Structured Disentangled Representations

Babak Esmaeili
Northeastern University
esmaeili.b@husky.neu.edu

Hao Wu
Northeastern University
wu.hao10@husky.neu.edu

Sarthak Jain
Northeastern University
jain.sar@husky.neu.edu

Alican Bozkurt
Northeastern University
alican@ece.neu.edu

N. Siddharth
University of Oxford
nsid@robots.ox.ac.uk

Brooks Paige
Alan Turing Institute
University of Cambridge
bpaige@turing.ac.uk

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

Jennifer Dy
Northeastern University
jdy@ece.neu.edu

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

Abstract

Deep latent-variable models learn representations of high-dimensional data in an unsupervised manner. A number of recent efforts have focused on learning representations that disentangle statistically independent axes of variation by introducing modifications to the standard objective function. These approaches generally assume a simple diagonal Gaussian prior and as a result are not able to reliably disentangle discrete factors of variation. We propose a two-level hierarchical objective to control relative degree of statistical independence between blocks of variables and individual variables within blocks. We derive this objective as a generalization of the evidence lower bound, which allows us to explicitly represent the trade-offs between mutual information between data and representation, KL divergence between representation and prior, and coverage of the support of the empirical data distribution. Experiments on a variety of datasets demonstrate that our objective can not only disentangle discrete variables, but that doing so also improves disentanglement of other variables and, importantly, generalization even to unseen combinations of factors.

Proceedings of the 22nd International Conference on Artificial Intelligence and Statistics (AISTATS) 2019, Naha, Okinawa, Japan. PMLR: Volume 89.

1 Introduction

Deep generative models represent data $\boldsymbol{x}$ using a low-dimensional variable $\boldsymbol{z}$ (sometimes referred to as a code). The relationship between $\boldsymbol{x}$ and $\boldsymbol{z}$ is described by a conditional probability distribution $p_\theta(\boldsymbol{x}|\boldsymbol{z})$ parameterized by a deep neural network. There have been many recent successes in training deep generative models for complex data types such as images [Gatys et al., 2015, Gulrajani et al., 2017], audio [Oord et al., 2016], and language [Bowman et al., 2016]. The latent code $\boldsymbol{z}$ can also serve as a compressed representation for downstream tasks such as text classification [Xu et al., 2017], Bayesian optimization [Gómez-Bombarelli et al., 2018, Kusner et al., 2017], and lossy image compression [Theis et al., 2017]. The setting in which an approximate posterior distribution $q_\phi(\boldsymbol{z}|\boldsymbol{x})$ is simultaneously learnt with the generative model via optimization of the evidence lower bound (ELBO) is known as a variational autoencoder (VAE), where $q_\phi(\boldsymbol{z}|\boldsymbol{x})$ and $p_\theta(\boldsymbol{x}|\boldsymbol{z})$ represent probabilistic encoders and decoders respectively. In contrast to VAE, inference and generative models can also be learnt jointly in an adversarial setting [Makhzani et al., 2015, Dumoulin et al., 2016, Donahue et al., 2016].

While deep generative models often provide high-fidelity reconstructions, the representation $\boldsymbol{z}$ is generally not directly amenable to human interpretation. In contrast to classical methods such as principal components or factor analysis, individual dimensions of $\boldsymbol{z}$ don't necessarily encode any particular semantically meaningful variation in $\boldsymbol{x}$. This has motivated a search for ways of learning *disentangled* representations, where perturbations of an individual dimension of the latent code $\boldsymbol{z}$ perturb the corresponding $\boldsymbol{x}$ in an interpretable manner. Various strategies for weak supervision have been employed, including semi-supervision of latent

$$\mathcal{L}(\theta,\phi) = \mathbb{E}_{q_\phi(\boldsymbol{z},\boldsymbol{x})}\left[\log\frac{p_\theta(\boldsymbol{x},\boldsymbol{z})}{p_\theta(\boldsymbol{x})p(\boldsymbol{z})} + \log\frac{q_\phi(\boldsymbol{z})q(\boldsymbol{x})}{q_\phi(\boldsymbol{z},\boldsymbol{x})} + \log\frac{p_\theta(\boldsymbol{x})}{q(\boldsymbol{x})} + \log\frac{p(\boldsymbol{z})}{q_\phi(\boldsymbol{z})}\right],$$

$$= \mathbb{E}_{q_\phi(\boldsymbol{z},\boldsymbol{x})}\left[\underbrace{\log\frac{p_\theta(\boldsymbol{x}\mid\boldsymbol{z})}{p_\theta(\boldsymbol{x})}}_{①} - \underbrace{\log\frac{q_\phi(\boldsymbol{z}\mid\boldsymbol{x})}{q_\phi(\boldsymbol{z})}}_{②}\right] - \underbrace{\mathrm{KL}\left(q(\boldsymbol{x})\mid\mid p_\theta(\boldsymbol{x})\right)}_{③} - \underbrace{\mathrm{KL}\left(q_\phi(\boldsymbol{z})\mid\mid p(\boldsymbol{z})\right)}_{④}.$$

