Improving Representation Learning in Autoencoders via Multidimensional Interpolation and Dual Regularizations

Sheng Qian[1*], **Guanyue Li**[2*], **Wen-Ming Cao**[3], **Cheng Liu**[4], **Si Wu**[2†] and **Hau San Wong**[3]
[1]Huawei Device Company Limited
[2]School of Computer Science and Engineering, South China University of Technology
[3]Department of Computer Science, City University of Hong Kong
[4]Department of Computer Science, Shantou University
qiansheng3@huawei.com, csliguanyue007@mail.scut.edu.cn, {wenmincao2-c,
cliu272-c}@my.cityu.edu.hk, cswusi@scut.edu.cn, cshswong@cityu.edu.hk

Abstract

Autoencoders enjoy a remarkable ability to learn data representations. Research on autoencoders shows that the effectiveness of data interpolation can reflect the performance of representation learning. However, existing interpolation methods in autoencoders do not have enough capability of traversing a possible region between datapoints on a data manifold, and the distribution of interpolated latent representations is not considered. To address these issues, we aim to fully exert the potential of data interpolation and further improve representation learning in autoencoders. Specifically, we propose a multidimensional interpolation approach to increase the capability of data interpolation by setting random interpolation coefficients for each dimension of the latent representations. In addition, we regularize autoencoders in both the latent and data spaces, by imposing a prior on the latent representations in the Maximum Mean Discrepancy (MMD) framework and encouraging generated datapoints to be realistic in the Generative Adversarial Network (GAN) framework. Compared to representative models, our proposed approach has empirically shown that representation learning exhibits better performance on downstream tasks on multiple benchmarks.

1 Introduction

Among unsupervised learning frameworks of representation learning, autoencoders (AEs) and variants [Vincent *et al.*, 2010; Kingma and Welling, 2014; Srivastava *et al.*, 2014; Makhzani *et al.*, 2016] have shown a remarkable ability to encode compressed latent representations by reconstructing datapoints. In particular, by decoding the interpolated latent representation of two datapoints, an autoencoder can generate an interpolated datapoint which approximates these two datapoints semantically. This indicates that autoencoders possess an effective capability of data interpolation. Recently, several research works have paid attention to exploring data interpolation methods in autoencoders [Larsen *et al.*, 2016; Berthelot *et al.*, 2019] and generative models [Radford *et al.*, 2015; White, 2016]. The research has demonstrated that there is a close relationship between representation learning and data interpolation. In other words, the effectiveness of data interpolation can reflect the performance of representation learning to some extent.

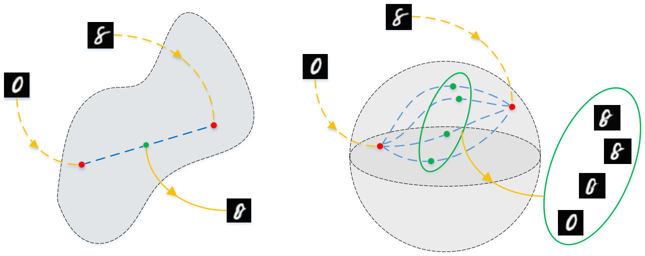


Figure 1: In ACAI (left), only one interpolation path (blue dashed line) between two latent representations (red points) is considered. However, our model (right) can traverse a possible region with many more paths, thus the diversity of interpolated samples will increase. In addition, the latent distribution is constrained to match with a prior (illustrated by a sphere).

Inspired by the relationship between data interpolation and representation learning, there is a possibility to further improve representation learning through explicitly considering data interpolation in autoencoders. The recent work [Berthelot *et al.*, 2019] proposes a generative adversarial regularization strategy referred to as Adversarially Constrained Autoencoder Interpolation (ACAI). ACAI promotes high-quality data interpolation by introducing regularization into the data reconstruction process, and improves representation learning for downstream tasks. However, as empirically shown in [White, 2016], generative models will generate inferior datapoints when performing linear interpolation, which may cause the distribution of interpolated latent representations to diverge from a prior. Besides, linear interpolation does not have enough capability to traverse a possible region between two datapoints on a data manifold, and the distribution of the latent representations is not considered when performing

[*]Equal contribution.

[†]Corresponding author.

interpolation.

