Bounded Information Rate Variational Autoencoders

Daniel Braithwaite
Victoria University of Wellington
School of Engineering and Computer Science
daniel.braithwaite@ecs.vuw.ac.nz

W. Bastiaan Kleijn
Victoria University of Wellington
School of Engineering and Computer Science
bastiaan.kleijn@ecs.vuw.ac.nz

ABSTRACT

This paper introduces a new member of the family of Variational Autoencoders (VAE) that constrains the rate of information transferred by the latent layer. The latent layer is interpreted as a communication channel, the information rate of which is bound by imposing a pre-set signal-to-noise ratio. The new constraint subsumes the mutual information between the input and latent variables, combining naturally with the likelihood objective of the observed data as used in a conventional VAE. The resulting Bounded-Information-Rate Variational Autoencoder (BIR-VAE) provides a meaningful latent representation with an information resolution that can be specified directly in bits by the system designer. The rate constraint can be used to prevent overtraining, and the method naturally facilitates quantisation of the latent variables at the set rate. Our experiments confirm that the BIR-VAE has a meaningful latent representation and that its performance is at least as good as state-of-the-art competing algorithms, but with lower computational complexity.

KEYWORDS

Machine Learning, Variational Autoencoder, Representation Learning

1 INTRODUCTION

Generative modelling is an area of machine learning that focuses on discovering the distribution of a data-set. Latent variable models assume there is some collection of underlying information that can characterise the data efficiently. For example, hair colour and facial expression might be a subset of features that describe images of faces. Good representations have numerous applications in machine learning. The effectiveness of machine learning techniques depends on the quality of the data being used as input. Consequently, feature construction/extraction is an important pre-processing step in many machine learning applications [5]. If the features learned by these generative latent feature models represent the essential components of the input data-set, then it may be possible to use them in place of the original data as the input to another machine learning model, such as a classifier. This paper aims to produce a generative model with features that are a meaningful representation of the data.

Variational Autoencoders [19, 28] (VAEs) and Generative Adversarial Networks [11] (GANs) are common latent feature models. However, the latent features produced by the GAN and VAE often are not a good summary of the input. From a representation learning [5] standpoint, these models leave much to be desired.

A GAN consists of two components, a generator and a discriminator; both are implemented with neural networks. The generator attempts to create fake data that is indistinguishable from real data, and the discriminator attempts to distinguish between the real and fake data, creating a game between the two networks. The generator input is noise $z \sim p(z)$ of a predefined distribution. GANs have been used in a wide variety of tasks, including image-to-image translation [18, 33] and image super-resolution [22]. An effort has been made to make the representation of z meaningful [7].

It has been shown that the GAN objective function is equivalent to minimising the variational lower bound [3] on the mutual information between the discriminator's input, x_{Dis} and the corresponding labels, y (whether the data is real or fake) [17, 23]. If $I(x_{Dis}, y) = 0$, then x_{Dis} carries no information about whether the samples are real or fake. However, minimising a lower bound on $I(x_{Dis}, y)$ does not guarantee the quantity will be equal to 0. This is likely a cause of the instability of the GAN paradigm [17, 23].

The Variational Autoencoder is a method for learning generative latent variable models that avoids the problems present with the GAN. The VAE model is defined as $p_\theta(x) = \int_z p(z) \cdot p_\theta(x|z)\, dz$ where $p_\theta(x|z)$ is a distribution implemented using a neural network with parameter θ, and the latent features, z, are assumed to be distributed according to $p(z)$, which is pre-defined. Maximising the likelihood of the data given the model is a natural way to train the parameters. However, because of the integral over z, the likelihood is typically intractable in a practical implementation. Instead, a lower bound to the likelihood is maximised, called the evidence lower bound, or ELBO. Optimisation of the ELBO induces another distribution $q_\phi(z|x)$, often called the "encoder", which is also implemented with a neural network. Maximising the ELBO corresponds to optimising the likelihood of the data under the model and minimising the Kullback–Leibler divergence between $q_\phi(z|x)$ and $p(z)$, where $p(z)$ is still the assumed distribution of latent features (a unit Gaussian is often used [19]). VAEs have been successfully applied to a variety of different problem domains, such as learning to generate handwritten digits [19], faces [19, 20] and CIFAR images [13].

Variational Autoencoders (VAEs) have been criticised because of their inability to learn latent features that are a meaningful representation of the data [8, 16, 27, 32]. However, the original formulation of the VAE was to learn a generative model, not to produce latent features that represent the salient information of the data. Recent work into improving the representation learning capabilities of the VAE have adjusted its objective function to reward models with meaningful features [16, 27, 32]. This paper proposes that it is not enough to modify the VAE architecture, because the

KDD'18 Deep Learning Day, August 2018, London, UK

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM.
https://doi.org/10.1145/nnnnnnn.nnnnnnn

original VAE formulation is not concerned with meaningful latent representations. Instead, a new approach is needed.

We propose the Bounded Information Rate Variational Autoencoder (BIR-VAE). The BIR-VAE maximises the likelihood of the data subject to a bound on the information rate that can be conveyed by the latent variables from encoder to decoder. The bound is straightforward to implement by forcing the conditional distribution of the encoder output given the input, $q_\phi(z|x)$, to have a Gaussian distribution with fixed, pre-determined standard deviation. In most scenarios, the bound will be reached, and this implies that the BIR-VAE approach subsumes the objective of mutual information maximisation between the input x and the latent variables z subject to the rate constraint.

The remainder of this paper first surveys recent works which build on the VAE to develop representation learning models, discussing the problems with each. Next, the BIR-VAE is derived; this is done by identifying the criterion that the model should satisfy and subsequently, converting these into a function that can be optimised. Lastly, the BIR-VAE is evaluated experimentally.

2 BACKGROUND

This section describes recent work towards creating meaningful latent features in VAEs; thus both motivating our work and providing a context for it. We first discuss in section 2.1 the basic VAE [19, 28], and then in section 2.2 some variants of this method that specifically aim to make the latent features more meaningful. This is generally done by considering the mutual information between the input and the latent variables. The variants include InfoVAE [32], the method of the "Fixing a Broken ELBO" paper [1], the Mutual Autoencoder [27] and the Adversarial Autoencoder [25].

As is common in work on VAEs, we abuse the formal notation of probability theory. We do not distinguish between random variables and their realisations, assuming that this is clear from the context. We also use the common convention that the argument of a density labels the density when it is not ambiguous. As an illustration, using both these conventions we can state that $p(z)$ and $p(x)$ describe the densities of the random variables x and z. Random variables are real-valued except where stated otherwise.