Figure 1: *ELBO decomposition*. The VAE objective can be defined in terms of KL divergence between a generative model $p_\theta(\boldsymbol{x},\boldsymbol{z}) = p_\theta(\boldsymbol{x}\mid\boldsymbol{z})p(\boldsymbol{z})$ and an inference model $q_\phi(\boldsymbol{z},\boldsymbol{x}) = q_\phi(\boldsymbol{z}\mid\boldsymbol{x})q(\boldsymbol{x})$. We can decompose this objective into 4 terms. Term ①, which can be intuitively thought of as the uniqueness of the reconstruction, is regularized by the mutual information ②, which represents the uniqueness of the encoding. Minimizing the KL in term ③ is equivalent to maximizing the marginal likelihood $\mathbb{E}_{q(\boldsymbol{x})}[\log p_\theta(\boldsymbol{x})]$. Combined maximization of ① + ③ is equivalent to maximizing $\mathbb{E}_{q_\phi(\boldsymbol{z},\boldsymbol{x})}[\log p_\theta(\boldsymbol{x}\mid\boldsymbol{z})]$. Term ④ matches the inference marginal $q_\phi(\boldsymbol{z})$ to the prior $p(\boldsymbol{z})$, which in turn ensures realistic samples $\boldsymbol{x}\sim p_\theta(\boldsymbol{x})$ from the generative model.

variables [Kingma et al., 2014, Siddharth et al., 2017], triplet supervision [Karaletsos et al., 2015, Veit et al., 2016], or batch-level factor invariances [Kulkarni et al., 2015, Bouchacourt et al., 2017]. There has also been a concerted effort to develop fully unsupervised approaches that modify the VAE objective to induce disentangled representations. A well-known example is β-VAE [Higgins et al., 2016]. This has prompted a number of approaches that modify the VAE objective by adding, removing, or altering the weight of individual terms [Kumar et al., 2017, Zhao et al., 2017, Gao et al., 2018, Achille and Soatto, 2018].

In this paper, we introduce hierarchically factorized VAEs (HFVAEs). The HFVAE objective is based on a two-level hierarchical decomposition of the VAE objective, which allows us to control the relative levels of statistical independence between groups of variables and for individual variables in the same group. At each level, we induce statistical independence by minimizing the *total correlation* (TC), a generalization of the mutual information to more than two variables. A number of related approaches have also considered the TC [Kim and Mnih, 2018, Chen et al., 2018, Gao et al., 2018], but do not employ the two-level decomposition that we consider here. In our derivation, we reinterpret the standard VAE objective as a KL divergence between a generative model and its corresponding inference model. This has the side benefit that it provides a unified perspective on trade-offs in modifications of the VAE objective.

We illustrate the power of this decomposition by disentangling discrete factors of variation from continuous variables, which remains problematic for many existing approaches. We evaluate our methodology on a variety of datasets including dSprites, MNIST, Fashion MNIST (F-MNIST), CelebA and 20NewsGroups. Inspection of the learned representations confirms that our objective uncovers interpretable features in an unsupervised setting, and quantitative metrics demonstrate improvement over related methods. Crucially, we show that the learned representations can recover combinations of latent features that were not present in any examples in the training set, which has long been an implicit goal in learning disentangled representations that is now considered explicitly.

2 A Unified View of VAE Objectives

Variational autoencoders jointly optimize two models. The generative model $p_\theta(\boldsymbol{x},\boldsymbol{z})$ defines a distribution on a set of latent variables $\boldsymbol{z}$ and observed data $\boldsymbol{x}$ in terms of a prior $p(\boldsymbol{z})$ and a likelihood $p_\theta(\boldsymbol{x}\mid\boldsymbol{z})$, which is often referred to as the *decoder* model. This distribution is estimated in tandem with an *encoder*, a conditional distribution $q_\phi(\boldsymbol{z}\mid\boldsymbol{x})$ that performs approximate inference in this model. The encoder and decoder together define a probabilistic autoencoder.

The VAE objective is traditionally defined as sum over datapoints $\boldsymbol{x}^n$ of the expected value of the per-datapoint ELBO, or alternatively as an expectation over an empirical distribution $q(\boldsymbol{x})$ that approximates an unknown data distribution with a finite set of data points,

$$\mathcal{L}^{\mathit{VAE}}(\theta,\phi) := \mathbb{E}_{q(\boldsymbol{x})}\left[\mathbb{E}_{q_\phi(\boldsymbol{z}\mid\boldsymbol{x})}\left[\log\frac{p_\theta(\boldsymbol{x},\boldsymbol{z})}{q_\phi(\boldsymbol{z}\mid\boldsymbol{x})}\right]\right], \qquad q(\boldsymbol{x}) := \frac{1}{N}\sum_{n=1}^{N}\delta_{\boldsymbol{x}^n}(\boldsymbol{x}) \tag{1}$$