To alleviate the above issues, we replace linear interpolation with a more effective interpolation method, and explicitly take the distribution of the latent representations into consideration. Specifically, we propose a new interpolation method referred to as multidimensional interpolation, and fuse dual regularizations in both the latent and data spaces, so that both data interpolation and representation learning in autoencoders can be further improved. Briefly speaking, multidimensional interpolation treats each dimension of the latent representation differently, and sets random interpolation coefficients for each dimension. As a result, it will enhance the quality and diversity of interpolation. Furthermore, we constrain the distribution of the latent representations to match with a prior using the MMD framework [Gretton *et al.*, 2007], so that interpolated representations can follow this prior. Then, datapoints generated by performing interpolation and sampling from this prior are encouraged to be perceptually realistic based on the GAN framework [Goodfellow *et al.*, 2014]. In other words, generated datapoints are expected to follow the real data distribution. In Figure 1, we provide a comparison of interpolations between ACAI and our proposed model. Overall, the main contributions of this work are as follows:

- We improve the ACAI model by extending linear interpolation to multidimensional interpolation, such that the latent space can be better explored and thus the diversity of interpolation can be significantly increased.
- On the other hand, we incorporate MMD-based regularization in the latent space with adversarial regularization in the data space to improve the realism of the interpolated datapoints. Compared with ACAI, the main advantage of this work lies in effective data generation from a prior.
- Due to the aforementioned strategies, representation learning in autoencoders can be enhanced significantly. As a result, better performance can be achieved on downstream tasks.

2 Related Work

In this section, we briefly introduce some related works on autoencoder variants and data interpolation.

2.1 Autoencoders

Based on regularizations introduced into autoencoders, these variants are generally divided into three categories: Normal-based, VAE-based and GAN-based (VAE refers to as variational AE [Kingma and Welling, 2014]). Specifically, in the Normal-based case, dropout AE (DoAE) [Srivastava *et al.*, 2014] regularizes the latent representation with dropout which randomly drops units in hidden layers. Denoising AE (DnAE) [Vincent *et al.*, 2010] reconstructs clean datapoints from their corrupted versions and can learn the data manifold implicitly. Sparse AE (SAE) [Xu *et al.*, 2016] adds a sparsity term into the objective function to learn high-level features. These three variants explicitly make data reconstruction more difficult by introducing some perturbations and can learn more robust representations.

Different from the above approaches, AEs are regularized to explicitly match the distribution of the latent representations with a predefined prior in the VAE-based and the GAN-based cases. In the VAE-based case, VAE introduces an additional loss into the reconstruction loss. This loss measures the KL divergence between the distribution of the latent representations and the prior. Besides the KL divergence, MMD is also used to measure the distribution divergence in [Tolstikhin *et al.*, 2018]. Along with learning continuous representations, vector-quantized VAE (VQVAE) [van den Oord *et al.*, 2017] applies the distributional regularization framework into learning discrete representations using a learned quantized codebook.

In the GAN-based case, adversarial AE (AAE) [Makhzani *et al.*, 2016] introduces an auxiliary subnetwork referred to as a discriminator based on the GAN framework, and forces this discriminator to determine whether latent representations are inferred from the encoder or sampled from the prior. Moreover, VAE-GAN [Larsen *et al.*, 2016] fuses both the VAE and GAN frameworks in the autoencoder. In VAE-GAN, the traditional pixel-wise reconstruction loss is replaced by an adversarial feature-wise reconstruction loss obtained from the GAN's discriminator.

2.2 Interpolation

In the latent space there are two basic types of interpolations: linear interpolation and spherical interpolation. Linear interpolation combines latent representations linearly, which is easily understood. Hence, it is frequently used to inspect the performance of representation learning [Larsen *et al.*, 2016; Radford *et al.*, 2015; Wu *et al.*, 2016; Dumoulin *et al.*, 2017]. When the latent space is high dimensional, interpolated representations on the linear interpolation path always suffer from change of magnitude. To address this issue, [White, 2016] introduces the spherical interpolation by applying nonlinear interpolation based on a great circle path on an n-dimensional hypersphere. Besides, [Laine, 2018] proposes to perform interpolation along geodesics in the data space rather than in the latent space. [Agustsson *et al.*, 2019] designs several interpolation operations to reduce mismatch between the distribution of interpolated data and a prior.

3 Background of ACAI

ACAI model aims to improve data interpolation and representation learning in autoencoders. Specifically, the network architecture of ACAI consists of an autoencoder and a discriminator. The former, in addition to serving as a basic autoencoder, also forms a GAN with the discriminator to propel the interpolated data to perceptually approximate real data to as realistic an extent as possible.

Formally, the encoder and the decoder are denoted as $enc(\cdot)$ and $dec(\cdot)$ in the autoencoder, respectively; the discriminator is denoted as $dis(\cdot)$. For two datapoints $\mathbf{x}_1$ and $\mathbf{x}_2$, $\mathbf{z}_1 = enc(\mathbf{x}_1)$ and $\mathbf{z}_2 = enc(\mathbf{x}_2)$ are their latent representations, respectively. Then, linear interpolation synthesizes a mixture representation $\mathbf{z}_\lambda$:

$$\mathbf{z}_\lambda = \lambda \cdot \mathbf{z}_1 + (1 - \lambda) \cdot \mathbf{z}_2, \tag{1}$$

where λ, constrained to the range $[0, 0.5]$, is an interpolation coefficient of two latent representations. By decoding $\mathbf{z}_\lambda$, an interpolated datapoint $\mathbf{x}_\lambda = dec(\mathbf{z}_\lambda)$ is generated. From another perspective, the degree of realism of $\mathbf{x}_\lambda$ can be controlled by adjusting λ. The larger λ is, the less realistic $\mathbf{x}_\lambda$ becomes.

