Rate-Regularization and Generalization in VAEs

Alican Bozkurt*
Northeastern University
alican@ece.neu.edu

Babak Esmaeili*
Northeastern University
esmaeili.b@northeastern.edu

Jean-Baptiste Tristan
Boston College
tristanj@bc.edu

Dana H. Brooks
Northeastern University
brooks@ece.neu.edu

Jennifer G. Dy
Northeastern University
jdy@ece.neu.edu

Jan-Willem van de Meent
Northeastern University
j.vandemeent@northeastern.edu

Abstract

Variational autoencoders optimize an objective that combines a reconstruction loss (the distortion) and a KL term (the rate). The rate is an upper bound on the mutual information, which is often interpreted as a regularizer that controls the degree of compression. We here examine whether inclusion of the rate also acts as an inductive bias that improves generalization. We perform rate-distortion analyses that control the strength of the rate term, the network capacity, and the difficulty of the generalization problem. Decreasing the strength of the rate paradoxically *improves* generalization in most settings, and reducing the mutual information typically leads to underfitting. Moreover, we show that generalization continues to improve even after the mutual information saturates, indicating that the gap on the bound (i.e. the KL divergence relative to the inference marginal) affects generalization. This suggests that the standard Gaussian prior is not an inductive bias that typically aids generalization, prompting work to understand what choices of priors improve generalization in VAEs.

1 Introduction

Variational autoencoders (VAEs) learn representations in an unsupervised manner by training an encoder, which maps high-dimensional data to a lower-dimensional latent code, along with a decoder, which parameterizes a manifold that is embedded in the data space (Kingma and Welling, 2013; Rezende et al., 2014). Much of the work on VAEs has been predicated on the observation that distances on the learned manifold can reflect semantically meaningful factors of variation in the data. This is commonly illustrated by visualizing interpolations in the latent space, or more generally, interpolations along geodesics (Chen et al., 2019).

The ability of VAEs to interpolate is often attributed to the variational objective (Ghosh et al., 2019). VAEs maximize a lower bound on the log-marginal likelihood, which comprises a reconstruction loss and a Kullback-Leibler (KL) divergence between the encoder and the prior (called the rate). Minimizing the reconstruction loss in isolation is equivalent to training a deterministic autoencoder. For this reason, the rate is often interpreted as a regularizer that induces a smoother representation (Chen et al., 2016; Berthelot et al., 2018).

In this paper, we ask the question of whether the inclusion of the rate term also improves generalization. That is, does this penalty reduce the reconstruction loss for inputs that were unseen during training? A known property of VAEs is that the optimal decoder will memorize training data in the limit of infinite capacity (Alemi et al., 2018; Shu et al., 2018), as will a deterministic autoencoder (Radhakrishnan et al., 2019). At the same time, there is empirical evidence that VAEs can underfit the training data, and that reducing the strength of the rate term can mitigate underfitting (Hoffman et al., 2017). Therefore, we might hypothesize that VAEs behave like any other model in machine learning; high-capacity VAEs will overfit the training data, but we can improve generalization by adjusting the strength of the KL term to balance overfitting and underfitting.

*Equal contribution.

Proceedings of the 24th International Conference on Artificial Intelligence and Statistics (AISTATS) 2021, San Diego, California, USA. PMLR: Volume 130.

To test this hypothesis, we performed experiments that systematically vary the strength of the rate term and the network capacity. In these experiments, we deliberately focus on comparatively simple network architectures in the form of linear and convolutional layers with standard spherical Gaussian priors. These architectures remain widely used in work on VAEs, particularly work that focuses on disentangled representations, and systematically investigating these cases provides us with results that can form a basis for understanding the wide variety of more sophisticated architectures that exist in the literature.

The primary aim of our experiments is to carefully control the difficulty of the generalization problem. Our goal in doing so is to disambiguate between apparent generalization that can be achieved by simply reconstructing the most similar memorized training examples and generalization that requires reconstruction of examples that differ substantially from those seen in the training set. To achieve this goal, we have created a dataset of J-shaped tetrominoes that vary in color, size, position, and orientation. This dataset gave us a sufficient variation of both the amount of training data and the density of data in the latent space, as well as sufficient sensitivity of reconstruction loss to variation in these factors, in order to evaluate out-of-domain generalization to unseen combinations of factors.

The surprising outcome of our experiments is that the rate term does not, in general, improve generalization in terms of the reconstruction loss. We find that VAEs memorize training data in practice, even for simple 3-layer fully-connected architectures. However, contrary to intuition, *reducing* the strength of the rate term *improves* generalization under most conditions, including in out-of-domain generalization tasks. The only case where an optimum level of rate-regularization emerges is when low-capacity VAEs are trained on data that are sparse in the latent space. We show that these results hold for both MLP and CNN-based architectures, as well as a variety of datasets.

These results suggest that we need to more carefully quantify the effect of each term in the VAE objective on the generalization properties of the learned representation. To this end, we decompose the KL divergence between the encoder and the prior into its constituent terms: the mutual information (MI) between data and the latent code and the KL divergence between the inference marginal and the prior. We find that the MI term saturates as we reduce the strength of the rate term, which indicates that it is in fact the KL between the inference marginal and prior that drives improvements in generalization in high-capacity models. This suggests that the standard spherical Gaussian prior in VAEs is not an inductive bias that aids generalization in most cases, and that more flexible learned priors may be beneficial in this context.

2 Variational Autoencoders

VAEs jointly train a generative model $p_\theta(\boldsymbol{x}, \boldsymbol{z})$ and an inference model $q_\phi(\boldsymbol{x}, \boldsymbol{z})$. The generative model comprises a prior $p(\boldsymbol{z})$, typically a spherical Gaussian, and a likelihood $p_\theta(\boldsymbol{x} \,|\, \boldsymbol{z})$ that is parameterized by a neural network known as the decoder. The inference model is defined in terms of a variational distribution $q_\phi(\boldsymbol{z} \,|\, \boldsymbol{x})$, parameterized by an encoder network, and a data distribution $q(\boldsymbol{x})$, which is typically an empirical distribution $q(\boldsymbol{x}) = \frac{1}{N} \sum_n \delta_{\boldsymbol{x}_n}(\boldsymbol{x})$ over training data $\{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N\}$. The two models are optimized by maximizing a variational objective (Higgins et al., 2017)

$$\begin{aligned} \mathcal{L}_\beta(\theta, \phi) = &\mathbb{E}_{q_\phi(\boldsymbol{z},\boldsymbol{x})}\big[\log p_\theta(\boldsymbol{x} \,|\, \boldsymbol{z})\big] \\ &- \beta\, \mathbb{E}_{q(\boldsymbol{x})}\big[\mathrm{KL}(q_\phi(\boldsymbol{z} \,|\, \boldsymbol{x}) \,\|\, p(\boldsymbol{z}))\big]. \end{aligned} \tag{1}$$

The multiplier β, which in a standard VAE is set to 1, controls the relative strength of the reconstruction loss and the KL loss. We will throughout this paper refer to these two terms $\mathcal{L}_\beta = -D - \beta R$ as the distortion D and the rate R. The distortion defines a reconstruction loss, whereas the rate constrains the encoder distribution $q_\phi(\boldsymbol{z} \,|\, \boldsymbol{x})$ to be similar to the prior $p(\boldsymbol{z})$. As β approaches 0, the VAE objective becomes similar to that of a deterministic autoencoder; in absence of the rate term, the distortion is minimized when the encoding is a delta-peak at the maximum-likelihood value $\mathrm{argmax}_{\boldsymbol{z}} \log p(\boldsymbol{x} \mid \boldsymbol{z})$. For this reason, a standard interpretation is that the rate serves to induce a smoother representation and ensures that samples from the generative model are representative of the data.

