Shape your Space: A Gaussian Mixture Regularization Approach to Deterministic Autoencoders

Amrutha Saseendran[1], Kathrin Skubch[1], Stefan Falkner[1] and Margret Keuper[2,3]
[1]Bosch Center for Artificial Intelligence,
[2]University of Siegen, [3]Max Planck Institute for Informatics, Saarland Informatics Campus

Motivation

- The variational formulation in VAEs poses considerable practical challenges.
- The over simplistic assumption of a unimodal Gaussian prior in VAEs lead to an unsatisfying trade-off between the quality of reconstructed samples and the prior regularization.
- VAEs trained with more expressive priors, like multimodal Gaussian mixture models (GMMs), improve in terms of generative performance, but often come with increased computational complexity and training instability.
- Recent work in deterministic autoencoders[1] offers a promising alternative to VAEs, but requires an additional ex-post density estimation for high quality sampling.
- ***We propose a generative model that elegantly combines novel training objectives for deterministic autoencoders with the extension to multi-modal priors without increasing training complexity or compromising sampling quality.***

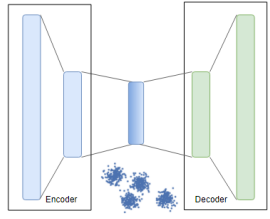


Figure: Proposed deterministic autoencoder trained with multi-modal GMM prior

[1] *Ghosh, Partha et al. "From Variational to Deterministic Autoencoders." ICLR2020*

Regularized Autoencoders (RAEs)

- RAEs[1] reinterpret deterministic autoencoders as variational models.
- The model is trained with a regularization loss that maximizes the negative log-likelihood of the latent samples under a unimodal Gaussian prior,

$$\mathcal{L}_{\text{RAE}} = \underbrace{\mathcal{L}_{\text{REC}}}_{\text{reconstrcution}} + \underbrace{\beta \mathcal{L}_{\mathbf{Z}}^{\text{RAE}} + \lambda \mathcal{L}_{\text{REG}}}_{\text{regularization}}$$

- For high quality sampling, the model requires an ex-post density estimation with a multi-modal GMM.
- Sampling quality can suffer significantly if the learned latent space can not be modeled well by a GMM.

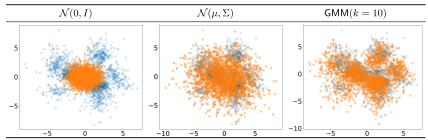


Figure: Aggregated posterior mismatch in VAEs - different density estimations of the latent space of a VAE learned on MNIST. The figure shows 2000 test set samples and the estimators; isotropic Gaussian (left), multivariate Gaussian (center) and a 10-component GMM (right) *(Ghosh,P et al., 2020)*

[1] *Ghosh, Partha et al. "From Variational to Deterministic Autoencoders." ICLR2020*

Proposed Latent Regularization

- We propose to shape the latent space during training to enable high quality sampling without employing additional density estimation.
- Our proposed regularization scheme can be extended readily from unimodal priors from the standard VAE formulation to expressive multimodal priors.
- Our training objective is inspired by the non-parametric statistical Kolmogorov-Smirnov (KS) test used to determine the equality of one-dimensional probability distributions.
- The KS test compares the cumulative distribution function (CDF) of the reference distribution with empirical CDF of the samples.
- Extension of KS distance to higher dimensions is challenging since it requires matching joint CDFs.
- ***To overcome this, we consider marginal CDFs and correlations in the target prior distribution separately.***

Uni-modal Latent Regularization

- In the unimodal case we consider the standard VAE prior, i.e. a multivariate Gaussian.
- Given d-dimensional latent samples $\mathbf{z}_1, \ldots, \mathbf{z}_N$, the empirical marginal CDFs $\tilde{F}$ is matched with the 1D CDFs of the marginal distributions of the uni-modal Gaussian prior Φ,

$$\mathcal{L}_{\text{KS}}(\mathbf{z}_1, \ldots, \mathbf{z}_N) = \frac{1}{d} \sum_{j=1}^{d} \text{MSE}\left(\tilde{F}_j^{(N)}(\mathbf{z}_j), \Phi(\bar{\mathbf{z}}_j)\right).$$

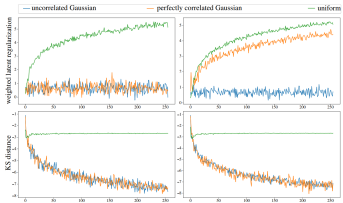


Figure: Uni-modal latent regularization in one and two dimensions for varying numbers of samples (x-axis) from different distributions. In two dimensions (right), the KS distance alone can not differentiate the target prior from other probability distributions.

