EVALUATING DISENTANGLEMENT IN GENERATIVE MODELS WITHOUT KNOWLEDGE OF LATENT FACTORS

Chester Holtz & Gal Mishne & Alexander Cloninger
University of California San Diego
San Diego, USA
{chholtz, gmishne, acloninger}@ucsd.edu

ABSTRACT

Probabilistic generative models provide a flexible and systematic framework for learning the underlying geometry of data. However, model selection in this setting is challenging, particularly when selecting for ill-defined qualities such as disentanglement or interpretability. In this work, we address this gap by introducing a method for ranking generative models based on the training dynamics exhibited during learning. Inspired by recent theoretical characterizations of disentanglement, our method does not require supervision of the underlying latent factors. We evaluate our approach by demonstrating the need for disentanglement metrics which do not require labels—the underlying generative factors. We additionally demonstrate that our approach correlates with baseline supervised methods for evaluating disentanglement. Finally, we show that our method can be used as an unsupervised indicator for downstream performance on reinforcement learning and fairness-classification problems.

1 INTRODUCTION

Generative models provide accurate models of data, without expensive manual annotation Bengio (2009); Kingma et al. (2014). However, in contrast to classifiers, fully unsupervised model selection within the class of generative models is far from a solved problem Locatello et al. (2019b). For example, simply computing and comparing likelihoods can be a challenge for some families of recently proposed models Goodfellow et al. (2014); Li et al. (2015). Given two models that exhibit similar loss-values on a held-out dataset, there is no computationally friendly way to determine whether one likelihood is significantly higher than the other. Permutation testing or other generic strategies are often computationally prohibitive and it is unclear if likelihood correlates with desirable qualities of generative models. We are motivated to address this problem.

In this work, we focus on the problem of unsupervised model selection, namely Variational Autoencoders (VAEs), for *disentanglement* Higgins et al. (2017). In this context, model selection without full or partial supervision of the ground truth generative process and/or attribute labels is currently an open problem and existing metrics exhibit high variance, even for models with the same hyperparameters trained on the same datasets Locatello et al. (2019b;a). Since ground truth generative factors are unknown or expensive to provide in most real-world tasks, it is important to develop efficient unsupervised methods.

To address the aforementioned issues, we propose a simple and flexible method for fully unsupervised model selection for VAE-based disentangled representation learning. Our approach is inspired by recent findings that attempt to explain why VAEs disentangle Rolinek et al. (2019). We characterize disentanglement quality by performing pairwise comparisons between the training dynamics exhibited by models during gradient descent. We validate our approach using baselines discussed in Locatello et al. (2019b;a) and demonstrate that the model rankings produced by our approach correlate well with performance on downstream tasks.

1.1 CONTRIBUTIONS

Our contributions can be summarized as follows:

1. We design a novel method for model selection for disentanglement based on the activation dynamics of the decoder observed throughout training.
2. Notably, our method is fully unsupervised—our method does not rely on class labels, training supervised models, or ground-truth generative factors.
3. We evaluate our proposed metric by demonstrating strong correlation with supervised baselines on the dSprites dataset Matthey et al. (2017) and downstream performance on reinforcement learning and classification tasks Watters et al. (2019); Locatello et al. (2019a).

2 BACKGROUND

In this section, we review the background notation, the basic variational autoencoder (VAE) framework, and the concept of *disentanglement* in the context of this framework.

2.1 VARIATIONAL AUTOENCODERS

Let $X = \{x_i\}_{i=1}^N$ be a dataset consisting of N i.i.d samples $x_i \in \mathbb{R}^n$ of a random variable x. An autoencoder framework is comprised of two mappings: the encoder $\text{Enc}_\phi : \mathbb{R}^n \to Z$, paramaterized by ϕ, and the decoder $\text{Dec}_\theta : Z \to \mathbb{R}^n$, paramterized by θ. Z is typically termed the *latent space*. In the *variational* autoencoder (VAE) framework, both mappings are taken to be probabalistic and a fixed prior distribution $p(z)$ over Z is assumed.