While there is evidence that the rate term indeed induces a smoother representation (Shamir et al., 2010), it is not clear whether this smoothness mitigates overfitting, or indeed to what extent VAEs are prone to overfitting in the first place. Several researchers (Bousquet et al., 2017; Rezende and Viola, 2018; Alemi et al., 2018; Shu et al., 2018) have pointed out that an infinite-capacity optimal decoder will memorize training data, which suggests that high-capacity VAEs will overfit. On the other hand, there is also evidence of underfitting; setting $\beta < 1$ can improve the quality of reconstructions in VAEs for images (Hoffman et al., 2017; Engel et al., 2017), natural language (Wen et al., 2017), and recommender systems (Liang et al., 2018).

More broadly, precisely what constitutes generalization and overfitting in this model class is open to interpretation. If we view the VAE objective primarily as a means of training a generative model, then it makes sense to evaluate model performance in terms of the log marginal likelihood $\log p_\theta(\boldsymbol{x})$. This view is coherent

for the standard VAE objective ($\beta = 1$), which defines a lower bound

$$\begin{aligned}\mathcal{L}(\theta, \phi) &= \mathbb{E}_{q(\boldsymbol{x})}\big[\log p_\theta(\boldsymbol{x}) - \mathrm{KL}(q_\phi(\boldsymbol{z} \mid \boldsymbol{x}) \parallel p_\theta(\boldsymbol{z} \mid \boldsymbol{x}))\big] \\ &\leq \mathbb{E}_{q(\boldsymbol{x})}\big[\log p_\theta(\boldsymbol{x})\big].\end{aligned}$$

The KL term indirectly regularizes the generative model when the encoder capacity is constrained (Shu et al., 2018). Note however that $\mathcal{L}_\beta$ is not a lower bound on $\log p_\theta(\boldsymbol{x})$ when $\beta < 1$. This means that it does not make sense to evaluate generalization in terms of $\log p_\theta(\boldsymbol{x})$ when $\beta \to 0$, or in deterministic autoencoders that do not define a generative model to begin with.

In this paper, we view the VAE primarily as a model for learning representations in an unsupervised manner. In this view, generation is more ancillary; The encoder and decoder serve to define a lossy compressor and decompressor, or equivalently to define a low-dimensional manifold that is embedded in the data space. Our hope is that the learned latent representation reflects semantically meaningful factors of variation in the data, whilst discarding nuisance variables.

The view of VAEs as lossy compressors can be formalized by interpreting the objective $\mathcal{L}_\beta$ as a special case of information-bottleneck (IB) objectives (Tishby et al., 2000; Alemi et al., 2017, 2018). This interpretation relies on the the observation that the decoder $p_\theta(\boldsymbol{x} \mid \boldsymbol{z})$ defines a lower bound on the MI in the inference model $q_\phi(\boldsymbol{z}, \boldsymbol{x})$ in terms of a distortion D and entropy H

$$H - D \leq I_q[\boldsymbol{x}; \boldsymbol{z}], \tag{2}$$

$$D = -\mathbb{E}_{q_\phi(\boldsymbol{x},\boldsymbol{z})}\big[\log p_\theta(\boldsymbol{x} \mid \boldsymbol{z})\big], \tag{3}$$

$$H = -\mathbb{E}_{q(\boldsymbol{x})}\big[\log q(\boldsymbol{x})\big]. \tag{4}$$

Similarly, the rate R is an upper bound on this same mutual information $R \geq I_q[\boldsymbol{x}; \boldsymbol{z}]$,

$$\begin{aligned}R &= \mathbb{E}_{q_\phi(\boldsymbol{x},\boldsymbol{z})}\big[\mathrm{KL}(q_\phi(\boldsymbol{z} \mid \boldsymbol{x}) \parallel p(\boldsymbol{z}))\big] \\ &= I_q[\boldsymbol{x}; \boldsymbol{z}] + \mathrm{KL}(q_\phi(\boldsymbol{z}) \parallel p(\boldsymbol{z})).\end{aligned} \tag{5}$$

Here the term $\mathrm{KL}(q_\phi(\boldsymbol{z}) \parallel p(\boldsymbol{z}))$ is sometimes called "the marginal KL" in the literature (Rezende and Viola, 2018). The naming of the rate and distortion terms originates from rate-distortion theory (Cover and Thomas, 2012), which seeks to minimize $I_q[\boldsymbol{x}; \boldsymbol{z}]$ subject to the constraint $D \leq D^*$. The connection to VAEs now arises from the observation that $\mathcal{L}_\beta$ is a Lagrangian relaxation of the rate-distortion objective

$$\mathcal{L}_\beta = -D - \beta\, R. \tag{6}$$

The appeal of this view is that it suggests an interpretation of the distortion D as an empirical risk and of $I_q[\boldsymbol{x}; \boldsymbol{z}]$ as a regularizer (Shamir et al., 2010). This leads to the hypothesis that VAEs may exhibit a classic bias-variance trade-off: In the limit $\beta \to 0$, we may expect low distortion on the training set but poor generalization to the test set, whereas increasing β may mitigate this form of overfitting.

At the same time, the rate-distortion view of VAEs gives rise to some peculiarities. Standard IB methods use a regressor or classifier $p_\theta(\boldsymbol{y} \mid \boldsymbol{x})$ to define a lower bound $H - D \leq I_q[\boldsymbol{y}; \boldsymbol{z}]$ on the MI between the code $\boldsymbol{z}$ and a target variable $\boldsymbol{y}$ (Tishby et al., 2000). The objective is to maximize $I_q[\boldsymbol{y}; \boldsymbol{z}]$, which serves to learn a representation $\boldsymbol{z}$ which is predictive of $\boldsymbol{y}$, whilst minimizing $I_q[\boldsymbol{x}; \boldsymbol{z}]$, which serves to compress $\boldsymbol{x}$ by discarding information irrelevant to $\boldsymbol{y}$. However, this interpretation does not translate to the special case of VAEs, where $\boldsymbol{x} = \boldsymbol{y}$. Here any compression will necessarily increase the distortion since $D \geq H - I_q[\boldsymbol{x}; \boldsymbol{z}]$.