2.1 The Variational Autoencoder

Variational Autoencoders [19, 28] (VAEs) are a type of generative latent feature model. That is, they learn a relationship between a set of latent features z, and the data x. The VAE model is written as $p_\theta(x) = \int_z p(z) \cdot p_\theta(x|z)\, dz$, where $p_\theta(x|z)$ is given by a neural network with parameter θ and the distribution $p(z)$ is assumed to be simple, e.g., a unit Gaussian [28]. Maximising the likelihood of the data under the model is a natural way to train the parameters. However, this is often intractable because of the integral over z. Instead, VAEs maximise a lower-bound on the likelihood called the evidence lower bound, or ELBO. The ELBO induces another probability distribution, $q_\phi(z|x)$, which is represented by a neural network with parameters ϕ. The objective function is

$$O_{ELBO} = -D_{KL}[q_\phi(z|x)||p(z)] + \mathbb{E}_{q_\phi(z|x)}[\log p_\theta(x|z)], \quad (1)$$

where D_{KL} is the Kulback-Leibler divergence. In the context of Autoencoders, $p_\theta(x|z)$ can be interpreted as the decoder and $q_\phi(z|x)$ as the encoder.

The VAE's similarity to an Autoencoder is deceptive. On the surface, one distribution encodes data points into a vectors of latent variables, and another distribution subsequently decodes the latent vectors back into data points. Optimising the ELBO then, in part, maximises the likelihood of the data under the model $p_\theta(x|z)$, and intuitively, z is expected to represent the salient information in the data. However, the notion of an encoder was not present in the original objective, which was to maximise the likelihood of $p_\theta(x)$. Let us write the ELBO in its most basic form, as the sum of a likelihood and a Kullback-Leibler divergence:

$$\begin{aligned} O_{ELBO} &= \log p_\theta(x) - D_{KL}[q_\phi(z|x)||p_\theta(z|x)] \\ &= \mathbb{E}_{z\sim q_\phi(z|x)}[\log p_\theta(x|z)] - D_{KL}[q_\phi(z|x)||p(z)]. \end{aligned} \quad (2)$$

(2) shows that when the ELBO is maximal, it is equal to the likelihood (assuming the network q_ϕ is of sufficient complexity). Consequently, both $D_{KL}[q_\phi(z|x)||p_\phi(z|x)]$ and $D_{KL}[q_\phi(z|x)||p(z)]$ must be 0, which can only occur when z is independent from x. If for a given model the ELBO cannot become maximal, then the KL divergence between $q_\phi(z|x)$ and $p(z)$ must be non-zero. Therefore z and x are dependent. When z does carry information about x, it is because the decoder is not of sufficient complexity. This phenomenon was identified when using an LSTM decoder [6], and recent works introduce it as the Information Preference Property [1, 8, 27, 32]. In the context of representation learning, the Information Preference Property is problematic. However, the above argument shows that learning salient features of the data was never the purpose of maximising the ELBO.

The second issue is called the Exploding Latent Space problem [32], which occurs when the model is sufficiently restrictive, and a larger ELBO can be achieved by maximising the likelihood regardless of the KL divergence term. Optimising $\mathbb{E}_{q_\phi(z|x)}[\log p_\theta(x|z)]$ maximises the likelihood of observing the data given its corresponding latent variables. Consequently, for a data-set, $\{x_1, ..., x_n\}$, minimising the probability of sampling any $x_j \neq x_i$ but x_i from $p_\theta(x|z)$, where $z \sim q_\phi(z|x_i)$ is a way to increase the likelihood. Therefore, if the distributions $q_\phi(z|x_i)$ have disjoint supports, the decoder can be selected as to map the support of $q_\phi(z|x_i)$ to a distribution centred on x_i. This observation shows that maximising the likelihood drives the distributions $q_\phi(z|x_i)$ apart. The KL divergence term regularises this behaviour by pushing the distributions $q_\phi(z|x_i)$ together (towards $p(z)$); however, it is not always successful [32].

The original VAE objective was to learn a generative model of the form $p_\theta(x) = \int_z p(z) \cdot p_\theta(x|z)\, dz$; however, the Exploding Latent Space problem causes the distributions $q_\phi(z|x)$ to diverge rather than converge on $p(z)$. Consequently, the latent variables are not distributed according to $p(z)$, meaning the generative model $p_\theta(x) = \int_z p(z) \cdot p_\theta(x|z)\, dz$ will not produce convincing samples.

This section has shown that a high likelihood (or ELBO) is not indicative of latent features that represent the salient information of the data. Moreover, maximising the ELBO is not suppose to learn a latent representation that is meaningful, because the encoder is a construct of the ELBO and not the original problem formulation.

When the model has learned latent features that have captured the data, it is because the decoder is sufficiently restrictive. On the other hand, a high likelihood (and ELBO) can also occur when a poor generative model has been learned. It is worth noting that the quality of z as a representation of x is controlled by the information between x and z, which is determined by the joint distribution $q_\phi(z, x)$, something that is not directly affected by maximising the likelihood (and ELBO) [27].

2.2 Representation Learning based on Mutual Information

Mutual information maximisation is an increasingly common method for representation learning, which has been recently applied to both VAEs [16, 27, 32] and GANs [7]. In general, mutual information is a measure that has a wide range of applications in neural networks.

In 1988, Linsker introduced InfoMax [24] as a paradigm for optimising Neural Networks. An InfoMax algorithm optimises a function as to maximise the mutual information between the input and output under specified constraints. Particularly well known is the Bell and Sejnovski algorithm [4] that uses InfoMax to perform Independent Component Analysis.

Recently, mutual information has been used to study the dynamics of learning in deep neural networks [29, 30]. The view is that in supervised learning each successive network layer attempts to reduce information about the input while retaining as much information about the desired output as possible. Therefore, the learning network is seen as implementing an approximation to the information bottleneck principle [31]. The information bottleneck principle simultaneously minimises the mutual information between the input and the current layer and maximises of the mutual information between the current network layer and the desired output, subject to a relative weighting.

In the context of VAEs, mutual information is used to ensure that the latent variables z provide useful information about the input x. In this subsection, we discuss some approaches to this paradigm in more detail, thus providing a context and a motivation for the BIR-VAE that we introduce in section 3.

2.2.1 Info Variational Autoencoders. The family of InfoVAE models [32] was proposed for solving both the Information Preference Property and the Exploding Latent Space problem that were discussed in section 2.1. Rearranging the ELBO objective function (1) gives the base formula that is modified to find the InfoVAE objective:

$$O_{ELBO} = -D_{KL}[q_\phi(z)||p(z)] + \mathbb{E}_{p(z)}[D_{KL}[q_\phi(x|z)||p_\theta(x|z)]]. \quad (3)$$

The InfoVAE objective function is constructed by adding a scaling term, λ to the divergence between $q_\phi(z)$ and $p(z)$ in (3), and adding the mutual information between x and z to the equation with regularisation parameter α:

$$\begin{aligned} O_{InfoVAE} &= -\lambda D_{KL}[q_\phi(z)||p(z)] \\ &\quad + \mathbb{E}_{q_\phi(z|x)}[\log D_{KL}[q_\phi(x|z)||p_\theta(x|z)]] \\ &\quad + \alpha I_q(x; z) \qquad (4) \\ &= -\mathbb{E}_{p_D(x)}\,\mathbb{E}_{q_\phi(z|x)}[\log p_\theta(x|z)] \\ &\quad - (1-\alpha)\,\mathbb{E}_{p_D(x)}\,D_{KL}[q_\phi(z|x)||p(z)] \\ &\quad - (\alpha + \lambda - 1)D_{KL}[q_\phi(z)||p(z)] \qquad (5) \end{aligned}$$

where $p_D(x)$ is the data distribution. (5) gives the InfoVAE's objective. It cannot be optimised directly because of the KL divergence term between $q_\phi(z)$ and $p(z)$. It is proven [32] that if $\alpha < 1$ and $\lambda > 0$, then $D_{KL}[q_\phi(z)||p(z)]$ can be replaced with any strict divergence between $q_\phi(z)$ and $p(z)$. Consequently, it is possible to use the Maximum Mean Discrepancy [14] as the divergence; this model is named the MMD-VAE.