The training objective is the marginalized log-likelihood:

$$\sum_{i=1}^{n} \log p(x_i) \tag{1}$$

In practice, the parameters of the model; ϕ and θ are jointly trained via gradient descent to minimize a more tractable surrogate: the Evidence Lower Bound (ELBO)

$$\mathbb{E}_{z\sim q(z|x_i)} \log p(x_i|z) - D_{\text{KL}}(q(z|x_i)||p(z)) \tag{2}$$

where the first term corresponds to the reconstruction loss and the second corresponds to the KL divergence between the latent representation $q(z|x_i)$ and the prior distribution $p(z)$, typically chosen to be the standard normal $\mathcal{N}(0, I)$. A significant extension, β-VAE, proposed by Higgins et al. (2017) introduces a weight parameter β on the KL term:

$$\mathbb{E}_{z\sim q(z|x_i)} \log p(x_i|z) - \beta D_{\text{KL}}(q(z|x_i)||p(z)). \tag{3}$$

The value of β is usually chosen to induce certain desireable qualities in the latent representation—e.g. interpretability or disentanglement Chen et al. (2018); Ridgeway & Mozer (2018). Recent work has also proposed methods for selecting β adaptively or according to pre-defined schedules during training Bowman et al. (2016); Fu et al. (2019).

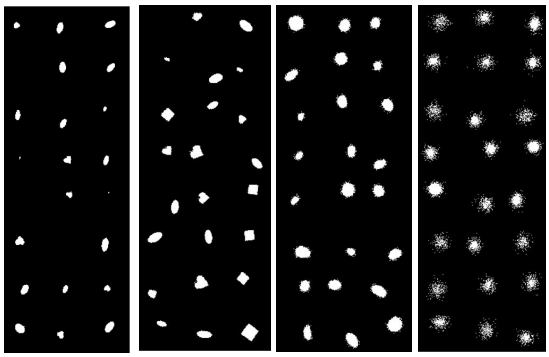
2.2 SUPERVISED METHODS FOR EVALUATING DISENTANGLEMENT

There have been recent efforts by the deep learning community towards learning intrinsic generative factors from data, commonly referred to as learning a disentangled representation. While there are few formalizations of disentanglement, an informal description is provided by Bengio (2009):

> *a representation where a change in one dimension corresponds to a change in one factor of variation, while being relatively invariant to changes in other factors.*

Recent work has shown that learning disentangled representations facilitate robustness Yang et al. (2021), interpretability Zhu et al. (2021), and other desireable characteristics Locatello et al. (2019a). A simple example of the difference between the quality of samples drawn from entangled and disentangled models is provided in Fig. 1. As a result, evaluating models for their ability to learn disentangled latent spaces has received a large amount of attention in recent years Locatello et al. (2019b).

β-VAE & FactorVAE. β-VAE Higgins et al. (2017) and FactorVAE Kim & Mnih (2018) are popular methods for evaluating disentanglement and encouraging learning disentangled representations. As



(a) $\beta = 0.1$ (b) $\beta = 1.0$ (c) $\beta = 10$ (d) $\beta = 100$

Figure 1: (**a—d.**) dSprite samples from *entangled* and *disentangled* latent spaces. Note the occurrence of noisy, missing, and unrealistic samples when β is set inappropriately ($\beta = 0.1, 10, 100$). We propose an unsupervised algorithm to find the appropriate choice of parameters such as β to encourage disentanglement.

previously mentioned, β-VAE uses a modified version of the VAE objective with a larger weight ($\beta > 1$) on the KL divergence between the variational posterior and the prior, and has proven to be effective for encouraging disentangled representations.

In addition to introducing a modification of the ELBO loss, Higgins et al. (2017) proposed a supervised metric that attempts to quantify disentanglement when the ground truth factors of a data set are given. The metric is the error rate of a linear classifier computed as follows:

1. Choose a factor and sample data x with the factor fixed
2. Obtain their representations (mean of $q(z|x)$)
3. Take the absolute value of the pairwise differences of these representations.
4. The mean of these statistics across the pairs are the variables and the index of the fixed factor is the corresponding response

Intuitively, if the learned representations were perfectly disentangled, the dimension of the encoding corresponding to the fixed generative factor would be exactly zero, and the linear classifier would map the index of the zero to the index of the factor. However this metric has several weaknesses:

1. The classifier requires labeled generative factors
2. The metric is sensitive to the classifier's parameters
3. The coefficients of the classifier may not be sparse
4. The classifier may give 100% accuracy even when only $k - 1$ factors out of k have been disentangled

In an attempt to resolve the issues resulting from the application of a parametric model, Kim & Mnih (2018) proposed to replace the linear predictor with a nonparametric majority-vote classifier applied to the empirical variances of the latent embeddings. In other words, the classifier predicts the generative factor k corresponding to the latent dimension with the smallest variance. However, although the drawbacks of linear classification are addressed, certain new limitations are introduced: (1.) an assumption of independence between generative factors (2.) necessity of factor labels.