To better understand the various modifications of the VAE objective, which have often been introduced in an ad hoc manner, we here consider an alternate but equivalent definition of the VAE objective as a KL divergence between the generative model $p_\theta(\boldsymbol{x},\boldsymbol{z})$ and inference model $q_\phi(\boldsymbol{z},\boldsymbol{x}) = q(\boldsymbol{z}\mid\boldsymbol{x})q(\boldsymbol{x})$,

$$\begin{aligned}\mathcal{L}(\theta,\phi) &:= -\mathrm{KL}\left(q_\phi(\boldsymbol{z},\boldsymbol{x})\mid\mid p_\theta(\boldsymbol{x},\boldsymbol{z})\right) \\ &= \mathbb{E}_{q_\phi(\boldsymbol{z},\boldsymbol{x})}\left[\log\frac{p_\theta(\boldsymbol{x},\boldsymbol{z})}{q_\phi(\boldsymbol{z}\mid\boldsymbol{x})}\right] - \mathbb{E}_{q(\boldsymbol{x})}[\log q(\boldsymbol{x})].\end{aligned} \tag{2}$$

This definition differs from the expression in Equation (1) only by a constant term $\log N$, which is the entropy of the empirical data distribution $q(\boldsymbol{x})$. The advantage of this interpretation as a KL divergence is that it becomes

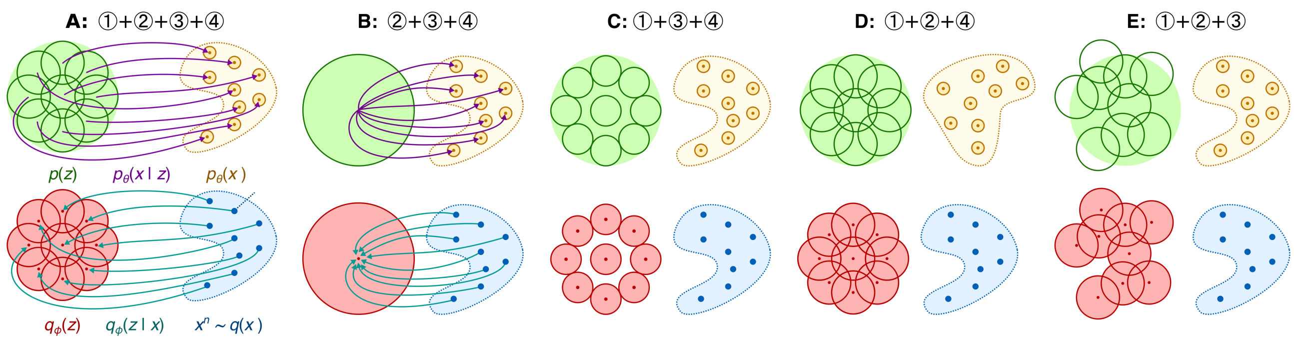

Figure 2: Illustration of the role of each term in the decomposition from Figure 1. **A:** Shows the default objective, whereas each subsequent figure shows the effect of removing one term from the objective. **B:** Removing ① means that we no longer require a unique $\boldsymbol{z}$ for each $\boldsymbol{x}^n$. Term ② will then minimize $I(\boldsymbol{x};\boldsymbol{z})$ which means that each $\boldsymbol{x}^n$ is mapped to the prior. **C:** Removing ② eliminates the constraint that $I(\boldsymbol{x};\boldsymbol{z})$ must be small under the inference model, causing each $\boldsymbol{x}^n$ to be mapped to a unique region in $\boldsymbol{z}$ space. **D:** Removing ③ eliminates the constraint that $p_\theta(\boldsymbol{x})$ must match $q(\boldsymbol{x})$. **E:** Removing ④ eliminates the constraint that $q_\phi(\boldsymbol{z})$ must match $p(\boldsymbol{z})$.

more apparent what it means to optimize the objective with respect to the generative model parameters θ and the inference model parameters ϕ. In particular, it is clear that the KL is minimized when $p_\theta(\boldsymbol{x},\boldsymbol{z}) = q_\phi(\boldsymbol{z},\boldsymbol{x})$, which in turn implies that marginal distributions on data $p_\theta(\boldsymbol{x}) = q(\boldsymbol{x})$ and latent code $q_\phi(\boldsymbol{z}) = p(\boldsymbol{z})$ must also match. We will refer to $q_\phi(\boldsymbol{z})$, as the *inference marginal*, which is the average over the data of the encoder distribution $q_\phi(\boldsymbol{z}) = \int q_\phi(\boldsymbol{z},\boldsymbol{x})\, d\boldsymbol{x} = \frac{1}{N}\sum_{n=1}^{N} q_\phi(\boldsymbol{z} \mid \boldsymbol{x}^n)$.

To more explicitly represent the trade-offs that are implicit in optimizing the VAE objective, we perform a decomposition (Figure 1) similar to the one obtained by Hoffman and Johnson [2016]. This decomposition yields 4 terms. Terms ③ and ④ enforce consistency between the marginal distributions over $\boldsymbol{x}$ and $\boldsymbol{z}$. Minimizing the KL in term ③ maximizes the marginal likelihood $\mathbb{E}_{q(\boldsymbol{x})}[\log p_\theta(\boldsymbol{x})]$, whereas minimizing ④ ensures that the inference marginal $q_\phi(\boldsymbol{z})$ approximates the prior $p(\boldsymbol{z})$. Terms ① and ② enforce consistency between the conditional distributions. Intuitively speaking, term ① maximizes the identifiability of the values $\boldsymbol{z}$ that generate each $\boldsymbol{x}^n$; when we sample $\boldsymbol{z} \sim q_\phi(\boldsymbol{z} \mid \boldsymbol{x}^n)$, then the likelihood $p_\theta(\boldsymbol{x}^n \mid \boldsymbol{z})$ under the generative model should be *higher* than the marginal likelihood $p_\theta(\boldsymbol{x}^n)$. Term ② regularizes term ① by minimizing the mutual information $I(\boldsymbol{z};\boldsymbol{x})$ in the inference model, which means that $q_\phi(\boldsymbol{z} \mid \boldsymbol{x}^n)$ maps each $\boldsymbol{x}^n$ to less identifiable values.