By explicitly regularizing the interpolation process, the discriminator is trained to distinguish between real datapoints and interpolated datapoints. It predicts the interpolation coefficients for interpolated datapoints, and consistently outputs 0 for real datapoints. Similar to GAN, the discriminator's loss function is given by

$$L_{dis} = \|dis(\mathbf{x}_\lambda) - \lambda\|^2 + \|dis(\mathbf{x})\|^2. \tag{2}$$

Meanwhile, the autoencoder's loss function is modified by adding a regularization term as follows:

$$L_{ae} = \|\mathbf{x} - dec(enc(\mathbf{x}))\|^2 + \|dis(\mathbf{x}_\lambda)\|^2. \tag{3}$$

ACAI has verified that inferred latent representations show more effectiveness on downstream tasks, which indicates that there is a possible link between good data interpolation and useful latent representation learning.

4 Proposed Model

By gradually changing the interpolation coefficient in ACAI, an interpolation path consisting of a set of interpolated datapoints will be formed. This interpolation path can be viewed as a direct line in a linear space. However, linear interpolation is not suitable in a high-dimensional space, and cannot well traverse a possible region between two datapoints. Besides, no explicit constraints have been imposed on the distribution of the latent representations. After interpolation, it leads to a situation where some interpolated representations and corresponding datapoints do not belong to the representation domain and the data domain, respectively. Overall, the capability of interpolation in improving representation learning has not been fully exploited.

To alleviate the above shortcomings, we propose multidimensional interpolation to enhance the capability of interpolation. In addition, we fuse dual regularizations in both the latent and data spaces to improve representation learning. More details will be described in the following subsections.

4.1 Multidimensional Interpolation

Although linear interpolation is straightforward and effective in ACAI, there exist three shortcomings: (1) only a single interpolation path is considered; (2) the interpolation path is confined to a direct line; (3) all dimensions of the latent representation are equally treated.

Hence, to enhance the capability of interpolation, we propose a new interpolation method called multidimensional interpolation. Formally, assuming that latent representations are d-dimension vectors, multidimensional interpolation aims to synthesize a mixture representation as follows:

$$\mathbf{z}_\mu = \boldsymbol{\mu} \odot \mathbf{z}_1 + (1 - \boldsymbol{\mu}) \odot \mathbf{z}_2, \tag{4}$$

where $\boldsymbol{\mu} = [\mu_1, \mu_2, ..., \mu_d]$ is an interpolation coefficient, and $\odot$ denotes the element-wise product between two vectors. What needs to be emphasized is that $\boldsymbol{\mu}$ is a vector with the same number of dimensions as the latent representations, and the value of each component of $\boldsymbol{\mu}$ is randomly sampled from the uniform distribution on $[0, 0.5]$.

Compared with the scalar coefficient λ in linear interpolation, each component of $\boldsymbol{\mu}$ is randomly set, so that each dimension of the latent representations will not be equally treated when performing interpolation. As a result, by varying the interpolation coefficient, an interpolation region will be formed between two latent representations, and the corresponding path can be any curve within this region. From the perspective of manifold learning, the interpolation possesses a capability of traversing a possible region between two datapoints on a data manifold. Overall, multidimensional interpolation can overcome the abovementioned shortcomings and enhance both the quality and diversity of interpolation paths.

4.2 Dual Regularizations

Regularization in Latent Space

When performing data interpolation, the interpolated datapoints are expected to be perceptually realistic. Take natural images as an example, although interpolated datapoints seem realistic, they may not belong to the original image domain in terms of structure and appearance. From the perspective of distribution matching, the corresponding interpolated representations may not well follow the real distribution of the latent representations.

To alleviate the negative effect on representation learning due to the above issue, we explicitly regularize the distribution of the latent representations to follow a prior Gaussian distribution. Then, the corresponding interpolated representations can be expected to follow this prior. Specifically, we propose to minimize the MMD distance between the distribution of the latent representations and this prior.

Distributional Regularization. In practice, the MMD distance can be estimated using a finite number of samples based on the kernel trick, and the square of MMD distance can be approximated by a two-sample test. Given $\mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_n\} \sim p_\mathbf{x}(\mathbf{x})$ and $\mathbf{Z} = \{\mathbf{z}_1, \mathbf{z}_2, ..., \mathbf{z}_n\} \sim p_\mathbf{z}(\mathbf{z})$, where $p_\mathbf{x}(\mathbf{x})$ and $p_\mathbf{z}(\mathbf{z})$ are the data distribution and the prior of latent space respectively, a two-sample test is performed to determine whether the latent representations are inferred from the encoder or sampled from the prior, and the square of MMD distance L_{mmd} is defined as

$$\begin{aligned} L_{mmd} =& \mathbb{E}[k(enc(\mathbf{x}_i), enc(\mathbf{x}_{i'}))] + \mathbb{E}[k(\mathbf{z}_j, \mathbf{z}_{j'})] - \\ & 2\mathbb{E}[k(enc(\mathbf{x}_i), \mathbf{z}_j)], \end{aligned} \tag{5}$$

where $k(\cdot, \cdot)$ is a kernel. By minimizing this distance, the distribution of interpolated latent representations can match with the prior.