In our experiments, we will explicitly investigate to what extent β controls a trade-off between overfitting and underfitting. To do so, we will compute *RD* curves that track the rate and distortion under varying β. While *RD* curves have been used to evaluate model performance on the training set (Alemi et al., 2018; Rezende and Viola, 2018), we are not aware of work that explicitly probes generalization to a test set.

To see how overfitting and underfitting may manifest in this analysis, we can consider the hypothetical case of infinite-capacity encoders and decoders. For such networks, both bounds will be tight at the optimum and $\mathcal{L}_\beta = (1 - \beta) I_q[\boldsymbol{x}; \boldsymbol{z}] - H$. Maximizing $\mathcal{L}_\beta$ with respect to ϕ will lead to an *autodecoding* limit when $\beta > 1$, which minimizes $I_q[\boldsymbol{x}; \boldsymbol{z}]$, and an *autoencoding* limit when $\beta < 1$, which maximizes $I_q[\boldsymbol{x}; \boldsymbol{z}]$ (Alemi et al., 2018). One hypothesis is that we will observe poor generalization to the test set in either limit, since maximizing $I_q[\boldsymbol{x}; \boldsymbol{z}]$ could lead to overfitting whereas minimizing $I_q[\boldsymbol{x}; \boldsymbol{z}]$ could lead to underfitting. Moreover, an infinite-capacity generator will fully memorize the training data, which could lead to poor generalization performance in terms of the log marginal likelihood.

In practice, it may well be that the decoder $p_\theta(\boldsymbol{x} \mid \boldsymbol{z})$ can be approximated as an infinite-capacity model. We present empirical evidence of this phenomenon in Appendix **??** that is consistent with recent analyses (Bousquet et al., 2017; Rezende and Viola, 2018; Alemi et al., 2018; Shu et al., 2018). However, it is typically not the case that the prior $p(\boldsymbol{z})$ has a high capacity. In fact, a standard spherical Gaussian prior effectively has 0 capacity, since its mean and variance define an affine transformation that can be trivially absorbed into the first linear layer of any encoder and decoder. This means that the upper bound will be loose and that the rate R may in practice represent a trade-off between $I_q[\boldsymbol{x}; \boldsymbol{z}]$ and $\mathrm{KL}(q_\phi(\boldsymbol{z}) \parallel p(\boldsymbol{z}))$, at least when

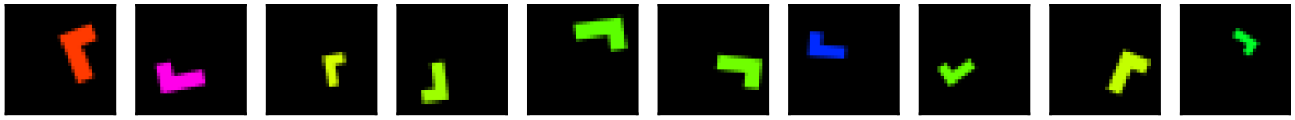

Figure 1: We simulate 164k tetrominoes that vary in position, orientation, size, and color.

the encoder capacity is limited. We present evidence of this trade-off in Section 4.5.

3 Related Work

Generalization in VAEs. Recent work that evaluates generalization in VAEs has primarily considered this problem from the perspective of VAEs as generative models. Shu et al. (2018) consider whether constraining encoder capacity can serve to mitigate data memorization, whereas Zhao et al. (2018) ask whether VAEs can generate examples that deviate from training data. Kumar and Poole (2020) derive a deterministic approximation to the β-VAE objective and show that β-VAE regularizes the generative by imposing a constraint on the Jacobian of the encoder. Whereas Kumar and Poole (2020) evaluate generalization in terms of FID scores (Heusel et al., 2017), we here focus on *RD* curves. Huang et al. (2020) also discuss evaluating deep generative models based on *RD* curves. They show that this type of analysis can be used to uncover some of the known properties of VAEs such as the "holes problem" (Rezende and Viola, 2018) by tracking the change in the curve for different sizes of latent space. In our work, we focus on the change of the *RD* curve as the generalization problem becomes more difficult.

Generalization and regularization in deterministic autoencoders. Zhang et al. (2019) and Radhakrishnan et al. (2019) study generalization in deterministic autoencoders, showing that these models can memorize training data if they are over-parameterized. We overall observed a similar behaviour in our experiments. However, for our experiments, we did not consider architectures as deep as the ones in Zhang et al. (2019) and Radhakrishnan et al. (2019). Ghosh et al. (2019) show that combining deterministic autoencoders with regularizers other than the rate can lead to competitive generative performance.

Generalization of disentangled representations. Our work is indirectly related to research on disentangled representations, in the sense that some of this work is motivated by the desire to learn representations that can generalize to unseen combinations of factors (Narayanaswamy et al., 2017; Kim and Mnih, 2018; Esmaeili et al., 2019; Chen et al., 2018; Locatello et al., 2019). There has been some work to quantify the effect of disentangling on generalization (Eastwood and Williams, 2018; Esmaeili et al., 2019; Locatello et al., 2019), but the extent of this effect remains poorly understood. In this paper, we explicitly design our experiments to test generalization to data with unseen combinations of factors, but we are not interested in disentanglement per se.

4 Experiments

To quantify the effect of rate-regularization on generalization, we designed a series of experiments that systematically control three factors in addition to the β-coefficient: the amount of training data, the density of training data relative to the true factors of variation, and the depth of the encoder and decoder networks. To establish baseline results, we begin with experiments that vary all three factors in fully-connected architectures on a simulated dataset of Tetrominoes. We additionally consider convolutional architectures, as well as other simulated and non-simulated datasets.

4.1 Tetrominoes Dataset

When evaluating generalization we have two primary requirements for a dataset. The first is that failures in generalization should be easy to detect. A good way to ensure this is to employ data for which we can achieve high-quality reconstructions for training examples, which makes it easier to identify degradations for test examples. The second requirement is that we need to be able to disambiguate effects that arise from a lack of data from those that arise from the difficulty of the generalization problem. When a dataset comprises a small number of examples, this may not suffice to train an encoder and decoder network. Conversely, even when employing a large training set, a network may not generalize when there are a large number of generative factors.

