIST dataset and remains competitively aligned with the performance on the CelebA dataset.

Table 4. Comparision of LoRAE against modern state-of-the-art deterministic autoencoders (FID Score is used for comparison).

	MNIST		CelebA	
	$\mathcal{N}$	GMM	$\mathcal{N}$	GMM
WAE	21.04	11.32	57.6	45.91
RAE	22.12	11.54	**50.31**	46.05
LoRAE	**19.50**	**11.09**	56.29	**45.43**

Moreover, we broaden our comparative analysis to include various other generative models on the CelebA dataset. This encompasses fundamental models like the *generative adversarial network* (GAN) [22], *least squares GAN* (LS-GAN) [23], *non-saturating GAN* (NS-GAN)] [24], along with the combined architecture of VAE and Flow models (VAE+Flow) [25].

The reported values for all GAN-based models in this Table 5 correspond to the optimal FID results from an extensive hyperparameter search, conducted separately for each dataset, as detailed by *Lucic et al.* in [26]. Instances involving severe mode collapse were excluded to prevent inflation of these FID scores. Our analysis clearly demonstrates that

Table 5. Comparison of LoRAE against well-known generative models on CelebA dataset. (FID Score is used for comparison)

GAN	LS-GAN	NS-GAN	VAE + Flow	**LoRAE**
65.2	**54.1**	57.3	65.7	56.3

LoRAE achieves FID scores that align closely with those attained by other established generative models.

To conclude Section 4.2, we offer empirical evidence that the incorporation of a nuclear norm penalty term into a simple AE framework, which enforces an explicit low-rank constraint on the latent space, leads to a substantial enhancement in the model's generative capabilities.

4.3. Downstream Classification Task

Latent variables play a crucial role in downstream tasks as they encapsulate the fundamental underlying structure of the data distribution [27–29]. The potential of these self-supervised learning methods to outperform purely-supervised models is particularly promising. We engage in downstream classification using the MNIST dataset. For each method - the simple AE, VAE, and LoRAE - we train a *multi-layer perceptron* (MLP) layer on top of the trained encoder (Parameters for classification are given in our *supplementary material*). The fine-tuning of the pre-trained encoder is excluded, except for instance involving the purely supervised version of the basic AE (tagged as *supervised* in

Table 6). As shown in Table 6, representation learned by LoRAE showcases significantly superior classification accuracy in comparison to simple AE and VAE. Hence, it is evident that LoRAE surpasses simple AE, VAE and the supervised version in the low-data regime while maintaining a similar level of performance when operating on full-length datasets.

Table 6. Classification accuracy (in %) obtained from a simple AE, VAE and LoRAE on MNIST dataset.

Size of training set	10	100	1000	10,000	60,000
AE	41.0	68.16	89.8	96.5	98.1
VAE	41.5	77.47	94.0	**98.5**	98.9
LoRAE	**46.6**	**89.02**	**95.4**	97.9	98.6
Supervised	37.8	73.59	94.2	98.3	**99.2**

We provide strong theoretical support for its improved performance in downstream classification tasks through Theorem 2 in Section 5.2.

4.4. Effect of Hyperparameters in LoRAE

In this section, we investigate two crucial hyperparameters: **(i)** penalty parameter λ and **(ii)** encoder output dimension (latent dimension). We seek insights into their impact on LoRAE's generative capacity.

4.4.1 Effect of Penalty Parameter λ on FID

Elevating the penalty term results in a more pronounced regularization effect, whereas reducing it lessens this impact. We investigate the effect of the regularization penalty parameter on FID scores. Intuitively, the rank of the latent space appears to have an inverse relationship with the penalty parameter. A more pronounced emphasis on nuclear norm minimization is likely to lead to a reduction in the rank of the latent space (more discussion in Section 5.3).

Table 7. Effect of varying penalty parameter for MNIST dataset.