When $\alpha \neq 1$, $D_{KL}[q_\phi(z|x)||p(z)]$ and $D_{KL}[q_\phi(z)||p(z)]$ are being simultaneously minimised. This pair is optimal only if z is independent from x. However, the original objective (4) was formulated to maximise the mutual information between z and x. Consequently, the InfoVAE objective is penalising the model when $I_{q_\phi}(x; z) > 0$, while maximising $I_{q_\phi}(x; z)$.

2.2.2 Adversarial Autoencoder. The Adversarial Autoencoder (AAE) [25], is structured like a VAE, except instead of minimising the KL divergence between $q_\phi(z|x)$ and $p(z)$, it uses an adversarial training technique to drive the distribution $q_\phi(z)$ towards $p(z)$, where $p(z)$ is a predetermined distribution, same as for a VAE. Samples taken from $p(z)$ are considered the real data and latent variable vectors produced by the encoder network are considered fakes, an additional neural network is constructed which is used to discriminate between samples from $p(z)$ and the latent variable vectors. The encoder network is penalised if the discriminator can tell that the vector of latent features did not come from $p(z)$, so it is driven to produce latent codes that are distributed according to $p(z)$.

The Adversarial Autoencoder is also part of the family of InfoVAEs. This can be seen by taking (5), letting $\alpha = \lambda = 1$ and using the Jensen-Shannon divergence between $q_\phi(z)$ and $p(z)$ [32]. Consequently, AAEs do not suffer from the same problems that a standard VAE does [32]. However, training GANs can be unstable, and with new methods to improve stability, it can be slow [2, 15]. Consequently, other methods are preferable to the AAE [32].

2.2.3 Fixing a Broken ELBO. As discussed previously, maximising the ELBO is not sufficient for representation learning as it gives no guarantees that z will contain any information about x. Moreover, maximising the ELBO encourages z to be independent of x. [1] makes use of variational upper and lower bounds on the mutual information to prevent this behaviour.

Consider again the data to be x. An "encoding" distribution $q_\phi(z|x)$ takes data vectors and produces a distribution over latent representations. The encoder induces two distributions of interest, $q_\phi(z)$ and $q_\phi(x|z)$, both of which cannot be computed. Consequently, $p_\theta(x|z)$ and $m_\omega(z)$ are introduced, which are approximations of $q_\phi(x|z)$ and $q_\phi(z)$ respectively.

The encoding channel represented by the distribution $q_\phi(z|x)$ has a maximum amount of information that can be transferred through it, denoted R. Consequently, $I(x; z) \leq R$ because z cannot contain more information about x than can be put through the encoding channel. The mutual information is bounded from below by the entropy of x minus reconstruction likelihood.

$$\begin{aligned} H(x) - \mathbb{E}_{p_D(x)}\,\mathbb{E}_{q_\phi(z|x)}[\log p_\theta(x|z)] \leq I_{q_\phi}(x; z) \\ \leq \mathbb{E}_{p_D(x)}[D_{KL}[q_\phi(z|x)||m_\omega(z)]]. \end{aligned} \quad (6)$$

(Following [1] we implicitly assume discrete variables.) (6) demonstrates the bounds on the mutual information. To train this model

either the upper or lower bound is regularised to stay at a predetermined value while the other is optimised [1]. However, this solution does require computing both the distortion and rate, increasing the complexity of the optimisation problem. Moreover, regularising the model to have a preferred rate does not guarantee that this condition will be met.

2.2.4 Mutual Autoencoder. The Mutual Autoencoder [27] is another approach which uses the mutual information to ensure a meaningful latent representation is learned. It regularises $I_\theta(x, z)$ to keep it at a pre-specified value. However, given the difficulty of computing the mutual information, an approximation (lower bound) is used instead. The approximation of $I_\theta(X, Z)$ is given by $I_\theta(X, Z) \geq \hat{I}_\theta(X, Z) = H(z) + \mathbb{E}[\log r(z|x)]$, where r(z|x) is any conditional distribution; this is called the Variational Infomax bound [3].

The distribution $r(z|x)$ is an auxiliary model that must also be trained, increasing the complexity of training and the number of parameters. Another concern is that if $r(z|x)$ is $p_\theta(z|x)$ then $I_\theta(X, Z) = \hat{I}_\theta(X, Z)$, but if $r(z|x)$ is not a good approximation of $p_\theta(z|x)$ then $\hat{I}_\theta(X, Z)$ is not a good measure of $I_\theta(X, Z)$. In other words, there are no guarantees on how tight the bound on the mutual information is. The Mutual Autoencoder has promising experimental results; however, the authors report that the Mutual Autoencoder is slow to train because it requires computing the additional mutual information term.

3 BOUNDED INFORMATION RATE VAE

This section first derives the Bounded Information Rate VAE (BIR-VAE) by describing the induced information rate bound on the encoding channel, and then deriving an objective function that can be optimised. A theoretical comparison between the BIR-VAE and other recent works is also given, describing the contribution of the BIR-VAE in relation to the existing work.

3.1 Theory

The fundamental principle of BIR-VAE is to define an objective function that maximises the likelihood that the input is observed at its output (similar to basic VAE [19, 28]), subject to a constraint on the information rate flowing through the latent layer. The objective and the constraint naturally lead to a meaningful representation with any desired resolution of information about the input. Importantly, as we will show below, this simple paradigm enforces the mutual information between the latent variables z and the input x, without requiring the computation of the mutual information.

A distinction is made between the output of the encoder prior to the channel, y, and the latent variables, z, that form the output of the channel. Hence y corresponds to the mean of the latent variable distribution in a conventional VAE. The desired distribution of the latent variables $q_\phi(z)$ is defined as iid Gaussian with unit variance in each dimension. If x is the random input vector then the deterministic network $\mu_\phi(\cdot)$ transforms x into y, the mean of the distribution $q_\phi(z|x)$. Noise is added to y, to throttle the information throughout the latent variables; this gives, z: $z = y + \epsilon$, where ϵ is iid Gaussian noise with variance $\sigma_\epsilon^2 < 1$ for each dimension. The variance σ_ϵ^2 is set by the system designer and determines the information rate. Note that this differs from the conventional VAE, where the variance σ_ϵ^2 is learned, and the information rate is unknown. The information rate of BIR-VAE across the channel is now bound to [9]

$$I = \frac{d}{2} \log_2(\frac{1}{\sigma_\epsilon^2}), \tag{7}$$

where d is the dimensionality of the latent layer.