Note that term ① is intractable in practice, since we are not able to pointwise evaluate $p_\theta(\boldsymbol{x})$. We can circumvent this intractability by combining ① + ③ into a single term, which recovers the likelihood

$$\begin{aligned}&\operatorname*{argmax}_{\theta,\phi} \mathbb{E}_{q_\phi(\boldsymbol{z},\boldsymbol{x})}\left[\log\frac{p_\theta(\boldsymbol{x} \mid \boldsymbol{z})}{p_\theta(\boldsymbol{x})} + \log\frac{p_\theta(\boldsymbol{x})}{q(\boldsymbol{x})}\right] \\ &= \operatorname*{argmax}_{\theta,\phi} \mathbb{E}_{q_\phi(\boldsymbol{z},\boldsymbol{x})}\left[\log p_\theta(\boldsymbol{x} \mid \boldsymbol{z})\right].\end{aligned}$$

To build intuition for the impact of each of these terms, Figure 2 shows the effect of removing each term from the objective. When we remove ③ or ④ we can learn models in which $p_\theta(\boldsymbol{x})$ deviates from $q(\boldsymbol{x})$, or $q_\phi(\boldsymbol{z})$ deviates from $p(\boldsymbol{z})$. When we remove ①, we eliminate the requirement that $p_\theta(\boldsymbol{x}^n \mid \boldsymbol{z})$ should be higher when $\boldsymbol{z} \sim q_\phi(\boldsymbol{z} \mid \boldsymbol{x}^n)$ than when $\boldsymbol{z} \sim p(\boldsymbol{z})$. Provided the decoder model is sufficiently expressive, we would then learn a generative model that ignores the latent code $\boldsymbol{z}$. This undesirable type of solution does in fact arise in certain cases, even when ① is included in the objective, particularly when using auto-regressive decoder architectures [Chen et al., 2016b].

When we remove ②, we learn a model that minimizes the overlap between $q_\phi(\boldsymbol{z} \mid \boldsymbol{x}^n)$ for different data points $\boldsymbol{x}^n$, in order to maximize ①. This maximizes the mutual information $I(\boldsymbol{x};\boldsymbol{z})$, which is upper-bounded by $\log N$. In practice ② often saturates to $\log N$, even when included in the objective, which suggests that maximizing ① outweighs this cost, at least for the encoder/decoder architectures that are commonly considered in present-day models.

3 Hierarchically Factorized VAEs

In this paper, we are interested in defining an objective that will encourage statistical independence between features. The β-VAE objective [Higgins et al., 2016] aims to achieve this goal by defining the objective

$$\begin{aligned}\mathcal{L}^{\beta\text{-VAE}}(\theta,\phi) = \mathbb{E}_{q(\boldsymbol{x})}\Big[&\mathbb{E}_{q_\phi(\boldsymbol{z}|\boldsymbol{x})}\left[\log p_\theta(\boldsymbol{x}|\boldsymbol{z})\right] \\ &- \beta \mathrm{KL}\left(q_\phi(\boldsymbol{z}|\boldsymbol{x}) \,||\, p(\boldsymbol{z})\right)\Big].\end{aligned}$$

We can express this objective in the terms of Figure 1 as ① + ③ + β (② + ④). In order to induce disentangled representations, the authors set $\beta > 1$. This works well in certain cases, but it has the drawback that it also increases the strength of ②, which means that the encoder model may discard more information about $\boldsymbol{x}$ in order to minimize the mutual information $I(\boldsymbol{x};\boldsymbol{z})$.

Looking at the β-VAE objective, it seems intuitive that

$$
\begin{aligned}
-\mathrm{KL}\left(q_\phi(\boldsymbol{z}) \,||\, p(\boldsymbol{z})\right) &= -\mathbb{E}_{q_\phi(\boldsymbol{z})}\left[\log \frac{q_\phi(\boldsymbol{z})}{\prod_d q_\phi(\boldsymbol{z}_d)} + \log \frac{\prod_d q_\phi(\boldsymbol{z}_d)}{\prod_d p(\boldsymbol{z}_d)} + \log \frac{\prod_d p(\boldsymbol{z}_d)}{p(\boldsymbol{z})}\right] \\
&= \underbrace{\mathbb{E}_{q_\phi(\boldsymbol{z})}\left[\log \frac{p(\boldsymbol{z})}{\prod_d p(\boldsymbol{z}_d)} - \log \frac{q_\phi(\boldsymbol{z})}{\prod_d q_\phi(\boldsymbol{z}_d)}\right]}_{\text{Ⓐ}} - \sum_d \underbrace{\mathrm{KL}\left(q_\phi(\boldsymbol{z}_d) \,||\, p(\boldsymbol{z}_d)\right)}_{\text{Ⓑ}}.
\end{aligned}
$$

$$
\text{Ⓑ} = \underbrace{\mathbb{E}_{q_\phi(\boldsymbol{z}_d)}\left[\log \frac{p(\boldsymbol{z}_d)}{\prod_e p(\boldsymbol{z}_{d,e})} - \log \frac{q_\phi(\boldsymbol{z}_d)}{\prod_e q_\phi(\boldsymbol{z}_{d,e})}\right]}_{\text{ⓘ}} - \sum_e \underbrace{\mathrm{KL}\left(q_\phi(\boldsymbol{z}_{d,e}) \,||\, p(\boldsymbol{z}_{d,e})\right)}_{\text{ⓘⓘ}}
$$