Regularization in Data Space

Data Generation. After performing interpolation and imposing regularization on the distribution of the latent representations, three types of data can be generated. First, as with ACAI, by performing interpolation and decoding the mixture representation $\mathbf{z}_\mu$ inferred from two datapoints, an interpolated datapoint $\mathbf{x}_\mu$ is generated as follows:

$$\mathbf{x}_\mu = dec(\mathbf{z}_\mu). \tag{6}$$

Second, an interpolated datapoint can be generated by decoding a mixture representation between a latent representation inferred from a datapoint and a stochastic latent representation sampled from the prior as follows:

$$\mathbf{x}_{\mu z} = dec(\boldsymbol{\mu} \odot enc(\mathbf{x}) + (1 - \boldsymbol{\mu}) \odot \mathbf{z}). \tag{7}$$

Third, similar to VAE, when the distribution of the latent representations is expected to match the prior, a datapoint can be generated by directly decoding a stochastic latent representation sampled from the prior as follows:

$$\mathbf{x}_z = dec(\mathbf{z}). \tag{8}$$

Besides the basic $\mathbf{x}_\mu$, we expect that with the introduction of two datapoints $\mathbf{x}_{\mu z}$ and $\mathbf{x}_z$, data interpolation and representation learning can be further improved.

Adversarial Regularization. The above three types of data are encouraged to be perceptually realistic and cannot be distinguished from real data. Thus, we also adopt adversarial regularization in ACAI. Owing to the introduction of two additional types of datapoints, the original loss functions in ACAI need to be modified to generalize to this case.

Specifically, for the datapoints $\mathbf{x}_\mu$ and $\mathbf{x}_{\mu z}$, the discriminator predicts the interpolation coefficients $\boldsymbol{\mu}$ for interpolated datapoints. Since the datapoint $\mathbf{x}_z$ is directly generated based on the prior, it would be less realistic than $\mathbf{x}_\mu$ and $\mathbf{x}_{\mu z}$, and does not effectively approximate real data. As mentioned in Section 3, the interpolation coefficient can reflect the realism of the interpolated datapoint. Therefore, the discriminator predicts a predefined maximum of the interpolation coefficient for $\mathbf{x}_z$. This predefined maximum is set as 0.5 here.

From another perspective, we incorporate data generation into the interpolation process, and formulate them in a unified framework. To sum up, the overall losses of the discriminator and the autoencoder are modified as follows:

$$\begin{aligned} L_{dis} =& \|dis(\mathbf{x})\|^2 + \|dis(\mathbf{x}_\mu) - \boldsymbol{\mu}\|^2 + \\ & \|dis(\mathbf{x}_{\mu z}) - \boldsymbol{\mu}\|^2 + \|dis(\mathbf{x}_z)) - 0.5\|^2, \end{aligned} \tag{9}$$

$$\begin{aligned} L_{ae} =& \|\mathbf{x} - dec(enc(\mathbf{x}))\|^2 + \omega_1 L_{mmd} + \omega_2 \|dis(\mathbf{x}_\mu)\|^2 + \\ & \omega_3 \|dis(\mathbf{x}_{\mu z})\|^2 + \omega_4 \|dis(\mathbf{x}_z)\|^2, \end{aligned} \tag{10}$$

where ω_1, ω_2, ω_3 and ω_4 are hyper-parameters for adjusting the weights of the above losses. We set ω_1 and ω_2 as 1.0 and 0.5 for all experiments, respectively. The remaining parameters ω_3 and ω_4 are tuned to be the best by experience.

5 Experiments

5.1 Experimental Setting

In our experiments, we evaluate our proposed models on the following datasets: MNIST [LeCun *et al.*, 1998], SVHN [Netzer *et al.*, 2011], CIFAR-10 [Krizhevsky and Hinton, 2009] and CelebA [Liu *et al.*, 2015]. For a fair comparison, we use the network architecture from ACAI for all discussed models.

To study the effect of multidimensional interpolation, we only replace the linear interpolation step of ACAI with the proposed interpolation operation. In this case, this model is referred to as Autoencoder with Multidimensional Interpolation (MI-AE). To study the effect of regularization in the latent space, we only add distributional regularization into ACAI. In this case, this model is referred to as Autoencoder with Latent Regularization (LR-AE). Furthermore, we consider both multidimensional interpolation and dual regularizations to verify their effectiveness. In this case, this model is referred to as Autoencoder with Multidimensional Interpolation and Dual Regularizations (MIDR-AE).