λ	10^{-1}	10^{-2}	10^{-3}	10^{-4}	10^{-5}
Rank of latent space	7	8	**18**	40	65
FID	23.65	22.18	**11.09**	25.60	40.75

In Table 7, it becomes evident that extremely low and high-dimensional latent spaces result in suboptimal generative capabilities for LoRAE. This highlights the significance of identifying an optimal penalty parameter (λ) that yields the most suitable latent space for generative tasks, as indicated by the best FID. Hence, the hyperparameter (λ) assumes a critical role and requires optimization in practical applications.

4.4.2 Comparison with Simple AE on Various Latent Dimension

Autoencoders that exhibit diverse latent dimensions or varying prior configurations inherently necessitate a balance between obtaining valuable representations. In this context, we explore the impact of latent dimensionality on FID scores for both LoRAE and a simple AE. From Table 8. Effect of varying latent dimensions on LoRAE compared to a simple AE using CelebA. (FID Score is used for comparison)

Table 8. Effect of varying latent dimensions on LoRAE compared to a simple AE using CelebA. (FID Score is used for comparison)

Latent Dimension	64	128	256	512	1024
AE	**69.08**	68.13	66.94	91.42	107.98
LoRAE	71.58	**56.29**	**62.42**	**62.93**	**63.89**

ble 8, it becomes apparent that LoRAE, when endowed with larger latent dimensions, demonstrates enhanced performance compared to the optimally dimensioned AE. Additionally, it is noticeable that as the dimensionality increases, the performance of LoRAE tends to plateau.

5. Theoretical Analysis

In this section, we offer an exhaustive theoretical examination of our proposed model. Our focus revolves around two key aspects: firstly, an exploration of the rate of convergence exhibited by our learning algorithm, and secondly, an exploration of the lower bound on the min-max distance ratio. Furthermore, we also establish a proof demonstrating the inverse correlation between the rank of LoRAE's latent space and the penalty parameter (λ). Collectively, these analyses highlight and confirm the effectiveness of LoRAE.. All proofs are given in our *supplementary materials*.

5.1. Convergence Analysis

In this section, we are going to present the convergence analysis for the loss function as described in Eq.(1), considering the ADAM iterations outlined in Algorithm 1. Before delving into the analysis, we will outline the assumptions that have been considered. **(i)** The loss function in Eq.(1) is $K - Lipchitz$, **(ii)** it has a $\sigma > 0$ bounded gradient, i.e $\|\nabla \mathcal{L}(\mathbf{E}, \mathbf{D}, \mathbf{M})\|_2 < \sigma < \infty$, **(iii)** it has a well-defined minima, i.e $\mathcal{L}(\mathbf{E}^*, \mathbf{D}^*, \mathbf{M}^*) \leq \mathcal{L}(\mathbf{E}, \mathbf{D}, \mathbf{M})$, where $\mathbf{E}^*, \mathbf{D}^*, \mathbf{M}^* = \arg\min_{\mathbf{E}, \mathbf{D}, \mathbf{M}} \mathcal{L}(\mathbf{E}, \mathbf{D}, \mathbf{M})$ and **(iv)** we prove convergence for deterministic version of ADAM.

Theorem 1. *Let the loss function $\mathcal{L}(\boldsymbol{E}, \boldsymbol{D}, \boldsymbol{M})$ be $K-Lipchitz$ and let $\gamma < \infty$ be an upper bound on the norm of the gradient of $\mathcal{L}$. Then the following holds for the deterministic version (when batch size = total dataset) of Algorithm (1):*

For any $\sigma > 0$ if we let $\alpha = \sqrt{2(\mathcal{L}(\boldsymbol{E}_0, \boldsymbol{D}_0, \boldsymbol{M}_0) - \mathcal{L}(\boldsymbol{E}^*, \boldsymbol{D}^*, \boldsymbol{M}^*))/K\delta^2 T}$, then

there exists a natural number $T(\sigma, \delta)$ (depends on σ and δ) such that $\|\mathcal{L}(\boldsymbol{E}_t, \boldsymbol{D}_t, \boldsymbol{M}_t)\|_2 \leq \sigma$ for some $t \geq T(\sigma, \delta)$, where $\delta^2 = \frac{\gamma^2}{\epsilon^2}$.