To satisfy both requirements, we begin with experiments on simulated data. This ensures that we can explicitly control the density of data in the space of generative factors, and that we can easily detect degradations in reconstruction quality. We initially considered the dSprites dataset (Matthey et al., 2017), which contains 3 shapes at 6 scales, 40 orientations, and 32^2 positions. Unfortunately, shapes in this dataset are close to convex. Varying either the shape or the rota-

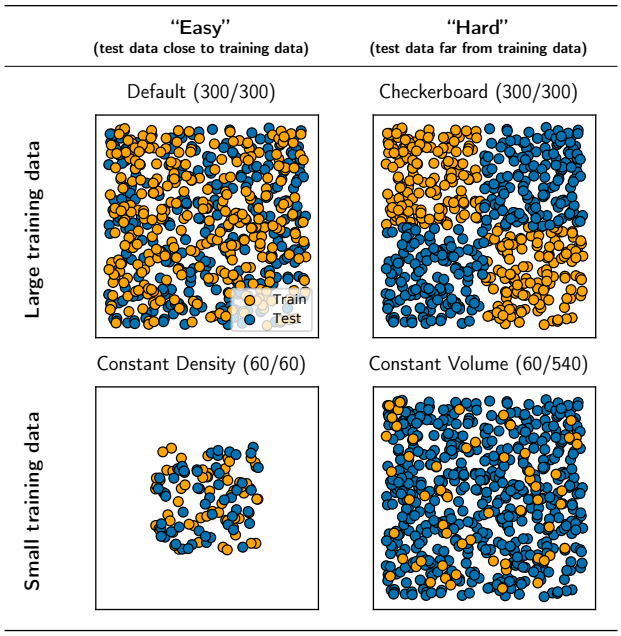

Figure 2: We define 4 train/test splits, which vary in the amount of training data and the typical distance between test data and their nearest neighbors in the training set. Here we show 600 samples with 2 generative factors for visualization.

tion results in small deviations in pixel space, which in practice makes it difficult to evaluate whether a model memorizes the training data.

To overcome this limitation, we created the Tetrominoes dataset. This dataset comprises 163,840 procedurally generated 32×32 color images of a J-shaped tetromino, which is concave and lacks rotational symmetry. We generate images based on five i.i.d. continuous generative factors, which are sampled uniformly at random: rotation (sampled from the [0.0, 360.0] range), color (hue, sampled from [0.0, 0.875] range), scale (sampled from [2.0, 5.0] range), and horizontal and vertical position (sampled from an adaptive range to ensure no shape is placed out of bounds). To ensure uniformity of the data in the latent space, we generate a stratified sample; we divide each feature range into bins and sample uniformly within bins. Examples from the dataset are shown in Figure 1.

4.2 Train/Test Splits

In our experiments, we compare 4 different train/test splits that are designed to vary two components: (1) the amount of training data, (2) the typical distance between training and test examples.

1. 50/50 random split (Default). The base case in our analysis (Figure 2, 1st from left) is a 82k/82k random train/test split of the full dataset. This case is designed to define an "easy" generalization problem, where similar training examples will exist for most examples in the test set.

2. Large data, (Checkerboard) split. We create a 82k/82k split in which a 5-dimensional "checkerboard" mask partitions the training and test set (Figure 2, 2nd from left). This split has the same amount of training data as the base case, as well as the same (uniform) marginal distribution for each of the feature values. This design ensures that for any given test example, there are 5 training examples that differ in one feature (e.g. color) but are similar in all other features (e.g. position, size, and rotation). This defines an out-of-domain generalization task, whilst at the same time ensuring that the model does not need to extrapolate to unseen feature values.

3. Small data, constant density (CD). We create train/test splits for datasets of {8k, 16k, 25k, 33k, 41k, 49k, 57k, 65k} examples by constraining the range of feature values (Figure 2, 2nd from right), ensuring that the density in the feature space remains constant as we reduce the amount of data.

4. Small data, constant volume (CV). Finally, we create train/test splits by selecting {8k, 16k, 25k, 33k, 41k, 49k, 57k, 65k} training examples at random without replacement (Figure 2, 1st from right). This reduces the amount of training data whilst keeping the volume fixed, which increases the typical distance between training and test examples.

4.3 Network Architectures and Training

We use ReLU activations for both fully-connected and convolutional networks with a Bernoulli likelihood in the decoder[1]. We use a 10-dimensional latent space and assume a spherical Gaussian prior. All models are trained for 257k iterations with Adam using a batch size of 128, with 5 random restarts. For MLP architectures, we keep the number of hidden units fixed to 512 across layers. For the CNN architectures, we use 64 channels with kernel size 4 and stride 2 across layers. See Appendix **??** for further details.

4.4 Results

Fully-Connected Architectures on Dense and Sparse Data. We begin with a comparison between 1-layer and 3-layer fully-connected architectures on a dense CV (82k/82k) split and a sparser CV (16k/147k) split. Based on existing work (Radhakrishnan et al., 2019), our hypothesis in this experiment is that the 3-

[1]The Bernoulli likelihood is a very common choice in the VAE literature even for input domain of $[0, 1]$. For a more detailed discussion, see Appendix **??**

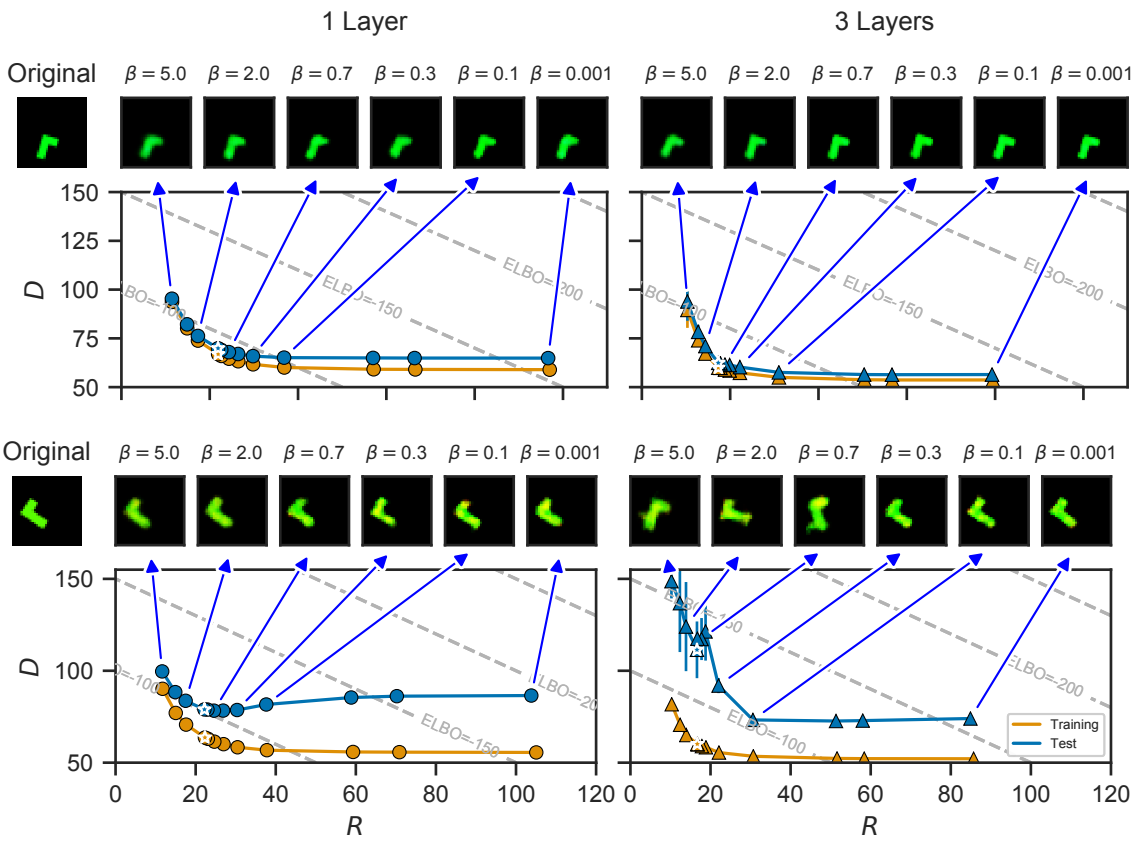

Figure 3: Training and test *RD* curves evaluated on the CV(82k/82k) split (*Top*) and CV(16k/147k) split (*bottom*), for a 1-layer and a 3-layer architecture. Each dot constitutes a β value (white stars indicate the β=1), averaged over 5 restarts. Images show reconstructions of a test example.