To show that BIR-VAE subsumes maximising the mutual information between the latent variables z and the input x we consider an objective function that maximises

(1) the likelihood of the input to be seen at the output (similar to basic VAE [19] [28]), and
(2) the mutual information between the latent variables and the input (similar to InfoVAE [32], and to the Mutual Autoencoder [27]),

subject to a constraint on the information rate flowing through the latent layer. The constrained information rate is induced by placing two restrictions on the latent configuration. Firstly, the distribution $q_\phi(z)$ is defined to be $N(0, I)$. Secondly, $q_\phi(z|x)$ is defined as a Gaussian distribution with arbitrary mean and a variance of σ_ϵ^2 in each dimension.

Let $p_D(x)$ be the data distribution over x, and let $q_\phi(z|x)$ and $p_\theta(x|z)$ be the encoder and decoder respectively. Furthermore, let $I_{q_\phi}(x; z)$ be the mutual information between x and z under the joint distribution $q_\phi(x, z)$. Then we have

$$\begin{aligned} \max_{\phi,\theta} \quad & \mathbb{E}_{p_D(x)} \mathbb{E}_{q_\phi(z|x)}[\log p_\theta(x|z)] + \omega I_{q_\phi}(x; z) \\ \text{subject to} \quad & q_\phi(z) = N(0, I), \\ & \mathbb{E}_{q_\phi(z|x)}[(z - \mathbb{E}_{q_\phi(z|x)}[z])^2] = \sigma_\epsilon^2 I, \end{aligned} \tag{8}$$

where σ_ϵ^2 is a variance set by the system designer that determines the rate constraint, and ω is a weighting. It is possible to satisfy the second constraint, $\mathbb{E}_{q_\phi(z|x)}[(z - \mathbb{E}_{q_\phi(z|x)}[z])^2] = \sigma_\epsilon^2 I$ by fixing the amount of noise introduced in the latent layer.

The first constraint, $q_\phi(z) = N(0, I) = p_\theta(z)$, is satisfied when the Maximum Mean Discrepancy [14] between $q_\phi(z)$ and $p(z)$ is 0. We can form a Lagrangian an write the BIR-VAE objective (8) as

$$\begin{aligned} \max_{\phi,\theta} \quad & \mathbb{E}_{p_D(x)} \mathbb{E}_{q_\phi(z|x)}[\log p_\theta(x|z)] + \omega I_{q_\phi}(x; z) \\ & -\lambda MMD[q_\phi(z)||N(0, I)] \\ \text{subject to} \quad & \mathbb{E}_{q_\phi(z|x)}[(z - \mathbb{E}_{q_\phi(z|x)}[z])^2] = \sigma_\epsilon^2 I. \end{aligned} \tag{9}$$

To optimise (9) the term $I_{q_\phi}(x; z)$ must be made tractable. A convenient form for the mutual information is

$$I_{q_\phi}(x; z) = h_{q_\phi(z)}(z) + \mathbb{E}_{p_D(x)}[h_{q_\phi(z|x)}(z)], \tag{10}$$

where h denotes differential entropy. We note that if the constraint $q_\phi(z) = N(0, I)$ is satisfied, then the differential entropy $h_{q_\phi}(z)$ is fixed. Hence the differential entropy $h_{q_\phi(z)}(z)$ can be omitted from the optimisation problem. The differential entropy $h_{q_\phi(z|x)}(z)$ is the entropy of the latent variable z for a given input x. This is also fixed as this conditioned variable has a Gaussian distribution with variance σ_ϵ^2. Consequently, the second term of the mutual information can also be omitted.

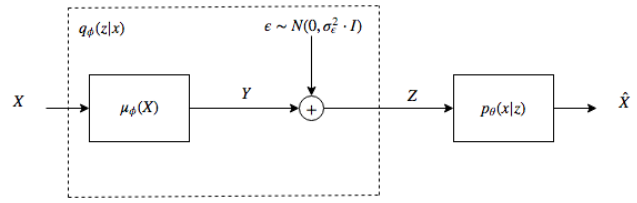


Figure 1: Architecture of the Bounded Information Rate VAE. The encoder network, $\mu_\phi(X) = y$, outputs the mean of the distribution $q_\phi(z|x)$, then noise $\epsilon \sim N(0, \sigma_\epsilon^2)$ is added to y to get the latent variables z.

We have now show that it is possible to write the BIR-VAE objective as

$$\begin{aligned} \max_{\phi,\theta} \quad & \mathbb{E}_{p_D(x)} \mathbb{E}_{q_\phi(z|x)}[\log p_\theta(x|z)] - \lambda MMD[q_\phi(z)||N(0,I)] \\ \text{subject to} \quad & \mathbb{E}_{q_\phi(z|x)}[(z - \mathbb{E}_{q_\phi(z|x)}[z])^2] = \sigma_\epsilon^2 I. \end{aligned} \tag{11}$$

Figure 1 shows the structure of the model, which is similar to the implementation of a VAE except that the variance of z given x is not computed by the BIR-VAE encoder, because it is a constant.

The BIR-VAE decoder ($P_\theta(x|z)$) outputs a distribution, however, if the output distribution is assumed to be an isotropic Gaussian (i.e. $N(0, \sigma^2 \cdot I)$), then the decoder produces as output simply the mean of an isotropic Gaussian with $\sigma^2 = 1$ [19]. This reduces the log likelihood to the negative mean square error:

$$\begin{aligned} & \log\left[\det(2\pi\Sigma) \cdot e^{-\frac{1}{2}(x-\mu)^T \Sigma^{-1}(x-\mu)}\right] \\ & = \log\left[\det(2\pi\Sigma)\right] + -\frac{1}{2}(x-\mu)^2, \end{aligned} \tag{12}$$

where Σ, μ are the covariance matrix and mean, respectively. In a practical implementation we maximise (12) over a batch and $\log[\det(2\pi\Sigma)]$ is ignored (because it is constant). The parameter λ can be simply set to a value that ensures the MMD between $q_\phi(z)$ and $N(0, I)$ is on a similar scale to the likelihood error.

Algorithm 1 summarises the method. In the Algorithm capital letters denote data sets equivalent to a minibatch and *MSE* denotes mean squared error.

Algorithm 1: The BIR-VAE algorithm.

Data: Input signal $\{x_i\}$
Result: Optimised parameters θ^*, ϕ^* for encoder and decoder
set variance σ_ϵ^2 of distribution $q_\phi(z|x)$;
set weight λ ;
initialise parameters θ, ϕ;
for *each epoch* $l \in \mathcal{L}$ **do**
  for *each minibatch* $n \in \mathcal{N}$ **do**
    $X_I \leftarrow$ current minibatch ;
    $Y \leftarrow \mu_\phi(X_I)$ % encoder;
    $Z \leftarrow Y + \epsilon, \ \epsilon \sim N(0, \sigma_\epsilon^2)$ % channel;
    $\hat{X} \leftarrow p_\theta(\cdot|Z)$ % decoder;
    $L \leftarrow MSE(\hat{X}, X_I) + \lambda MMD[q_\phi(Z)||N(0, I)]$;
    $(\theta, \phi) \leftarrow$ +Adam update of θ, ϕ to minimise L;
  end
end

3.2 Discussion

The approach described in 3.1 can be seen as a more tractable method to reach the goals of the Mutual Autoencoder [27]. Instead of attempting to fix the mutual information of z and x through a penalty term, the BIR-VAE physically restricts the information rate of the encoding channel. Where the Mutual Autoencoder requires computing the mutual information of z and x (so that it can be regularised), the BIR-VAE avoids this by building the information rate restriction into the model. By this same argument, the BIR-VAE also improves upon the solution presented in [1].