With our analysis, we showed that our algorithm attains convergence to a stationary point with rate $\mathcal{O}(\frac{1}{T^{1/4}})$ when proper learning rate $\alpha > 0$ is set.

Therefore, we are in a position to confidently assert that the integration of the supplementary linear layer $\mathbf{M}$ and the utilization of nuclear norm regularization indeed ensure the guaranteed convergence of our learning algorithm.

5.2. Lower Bound on min-max Distance Ratio

Considering a given set of independent and identically distributed (i.i.d) random variables, denoted as $x, x_1, x_2, \ldots, x_N$, where each of these variables lies in the Euclidean space $\mathbb{R}^d$ (with x and x_i applicable for all $i \in \{1, 2, \ldots, N\}$), the corresponding embeddings $f^*(x)$ and $f^*(x_{1 \leq i \leq N})$ also demonstrate an i.i.d nature, irrespective of the method used for their learning as highlighted by *Dasgupta et al.* in [30]. *Beyer et al.* in [31] postulated that, in high-dimensional spaces, the minimum distance $\left(d_{min}^{f^*} = \min_{1 \leq j \leq N} d^{f^*}(x, x_j) \right)$[4] and the maximum distance $\left(d_{max}^{f^*} = \max_{1 \leq j \leq N} d^{f^*}(x, x_j) \right)$ tend to be similar when considering $f(x)$ and $f(x_{1 \leq i \leq N})$ as i.i.d. Hence, the notion of similarity and dissimilarity with respect to the distance function between data points in embedding space is completely lost in such cases. In autoencoders, to avoid learning high-dimensional latent features, one may directly reduce the dimensionality of the output layer of the encoder, but this will cause dimensional collapse. In our model, we consider the matrix $\mathbf{M}$ to transform the encoder embedding result $\mathbf{E}(x)$ into the latent vector $\mathbf{M}(\mathbf{E}(x))$. When we further introduce the low-rank constraint for $\mathbf{M}$, we can obtain a low-dimensional latent space for a simple autoencoder.

Given that our approach explicitly imposes a constraint on the dimensionality of the learned latent space, it follows intuitively that the min-max distance ratio, i.e., $(d_{max}^{\mathbf{E}^*\mathbf{M}^*} - d_{min}^{\mathbf{E}^*\mathbf{M}^*})/d_{min}^{\mathbf{E}^*\mathbf{M}^*}$, should invariably possess a lower-bound.

Theorem 2. *Given any set of i.i.d* x, $x_1, x_2, \ldots, x_N \in \mathbb{R}^l$, *we denote* $d_{max}^{\mathbf{E}^*\mathbf{M}^*} = \max_{1 \leq j \leq N} d^{\mathbf{E}^*\mathbf{M}^*}(x, x_j)$ *and* $d_{min}^{\mathbf{E}^*\mathbf{M}^*} = \min_{1 \leq j \leq N} d^{\mathbf{E}^*\mathbf{M}^*}(x, x_j)$, *then we always have the conditional probability:*

$$\mathbb{P}\left(\frac{d_{max}^{\mathbf{E}^*\mathbf{M}^*} - d_{min}^{\mathbf{E}^*\mathbf{M}^*}}{d_{min}^{\mathbf{E}^*\mathbf{M}^*}} \geq \Theta(\mathcal{D}, \lambda) \middle| \lambda > 0 \right) = 1 \quad (3)$$

where $d^{\mathbf{E}^*\mathbf{M}^*}(x, x_j) = \frac{\|\mathbf{M}^*(\mathbf{E}^*(x)) - \mathbf{M}^*(\mathbf{E}^*(x_i))\|_2}{rank(\mathbf{M}^*)}$, $\mathcal{D}$ *denotes the training dataset and* $\Theta(\mathcal{D}, \lambda)$ *depends on the training set and regularization penalty parameter* λ.