layer architecture will be more prone to overfitting the training data (particularly in the sparser case), and our goal is to establish to what extent rate-regularization affects the degree of overfitting.

Figure 3 shows *RD* curves on the training and test set. We report the mean across 5 restarts, with bars indicating the standard deviation, for 12 β values[2]. White stars mark the position of the standard VAE (β=1) on the *RD* plane. Diagonal lines show iso-contours of the evidence lower bound $\mathcal{L}_{\beta=1} = -D - R$. Above each panel, we show reconstructions for a test-set example that is difficult to reconstruct, in the sense that it falls into the 90$^{\text{th}}$ percentile in terms of the ℓ_2-distance between its nearest neighbor in the training set.

For the dense CV (82k/82k) split (*top*), we observe no evidence of memorization. Moreover, increasing model capacity uniformly improves generalization, in the sense that it decreases both the rate and the distortion, shifting the curve to the bottom left.

For the sparse CV (16k/147k) split (*bottom*), we see a different pattern. In the 1-layer model, we observe a trend that appears consistent with a classic bias-variance trade-off. The distortion on the training set decreases monotonically as we reduce β, whereas the distortion on the test set initially decreases, achieves a minimum, and somewhat increases afterwards. This suggests that β may control a trade-off between overfitting and underfitting, although there is no indication of data memorization. When we perform early stopping (see Appendix **??**), the RD curve once again becomes monotonic, which is consistent with this interpretation in terms of overfitting.

In the 3-layer architecture, we observe a qualitatively different trend. Here we see evidence of data memorization; some reconstructions resemble memorized neighbors in the training set. However, counterintuitively, no memorization is apparent at smaller β values. When looking at the iso-contours, we observe that the test-set lower bound $\mathcal{L}_{\beta=1} \leq \log p_\theta(\boldsymbol{x})$ achieves a maximum at $\beta = 0.1$. Additional analysis (see Appendix **??**) shows that this maximum also corresponds to the maximum of the log marginal likelihood $\log p_\theta(\boldsymbol{x})$. In short, high-capacity networks are capable of memorizing the

[2] $\beta \in \{0.001, 0.005, 0.01, 0.1, 0.3, 0.5, 0.7, 0.9, 1., 2., 3., 5.\}$

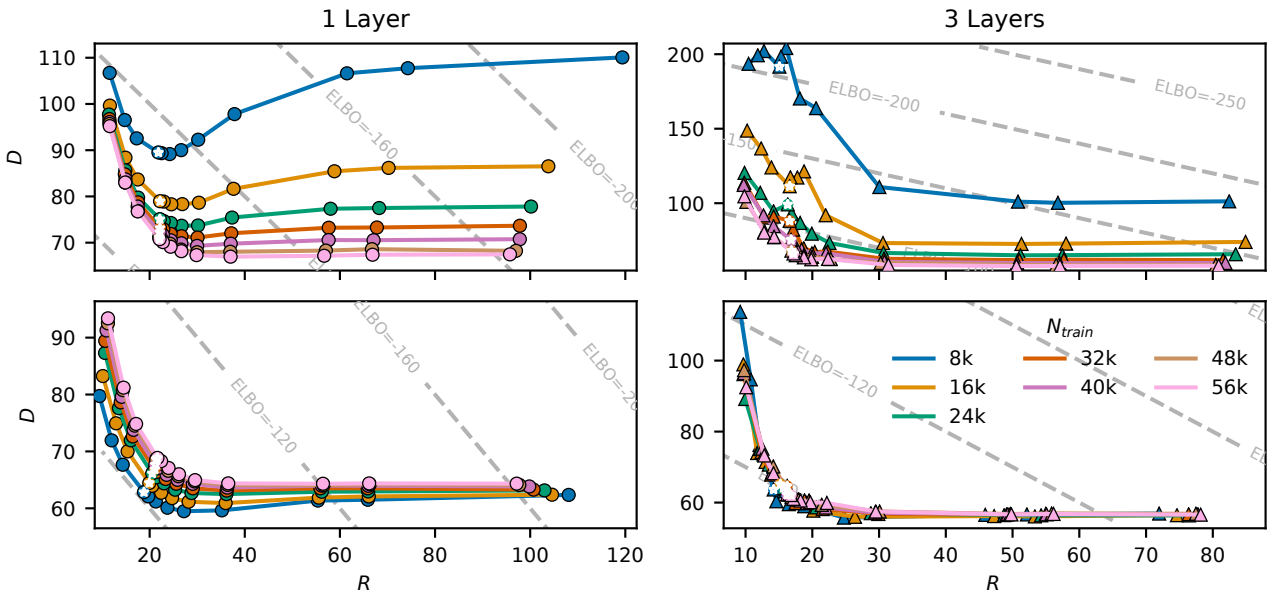

Figure 4: Test-set *RD* curves for constant volume (*top*) and constant density splits (*bottom*) with varying training set sizes. White stars indicate the *RD* value for a standard VAE (β=1).

training data, as expected. However, paradoxically, this memorization occurs when β is large, where we would expect underfitting based on the 1-layer results, and the generalization gap, in terms of both D and $\log p_\theta(\boldsymbol{x})$, is smallest at $\beta = 0.1$.

Role of the Training Set Size. The qualitative discrepancy between training and test set *RD* curves in Figure 3 has to our knowledge not previously been reported. One possible reason for this is that this behavior would not have been apparent in other experiments; there is virtually no generalization gap in the dense CV (82k/82k) split. The differences between 1-layer and 3-layer architectures become visible in the sparse CV (16k/147k) split. Whereas the dense CV (82k/82k) split is representative of typically simulated datasets in terms of the number of examples and density in the latent space, the CV (16k/147k) split has a training set that is tiny by deep learning standards. Therefore, we need to verify that the observed effects are not simply attributable to the size of the training set.

To disambiguate between effects that arise from the size of the data and effects that arise due to the density of the data, we compare CV and CD splits with training set sizes $N_{\text{train}} = \{8\text{k}, 16\text{k}, 32\text{k}, 56\text{k}\}$. Since CD splits have a fixed density rather than a fixed volume, the examples in the test set will be closer to their nearest neighbors in the training set, resulting in an easier generalization problem.