Mutual Information Gap. The Mutual Information Gap (MIG) Chen et al. (2018) metric involves estimating the mutual information between generative factors each latent dimensions. For each factor, Chen et al. (2018) consider the pair of latent dimensions with the highest MI scores. It is assumed that in a disentangled representation the difference between these two scores would be large. The MIG score is the average normalized difference between pairs of MI scores. Chen et al. (2018) claim that the MIG score is more general compared to the β-VAE and FactorVAE metrics. However, as with β-VAE and FactorVAE, the labels of the underlying generative factors are required.

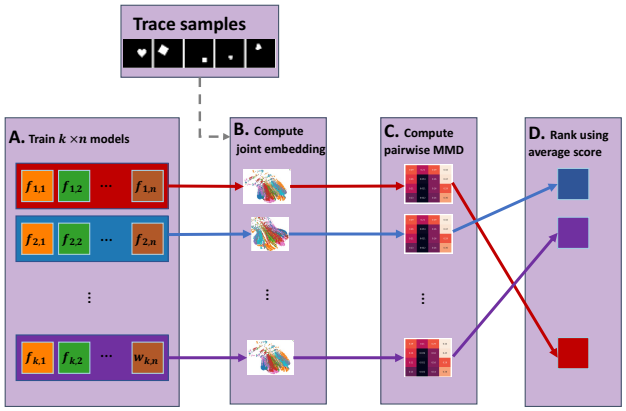


Figure 2: Framework. (**A**) Step 1: Model training: n networks are trained from different initializations for each model specification. (**B**) Step 2: Networks are jointly embedded according to training dynamics. (**C**) Step 3: Pairwise MMD scores are computed from the joint embeddings calculated in Step 2. (**D**) Step 4: The network specifications are sorted using the mean MMD score.

2.3 Intrinsic Indicators of Disentanglement

In this section, we review recent work on identifying fundamental indicators of disentangled models. Recent work Duan et al. (2020); Zhou et al. (2021); Rotman et al. (2022); Bounliphone et al. (2015); Khrulkov & Oseledets (2018) has explored various unsupervised scoring functions to evaluate and compare generative models for similarity, completeness, and disentanglement from the perspective of the latent space. In particular, Duan et al. (2020) and Zhou et al. (2021) propose unsupervised methods for measuring disentanglement based on computing notions of similarity between generative models. Zhou et al. (2021) utilize persistence homolgy to motivate a topological dissimilarity between latent embedding spaces, while Zhou et al. (2021) propose a statistical test between simpling distributions of latent activations. As far as we are aware, we are the first to exploit the dynamics of activations observed during training.

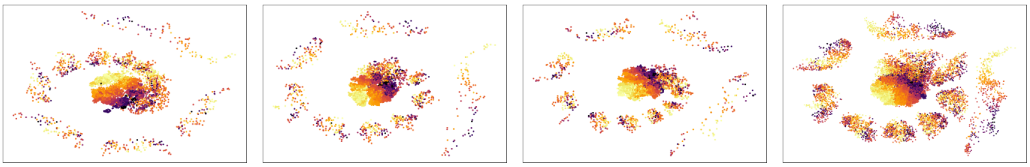
He et al. (2019) investigate disentanglement by studying the dynamics of the deep VAE ELBO loss observed during gradient descent. Their conclusions suggest that artifacts of poor hyperparameter selection or architecture design, e.g., posterior collapse, are a direct product of a "mismatch" between the variational distribution and true posterior. Chechik et al. (2005); Kunin et al. (2019) showed that suitable regularization allows linear autoencoders to recover principal components up to rotation. Lucas et al. (2019) explicitly show that linear VAEs with a diagonal covariance structure recover the principal components exactly.

Significantly, Rolinek et al. (2019) observed that the diagonal covariance used in the variational distribution of VAEs encourages orthogonal representation. They utilize linearaizations of deep networks to rigorously motivate these observations, along with an assumption to handle the presence of posterior collapse. Following up on this work, Kumar & Poole (2020) empirically demonstrate a more general relationship between the variational distribution covariance and the Jacobian of the decoder. In particular, Kumar & Poole (2020) show that a block diagonal covariance structure implies a block structure of the Jacobian of the decoder.