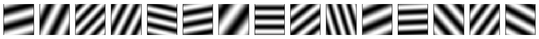
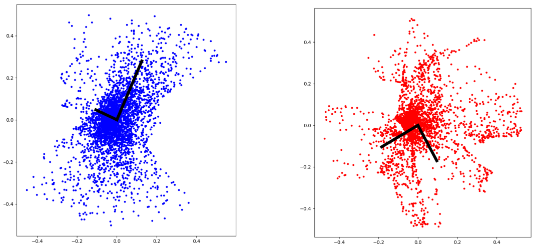


Figure 2: Examples from the synthetic dataset.



(a) ACAI (b) MI-AE

Figure 3: Distributions of the latent representations from ACAI (left) and MI-AE (right). The principal components of the latent representations are illustrated by black solid lines.

5.2 Interpolations on Synthetic Data

In this subsection, we investigate the difference between linear and multidimensional interpolations on a synthetic dataset. Specifically, this dataset consists of 10k gray images with 32*32 resolution. An image is generated using a sinusoidal function, which looks like parallel white stripes with a black background. The value of the pixel (x, y) in an image is calculated as follows:

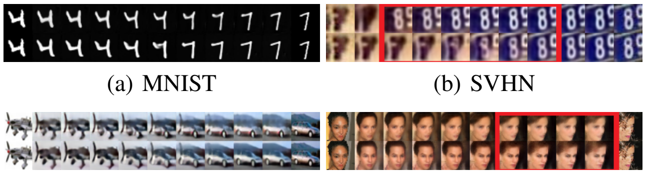
$$g(x, y) = \sin(\alpha(\Delta y \cos\beta - \Delta x \sin\beta)), \tag{11}$$

where $\alpha \in [0.3, 0.7]$ controls the number of stripes, and $\beta \in [0, 2\pi]$ determines the angle of stripes. $\Delta x = x - x_c$ and $\Delta y = y - y_c$ denote the relative positions between the pixel (x, y) and the central pixel (x_c, y_c). We set $\Delta\alpha = 0.004$ and $\Delta\beta = 0.02\pi$, and randomly choose α and β to generate the whole dataset. Some samples are shown in Figure 2.

We use the synthesized data to train ACAI and MI-AE, and compare the difference between the distributions of learned latent representations. For better visualization, we set the number of dimensions of the latent representations as $d_z = 2$ in the experiments.

Model d_z	MNIST 32/256	SVHN 32/256	CIFAR-10 256/1024	CelebA 32/256
ACAI	0.24 / 0.22	0.27 / 0.13	0.08 / 0.06	0.18 / 0.16
MI-AE	0.14 / 0.12	0.11 / 0.07	0.04 / 0.03	0.04 / 0.06

Table 1: The mean of the absolute off-diagonal values in the normalized covariance matrix for different datasets.



(a) MNIST (b) SVHN

(c) CIFAR-10 (d) CelebA

Figure 4: Interpolated samples on real datasets based on linear interpolation in ACAI (the first row) and multidimensional interpolation in MI-AE (the second row). The major difference between them can be viewed from images encapsulated by red rectangles (subfigures (b) and (d)) with zooming in for a better view.

In Figure 3, both ACAI and MI-AE gather most of the latent representations close to the origin. We display two principal components of the learned latent representations for ACAI and MI-AE. From Figure 3, we can see that the difference between two components from ACAI is larger than that from MI-AE. We calculate the mean of the absolute off-diagonal values in the normalized covariance matrix [Lee Rodgers and Nicewander, 1988] to quantify the correlation between dimensions. The correlation coefficient values between dimensions are 0.523 and 0.159 for ACAI and MI-AE, respectively, which indicates that the correlation between dimensions of the latent representations from MI-AE is much less than that from ACAI. This indicates that autoencoders with multidimensional interpolation learn latent representations with lower correlation.

In addition, we perform experiments on real datasets and measure the resulting correlation between dimensions of the learned latent representations. As shown in Table 1, the values for ACAI are much larger than those for MI-AE on all datasets. Overall, using multidimensional interpolation helps autoencoders to disentangle the correlation between dimensions of the latent representations.

5.3 Interpolations on Real Data

For a fair comparison with ACAI, we adopt the proposed MI-AE and set $\boldsymbol{\mu} = [\lambda, \lambda, ..., \lambda]$, and further compare the performance of their interpolations on real data. The interpolated samples are shown in Figure 4.