Figure 4 shows the test-set *RD* curves for this experiment. In the CV splits, the qualitative discrepancy between 1-layer and 3-layer networks becomes more pronounced as we decrease the size of the training set. However, in the CD splits, discrepancies are much less pronounced. *RD* curves for 3-layer networks are virtu-

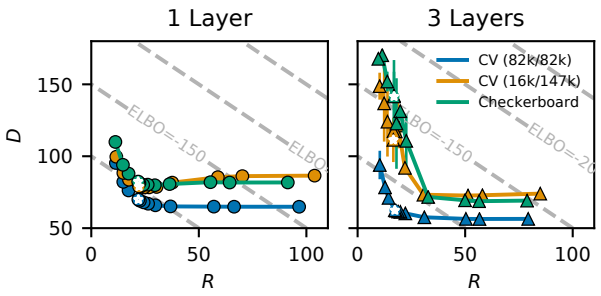

Figure 5: *RD* Curves for the CV(82k/82k), CV(16k/147k), and Checkerboard Splits.

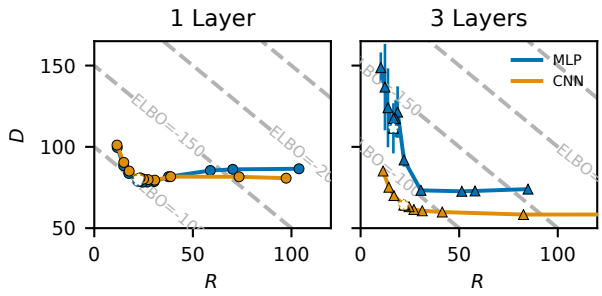

Figure 6: *RD* Curves for the CV(16k/147k) for MLP and CNN Architectures.

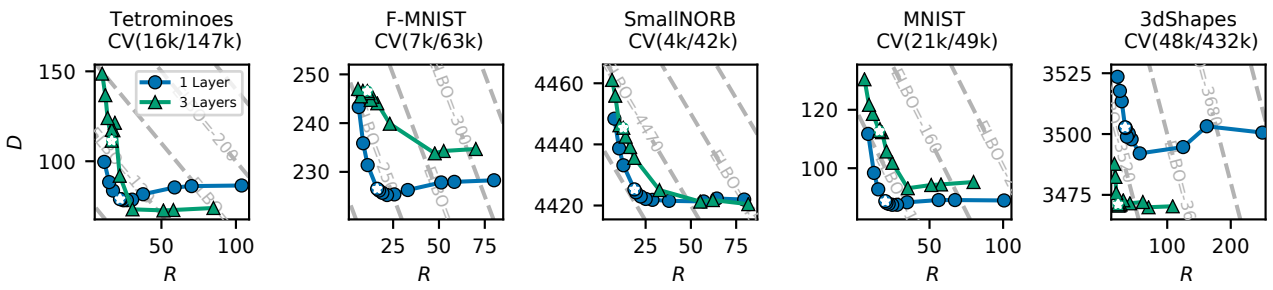

Figure 7: *RD* curves shown on various datasets trained with 1 and 3 layers.

ally indistinguishable. *RD* curves for 1-layer networks still exhibit a minimum, but there is a much weaker dependence on the training set size. Moreover, generalization performance marginally improves as we decrease the size of the training set. This may be attributable to the manner in which we construct the splits. Because we simulate data using a 5-dimensional hypercube of generative factors, limiting the volume has the effect of decreasing the surface to volume ratio, which would mildly reduce the typical distance between training and test set examples.

In-Sample and Out-of-Sample Generalization. A possible takeaway from the results in Figure 4 is that the amount of training data itself does not strongly affect generalization performance, but that the similarity between test and training set examples does. To further test this hypothesis, we compare the CV (82k/82k) and CV (16k/147k) splits to the Checkerboard (82k/82k) split, which allows us to evaluate out-of-sample generalization to unseen combinations of factors. *RD* curves in Figure 5 show similar generalization performance for the Checkerboard and CV (16k/147k) splits. This is consistent with the fact that these splits have a similar distribution over pixel-distances between test set and nearest training set examples (Figure **??**).

Convolutional architectures. A deliberate limitation of our experiments is that we have considered fully-connected networks, which are an extremely simple architecture. There are of course many other encoder and decoder architectures for VAEs (Kingma et al., 2016; Gulrajani et al., 2017; Van den Oord et al., 2016). In Figure 6, we compare *RD* curves for MLPs with those for 1-layer and 3-layer CNNs (see Table **??** for details). We observe a monotonic curve for 3-layer CNNs and only a small degree of non-monotonicity in the 1-layer CNN. Since most architectures will have a higher capacity than a 3-layer MLP or CNN, we can interpret the results for 3-layer networks as the most representative of other architectures.

Additional Datasets. Our analysis thus far shows that the generalization gap grows when we increase the difficulty of a generalization problem, which is expected. The unexpected result is that, depending on model capacity, we either observe U-shaped *RD* curves that are consistent with a bias-variance trade-off, or L-shaped curves in which generalization improves as we reduce β. To test whether both phenomena also occur in other datasets, we perform experiments on the Fashion-MNIST (Xiao et al., 2017), SmallNORB (LeCun et al., 2004), MNIST (LeCun et al., 1998), and 3dShapes (Burgess and Kim, 2018) datasets.

We show the full results of this analysis for a range of CV splits in Appendix **??**. In Figure 7 we compare 1-layer and 3-layer networks for a single split with a small training set for each dataset. We see that the *RD* curves for the 1-layer network exhibits a local minimum in most datasets. Curves for the 3-layer network are generally closer to monotonic, although a more subtle local minimum is visible in certain cases. The one exception is the 3dShapes dataset, where the 3-layer network exhibits a more pronounced local minimum than the 1-layer network.