The BIR-VAE objective function, (11), can be seen as a special case of the InfoVAE objective, i.e. for the case $\alpha = 1$. However, it is not possible to use the InfoVAE objective when $\alpha = 1$ without restricting the mutual information between x and z under the encoding distribution. Without any restriction in place, the mutual information can be maximised to infinity by making $q_\phi(z|x)$ a deterministic mapping. The authors of the InfoVAE paper note this, and state that ensuring that the variance of $q_\phi(z|x)$ does not approach 0 is sufficient to prevent this behaviour. They do not discuss how this is achieved. Furthermore, by making this restriction, they are inducing a maximum information rate on the encoding channel, something that is not identified.

In contrast to existing VAEs [19, 28], the variance of the noise ϵ is pre-determined. In traditional VAEs, the variance of the noise was allowed to vary across the domain of the latent layer variable. However, by scaling the means (or y values in the case of the BIR-VAE) across their domain, the same result is obtained. Hence, for a sufficiently flexible encoder and decoder, the ability to vary the noise variance across the domain is unlikely to affect performance significantly. Conventional VAEs can create a low SNR across the channel for all latent variables; this is the reason why a VAE can ignore the information arriving through the channel and maximise the likelihood using only the decoder.

The ability to set the channel rate of the BIR-VAE clarifies a disadvantage of the basic VAE structure. While VAEs attempt to provide a good likelihood for observations, it has no good reason to provide good performance between observed data other than that it provides a reasonable interpolation across the latent variables. In a BIR-VAE, the quality of this interpolation is dependent on the rate. In contrast, in a conventional VAE, the quality of the interpolation is uncertain. From a generative perspective, it is advantageous to set the rate of the BIR-VAE high. However, a high rate requires a larger database for training. For example, to get texture correct, a very high rate likely is required. Note that this differentiates VAEs from GANs. In GANs the generator performance is judged by a discriminator independently of data points seen. As the discriminator can rely on feature extraction for its judgement (for example for texture), it is less dependent on having seen similar data before.

Finally we note that the BIR-VAE naturally leads to an encoder-decoder system with a quantised bit stream that can be stored or transmitted. In this case the channel is replaced with a vector quantiser, e.g., [12] with the noise characteristics that approximate the Gaussian distribution of the additive noise ϵ. As the data has a well-defined Gaussian distribution, a lattice quantiser, e.g., [10] is particularly natural.

4 EXPERIMENTS

In this section, the performance of the BIR-VAE algorithm is evaluated on the MNIST [21] and SVHN [26] datasets. The meaningfulness of the latent variables and the effect of the information rate is investigated. As a reference system, the InfoVAE algorithm [32] is used. In this section the unit bpi refers to bits/image.

4.1 Experimental Setup and Reference System

The implementation of the BIR-VAE algorithm follows Algorithm 1. The Maximum Mean Discrepancy measure uses a Gaussian kernel,

$$k(x, x') = e^{-\frac{||x-x'||^2}{2\sigma_k^2}}, \tag{13}$$

with a variance $\sigma_k = \sqrt{d}$. The information rate of the BIR-VAE is set using (7). That is, the variance of $q_\phi(z|x)$ is set to $\sigma_\epsilon^2 = 1/(4^{\frac{I}{d}})$, where d is the dimensionality of z and I the information rate. The dimensionality d of the latent variables was varied in the experiments.

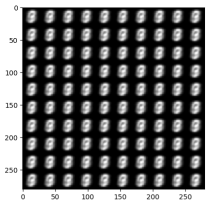
As a reference system we used the InfoVAE algorithm [32] described in section 2.2.1. It was selected as representative of state-of-the-art performance and because code written by the authors is available.[1] We used parameter settings $\alpha = 0$ and $\lambda = 1000$ as this is what the InfoVAE authors used when training on the MNIST dataset. Out of interest, $\alpha = 0.9$ will also be used, as this setting of α means the model prefers a larger mutual information. The setting of $\lambda = 1000$ was also used for the BIR-VAE model, except in one experiment where it is necessary to increase the regularisation constant to enforce the constraint that $q_\phi(z) = N(0, I)$. The InfoVAE does not explicitly define an upper bound on the mutual information. However, the authors of the InfoVAE note that ensuring $q_\phi(z|x)$ does not have vanishing variance is sufficient to regularise the behaviour of the model [32]. In the code provided by the authors the standard deviation of the conditional latent variable distribution, $q_\phi(z|x)$, is bounded by $\sigma_\epsilon \geq 0.01$; this regularisation was also used in our implementation of InfoVAE.

From comparison to the BIR-VAE algorithm, we note that the bounding of σ_ϵ in InfoVAE corresponds to an implicit bounding of the information rate. According to (7), this maximum rate is approximately 13.3 bpi with a two-dimensional latent space. Importantly, if this bound is set without consideration of the database size, over-fitting may result.

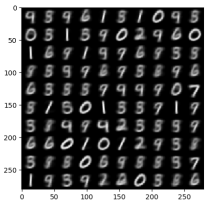
We used the MNIST database of hand-written digits [21] and the Street View House Numbers [26] (SVHN) dataset for our experiments. The MNIST database has 60,000 training images and 10,000 testing images. The data were used in their native form of 28×28 images with 32-bit intensities.

The Street View House Numbers [26] (SVHN) dataset is a collection of house numbers that have been segmented into individual digits. The SVHN database consists of 73,257 training samples and 26,032 testing samples. SVHN is more complex than MNIST because the images are larger (32×32) and in colour. It is worth noting that the SVHN images contain more than just the subject (number), they have distracting components around the edges.

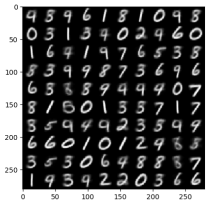
[1] https://github.com/ShengjiaZhao/InfoVAE/blob/master/mmd_vae_eval.py



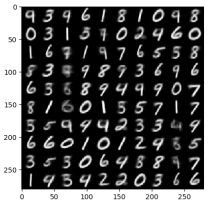
(a) Rate: 0 bpi.



(b) Rate: ≈ 3 bpi.

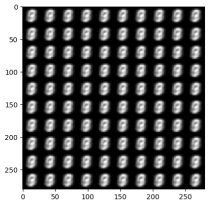


(c) Rate: ≈ 7 bpi.

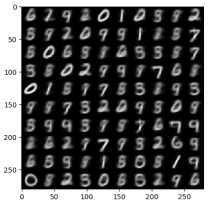


(d) Rate: ≈ 13.3 bpi.