The proposed method is motivated by these central results.

1. Local minima are global
2. The local linearization of the Jacobian, $J_i = \frac{\partial \text{Dec}_\theta(\mu_\phi(x_i))}{\partial \mu_\phi(x_i)}$ is orthogonal

In summary, we design a method to quantify disentaglement according to a novel notion of disagreement between decoder dynamics for multiple instantiations of VAE specifications during learning.



(a) (b) (c) (d)

Figure 3: (**a—c.**) 2d embeddings of training dynamics of individual network realizations. (**d.**) Joint 2d embeddings of training dynamics. x and y coordinates are integers between 0 and 64. Pairs of coordinates are mapped to single values via row-major order—i.e. samples are colored according to the value $x + y \cdot 64$.

3 Comparing VAEs via Learning Dynamics

The two results mentioned above imply that stability of the activation dynamics of the decoder with respect to different initializations may correlate with disentanglement. In this section, we propose a method for computing a similarity score between two decoders according to their activation dynamics. We hypothesize that realizations of a particular specification of a VAE (its architecture and various hyperparameters) which encourages disentangled representation learning will exhibit similar activation dynamics during training, regardless of initialization.

3.1 Finding a Common Representation

To compare the dynamics of multiple VAEs, we define a *multislice* kernel defined on the per-epoch activations between a fixed set of samples.

The Multislice Kernel. We construct a *Multislice Kernel* Gigante et al. (2019) defined over a fixed set of *trace* samples. The entries of the kernel—i.e. the similarities between input samples—are computed according to the *intermediate activations* exhibited by the decoder.

Following the notation of Gigante et al. (2019), the time trace $\mathbf{T}$ of the decoder is an $n \times m \times p$ tensor encoding the activations at each epoch $\tau \in [1, n]$ of p hidden units Dec_{θ_τ} with respect to each of m trace samples. A pair of kernels are constructed using $\mathbf{T}$: $\mathbf{K}_{\text{intraslice}}$, which encodes the affinity between pairs of trace samples indexed by i and j at epoch τ according to the activation patterns they induce when encoded and passed to Dec_{θ_τ}, and $\mathbf{K}_{\text{interslice}}$, which encodes the "self-affinity" between a sample i at time τ and itself at time ν:

$$\mathbf{K}^{(\tau)}_{\text{intraslice}}(i,j) = \exp(-||\mathbf{T}(\tau,i) - \mathbf{T}(\tau,j)||_2^{\alpha}/\sigma^2_{(\tau,i)})$$
$$\mathbf{K}^{(i)}_{\text{interslice}}(i,j) = \exp(-||\mathbf{T}(\tau,i) - \mathbf{T}(\nu,i)||_2^{2}/\epsilon^2),$$

where $\sigma_{\tau,i}$ and ϵ correspond to intraslice and interslice kernel bandwidth parameters. The multislice kernel matrix $\mathbf{K}$ and its symmetrization $\mathbf{K}'$ are then defined:

$$\mathbf{K}((\tau,i),(\nu,j)) = \begin{cases} \mathbf{K}_{\text{intraslice}} & \text{if } \tau = \nu \\ \mathbf{K}_{\text{interslice}} & \text{if } i = j \\ 0 & \text{otherwise} \end{cases} \qquad \mathbf{K}' = \frac{1}{2}(\mathbf{K} + \mathbf{K}^\top)$$

$\mathbf{K}'$ is row-normalized to obtain $\mathbf{P} = \mathbf{D}^{-1}\mathbf{K}'$. The row-stochastic matrix $\mathbf{P}$ represent a random walk over the samples across all epochs, where propagating from (τ, i) to (ν, j) is conditional on the transition probabilities between epochs τ and ν (Gigante et al., 2019). Powers of the matrix, the *diffusion kernel* $\mathbf{P}^t$, represents running the chain forward t steps. Gigante et al. (2019) define a distance based on $\mathbf{P}^t$ and a corresponding distance preserving embedding.

Joint embeddings. It is important to note that Gigante et al. (2019) originally proposed and applied the Multislice Kernel to characterize and differentiate the behavior of different classifiers by constructing visualizations of the network's *hidden units* based on their activations on a fixed set of training samples. In contrast, we propose to apply the Multislice Kernel to directly compare variational models according to the activation response of the decoder on a fixed set of *trace samples*.