Uni-modal Latent Regularization

- In the unimodal case we consider the standard VAE prior, i.e. a multivariate Gaussian.
- Given d-dimensional latent samples $\mathbf{z}_1, \ldots, \mathbf{z}_N$, the empirical marginal CDFs $\tilde{F}$ is matched with the 1D CDFs of the marginal distributions of the uni-modal Gaussian prior Φ,

$$\mathcal{L}_{\text{KS}}(\mathbf{z}_1, \ldots, \mathbf{z}_N) = \frac{1}{d} \sum_{j=1}^{d} \text{MSE}\left(\tilde{F}_j^{(N)}(\mathbf{z}_j), \Phi(\tilde{\mathbf{z}}_j)\right).$$

- The empirical covariance $\tilde{\Sigma}$ is explicitly matched with the target covariance Σ,

$$\mathcal{L}_{\text{CV}}(\mathbf{z}_1, \ldots, \mathbf{z}_N) = \frac{1}{d^2} \sum_{l,j=1}^{d} \left([\tilde{\Sigma}]_{l,j} - [\Sigma]_{l,j}\right)^2.$$

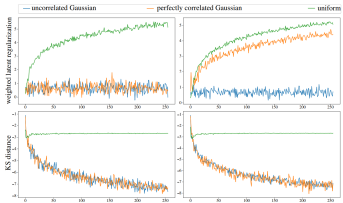


Figure: Uni-modal latent regularization in one and two dimensions for varying numbers of samples (x-axis) from different distributions. In two dimensions (right), the KS distance alone can not differentiate the target prior from other probability distributions.

Multi-modal Latent Regularization

- Encouraging a multi-modal latent representation enables effective modelling of complex input spaces.
- Our regularization scheme can be readily applied to expressive multi-modal prior distributions (GMM) since linear combination of Gaussians allows for closed form computations of CDFs and covariances.
- The total loss of the proposed model is the weighted combination of a simple reconstruction loss and the latent regularization,

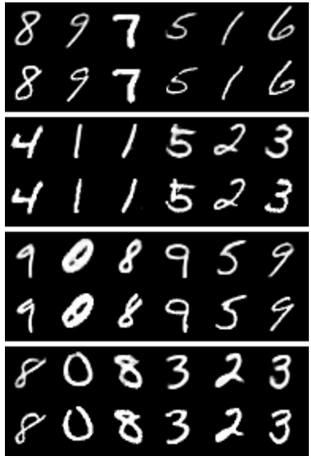
$$\mathcal{L}(\mathbf{x}) = \underbrace{\lambda_{\text{REC}}\mathcal{L}_{\text{REC}}(\mathbf{x}')}_{\text{mean squared error}} + \underbrace{\lambda_{\text{KS}}\mathcal{L}_{\text{KS}}(\mathbf{z}) + \lambda_{\text{CV}}\mathcal{L}_{\text{CV}}(\mathbf{z})}_{\text{multi-modal latent regularization}},$$

  where $\mathbf{x}'$ are reconstructions of samples $\mathbf{x}$ and $\mathbf{z}$ their latent representations.

- We propose a concise way to set λ_{KS} and λ_{CV} and a simple heuristic to estimate λ_{REC}.