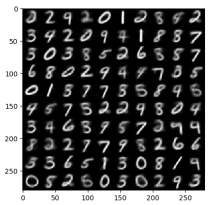
Figure 2: Digit reconstructions for the BIR-VAE model with varying information rates.



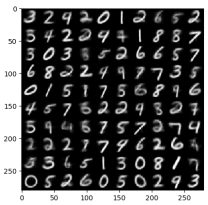
(a) Rate: 0 bpi.



(b) Rate: ≈ 3 bpi.



(c) Rate: ≈ 7 bpi.



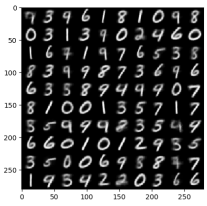
(d) Rate: ≈ 13.3 bpi.

Figure 3: Digits generated from BIR-VAEs with varying information rates.

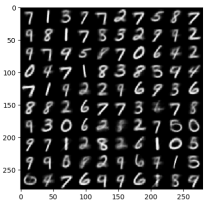
4.2 Results

We studied the relation between information rate and reconstructive and generative performance. We also studied the descriptiveness of the latent variables with respect to the input.

4.2.1 Effect of Information Rate. Figures 2 and 3 respectively show the reconstructions and generations from a BIR-VAE with different information rate limits on the encoding channel for two latent features ($d = 2$). As the information rate increases, the quality of both the reconstructions and generations increases. It should



(a) Reconstructed digits.



(b) Generated digits.

Figure 4: Figures generated using an InfoVAE model trained on the MNIST dataset.

Model	Training MSE	Testing MSE
BIR-VAE (0 bpi)	52.73	52.89
BIR-VAE (≈ 3 bpi)	37.04	36.98
BIR-VAE (≈ 7 bpi)	27.78	29.79
BIR-VAE (≈ 13.3 bpi)	26.01	31.38
InfoVAE ($\alpha = 0$)	26.97	29.85
InfoVAE ($\alpha = 0.9$)	28.68	30.74

Table 1: Mean Square Error (MSE) for various models trained on the MNIST dataset.

be noted that when the BIR-VAE's information rate was restricted to ≈ 3 bpi, the hyperparameter λ had to be increased to 10,000 to enforce the constraint that $q_\phi(z)$ is a unit Gaussian.

Reconstructions and generations produced by the InfoVAE model are shown in figures 4a and 4b, respectively. The BIR-VAE and InfoVAE are indistinguishable in quality when BIR-VAE has a rate that is identical to the maximum rate of the InfoVAE (as noted, ≈ 13.3 bpi if $d = 2$).

Table 1 shows the training and testing MSE for each of the models trained on the MNIST dataset. As expected, when the information rate is increased, the reconstruction MSE decreases. Both the InfoVAE models have similar performance in this situation.

As the aim of BIR-VAE is to obtain a meaningful latent representation it is useful to inspect how the information rate affects the organisation of the latent layer. This is particularly straightforward for the case with only two latent dimensions. Figure 5 shows the latent variables for the InfoVAE and BIR-VAE with ≈ 3 and ≈ 13.3 bits of information per image. The figure shows that the BIR-VAE with ≈ 13.3 bits of information per image has sharper boundaries between classes than the BIR-VAE with a information rate of ≈ 3 bpi. As might be expected, the InfoVAE has a latent representation similar to that of the BIR-VAE with a ≈ 13.3 bpi information rate. Sharper boundaries between classes mean that the model better understands the differences between the digit classes.

4.2.2 Avoiding Overfitting when Data is Limited. To study overfitting, we used the first 600 elements of the MNIST training data for training only. Again, we use a two-dimensional latent space.

Table 2 shows the MSE for five different models trained on the Reduced MNIST dataset. The table shows how adjusting the information rate of the BIR-VAE can be used to control the overfitting. The discrepancy between the training and test MSE is lowest for

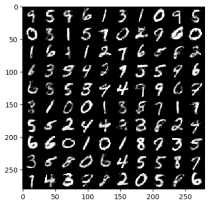
Model	Training MSE	Testing MSE
BIR-VAE(≈ 2 bpi)	35.23	42.82
BIR-VAE (≈ 3 bpi)	29.38	41.46
BIR-VAE (≈ 5 bpi)	22.10	42.97
InfoVAE ($\alpha = 0$)	10.88	60.54
InfoVAE ($\alpha = 0.9$)	12.08	61.68

Table 2: Mean Square Error (MSE) for the models trained on the reduced MNIST problem.

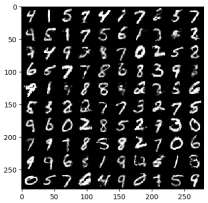
the BIR-VAE model with an information rate of ≈ 2 bits/image. In contrast, the InfoVAE shows clear signs of overfitting, with a large discrepancy between the performance for the training and testing databases.

Figures 6a and 7a show the digit reconstructions for the InfoVAE and BIR-VAE (≈ 2 bpi) respectively. The InfoVAE reconstructions are significantly sharper, but artefacts can be observed in the images. A similar observation can be made in the generated samples, shown in figures 6b and 7b.

Using the BIR-VAE allows the information rate of the encoding channel to be set judiciously. Restricting the information rate reduces the likelihood (increasing the error of the reconstructions), but the goal is to learn a good generative model as well as achieve good reconstructions. The InfoVAE produces crisp reconstructions and generations, whereas the BIR-VAE models produce images that are blurry. However, overall the quality of the ≈ 2 bpi BIR-VAE is better than the InfoVAE, confirming the results of table 2. The BIR-VAE produces images with fewer artefacts.



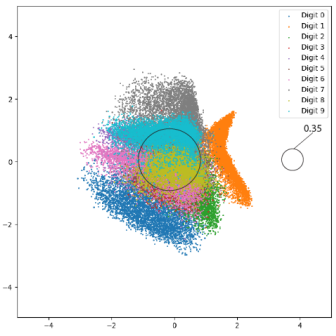
(a) Reconstructed digits.



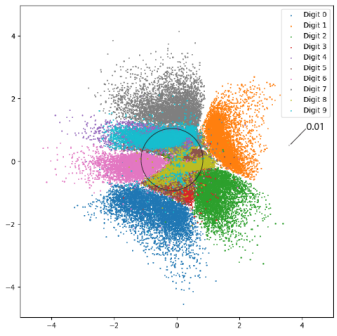
(b) Generated digits.

Figure 6: Figures taken from an InfoVAE model trained on a 600 element subset of MNIST.