From Figure 4, on MNIST and CIFAR-10, there is no obvious perceptual difference between interpolated samples. However, by zooming on SVHN and CelebA, more details can be observed. On SVHN, samples generated by ACAI are easily dominated by one of the real datapoints (highlighted with a red rectangle at the middle part), while our proposed MI-AE can generate more reasonable and balanced interpolated samples. On CelebA, samples generated by MI-AE at the right part are much clearer than those by ACAI (highlighted with a red rectangle at the right part), and the colors of lips generated by MI-AE can still be easily recognized. This indicates that our proposed MI-AE can learn latent representations while preserving more detailed patterns. Overall, the proposed multidimensional interpolation process shows a better performance than the linear interpolation in ACAI.

5.4 Improved Representation Learning

In this subsection, we examine whether multidimensional interpolation and dual regularizations are more beneficial to representation learning. In particular, we perform experiments on image classification and clustering based on inferred latent representations. By evaluating the performance of these downstream tasks and comparing our models with other models, the advantage of our models can be verified.

Here, we consider several representative models in Section 2 as follows: basic AE (BAE), DoAE, DnAE, VAE, AAE, VQVAE and ACAI. In addition, the optimal hyperparameters of our models are as follows: $\omega_3 = 0.5$ and $\omega_4 = 0.1$ for MNIST; $\omega_3 = 0.5$ and $\omega_4 = 0.05$ for SVHN; $\omega_3 = 0.01$ and $\omega_4 = 0.01$ for CIFAR-10, and $\omega_3 = 0.1$ and $\omega_4 = 0.05$ for CelebA.

Image Classification

In order to measure the quality of representation learning in autoencoders, a supervised classification task is also conducted based on the inferred latent representations. If the inferred representations have distilled the important class information, a classifier can achieve a reasonable performance regardless of its simplicity. Therefore, as with ACAI, we train the additional single-layer classifier [Coates *et al.*, 2011] by feeding inferred representations of the encoder along with the whole model. We summarize the classification results in Table 2.

From Table 2, MI-AE and LR-AE show a better performance than other models on all datasets. Especially on SVHN (d_z=32) and CIFAR-10 (d_z=256/1024), MI-AE and LR-AE obviously outperform the second best ACAI. Moreover, MIDR-AE further boosts the performance of image classification compared to MI-AE, and achieves the best performance. It is worth noting that ACAI only achieves an average accuracy of 34.47%, while MIDR-AE achieves 92.28% on SVHN when the number of dimensions of the latent representations is set as 32. These experimental results show that our autoencoder variants can distill more class information.

Viewed from the perspective of manifold learning, multidimensional interpolation is more beneficial to traversing a data manifold. By regularizing the distribution of the latent representations, the interpolated representations can also follow a prior. Therefore, both of them will improve representation learning.

Image Clustering

Since clustering groups datapoints with similar attributes together and separates datapoints with different attributes apart in an unsupervised way, it is another challenging task for evaluating the performance of representation learning. Therefore, if the autoencoder has uncovered the important class characteristics of the data, then performing clustering on the latent representations should yield reasonable results.

As with ACAI, we use PCA whitening on the latent representations, and randomly perform K-Means [MacQueen, 1967] 1k times to choose the clustering solution with the best objective value during training. For evaluation, we adopt the methodology of [Xie *et al.*, 2016; Hu *et al.*, 2017] to evaluate

Dataset	d_z	BAE	DoAE	DnAE	VAE	AAE	VQVAE	ACAI	MI-AE	LR-AE	MIDR-AE
MNIST	32	94.90±0.14	96.45±0.42	96.00±0.27	96.56±0.31	70.74±3.27	97.50±0.18	98.25±0.11	98.32±0.36	98.09±0.13	**99.17±0.03**
	256	93.94±0.13	94.50±0.29	98.51±0.04	98.74±0.14	90.03±0.54	97.25±1.42	99.00±0.08	99.07±0.10	99.10±0.02	**99.19±0.04**
SVHN	32	26.21±0.42	26.09±1.48	25.15±0.78	29.58±3.22	23.43±0.79	24.53±1.33	34.47±1.14	46.42±1.61	43.81±1.10	**92.28±0.62**
	256	22.74±0.05	25.12±1.05	77.89± 0.35	66.30±1.06	22.81±0.24	44.94±20.42	85.14±0.20	85.70±0.39	88.38±0.14	**94.30±0.18**
CIFAR-10	256	47.92±0.20	40.99±0.41	53.78±0.36	47.49±0.22	40.65±1.45	42.80±0.44	52.77±0.45	55.93±0.04	58.46±0.40	**63.28±0.13**
	1024	51.62±0.25	49.38±0.77	60.65± 0.14	51.39±0.46	42.86±0.88	16.22±12.44	63.99±0.47	67.56±0.03	66.73±0.03	**68.27±0.48**

Table 2: Classification accuracy using single-layer classifiers. Except for our models, the results of other models are excerpted from ACAI.