4.5 Is the Rate a Regularizer?

Our experiments suggest that the rate is not an inductive bias that typically reduces the reconstruction loss in high-capacity models. One possible explanation for these findings is that we should consider both terms in

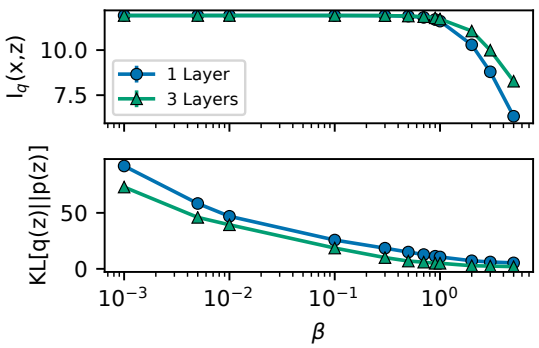

Figure 8: $I_q(\boldsymbol{x}, \boldsymbol{z})$ and $\mathrm{KL}(q_\phi(\boldsymbol{z}) \| p(\boldsymbol{z}))$ vs β for β-VAE Trained on CV(16k/147k).

the rate $R = I_q[\boldsymbol{x};\boldsymbol{z}] + \mathrm{KL}(q_\phi(\boldsymbol{z}) \,\|\, p(\boldsymbol{z}))$ when evaluating the effect of rate-regularization. The term $I_q[\boldsymbol{x};\boldsymbol{z}]$ admits a clear interpretation as a regularizer (Shamir et al., 2010). However, $\mathrm{KL}(q_\phi(\boldsymbol{z}) \,\|\, p(\boldsymbol{z}))$ is not so much a regularizer as a constraint that the aggregate posterior $q_\phi(\boldsymbol{z})$ should resemble the prior $p(\boldsymbol{z})$, which may require a less smooth encoder and decoder when learning a mapping from a multimodal data distribution to a unimodal prior. While we have primarily concerned ourselves with continuous factors for a single Tetramino shape, it is of course common to fit VAEs to multimodal data, particularly when the data contains distinct classes. A unimodal prior forces the VAE to learn a decoder that "partitions" the contiguous latent space into regions associated with each class, which will give rise to sharp gradients near class boundaries.

To understand how each of these two terms contributes to the rate, we compute estimates of $I_q(\boldsymbol{x};\boldsymbol{z})$ and $\mathrm{KL}(q_\phi(\boldsymbol{z}) \,\|\, p(\boldsymbol{z}))$ by approximating $q_\phi(\boldsymbol{z})$ with a Monte Carlo estimate over batches of size 512 (see Esmaeili et al. (2019)). Figure 8 shows both estimates as a function of β for the CV (16k/147k) split. As expected, $I_q(\boldsymbol{x};\boldsymbol{z})$ decreases when $\beta > 1$ but saturates to its maximum $\log N_{\mathrm{train}}$ when $\beta < 1$. Conversely, the term $\mathrm{KL}(q_\phi(\boldsymbol{z}) \,\|\, p(\boldsymbol{z}))$ is small when $\beta > 1$ but increases when $\beta < 1$. Based on the fact that the generalization gap in terms of both the reconstruction loss and $\log p_\theta \boldsymbol{x}$ is minimum at $\beta = 0.1$, it appears that the $\mathrm{KL}(q_\phi(\boldsymbol{z}) \,\|\, p(\boldsymbol{z}))$ term can have a significant effect on generalization performance. Additional experiments where we train VAEs with either the marginal KL or the MI term removed from the loss function confirm this effect of the marginal KL term on the generalization performance of VAEs (see Appendix **??**).

Our reading of these results is that it is reasonable to interpret the rate as an approximation of the MI when β is large. However, our experiments suggest that VAEs typically underfit in this regime, and therefore do not benefit from this form of regularization. When β is small, the MI saturates and we can approximate the rate as $R = \log N_{\mathrm{train}} + \mathrm{KL}(q_\phi(\boldsymbol{z}) \,\|\, p(\boldsymbol{z}))$. In this regime, we should not interpret the rate as a regularizer, but as a constraint on the learned representation, and there can be a trade-off between this constraint and the reconstruction accuracy.

5 Discussion

In this empirical study, we trained over 6000 VAE instances to evaluate how rate-regularization in the VAE objective affects generalization to unseen examples. Our results demonstrate that high-capacity VAEs can and do overfit the training data. However, paradoxically, memorization effects can be mitigated by decreasing β. These effects are more pronounced when test-set examples differ substantially from their nearest neighbors in the training set. For real-world datasets, this is likely to be the norm rather than the exception; few datasets have a small number of generative factors.

Based on these results, we argue that we should give the role of priors as inductive biases in VAEs more serious consideration. The KL relative to a standard Gaussian prior does not improve generalization performance in the majority of cases. With the benefit of hindsight, this is unsurprising; When we use a VAE to model a fundamentally multimodal data distribution, then mapping this data onto a contiguous unimodal Gaussian prior may not yield a smooth encoder, semantically meaningful distances in the latent space, or indeed a representation that generalizes to unseen data. This motivates future work to determine to what extent other priors, including priors that attempt to induce structured or disentangled representations, can aid generalization performance.

While these experiments are comprehensive, we have explicitly constrained ourselves to comparatively simple architectures and datasets. These architectures are not representative of the state of the art (Vahdat and Kautz; Maaløe et al., 2019; Razavi et al., 2019; Gulrajani et al., 2017; Van den Oord et al., 2016), particularly when we are primarily interested in generation. It remains an open question to what extent rate-regularization affects generalization in much higher-capacity architectures that are trained on larger datasets of natural images. Moreover, there are other factors that could potentially impact our results which we do not study here, including but not limited to: dimensionality of the latent space, the choice of prior, and the choice of training method. We leave the investigation of these factors in *RD* analysis to future work.

Acknowledgements

We would like to thank reviewers of a previous version of this manuscript for their detailed comments, as well as Sarthak Jain and Heiko Zimmermann for helpful discussions. This project was supported by the Intel Corporation, the 3M Corporation, the Air Force Research Laboratory (AFRL) and DARPA, NSF grant 1901117, NIH grant R01CA199673 from NCI, and startup funds from Northeastern University.

References

A. Alemi, B. Poole, I. Fischer, J. Dillon, R. A. Saurous, and K. Murphy. Fixing a broken ELBO. In *International Conference on Machine Learning*, pages 159–168, 2018.

A. A. Alemi, I. Fischer, J. V. Dillon, and K. Murphy.

Deep Variational Information Bottleneck. *International Conference on Learning Representations*, 2017.

D. Berthelot, C. Raffel, A. Roy, and I. Goodfellow. Understanding and improving interpolation in autoencoders via an adversarial regularizer. *arXiv preprint arXiv:1807.07543*, 2018.

O. Bousquet, S. Gelly, I. Tolstikhin, C.-J. Simon-Gabriel, and B. Schoelkopf. From optimal transport to generative modeling: the VEGAN cookbook. *arXiv preprint arXiv:1705.07642*, 2017.

C. Burgess and H. Kim. 3D shapes dataset. https://github.com/deepmind/3dshapes-dataset/, 2018.

N. Chen, F. Ferroni, A. Klushyn, A. Paraschos, J. Bayer, and P. van der Smagt. Fast approximate geodesics for deep generative models. In *International Conference on Artificial Neural Networks*, pages 554–566. Springer, 2019.

T. Q. Chen, X. Li, R. B. Grosse, and D. K. Duvenaud. Isolating sources of disentanglement in variational autoencoders. In *Advances in Neural Information Processing Systems*, pages 2610–2620, 2018.

X. Chen, D. P. Kingma, T. Salimans, Y. Duan, P. Dhariwal, J. Schulman, I. Sutskever, and P. Abbeel. Variational lossy autoencoder. *arXiv preprint arXiv:1611.02731*, 2016.