Image Generation

Generated Samples



Reconstructed Samples

Randomly Interpolated Samples

Image Generation

- We evaluate the FID of the generated samples from prior distribution (**Samp.**),

Dataset	FASHION MNIST	SVHN	CELEBA
	Samp.	Samp.	Samp.
VAE	50.50	61.01	68.01
WAE	39.66	58.08	58.91
CV-VAE	57.57	51.01	57.61
2sVAE	46.47	45.84	53.12
RAE	47.26	42.35	52.33
Ours	**33.70**	**37.42**	**49.79**

Table: Quantitative evaluation

Image Generation

- We evaluate the FID of the generated samples from prior distribution (**Samp.**), generated samples by fitting a GMM on the learned model (**GMM.**),

Dataset	FASHION MNIST		SVHN		CELEBA	
	Samp.	GMM	Samp.	GMM	Samp.	GMM
VAE	50.50	36.22	61.01	58.23	68.01	61.63
WAE	39.66	28.01	58.08	**34.87**	58.91	49.17
CV-VAE	57.57	38.28	51.01	54.19	57.61	52.72
2sVAE	46.47	–	45.84	–	53.12	–
RAE	47.26	29.59	42.35	35.12	52.33	48.23
Ours	**33.70**	**26.62**	**37.42**	36.46	**49.79**	**44.79**

Table: Quantitative evaluation

Image Generation

- We evaluate the FID of the generated samples from prior distribution (**Samp.**), generated samples by fitting a GMM on the learned model (**GMM.**), the reconstructed samples (**Rec.**)

Dataset	FASHION MNIST			SVHN			CELEBA		
	Samp.	GMM	Rec.	Samp.	GMM	Rec.	Samp.	GMM	Rec.
VAE	50.50	36.22	33.33	61.01	58.23	59.13	68.01	61.63	52.55
WAE	39.66	28.01	24.84	58.08	**34.87**	29.62	58.91	49.17	41.14
CV-VAE	57.57	38.28	35.10	51.01	54.19	48.53	57.61	52.72	45.32
2sVAE	46.47	–	31.93	45.84	–	44.27	53.12	–	44.78
RAE	47.26	29.59	24.54	42.35	35.12	**31.04**	52.33	48.23	41.61
Ours	**33.70**	**26.62**	**19.56**	**37.42**	36.46	31.27	**49.79**	**44.79**	**39.48**

Table: Quantitative evaluation

Image Generation

- We evaluate the FID of the generated samples from prior distribution (**Samp.**), generated samples by fitting a GMM on the learned model (**GMM.**), the reconstructed samples (**Rec.**) and the interpolated samples (**Inter.**).

Dataset	FASHION MNIST				SVHN				CELEBA			
	Samp.	GMM	Rec.	Inter.	Samp.	GMM	Rec.	Inter.	Samp.	GMM	Rec.	Inter.
VAE	50.50	36.22	33.33	44.12	61.01	58.23	59.13	50.29	68.01	61.63	52.55	58.39
WAE	39.66	28.01	24.84	35.01	58.08	**34.87**	29.62	27.16	58.91	49.17	41.14	47.08
CV-VAE	57.57	38.28	35.10	47.73	51.01	54.19	48.53	47.65	57.61	52.72	45.32	50.87
2sVAE	46.47	–	31.93	41.06	45.84	–	44.27	40.23	53.12	–	44.78	47.64
RAE	47.26	29.59	24.54	34.77	42.35	35.12	**31.04**	27.30	52.33	48.23	41.61	**46.58**
Ours	**33.70**	**26.62**	**19.56**	**29.17**	**37.42**	36.46	31.27	**24.87**	**49.79**	**44.79**	**39.48**	47.13

Table: Quantitative evaluation

Image Generation

- We evaluate the FID of the generated samples from prior distribution (**Samp.**), generated samples by fitting a GMM on the learned model (**GMM.**), the reconstructed samples (**Rec.**) and the interpolated samples (**Inter.**).
- ***Our model performs comparably or even better without employing the ex-post GMM fit.***

Dataset	FASHION MNIST				SVHN				CELEBA			
	Samp.	GMM	Rec.	Inter.	Samp.	GMM	Rec.	Inter.	Samp.	GMM	Rec.	Inter.
VAE	50.50	36.22	33.33	44.12	61.01	58.23	59.13	50.29	68.01	61.63	52.55	58.39
WAE	39.66	28.01	24.84	35.01	58.08	**34.87**	29.62	27.16	58.91	49.17	41.14	47.08
CV-VAE	57.57	38.28	35.10	47.73	51.01	54.19	48.53	47.65	57.61	52.72	45.32	50.87
2sVAE	46.47	–	31.93	41.06	45.84	–	44.27	40.23	53.12	–	44.78	47.64
RAE	47.26	29.59	24.54	34.77	42.35	35.12	**31.04**	27.30	52.33	48.23	41.61	**46.58**
Ours	**33.70**	**26.62**	**19.56**	**29.17**	**37.42**	36.46	31.27	**24.87**	**49.79**	**44.79**	**39.48**	47.13

Table: Quantitative evaluation

Unsupervised Image Clustering

- The goal is to naturally cluster the data points in the learned latent space with the multi-modal GMM prior.
- The model is trained with MNIST and FASHIONMNIST images with a 10 component GMM prior.
- The different components of the prior are considered as different classes/clusters to which the data points are mapped by the encoder.