Dataset	d_z	BAE	DoAE	DnAE	VAE	AAE	VQVAE	ACAI	MI-AE	LR-AE	MIDR-AE	KMeansData	DEC	IMSAT
MNIST	32	77.56	82.67	82.59	75.74	79.19	82.39	94.38	95.66	94.37	**96.94**	53.2	84.3	98.4
	256	53.70	61.35	70.89	83.44	81.00	96.80	96.17	96.38	96.70	**97.05**			
SVHN	32	19.38	21.42	17.91	16.83	17.35	15.19	20.86	36.91	29.45	**56.64**	17.9	11.9	57.3
	256	15.62	15.19	31.49	11.36	13.59	18.84	24.98	38.28	34.88	**51.14**			

Table 3: Clustering accuracy for using K-Means. Except for our models, the results of other models are excerpted from ACAI.

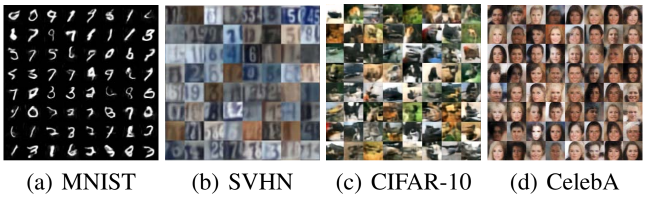
the "clustering accuracy" on the test dataset. We summarize the clustering results in Table 3.

From Table 3, MI-AE and LR-AE can produce reasonable performance gains than other models on both MNIST and SVHN. In particular, on SVHN, MI-AE and LR-AE significantly outperform ACAI. Compared to MI-AE and LR-AE, MIDR-AE further boosts the performance of image clustering and achieves the best performance. MIDR-AE is always superior to ACAI with a large performance gain on both datasets. Especially on SVHN, MIDR-AE almost has a clustering accuracy improvement of 20% above ACAI when the number of dimensions of the latent representations is set as 32 or 256. Similar to image classification, we can conclude that multidimensional interpolation and dual regularizations enhance representation learning.

Besides, we evaluate the clustering performance of three other models: KMeansData, DEC [Xie *et al.*, 2016] and IMSAT [Hu *et al.*, 2017]. These models are different from the autoencoder variants. Specifically, KMeansData refers to performing K-Means directly on the data; DEC simultaneously learns feature representations and cluster assignments using deep neural networks; and IMSAT learns invariant representations for clustering by adopting data augmentation and mutual information maximization [Deng *et al.*, 2009]. Compared to DEC, all our models are superior, while IMSAT just slightly outperforms MIDR-AE on two datasets. This demonstrates that our proposed models show a competitive performance on representation learning.

5.5 Data Generation

Since the distribution of the latent representations is enforced to match with a prior, one main advantage of our model is its capability of generating data by sampling from this prior. Here, we show some samples generated by our proposed model in Figure 5. Compared to our model, ACAI does not possess this capability since it does not consider the distribution of the latent representations. This indicates that our model can learn semantically meaningful latent representations which well reveal the data distribution.



(a) MNIST (b) SVHN (c) CIFAR-10 (d) CelebA

Figure 5: Samples generated by MIDR-AE on multiple datasets.

6 Conclusion

In this work, we have comprehensively studied both data interpolation and representation learning in autoencoders. We observe that existing interpolation methods in autoencoders do not have enough capability of traversing a data manifold, and the distribution of the latent representations is not considered when performing interpolation. To improve both data interpolation and representation learning, we propose a model referred to as MIDR-AE. In MIDR-AE, on one hand, a multidimensional interpolation process is proposed to enhance the capability of data interpolation. On the other hand, autoencoders are further regularized in both the latent and data spaces, which imposes a prior on the latent representations and encourages datapoints generated by performing interpolation and sampling from this prior to be realistic. Through extensive experiments, we verify that representation learning of our model exhibits better performance on downstream tasks compared to competing models. For the purpose of reproduction and extensions, our code is publicly available. [1]

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China (Project No. 61502173, U1611461), in part by the Research Grants Council of the Hong Kong Special Administration Region (Project No. CityU 11300715), in part by City University of Hong Kong (Project No. 7005055), in part by the Natural

[1] https://github.com/guanyuelee/midrae

Science Foundation of Guangdong Province (Project No. 2016A030310422), and in part by Fundamental Research Funds for the Central Universities (Project No. 2018ZD33).

References

[Agustsson *et al.*, 2019] Eirikur Agustsson, Alexander Sage, Radu Timofte, and Luc Van Gool. Optimal transport maps for distribution preserving operations on latent spaces of generative models. In *International Conference on Learning Representations*, 2019.

[Berthelot *et al.*, 2019] David Berthelot, Colin Raffel, Aurko Roy, and Ian Goodfellow. Understanding and improving interpolation in autoencoders via an adversarial regularizer. In *International Conference on Learning Representations*, 2019.

[Coates *et al.*, 2011] Adam Coates, Andrew Ng, and Honglak Lee. An analysis of single-layer networks in unsupervised feature learning. In *International Conference on Artificial Intelligence and Statistics*, pages 215–223, 2011.