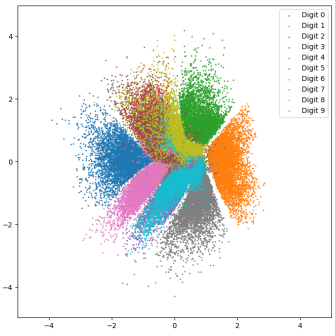
4.2.3 Sharpness of Generated Digits. The BIR-VAE results produce blurry reconstructions and generations at lower information rates. This is natural given the usage of a likelihood measure in combination with the assumption of a Gaussian distribution at the output (which leads to a squared error criterion). Increasing the information rate of the encoder channel (if enough data is present) improves the sharpness of the resulting images. It is to be expected that a higher dimensionality of the latent variable space performs better for higher rates. In a space of higher dimensionality the range of a particular digit (or subclass of a digit) has more neighboring digits (subclasses), facilitating re-arrangement and, hence learning. In a five-dimensional space, it is possible to achieve an information rate of ≈ 33 bpi with $\sigma_\epsilon = 0.01$. We display only the generated digits



(a) BIR-VAE with Information Rate of ≈ 3 bpi. The radius of large and small circles represent σ_z for $q_\phi(z)$ and σ_ϵ for $q_\phi(z|x)$ respectively.

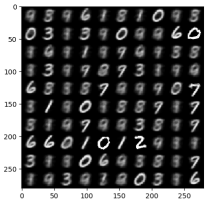


(b) BIR-VAE with Information Rate of ≈ 13.3 bpi. The radius of large and small circles represent σ_z for $q_\phi(z)$ and σ_ϵ for $q_\phi(z|x)$ respectively. It is difficult to see the circle representing σ_ϵ as it is very small.

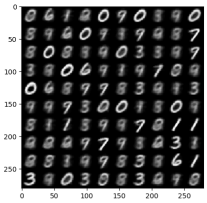


(c) InfoVAE.

Figure 5: Latent space plots for BIR-VAE and InfoVAE models.

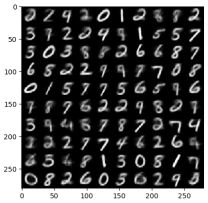


(a) Reconstructed digits.

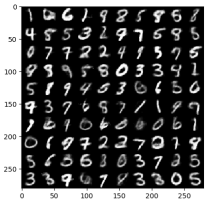


(b) Generated digits.

Figure 7: Figures taken from a BIR-VAE model trained on a 600 element subset of MNIST, the information rate is ≈ 2 bpi.



(a) Two latent dimensions.



(b) Five latent dimensions.

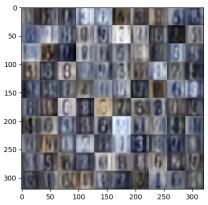
Figure 8: Generated samples taken from the BIR-VAE model trained on the MNIST dataset with two and five latent dimensions. Both models have an information rate of ≈ 33 bpi.

and not the reconstructions as the reconstructions are of higher quality.

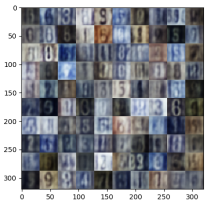
Figure 8 shows the generated digits for a BIR-VAE with a rate of ≈ 33 bpi for the dimensionalities of two and five of the latent space. It is seen that for the two-dimensional latent space (figure 8a) the performance does not increase significantly over the ≈ 13.3 bpi case. However, for the BIR-VAE with a five-dimensional latent space, shown in 8b, the degree of sharpness is increased significantly. This indicates that the degrees of freedom in the model affects learning, and hence the generative model quality independently of the information rate.

4.2.4 Performance on Street View House Numbers. Figure 9 compares the generated images from two BIR-VAEs on the SVHN dataset, both with a 20-dimensional latent space. The model with a higher information rate produces sharper and more convincing results.

Figure 10 shows the InfoVAE model trained on the SVHN dataset. When the information preference property of the the InfoVAE is set to 0, then the model generates simplistic samples which are not as detailed as either the BIR-VAE models. In contrast, the InfoVAE



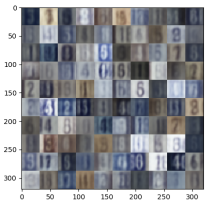
(a) Rate: ≈ 66 bpi.



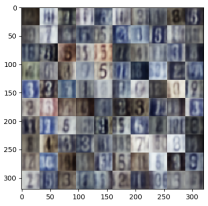
(b) Rate: ≈ 132 bpi.

Figure 9: Generated samples taken from the BIR-VAE model trained on the SVHN dataset with 20 latent features and varying information rate.

with $\alpha = 0.9$ has a similar level of generative quality as the BIR-VAE with an information rate of ≈ 132 bpi.



(a) $\alpha = 0$.



(b) $\alpha = 0.9$.

Figure 10: Generated samples taken from the InfoVAE model trained on the SVHN dataset with 20 latent features and varying information preference parameter, α.

Model	Training MSE	Testing MSE
BIR-VAE ($\approx$ 66 bpi)	16.72	17.97
BIR-VAE ($\approx$ 132 bpi)	11.16	14.92
InfoVAE ($\alpha = 0$)	17.32	20.50
InfoVAE ($\alpha = 0.9$)	11.24	14.87

Table 3: Mean Square Error (MSE) for the models trained on the SVHN dataset.

Table 3 shows the MSE performance for each model trained on the SVHN dataset; comparing the performance of the two models further demonstrates that the InfoVAE with $\alpha = 0.9$ and the BIR-VAE with an information rate of $\approx$ 132 bpi have equivalent performance.

4.3 Discussion of the Experimental Results

This section has shown that the ability to set the information rate of the encoder channel allows the quality of the model to be controlled precisely. While it is possible to set a similar bound on the information rate of an InfoVAE, this was not proposed as part of the InfoVAE model. The InfoVAE paradigm also does not guarantee that it will use the available information.

The BIR-VAE was shown to perform at least as well as the InfoVAE, with additional ability to use the information rate to prevent over-fitting. To facilitate learning, and to obtain sharply defined samples at high rates, the dimensionality of the BIR-VAE must be set appropriately.

5 CONCLUSION

The Bounded Information Rate Variational Autoencoder (BIR-VAE) is a new method for learning generative models with meaningful latent representations. By restricting the information rate of the encoding channel, the generative capacity of the BIR-VAE is constrained in a principled way. An important attribute of BIR-VAE is that in situations with limited data, restricting the channel capacity of the BIR-VAE prevents the model from overfitting.

The idea of using the mutual information between the input and the latent representation to learn meaningful representations has been used by other models, e.g. [1, 27]. Our experimental results show that the performance of the BIR-VAE is at least as good as that of competing algorithms. However, in contrast to competing methods, the BLIR-VAE does not require the explicit approximation or evaluation of the mutual information, thus reducing the computational complexity of the training.

The BIR-VAE paradigm is simple and intuitive. It trains an encoder-decoder network where the output of the encoder is subject to the addition of iid Gaussian noise with a fixed variance σ_ϵ^2, and the input to the decoder is enforced to be unit Gaussian. The choice of σ_ϵ^2 determines the information rate conveyed. To obtain a desired information rate I, the variance of the additive noise is set to $\sigma_\epsilon^2 = 4^{-\frac{I}{d}}$, where d is the dimensionality of the latent variables.

While not discussed in detail, the additive noise channel in the BIR-VAE algorithm can be replaced by a generic vector quantiser with similar statistics of its quantisation noise. The resulting bit-stream can be entropy-coded, to obtain a rate that closely approximates the set rate of BIR-VAE. Thus, BIR-VAE can be used as a trainable encoder-decoder system for storage or transmission.