In other words, in the work of Gigante et al. (2019), the entries of $\mathbf{K}'$ correspond to similarities between *hidden units*, but in our method, the entries of $\mathbf{K}'$ correspond to similarities between *trace samples*. This subtle difference is key and implies a simple and direct method to compare different VAEs with *different* architectures by comparing their associated multi-slice kernels computed on the *same* set of trace samples. We accomplish this by concatenating the rows of diffusion kernels associated with each set of realizations per-specification and computing the left singular vectors of this tall matrix. Inspired by Gigante et al. (2019), we expect that each individual the kernel encode some average sense of affinity across the data, and by extension that the matrix derived by concatenating these kernels is similarly meaningful.

In Fig. 3, we plot the embeddings associated with a single realizations (initializations) of a given model specification (Fig. 3 a—c) and the aligned embeddings (left singular vectors of concatenated kernels, Fig. 3 d). Note that each embedding is a slight perturbation of the others, but sample embeddings in the joint space roughly align according to a generative factor that explains a significant amount of sample variance (coordinate position).

3.2 Maximum Mean Discrepancy (MMD)

Recall that we propose to consider the cumulative kernel similarity between independent realizations of a VAE specification as a proxy for the disentanglement. To compute a similarity between joint embeddings, we apply the Maximum Mean Discrepancy (MMD) test statistic Gretton et al. (2012) to the left singular vectors of the matrix formed by concatenating diffusion kernels.

Definition 3.1. Maximum Mean Discrepancy (MMD; Gretton et al. (2012)) Let $\mathcal{F}$ be a Reproducing Kernel Hilbert space (RKHS), with continuous feature mapping $\phi(x) \in \mathcal{F}$ from each $x \in \mathcal{X}$, such that the inner product between the features is given by the kernel function $k(x, x') := \langle \phi(x), \phi(x') \rangle$. Then the squared population MMD is

$$\text{MMD}^2(\mathcal{F}, P_x, P_y) = \mathbb{E}_{x,x'}[k(x, x')] - 2\,\mathbb{E}_{x,y}\,[k(x, y\,)] + \mathbb{E}_{y,y'}[k(y, y')].$$

To summarize, distances between distributions are represented as distance between mean embeddings of features characterized by the map ϕ.

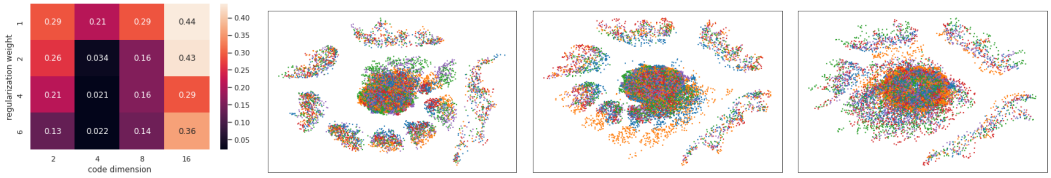
3.3 Unsupervised Model Selection for Disentanglement

As mentioned previously, we are motivated by the observation that networks which disentangle well exhibit "stable" learning dynamics under different initializations. We approximate this stability with the similarity between learning dynamics for networks that differ *only* in their initial weights. We characterize the learning dynamics according the principles proposed by Gigante et al. (2019)

Our method consists of four steps below and in Fig. 2.

1. Train $k \times n$ different models (k different "specifications": architectures / hyperparameters, n different random "realizations" per instance)
2. Jointly embed each group of n models using the left singular vectors of the concatenated multislice kernels
3. For each group, calculate the pairwise MMD metric between each of pair of n models
4. Report the average of the MMD metric over each group as the score for the corresponding realization

An example of the above algorithm applied to a set of VAEs which differ in architecture and regularization weight β is provided in Fig. 4 a. Note that the scores are smallest for networks with a latent space whose dimension is equal to the number of generative factors (4), and for fixed dimension, the scores generally increase as the regularization weight increases—agreeing with previous work Higgins et al. (2017). In Fig. 4 b—d we provide the 2-d restriction of our algorithm to a set of networks with fixed $\beta = 1$ with latent dimension chosen from $4, 8, 16$.



(a) (b) (c) (d)

Figure 4: (**a.**) Average MMD values for different choices of the dimension of z (dimension of the latent space). A lower value denotes more stable learning dynamics. Note that 4 is the ground truth number of generative factors. (**b—d.**) Joint 2-d embeddings of network dynamics as the number of latent dimensions ranges from $4, 8, 16$. Colors denote weight initializations.