[Deng *et al.*, 2009] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Fei-Fei Li. Imagenet: A large-scale hierarchical image database. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, 2009.

[Dumoulin *et al.*, 2017] Vincent Dumoulin, Ishmael Belghazi, Ben Poole, Olivier Mastropietro, Alex Lamb, Martin Arjovsky, and Aaron Courville. Adversarially learned inference. In *International Conference on Learning Representations*, 2017.

[Goodfellow *et al.*, 2014] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Neural Information Processing Systems*, pages 2672–2680, 2014.

[Gretton *et al.*, 2007] Arthur Gretton, Karsten Borgwardt, Malte Rasch, Bernhard Schölkopf, and Alex J Smola. A kernel method for the two-sample-problem. In *Neural Information Processing Systems*, pages 513–520, 2007.

[Hu *et al.*, 2017] Weihua Hu, Takeru Miyato, Seiya Tokui, Eiichi Matsumoto, and Masashi Sugiyama. Learning discrete representations via information maximizing self-augmented training. In *International Conference on Machine Learning*, pages 1558–1567, 2017.

[Kingma and Welling, 2014] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. In *International Conference on Learning Representations*, 2014.

[Krizhevsky and Hinton, 2009] Alex Krizhevsky and Geoffrey Hinton. Learning multiple layers of features from tiny images. Technical report, 2009.

[Laine, 2018] Samuli Laine. Feature-based metrics for exploring the latent space of generative models. In *International Conference on Learning Representations*, 2018.

[Larsen *et al.*, 2016] Anders Boesen Lindbo Larsen, Søren Kaae Sønderby, Hugo Larochelle, and Ole Winther. Autoencoding beyond pixels using a learned similarity metric. In *International Conference on Machine Learning*, pages 1558–1566, 2016.

[LeCun *et al.*, 1998] Yann LeCun, Léon Bottou, Yoshua Bengio, Patrick Haffner, et al. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

[Lee Rodgers and Nicewander, 1988] Joseph Lee Rodgers and W Alan Nicewander. Thirteen ways to look at the correlation coefficient. *The American Statistician*, 42(1):59–66, 1988.

[Liu *et al.*, 2015] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *IEEE International Conference on Computer Vision*, pages 3730–3738, 2015.

[MacQueen, 1967] James MacQueen. Some methods for classification and analysis of multivariate observations. In *Berkeley Symposium on Mathematical Statistics and Probability*, pages 281–297, 1967.

[Makhzani *et al.*, 2016] Alireza Makhzani, Jonathon Shlens, Navdeep Jaitly, Ian Goodfellow, and Brendan Frey. Adversarial autoencoders. In *International Conference on Learning Representations*, 2016.

[Netzer *et al.*, 2011] Yuval Netzer, Tao Wang, Adam Coates, Alessandro Bissacco, Bo Wu, and Andrew Y Ng. Reading digits in natural images with unsupervised feature learning. In *Neural Information Processing Systems*, 2011.

[Radford *et al.*, 2015] Alec Radford, Luke Metz, and Soumith Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. In *International Conference on Learning Representations*, 2015.

[Srivastava *et al.*, 2014] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: a simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, 15(1):1929–1958, 2014.

[Tolstikhin *et al.*, 2018] Ilya Tolstikhin, Olivier Bousquet, Sylvain Gelly, and Bernhard Schoelkopf. Wasserstein auto-encoders. In *International Conference on Learning Representations*, 2018.

[van den Oord *et al.*, 2017] Aaron van den Oord, Oriol Vinyals, et al. Neural discrete representation learning. In *Neural Information Processing Systems*, pages 6306–6315, 2017.

[Vincent *et al.*, 2010] Pascal Vincent, Hugo Larochelle, Isabelle Lajoie, Yoshua Bengio, and Pierre-Antoine Manzagol. Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion. *Journal of Machine Learning Research*, 11(12):3371–3408, 2010.

[White, 2016] Tom White. Sampling generative networks. *arXiv preprint arXiv:1609.04468*, 2016.

[Wu *et al.*, 2016] Jiajun Wu, Chengkai Zhang, Tianfan Xue, Bill Freeman, and Josh Tenenbaum. Learning a probabilistic latent space of object shapes via 3d generative-adversarial modeling. In *Neural Information Processing Systems*, pages 82–90, 2016.

[Xie *et al.*, 2016] Junyuan Xie, Ross Girshick, and Ali Farhadi. Unsupervised deep embedding for clustering analysis. In *International Conference on Machine Learning*, pages 478–487, 2016.

[Xu *et al.*, 2016] Jun Xu, Lei Xiang, Qingshan Liu, Hannah Gilmore, Jianzhong Wu, Jinghai Tang, and Anant Madabhushi. Stacked sparse autoencoder (ssae) for nuclei detection on breast cancer histopathology images. *IEEE Transactions on Medical Imaging*, 35(1):119–130, 2016.