T. M. Cover and J. A. Thomas. *Elements of information theory*. John Wiley & Sons, 2012.

C. Eastwood and C. K. I. Williams. A Framework for the Quantitative Evaluation of Disentangled Representations. In *International Conference on Learning Representations*, Feb. 2018.

J. Engel, M. Hoffman, and A. Roberts. Latent Constraints: Learning to Generate Conditionally from Unconditional Generative Models. *arXiv:1711.05772 [cs, stat]*, Nov. 2017.

B. Esmaeili, H. Wu, S. Jain, A. Bozkurt, N. Siddharth, B. Paige, D. H. Brooks, J. Dy, and J.-W. van de Meent. Structured disentangled representations. In K. Chaudhuri and M. Sugiyama, editors, *Proceedings of Machine Learning Research*, volume 89 of *Proceedings of Machine Learning Research*, pages 2525–2534. PMLR, 16–18 Apr 2019.

P. Ghosh, M. S. Sajjadi, A. Vergari, M. Black, and B. Schölkopf. From variational to deterministic autoencoders. *arXiv preprint arXiv:1903.12436*, 2019.

I. Gulrajani, K. Kumar, F. Ahmed, A. A. Taiga, F. Visin, D. Vazquez, and A. Courville. Pixelvae: A latent variable model for natural images. In *International Conference on Representations*, 2017.

M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter. GANs trained by a two time-scale update rule converge to a local Nash equilibrium. In *Advances in neural information processing systems*, pages 6626–6637, 2017.

I. Higgins, L. Matthey, A. Pal, C. Burgess, X. Glorot, M. Botvinick, S. Mohamed, and A. Lerchner. beta-VAE: Learning basic visual concepts with a constrained variational framework. In *International Conference on Learning Representations*, 2017.

M. D. Hoffman, C. Riquelme, and M. J. Johnson. The β-VAE's Implicit Prior. In *Workshop on Bayesian Deep Learning, NIPS*, pages 1–5, 2017.

S. Huang, A. Makhzani, Y. Cao, and R. Grosse. Evaluating lossy compression rates of deep generative models. *arXiv preprint arXiv:2008.06653*, 2020.

H. Kim and A. Mnih. Disentangling by factorising. In *International Conference on Machine Learning*, pages 2654–2663, 2018.

D. P. Kingma and M. Welling. Auto-encoding variational bayes. *International Conference on Learning Representations*, 2013.

D. P. Kingma, T. Salimans, R. Jozefowicz, X. Chen, I. Sutskever, and M. Welling. Improved variational inference with inverse autoregressive flow. In *Advances in neural information processing systems*, pages 4743–4751, 2016.

A. Kumar and B. Poole. On implicit regularization in β-vaes. *arXiv preprint arXiv:2002.00041*, 2020.

Y. LeCun, L. Bottou, Y. Bengio, P. Haffner, et al. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

Y. LeCun, F. J. Huang, and L. Bottou. Learning methods for generic object recognition with invariance to pose and lighting. In *Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2004. CVPR 2004.*, volume 2, pages II–104. IEEE, 2004.

D. Liang, R. G. Krishnan, M. D. Hoffman, and T. Jebara. Variational Autoencoders for Collaborative Filtering. In *Proceedings of the 2018 World Wide Web Conference*, WWW '18, pages 689–698, Lyon, France, Apr. 2018. International World Wide Web Conferences Steering Committee. ISBN 978-1-4503-5639-8. doi: 10.1145/3178876.3186150.

F. Locatello, S. Bauer, M. Lucic, G. Raetsch, S. Gelly, B. Schölkopf, and O. Bachem. Challenging common assumptions in the unsupervised learning of disentangled representations. In *International Conference on Machine Learning*, pages 4114–4124, 2019.

L. Maaløe, M. Fraccaro, V. Lievin, and O. Winther. Biva: A very deep hierarchy of latent variables for generative modeling. In *33rd Conference on Neural*

Information Processing Systems, page 8882. Neural Information Processing Systems Foundation, 2019.

L. Matthey, I. Higgins, D. Hassabis, and A. Lerchner. dsprites: Disentanglement testing sprites dataset. https://github.com/deepmind/dsprites-dataset/, 2017.

S. Narayanaswamy, T. B. Paige, J.-W. Van de Meent, A. Desmaison, N. Goodman, P. Kohli, F. Wood, and P. Torr. Learning disentangled representations with semi-supervised deep generative models. In *Advances in Neural Information Processing Systems*, pages 5925–5935, 2017.

A. Radhakrishnan, K. Yang, M. Belkin, and C. Uhler. Memorization in overparameterized autoencoders. *arXiv preprint arXiv:1810.10333v3*, 2019.

A. Razavi, A. van den Oord, and O. Vinyals. Generating Diverse High-Fidelity Images with VQ-VAE-2. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d. Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL `https://proceedings.neurips.cc/paper/2019/file/5f8e2fa1718d1bbcadf1cd9c7a54fb8c-Paper.pdf`.

D. J. Rezende and F. Viola. Taming VAEs. *arXiv preprint arXiv:1810.00597*, 2018.

D. J. Rezende, S. Mohamed, and D. Wierstra. Stochastic backpropagation and approximate inference in deep generative models. In *Proceedings of The 31st International Conference on Machine Learning*, pages 1278–1286, 2014.

O. Shamir, S. Sabato, and N. Tishby. Learning and generalization with the information bottleneck. *Theoretical Computer Science*, 411(29):2696–2711, June 2010. ISSN 0304-3975. doi: 10.1016/j.tcs.2010.04.006.

R. Shu, H. H. Bui, S. Zhao, M. J. Kochenderfer, and S. Ermon. Amortized inference regularization. In *Advances in Neural Information Processing Systems*, pages 4393–4402, 2018.

N. Tishby, F. C. Pereira, and W. Bialek. The information bottleneck method. *arXiv:physics/0004057*, Apr. 2000.

A. Vahdat and J. Kautz. NVAE: A Deep Hierarchical Variational Autoencoder. page 13.

A. Van den Oord, N. Kalchbrenner, L. Espeholt, O. Vinyals, A. Graves, et al. Conditional image generation with pixelcnn decoders. In *Advances in neural information processing systems*, pages 4790–4798, 2016.

T.-H. Wen, Y. Miao, P. Blunsom, and S. Young. Latent Intention Dialogue Models. In *International Conference on Machine Learning*, pages 3732–3741, July 2017.

H. Xiao, K. Rasul, and R. Vollgraf. Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms. *arXiv preprint arXiv:1708.07747*, 2017.

C. Zhang, S. Bengio, M. Hardt, and Y. Singer. Identity crisis: Memorization and generalization under extreme overparameterization. *arXiv preprint arXiv:1902.04698*, 2019.

S. Zhao, H. Ren, A. Yuan, J. Song, N. Goodman, and S. Ermon. Bias and generalization in deep generative models: An empirical study. In *Advances in Neural Information Processing Systems*, pages 10815–10824, 2018.