---

[4]Here, $d^{f^*}(., .)$ is some distance metric/measure.

By examining Eq.(3), it becomes evident that the min-max distance ratio inherently possesses a lower bound due to the incorporation of the nuclear norm regularization term. This suggests that the penalty parameter λ governs the establishment of the lower bound. Given that the min-max distance ratio maintains a consistent lower bound, LoRAE is equipped to effectively discriminate between similar and dissimilar data points. Consequently, the learned embeddings capture inherent similarities, leading to enhanced performance on downstream tasks.

5.3. Rank of Latent Space of LoRAE

As mentioned in Section 4.4.1, it is intuitively understood that stronger regularization in LoRAE leads to a reduction in the rank of the latent space. Additionally, within the same section, we presented empirical evidence of an inverse correlation between the rank of the latent space and the parameter λ. To provide robust mathematical support, we propose the following proposition.

Proposition 1. Rank of the latent space follows $\mathcal{O}(1/\lambda)$.

6. Conclusion

A pivotal element within autoencoder methodologies revolves around minimizing the information capacity of the latent space it learns. In this study, we extend beyond mere implicit measures and actively minimize the latent capacity of an autoencoder. Our approach involves minimizing the rank of the empirical covariance matrix associated with the latent space. This is achieved by incorporating a nuclear norm penalty term into the loss function. This addition aids the autoencoder in acquiring representations characterized by significantly reduced dimensions.

Incorporating a nuclear norm penalty alongside the vanilla reconstruction loss introduces various mathematical assurances. We provide robust mathematical assurances regarding the convergence of our algorithm and elucidate the factors contributing to its commendable performance in downstream tasks. Comparison experiments across multiple domains involving image generation and representation learning indicated that our learning algorithm acquires more reliable feature embedding than baseline methods. Both the theoretical and experimental results clearly demonstrated the necessity/significance of learning low-dimensional embeddings in autoencoders.

Acknowledgement

References

[1] Yoshua Bengio, Aaron Courville, and Pascal Vincent. Representation learning: A review and new perspectives. *IEEE transactions on pattern analysis and machine intelligence*, 35(8):1798–1828, 2013. 1

[2] David E Rumelhart, Geoffrey E Hinton, Ronald J Williams, et al. Learning internal representations by error propagation, 1985. 1

[3] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013. 1, 2

[4] Pascal Vincent, Hugo Larochelle, Yoshua Bengio, and Pierre-Antoine Manzagol. Extracting and composing robust features with denoising autoencoders. In *Proceedings of the 25th international conference on Machine learning*, pages 1096–1103, 2008. 1

[5] Marc'Aurelio Ranzato, Y-Lan Boureau, Yann Cun, et al. Sparse feature learning for deep belief networks. *Advances in neural information processing systems*, 20, 2007. 1

[6] Aaron Van Den Oord, Oriol Vinyals, et al. Neural discrete representation learning. *Advances in neural information processing systems*, 30, 2017. 1, 3

[7] Salah Rifai, Pascal Vincent, Xavier Muller, Xavier Glorot, and Yoshua Bengio. Contractive auto-encoders: Explicit invariance during feature extraction. In *Proceedings of the 28th international conference on international conference on machine learning*, pages 833–840, 2011. 1

[8] Li Jing, Jure Zbontar, et al. Implicit rank-minimizing autoencoder. *Advances in Neural Information Processing Systems*, 33:14736–14746, 2020. 1, 3

[9] Benjamin Recht, Maryam Fazel, and Pablo A Parrilo. Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization. *SIAM review*, 52(3):471–501, 2010. 2

[10] Sanjeev Arora, Nadav Cohen, Wei Hu, and Yuping Luo. Implicit regularization in deep matrix factorization. *Advances in Neural Information Processing Systems*, 32, 2019. 2, 3