Figure 3: *Hierarchical KL decomposition*. We can decompose term ④ into subcomponents Ⓐ and Ⓑ. Term Ⓐ matches the total correlation between variables in the inference model relative to the total correlation under the generative model. Term Ⓑ minimizes the KL divergence between the inference marginal and prior marginal for each variable $\boldsymbol{z}_d$. When the variable $\boldsymbol{z}_d$ contains sub-variables $\boldsymbol{z}_{d,e}$, we can recursively decompose the KL on the marginals $\boldsymbol{z}_d$ into term ⓘ, which matches the total correlation, and term ⓘⓘ, which minimizes the per-dimension KL divergence.

increasing the weight of term ④ is likely to aid disentanglement. One notion of disentanglement is that there should be a low degree of correlation between different latent variables $\boldsymbol{z}_d$. If we choose a mean-field prior $p(\boldsymbol{z}) = \prod_d p(\boldsymbol{z}_d)$, then minimizing the KL term should induce an inference marginal $q_\phi(\boldsymbol{z}) = \prod_d q_\phi(\boldsymbol{z}_d)$ in which $\boldsymbol{z}_d$ are also independent. However, in addition to being sensitive to correlations, the KL will also be sensitive to discrepancies in the shape of the distribution. When our primary interest is to disentangle representations, then we may wish to relax the constraint that the shape of the distribution matches the prior in favor of enforcing statistical independence.

To make this intuition explicit, we decompose ④ into two terms Ⓐ and Ⓑ (Figure 3). As with term ① + ②, term Ⓐ consists of two components. The second of these takes the form of a total correlation, which is the generalization of the mutual information to more than two variables,

$$
\begin{aligned}
TC(\boldsymbol{z}) &= \mathbb{E}_{q_\phi(\boldsymbol{z})}\left[\log \frac{q_\phi(\boldsymbol{z})}{\prod_d q_\phi(\boldsymbol{z}_d)}\right] \\
&= \mathrm{KL}\Big(q_\phi(\boldsymbol{z}) \Big|\Big| \prod_d q_\phi(\boldsymbol{z}_d)\Big)
\end{aligned} \tag{3}
$$

Minimizing the total correlation yields a $q_\phi(\boldsymbol{z})$ in which different $\boldsymbol{z}_d$ are statistically independent, hereby providing a possible mechanism for inducing disentanglement. In cases where $\boldsymbol{z}_d$ itself represents a group of variables, rather than a single variable, we can continue to decompose to another set of terms ⓘ and ⓘⓘ which match the total correlation for $\boldsymbol{z}_d$ and the KL divergences for constituent variables $\boldsymbol{z}_{d,e}$. This provides an opportunity to induce hierarchies of disentangled features. We can in principle continue this decomposition for any number of levels to define an hierarchically factorized VAE (HFVAE) objective. We here restrict ourselves to the two-level case, which corresponds to an objective of the form

$$
\mathcal{L}^{\text{HFVAE}}(\theta, \phi) = ① + ③ + ⓘⓘ + \alpha\, ② + \beta\, Ⓐ + \gamma\, ⓘ. \tag{4}
$$

In this objective, α controls the $I(\boldsymbol{x}; \boldsymbol{z})$ regularization, β controls the TC regularization between groups of variables, and γ controls the TC regularization within groups. This objective is similar to, but more general than, the one recently proposed by Kim and Mnih [2018] and Chen et al. [2018]. Our objective admits these objectives as a special case corresponding to a non-hierarchical decomposition in which $\beta = \gamma$. The first component of Ⓐ is not present in these objectives, which implicitly assume that $p(\boldsymbol{z}) = \prod_d p(\boldsymbol{z}_d)$. In the more general case where $p(\boldsymbol{z}) \neq \prod_d p(\boldsymbol{z}_d)$, maximizing Ⓐ with respect to ϕ will match the total correlation in $q(\boldsymbol{z})$ to that in $p(\boldsymbol{z})$.