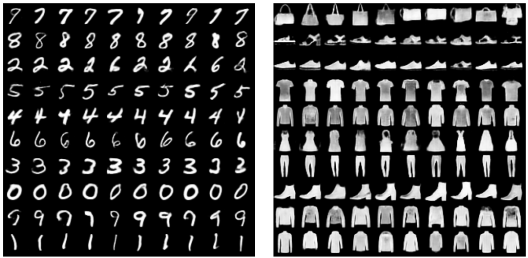


Figure: Qualitative evaluation (Each row in the figure shows randomly generated images from each component of the GMM prior)

Method	Acc(↑)	
	MNIST	FASHION-MNIST
JointVAE	78.33	51.51
CascadeVAE	84.19	**57.72**
Ours	**85.53**	56.24

Table: Quantitative evaluation - Unsupervised classification accuracy

Modelling discrete data structures

- The goal is to model complex discrete data structures such as ***arithmetic expressions*** and ***molecules***.
- We extend the GVAE[1] architecture and experimental settings to include our novel loss.
- Bayesian Optimization is performed in the learned latent space to generate samples with desired properties.
- In Chemical design experiments, we generate new drug like molecules by optimizing the ***water octanol partition coefficient score***.

Method	Score		
	1st(↑)	2nd(↑)	3rd(↑)
GVAE	3.13	3.10	2.37
CVAE	2.75	0.82	0.63
GCVVAE	3.22	2.83	2.63
GRAE	3.74	3.52	**3.14**
Ours	**4.15**	**3.84**	3.12

Table: Quantitative analysis - Top 3 best scores observed for generated molecules across methods

Number	SMILE	Score(↑)
1	C(CCC)CCCCCCCC	**4.15**
2	CCCCCCCCCCC	**3.84**
3	CCCCCc1cccc(c1)	**3.12**

Table: Qualitative analysis - The generated molecules corresponding to the observed best three scores

[1] *Kusner, Matt J. et al. "Grammar Variational Autoencoder." ICML (2017)*

Modelling discrete data structures

- The goal is to model complex discrete data structures such as ***arithmetic expressions*** and ***molecules***.
- We extend the GVAE[1] architecture and experimental settings to include our novel loss.
- Bayesian Optimization is performed in the learned latent space to generate samples with desired properties.
- In Chemical design experiments, we generate new drug like molecules by optimizing the ***water octanol partition coefficient score***.
- A well-structured latent space should yield ***valid sampling*** following the defined grammar rules.

Method	Validity	
	Frac. valid (↑)	Avg. score (↑)
GVAE	0.28 ± 0.04	-7.89 ± 1.90
CVAE	0.16 ± 0.04	-25.64 ± 6.35
GCVVAE	0.76 ± 0.06	-6.40 ± 0.80
GRAE	0.72 ± 0.09	-5.62 ± 0.71
Ours	**0.72 ± 0.03**	**-5.08 ± 1.30**

Table: Quantitative analysis - fraction of valid samples and corresponding average score of the generated molecules across methods

[1] *Kusner, Matt J. et al. "Grammar Variational Autoencoder." ICML (2017)*

Conclusion

- We propose an efficient end-to-end trainable deterministic autoencoder that allows high quality sampling from latent space.
- We introduce a novel deterministic regularization scheme derived from a strong metric on probability distributions to accommodate for expressive multi-modal priors.
- The proposed model achieves good sampling quality even without a ex-post GMM fit.
- Our experimental analysis shows the potential of the model to effectively structure the latent space of both continuous (images) and complex discrete domains (chemical molecules).
- The use of our multi-modal prior distributions significantly improved the optimization performance in the learned latent space.
- We also observed good clustering performance in the learned latent space.