[11] Suriya Gunasekar, Blake E Woodworth, Srinadh Bhojanapalli, Behnam Neyshabur, and Nati Srebro. Implicit regularization in matrix factorization. *Advances in neural information processing systems*, 30, 2017. 2

[12] Suriya Gunasekar, Jason D Lee, Daniel Soudry, and Nati Srebro. Implicit bias of gradient descent on linear convolutional networks. *Advances in neural information processing systems*, 31, 2018. 2

[13] Daniel Soudry, Elad Hoffer, Mor Shpigel Nacson, Suriya Gunasekar, and Nathan Srebro. The implicit bias of gradient descent on separable data. *The Journal of Machine Learning Research*, 19(1):2822–2878, 2018. 2

[14] Andrew M Saxe, James L McClelland, and Surya Ganguli. A mathematical theory of semantic development in deep neural networks. *Proceedings of the National Academy of Sciences*, 116(23):11537–11546, 2019. 2

[15] Gauthier Gidel, Francis Bach, and Simon Lacoste-Julien. Implicit regularization of discrete gradient dynamics in linear neural networks. *Advances in Neural Information Processing Systems*, 32, 2019. 2

[16] Irina Higgins, Loic Matthey, Arka Pal, Christopher Burgess, Xavier Glorot, Matthew Botvinick, Shakir Mohamed, and Alexander Lerchner. beta-vae: Learning basic visual concepts with a constrained variational framework. In *International conference on learning representations*, 2016. 2

[17] Partha Ghosh, Mehdi SM Sajjadi, Antonio Vergari, Michael Black, and Bernhard Schölkopf. From variational to deterministic autoencoders. *arXiv preprint arXiv:1903.12436*, 2019. 3, 6

[18] Ilya Tolstikhin, Olivier Bousquet, Sylvain Gelly, and Bernhard Schoelkopf. Wasserstein auto-encoders. *arXiv preprint arXiv:1711.01558*, 2017. 3, 6

[19] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 3

[20] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017. 4

[21] Gaurav Parmar, Richard Zhang, and Jun-Yan Zhu. On aliased resizing and surprising subtleties in gan evaluation. In *CVPR*, 2022. 4

[22] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014. 6

[23] Xudong Mao, Qing Li, Haoran Xie, Raymond YK Lau, Zhen Wang, and Stephen Paul Smolley. Least squares generative adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pages 2794–2802, 2017. 6

[24] William Fedus, Mihaela Rosca, Balaji Lakshminarayanan, Andrew M Dai, Shakir Mohamed, and Ian Goodfellow. Many paths to equilibrium: Gans do not need to decrease a divergence at every step. *arXiv preprint arXiv:1710.08446*, 2017. 6

[25] Danilo Rezende and Shakir Mohamed. Variational inference with normalizing flows. In *International conference on machine learning*, pages 1530–1538. PMLR, 2015. 6

[26] Mario Lucic, Karol Kurach, Marcin Michalski, Sylvain Gelly, and Olivier Bousquet. Are gans created equal? a large-scale study. *Advances in neural information processing systems*, 31, 2018. 6

[27] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9729–9738, 2020. 6

[28] Ishan Misra and Laurens van der Maaten. Self-supervised learning of pretext-invariant representations. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6707–6717, 2020. 6

[29] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR, 2020. 6

[30] Shib Dasgupta, Michael Boratko, Dongxu Zhang, Luke Vilnis, Xiang Li, and Andrew McCallum. Improving local identifiability in probabilistic box embeddings. *Advances in Neural Information Processing Systems*, 33:182–192, 2020. 8

[31] Kevin Beyer, Jonathan Goldstein, Raghu Ramakrishnan, and Uri Shaft. When is "nearest neighbor" meaningful? In *Database Theory—ICDT'99: 7th International Conference Jerusalem, Israel, January 10–12, 1999 Proceedings 7*, pages 217–235. Springer, 1999. 8