4 EXPERIMENTS

We evaluate the proposed method on the dSprites dataset Matthey et al. (2017). This dataset consists of binary images of individual shapes. Each image in the dataset can be fully described by four generative factors: shape (3 values), x and y position (32 values), size (6 values), and rotation (40 values). The generative process for this dataset is fully deterministic, resulting in $737,280$ images. We adopt the same convolutional encoder-decoder architecture presented in Higgins et al. (2017). Network instances vary with respect to the dimension of the code and the β-factor used during training with β chosen from the set $[1, 2, 4, 8, 16]$ and latent space dimension in $[2, 4, 8, 16, 32]$.

Table 1: Spearman rank correlations between rankings produced by our method and those produced by three supervised methods for increasing number of initializations per model.

# random seeds	$n = 5$	$n = 10$	$n = 25$
β-VAE	0.54 ± 0.12	0.57 ± 0.06	0.61 ± 0.02
FactorVAE	0.56 ± 0.21	0.60 ± 0.09	0.60 ± 0.04
MIG	0.60 ± 0.07	0.65 ± 0.03	0.69 ± 0.02

Correlation with supervised disentanglement metrics. We first demonstrate that our method produces rankings that are correlated with those produced using *supervised* baseline methods—methods that exploit supervision of latent factors.

Although the proposed method does not require supervision, it does necessitate training multiple realizations of each model specification. To make a fair comparison with existing methods, we compute the mean of the supervised disentanglement scores over each set of networks.

In Table 1, we show that rankings produced using our score correlate positively with rankings produced using the supervised methods. Furthermore, we observe that as the number of networks used to compute the score increases, the correlation and standard deviation improves.

Table 2: Parameters of noise regimes.

Noise Model	p_{shape}	p_{scale}	$p_{\text{orient.}}$	$p_{\text{pos.}}$
Noise model 1	0.05	0.1	0.05	0.05
Noise model 2	0.1	0.2	0.1	0.1
Noise model 3	0.3	0.3	0.3	0.3

A failure mode of supervised methods. In many real world datasets, the generative factors are unknown or are unreliably labeled. We demonstrate that methods which rely on supervision of the latent factors are brittle to label noise. We propose a new instance of dSprites: *noisy-dSprites*. We compare the robustness of various metrics on the dSprites dataset with labels diluted with different amounts of noise. More concretely, when selecting samples with fixed generative k (i.e. step 1. of the β-VAE metric), we perturb factor k, either uniformly if the factor is discrete (e.g. shape) or according to Gaussian noise with mean 0 (size, orientation, position). We introduce three *noise regimes* in Table 2. In table 3, for various disentanglement metrics, we show that as the amount of noise increases, the quality of the metric decays. However, since the proposed method does not require labeled generative factors, it is robust to label noise.

Table 3: Pearson's correlation between disentanglement scores of models evaluated on data with noisy factors and scores of models evaluated using true labels. A larger correlation implies robustness.

noise model	model 1	model 2	model 3
Ours	1.0	1.0	1.0
β-VAE	0.69	0.54	0.47
FactorVAE	0.84	0.75	0.63
MIG	0.86	0.76	0.67

Table 4: Correlations between disentanglement metrics and unfairness score from Locatello et al. (2019a) (smaller is better) and sample efficiency of a reinforcement learning agent on a toy clustering task Watters et al. (2019) (smaller is better).

# random seeds	unfairness	clustering
Ours	-0.72	-0.59
β-VAE	-0.75	-0.51
FactorVAE	-0.80	-0.56
MIG	-0.66	-0.48

Correlation with downstream task performance. It has been shown that learning (e.g. classifiers or RL agents) with disentangled representations Watters et al. (2019); Locatello et al. (2019a) is easier in some sense—i.e. online decision making can be done more efficiently (with respect to sample complexity) when the state-space is disentangled Watters et al. (2019). Here, we demonstrate that our method can be used to identify VAEs that are useful for downstream classification tasks where training data efficiency is important. More precisely, we evaluate efficiency as the number of steps needed to achieve 90% accuracy on a clustering task.

The agent is provided with a pre-trained encoder trained with $\beta \in \{0, 0.01, 0.1, 1\}$, an exploration policy and a transition model. The goal is to learn a reward predictor to cluster shapes by various generative factors. We use 5 random initialisations of the reward predictor for each possible MONet model, and train them to perform the clustering task detailed in Watters et al. (2019).