3.1 Approximation of the Objective

In order to optimize this objective, we need to approximate the inference marginals $q_\phi(\boldsymbol{z})$, $q_\phi(\boldsymbol{z}_d)$, and $q_\phi(\boldsymbol{z}_{d,e})$. Computing these quantities exactly requires a full pass over the dataset, since $q_\phi(\boldsymbol{z})$ is a mixture over all data points in the training set. We approximate $q_\phi(\boldsymbol{z})$ with a Monte Carlo estimate $\hat{q}_\phi(\boldsymbol{z})$ over the same batch of samples that we use to approximate all other terms in the objective $\mathcal{L}^{\text{HFVAE}}(\theta, \phi)$. For simplicity we will consider the term

$$
\begin{gathered}
\mathbb{E}_{q_\phi(\boldsymbol{z},\boldsymbol{x})}[\log q_\phi(\boldsymbol{z})] \simeq \frac{1}{B}\sum_{b=1}^{B} \log q_\phi(\boldsymbol{z}^b), \\
\boldsymbol{z}^b \sim q_\phi(\boldsymbol{z} \mid \boldsymbol{x}^b), \quad b \sim \text{Uniform}(1, \ldots, N)
\end{gathered} \tag{5}
$$

We define the estimate of $\hat{q}_\phi(\boldsymbol{z}^b)$ as (see Appendix A)

$$
\hat{q}_\phi(\boldsymbol{z}^b) := \frac{1}{N} q_\phi(\boldsymbol{z}^b \mid \boldsymbol{x}^b) + \frac{N-1}{N(B-1)} \sum_{b' \neq b} q_\phi(\boldsymbol{z}^b \mid \boldsymbol{x}^{b'}).
$$

This estimator differs from the one in Kim and Mnih [2018], which is based on adversarial-style estimation of the density ratio, and is also distinct from the estimators in Chen et al. [2018] who employ different approximations. We can think of this approximation as a partially stratified sample, in which we deterministically include the term $\boldsymbol{x}^n = \boldsymbol{x}^b$ and compute a Monte Carlo estimate over the remaining terms, treating indices $b' \neq b$ as samples from the distribution $q(\boldsymbol{x} \mid \boldsymbol{x} \neq \boldsymbol{x}^b)$. We now substitute $\log \hat{q}_\phi(\boldsymbol{z})$ for $\log q_\phi(\boldsymbol{z})$ in Equation (5). By Jensen's inequality this yields a lower bound on the original expectation. While this induces a bias, the estimator is consistent. In practice, the bias is likely to be small given the batch sizes (512-1024) needed to approximate the inference marginal.

4 Related Work

In addition to the work of Kim and Mnih [2018] and Chen et al. [2018], our objective is also related to, and generalizes, a number of recently-proposed modifications to the VAE objective (see Table 1 for an overview). Zhao et al. [2017] considers an objective that eliminates the mutual information in ② entirely and assigns an additional weight to the KL divergence in ④. Kumar et al. [2017] approximate the KL divergence in ④ by matching the covariance of $q_\phi(\boldsymbol{z})$ and $p(\boldsymbol{z})$. Recent work by Gao et al. [2018] connects VAEs to the principle of correlation explanation, and defines an objective that reduces the mutual information regularization in ② for a subset of "Anchor" variables $\boldsymbol{z}_a$. Achille and Soatto [2018] interpret VAEs from an information-bottleneck perspective and introduce an additional TC term into the objective. In addition to VAEs, generative adversarial networks (GANs) have also been used to learn disentangled representations. The InfoGAN [Chen et al., 2016a] achieves disentanglement by *maximizing* the mutual information between individual features and the data under the *generative model*.

In settings where we are not primarily interested in inducing disentangled representations, the β-VAE objective has also been used with $\beta < 1$ in order to improve the quality of reconstructions [Alemi et al., 2016, Engel et al., 2017, Liang et al., 2018]. While this also decreases the relative weight of ②, in practice it does not influence the learned representation in cases where $I(\boldsymbol{x}; \boldsymbol{z})$ saturates anyway. The tradeoff between likelihood and the KL term and the influence of penalizing the KL term on mutual information has been studied more in depth in Alemi et al. [2018], Burgess et al. [2018]. In other recent work, Dupont [2018] considered models containing both Concrete and Gaussian variables. However, the objective was not decomposed to get Ⓐ, but was based on the objective proposed by Burgess et al. [2018].

5 Experiments

To assess the quality of disentangled representations that the HFVAE induces, we evaluate a number of tasks and datasets. We consider *CelebA* [Liu et al., 2015] and *dSprites* [Higgins et al., 2016] as exemplars of datasets that are typically used to demonstrate general-purpose disentangling. As specific examples of datasets that require a discrete variable, we consider *MNIST* [LeCun et al., 2010] and *F-MNIST* [Xiao et al., 2017]. Finally, we consider an example that extends beyond image-based domains by using the HFVAE objective to train neural topic models on the *20NewsGroups* [Lang, 2007] dataset. We compare a number of objectives and priors on these datasets, including the standard VAE objective [Kingma and Welling, 2013, Rezende et al., 2014], the β-VAE objective [Higgins et al., 2016], the β-TCVAE objective [Chen et al., 2018], and our HFVAE objective.

Paper	**Objective**
Kingma and Welling [2013], Rezende et al. [2014]	① + ② + ③ + ④
Higgins et al. [2016]	① + ③ + β (② + ④)
Kumar et al. [2017]	① + ② + ③ + λ ④
Zhao et al. [2017]	① + ③ + λ ④
Alemi et al. [2018], Burgess et al. [2018]	① + ③ + γ|(② + ④) $- C$|
Gao et al. [2018]	① + ② + ③ + ④ $- \lambda$ ②[a]
Achille and Soatto [2018]	① + ③ + β ② + γ Ⓐ*
Kim and Mnih [2018], Chen et al. [2018]	① + ② + ③ + Ⓑ + β Ⓐ*
HFVAE (this paper)	① + ③ + ⓘⓘ + α ② + β Ⓐ + γ ⓘ

Table 1: Comparison of objectives in autoencoding deep generative models. The asterisk Ⓐ* indicates that the prior factorizes, i.e. $p(\boldsymbol{z}) = \prod_d p(\boldsymbol{z}_d)$. The notation ②[a] refers to restriction of the mutual information ② to a subset of "Anchor" variables $\boldsymbol{z}_a$.