REFERENCES

[1] A. A. Alemi, B. Poole, I. Fischer, J. V. Dillon, R. A. Saurous, and K. Murphy. 2018. Fixing a Broken ELBO. In *Proceedings of the 35th International Conference on Machine Learning (Proceedings of Machine Learning Research)*, J. Dy and A. Krause (Eds.), Vol. 80. PMLR, Stockholmsmässan, Stockholm Sweden, 159–168.

[2] Martin Arjovsky, Soumith Chintala, and Léon Bottou. 2017. Wasserstein Generative Adversarial Networks. In *Proceedings of the 34th International Conference on Machine Learning (Proceedings of Machine Learning Research)*, D. Precup and Y. W. Teh (Eds.), Vol. 70. PMLR, International Convention Centre, Sydney, Australia, 214–223.

[3] D. Barber and F. Agakov. 2004. The IM algorithm: a variational approach to information maximization. In *Advances in Neural Information Processing Systems 16*, S. Thrun, L. K. Saul, and B. Schölkopf (Eds.). 201–208.

[4] A. J. Bell and T. J. Sejnowski. 1995. An information-maximization approach to blind separation and blind deconvolution. *Neural computation* 7, 6 (1995), 1129–1159.

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

[6] S. R. Bowman, L. Vilnis, O. Vinyals, A. M. Dai, R. Jozefowicz, and S. Bengio. 2016. Generating sentences from a continuous space. In *Conference on Computational Natural Language Learning*.

[7] X. Chen, Y. Duan, R. Houthooft, J. Schulman, I. Sutskever, and P. Abbeel. 2016. InfoGAN: Interpretable representation learning by information maximizing generative adversarial nets. In *Advances in Neural Information Processing Systems 29*, D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett (Eds.). 2172–2180.

[8] X. Chen, D. P. Kingma, T. Salimans, Y. Duan, P. Dhariwal, J. Schulman, I. Sutskever, and P. Abbeel. 2017. Variational lossy autoencoder. In *International Conference on Learning Representations*.

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

[10] U. Erez, S. Litsyn, and R. Zamir. 2005. Lattices which are good for (almost) everything. *IEEE Transactions on Information Theory* 51, 10 (2005), 3401–3416.

[11] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. 2014. Generative Adversarial Nets. In *Advances in Neural Information Processing Systems 27*, Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger (Eds.). 2672–2680.

[12] R. M. Gray. 2012. *Source coding theory*. Vol. 83. Springer Science & Business Media.

[13] K. Gregor, I. Danihelka, A. Graves, D. J. Rezende, and D. Wierstra. 2015. DRAW: A Recurrent Neural Network For Image Generation. In *Proceedings of the 32nd International Conference on Machine Learning (Proceedings of Machine Learning Research)*, Francis Bach and David Blei (Eds.), Vol. 37. PMLR, Lille, France, 1462–1471.

[14] A. Gretton, K. M. Borgwardt, M. J. Rasch, B. Schölkopf, and A. Smola. 2012. A kernel two-sample test. *Journal of Machine Learning Research* 13, Mar (2012), 723–773.

[15] I. Gulrajani, F. Ahmed, M. Arjovsky, V. Dumoulin, and A. C. Courville. 2017. Improved Training of Wasserstein GANs. In *Advances in Neural Information Processing Systems 30*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (Eds.). 5767–5777.

[16] I. Higgins, L. Matthey, A. Pal, C. Burgess, X. Glorot, M. Botvinick, S. Mohamed, and A. Lerchner. 2017. Beta-VAE: Learning basic visual concepts with a constrained

variational framework. In *International Conference on Learning Representations*.
[17] F. Huszar. 2016. InfoGAN: using the variational bound on mutual information (twice). http://www.inference.vc/infogan-variational-bound-on-mutual-information-twice/
[18] P. Isola, J. Zhu, T. Zhou, and A. A. Efros. 2017. Image-to-image translation with conditional adversarial networks. In *Conference on Computer Vision and Pattern Recognition*.
[19] D. P. Kingma and M. Welling. 2013. Auto-encoding variational Bayes. *arXiv preprint arXiv:1312.6114* (2013).
[20] T. D. Kulkarni, W. F. Whitney, P. Kohli, and J. Tenenbaum. 2015. Deep convolutional inverse graphics network. In *Advances in Neural Information Processing Systems 28*, C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett (Eds.). 2539–2547.
[21] Y. LeCun. 1998. The MNIST database of handwritten digits. *http://yann. lecun. com/exdb/mnist/* (1998).
[22] C. Ledig, L. Theis, F. Huszár, J. Caballero, A. Cunningham, A. Acosta, A. Aitken, A. Tejani, J. Totz, Z. Wang, et al. 2017. Photo-realistic single image super-resolution using a generative adversarial network. In *Conference on Computer Vision and Pattern Recognition*.
[23] Y. Li. 2016. GANs, mutual information, and possibly algorithm selection? http://www.yingzhenli.net/home/blog/?p=421
[24] R. Linsker. 1988. Self-organization in a perceptual network. *Computer* 21, 3 (1988), 105–117.
[25] A. Makhzani, J. Shlens, N. Jaitly, I. Goodfellow, and B. Frey. 2016. Adversarial Autoencoders. In *International Conference on Learning Representations*.
[26] Y. Netzer, T. Wang, A. Coates, A. Bissacco, B. Wu, and A. Y. Ng. 2011. Reading digits in natural images with unsupervised feature learning. In *Advances in Neural Information Processing Systems workshop on deep learning and unsupervised feature learning*, Vol. 2011. 5.
[27] M. Phuong, M. Welling, N. Kushman, R. Tomioka, and S. Nowozin. 2018. The Mutual Autoencoder: Controlling Information in Latent Code Representations. https://openreview.net/forum?id=HkbmWqxCZ
[28] D. J. Rezende, S. Mohamed, and D. Wierstra. 2014. Stochastic Backpropagation and Approximate Inference in Deep Generative Models. In *Proceedings of the 31st International Conference on Machine Learning (Proceedings of Machine Learning Research)*, E. P. Xing and T. Jebara (Eds.), Vol. 32. PMLR, Bejing, China, 1278–1286.
[29] A. M. Saxe, Y. Bansal, J. Dapello, N. Advani, A. Kolchinsky, B. D. Tracey, and D. D. Cox. 2018. On the information bottleneck theory of deep learning. In *International Conference on Learning Representations*.
[30] R. Shwartz-Ziv and N. Tishby. 2017. Opening the black box of deep neural networks via information. *arXiv preprint arXiv:1703.00810* (2017).
[31] N. Tishby, F. C. Pereira, and W. Bialek. 2000. The information bottleneck method. *arXiv preprint physics/0004057* (2000).
[32] S. Zhao, J. Song, and S. Ermon. 2018. InfoVAE: Balancing Learning and Inference in Variational Autoencoders. *arXiv preprint arXiv:1706.02262v3* (2018).
[33] J. Zhu, T. Park, P. Isola, and A. A. Efros. 2017. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *International Conference on Computer Vision*.