We additionally evaluate our method according to it's fairness as defined in Locatello et al. (2019a):

$$\text{unfairness}(\hat{y}) = \frac{1}{|S|} \sum_{s} D_{\text{TV}}(p(\hat{y}), p(\hat{y}|s = s)) \quad \forall y$$

We adopted a similar setup described in Locatello et al. (2019a). A gradient boosted classifier is trained over 10000 labelled examples. The fairness score is computed by taking the mean of the fairness scores across all targets and all sensitive variables where the fairness scores are computed by measuring the total variation.

In Table 4, we see that our method exhibits high Spearman correlation with the fairness score and superior correlation with sample efficiency on the reinforcement learning task.

5 Conclusion and Future Work

We have introduced a method for unsupervised model selection for variational disentangled representation learning. We demonstrated that our metric is reliably correlated with three baseline supervised disentanglement metrics and with performance on two downstream tasks. Crucially, our method does not rely on supervision of the ground-truth generative factors and is therefore robust to nonexistent or noisily labeled generative factors. Future work includes exploring more challenging datasets, addressing scalability, and integrating labels and adapting our framework to other contexts by exploring qualities of neural networks correlate well with training dynamic stability.

REFERENCES

Yoshua Bengio. Learning deep architectures for AI. *Foundations and Trends in Machine Learning*, 2 (1):1–127, 2009. ISSN 1935-8237. doi: 10.1561/2200000006.

Wacha Bounliphone, Eugene Belilovsky, Matthew B. Blaschko, Ioannis Antonoglou, and Arthur Gretton. A test of relative similarity for model selection in generative models, 2015. URL https://arxiv.org/abs/1511.04581.

Samuel R. Bowman, Luke Vilnis, Oriol Vinyals, Andrew M. Dai, Rafal Józefowicz, and Samy Bengio. Generating sentences from a continuous space. In *CoNLL*, 2016.

Gal Chechik, Amir Globerson, Naftali Tishby, and Yair Weiss. Information bottleneck for gaussian variables. *Journal of Machine Learning Research*, 6(6):165–188, 2005. URL http://jmlr.org/papers/v6/chechik05a.html.

Ricky T. Q. Chen, Xuechen Li, Roger B Grosse, and David K Duvenaud. Isolating sources of disentanglement in variational autoencoders. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018. URL https://proceedings.neurips.cc/paper/2018/file/1ee3dfcd8a0645a25a35977997223d22-Paper.pdf.

Sunny Duan, Loic Matthey, Andre Saraiva, Nick Watters, Chris Burgess, Alexander Lerchner, and Irina Higgins. Unsupervised model selection for variational disentangled representation learning. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=SyxL2TNtvr.

Hao Fu, Chunyuan Li, Xiaodong Liu, Jianfeng Gao, Asli Celikyilmaz, and Lawrence Carin. Cyclical annealing schedule: A simple approach to mitigating KL vanishing. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 240–250, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1021. URL https://aclanthology.org/N19-1021.

Scott Gigante, Adam S Charles, Smita Krishnaswamy, and Gal Mishne. Visualizing the PHATE of Neural Networks. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper/2019/file/b4d168b48157c623fbd095b4a565b5bb-Paper.pdf.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, and K. Q. Weinberger (eds.), *Advances in Neural Information Processing Systems*, volume 27. Curran Associates, Inc., 2014. URL https://proceedings.neurips.cc/paper/2014/file/5ca3e9b122f61f8f06494c97b1afccf3-Paper.pdf.

Arthur Gretton, Karsten M. Borgwardt, Malte J. Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel two-sample test. *Journal of Machine Learning Research*, 13(25):723–773, 2012. URL http://jmlr.org/papers/v13/gretton12a.html.

Junxian He, Daniel Spokoyny, Graham Neubig, and Taylor Berg-Kirkpatrick. Lagging inference networks and posterior collapse in variational autoencoders. In *Proceedings of ICLR*, 2019.

Irina Higgins, Loïc Matthey, Arka Pal, Christopher P. Burgess, Xavier Glorot, Matthew M. Botvinick, Shakir Mohamed, and Alexander Lerchner. β-VAE: Learning basic visual concepts with a constrained variational framework. In *ICLR*, 2017.