A crucial feature of our experiments is the realization that HFVAE objective serves to induce representations in which correlations in the inference marginal q(z) match correlations in the prior p(z). The two-level decomposition allows us to control the level at which we induce independence. For example, when considering appropriate priors for the MNIST and F-MNIST data, we can take into account the fact that they each contain 10 explicit classes. Likewise for dSprites, containing 3 explicit shape classes. We model these distinct classes using a Concrete distribution [Maddison et al., 2017, Jang et al., 2017] of appropriate dimension. We can also assume that these datasets have a multidimensional continuous style variable. HFVAE allows us to induce a stronger independence between the style and the class, while allowing some correlation between the individual style dimensions. We can also use the two-level decomposition to *induce* correlations between variables by *decreasing* the strength of Ⓐ. As an example, we consider the task of uncovering correlations between topics.

A full list of priors employed is given in Table 2, and the associated model architectures are described in Appendix C. Note that in connection to Figure 3, subscript 'd' refers to the variables class or style represented by Concrete and normal distributions respectively, while subscript 'e' refers to individual dimension *within* the normal variable. In all models, we use a single implementation of the objective based on the Probabilistic Torch [Siddharth et al., 2017]

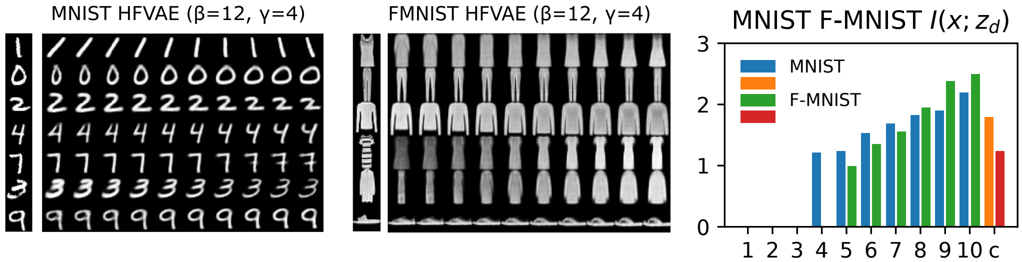

Figure 4: Left: MNIST and F-MNIST reconstructions for $\boldsymbol{z}_d \in (-3, 3)$. Rows contain both different samples from data and different dimensions d. Right: The mutual information ② for each individual dimension $I(\boldsymbol{x}; \boldsymbol{z}_d)$, ranked in ascending order, with the Concrete variable shown last. The HFVAE *prunes* 3 continuous dimensions in MNIST and 4 in F-MNIST.

library for deep generative models[1].

5.1 Qualitative and Quantitative Evaluation

We begin with a qualitative evaluation of the features that are identified when training with the HFVAE objective. Figure 4 shows results for MNIST and F-MNIST datasets. For the MNIST dataset the representation recovers 7 distinct interpretable features—slant, width, height, openness, stroking, thickness, and roundness, while choosing to ignore the remaining dimensions available. Observing the mutual information $I(\boldsymbol{x}; \boldsymbol{z}_d)$ between the data $\boldsymbol{x}$ and individual dimensions of the latent space $\boldsymbol{z}_d$ confirms the separation of latent space into useful and ignored subspaces.

As can be seen from Figure 4(right), the mutual information drops to zero for the unused dimensions. We observe a similar trend for the F-MNIST and CelebA datasets, recovering distinct interpretable factors. For F-MNIST (Figure 4(mid)), this can be width, length, brightness, etc. For CelebA (Appendix Figure 8), we uncover interpretable features such as the orientation of the face, variation from smiling to non-smiling, and the presence of sunglasses.

As a quantitative assessment of the quality of learned representations, we evaluate the metrics proposed by Kim and Mnih [2018] and Eastwood and Williams [2018] on the dSprites dataset with 10 random restarts for each model. For the Eastwood and Williams metric, we used a random forest to regress from features to the (known) ground-truth factors for the data. In Table 3, we list these metrics on the dSprites dataset for each of the model types and objectives defined above, noting that the HFVAE performs similar to other approaches.

[1]https://github.com/probtorch/probtorch

	VAE	HFVAE	
	Normal	Normal	Concrete
MNIST	10	10	10
F-MNIST	10	10	10
dSprites	10	10	3
CelebA	20	20	2

Table 2: Dimensionality of latent variables.

We found in our experiments that HFVAE, similar to other disentanglement objectives, is also fairly sensitive to the choice of hyperparameters and the starting random seed. While it is possible to achieve some level of disentanglement with a variety of β and γ values, achieving a clean disentanglement between the discrete and continuous variables remains a fairly difficult task and is mostly influenced by the starting random seed rather than the choice of hyperparameters. For more details on the issue of hyperparameter sensitivity, see Appendix Section E.