Valentin Khrulkov and Ivan Oseledets. Geometry score: A method for comparing generative adversarial networks. In Jennifer Dy and Andreas Krause (eds.), *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 2621–2629. PMLR, 10–15 Jul 2018. URL https://proceedings.mlr.press/v80/khrulkov18a.html.

Hyunjik Kim and Andriy Mnih. Disentangling by factorising. In Jennifer Dy and Andreas Krause (eds.), *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 2649–2658. PMLR, 10–15 Jul 2018. URL https://proceedings.mlr.press/v80/kim18b.html.

Diederik P. Kingma, Danilo J. Rezende, Shakir Mohamed, and Max Welling. Semi-supervised learning with deep generative models. In *Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 2*, NIPS'14, pp. 3581–3589, Cambridge, MA, USA, 2014. MIT Press.

Abhishek Kumar and Ben Poole. On Implicit Regularization in β-VAEs. In *Proceedings of the 37th International Conference on Machine Learning*, ICML'20. JMLR.org, 2020.

Daniel Kunin, Jonathan Bloom, Aleksandrina Goeva, and Cotton Seed. Loss landscapes of regularized linear autoencoders. In Kamalika Chaudhuri and Ruslan Salakhutdinov (eds.), *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 3560–3569. PMLR, 09–15 Jun 2019. URL https://proceedings.mlr.press/v97/kunin19a.html.

Yujia Li, Kevin Swersky, and Richard Zemel. Generative moment matching networks. In *Proceedings of the 32nd International Conference on International Conference on Machine Learning - Volume 37*, ICML'15, pp. 1718–1727. JMLR.org, 2015.

Francesco Locatello, Gabriele Abbati, Thomas Rainforth, Stefan Bauer, Bernhard Schölkopf, and Olivier Bachem. On the fairness of disentangled representations. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019a. URL https://proceedings.neurips.cc/paper/2019/file/1b486d7a5189ebe8d8c46afc64b0d1b4-Paper.pdf.

Francesco Locatello, Stefan Bauer, Mario Lucic, Gunnar Raetsch, Sylvain Gelly, Bernhard Schölkopf, and Olivier Bachem. Challenging common assumptions in the unsupervised learning of disentangled representations. In Kamalika Chaudhuri and Ruslan Salakhutdinov (eds.), *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 4114–4124. PMLR, 09–15 Jun 2019b. URL https://proceedings.mlr.press/v97/locatello19a.html.

James Lucas, George Tucker, Roger Grosse, and Mohammad Norouzi. *Don't Blame the ELBO! A Linear VAE Perspective on Posterior Collapse*. Curran Associates Inc., Red Hook, NY, USA, 2019.

Loic Matthey, Irina Higgins, Demis Hassabis, and Alexander Lerchner. dsprites: Disentanglement testing sprites dataset. https://github.com/deepmind/dsprites-dataset/, 2017.

Karl Ridgeway and Michael C. Mozer. Learning deep disentangled embeddings with the f-statistic loss. In *NeurIPS*, 2018.

Michal Rolinek, Dominik Zietlow, and Georg Martius. Variational autoencoders pursue pca directions (by accident). *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 12398–12407, 2019.

Michael Rotman, Amit Dekel, Shir Gur, Yaron Oz, and Lior Wolf. Unsupervised disentanglement with tensor product representations on the torus. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=neqU3HWDgE.

Nicholas Watters, Loïc Matthey, Matko Bosnjak, Christopher P. Burgess, and Alexander Lerchner. COBRA: Data-Efficient Model-Based RL through Unsupervised Object Discovery and Curiosity-Driven Exploration. *ArXiv*, abs/1905.09275, 2019.

Shuo Yang, Tianyu Guo, Yunhe Wang, and Chang Xu. Adversarial robustness through disentangled representations. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(4):3145–3153, May 2021. URL https://ojs.aaai.org/index.php/AAAI/article/view/16424.

Sharon Zhou, Eric Zelikman, Fred Lu, Andrew Y. Ng, Gunnar E. Carlsson, and Stefano Ermon. Evaluating the disentanglement of deep generative models through manifold topology. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=djwS0m4Ft_A.

X. Zhu, C. Xu, and D. Tao. Where and what? examining interpretable disentangled representations. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5857–5866, Los Alamitos, CA, USA, jun 2021. IEEE Computer Society. doi: 10.1109/CVPR46437.2021.00580. URL https://doi.ieeecomputersociety.org/10.1109/CVPR46437.2021.00580.