5.2 Disentangled Representations for Text

Research on disentangled representations has thus far almost exclusively considered visual (image) data. Exploration of disentangled representations for text is still in its infancy, and generally relies on weak supervision [Jain et al., 2018]. To assess whether the HFVAE objective can aid interpretability in textual domains, we consider documents in the 20NewsGroups dataset, containing e-mail messages from a number of internet newsgroups; some groups are more closely related (e.g. religion vs politics) whereas others are not related at all (e.g. science vs sports). We analyze this data using ProdLDA [Srivastava and Sutton, 2017] and neural variational document model (NVDM) [Miao et al., 2016], both of which are autoencoding neural topic models. ProdLDA approximates a Dirichlet prior using samples from a Gaussian that are normalized with a softmax function, and then combines this prior with log-linear likelihood model for the words. NVDM includes a MLP encoder and a log-linear decoder with the assumption of a Gaussian prior.

To evaluate whether HFVAEs are able to uncover these correlations between topics, we compare a normal ProdLDA model with 50 topics, which we train with a standard VAE

Model	Kim	Eastwood
VAE	0.63 ±0.06	0.30 ±0.10
β-VAE ($\beta = 4.0$)	0.63 ±0.10	0.41 ±0.11
β-TCVAE ($\beta = 4.0$)	0.62 ±0.07	0.29 ±0.10
HFVAE ($\beta = 4.0$, $\gamma = 3.0$)	0.63 ±0.08	0.39 ±0.16

Table 3: Disentanglement scores (± one standard deviation) for the dSprites dataset using the metrics proposed by Kim and Mnih [2018] and Eastwood and Williams [2018].

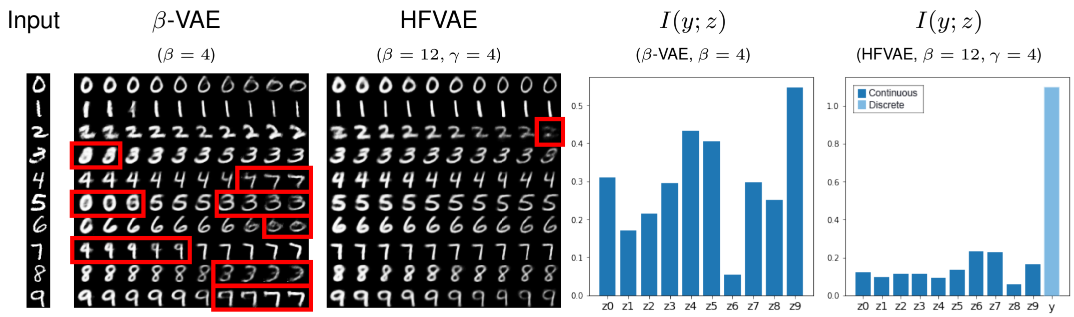

Figure 5: Left: Manipulation of the thickness variable over the range -3 to 3. The β-VAE is not able to maintain digit identity as we vary thickness. The HFVAE, which incorporates a discrete variable into the prior is able to maintain the digit identity across the entire range. Right: Mutual information between label $\boldsymbol{y}$ and individual dimensions of $\boldsymbol{z}$.

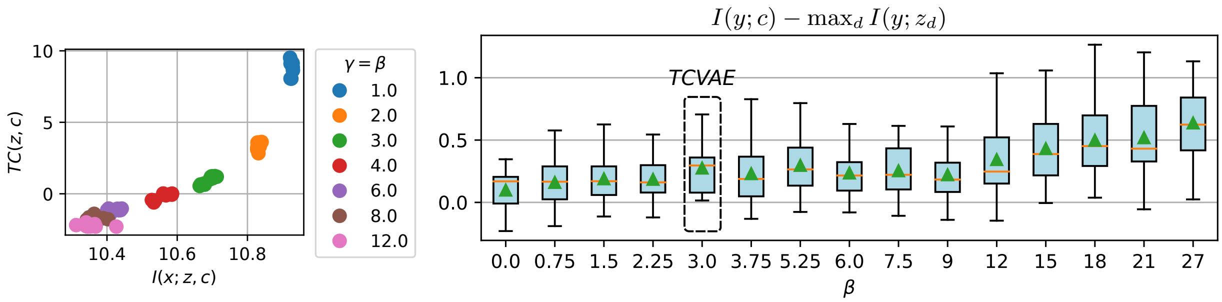

Figure 6: On the left, we plot $I(x; \{z, c\})$ vs $TC(\{z, c\})$ for 50 restarts with 7 values $\gamma = \beta$. On the right, we keep $\gamma = 3$ fixed and vary β, plotting mutual information gap (as proposed by Chen et al. [2018]) between the Concrete and the largest continuous variable. While there is generally no "best" value for γ, $TC(\{z, c\})$ decreases by 5 nats over range $\gamma = \beta = \{2, 3, 4, 6\}$ while $I(x; \{z, c\})$ decreases by 0.5 nats. Moreover, increasing β relative to γ also increases the mutual information gap.

objective, to a HFVAE implementation in which we assume two latent variables with 25 dimensions each. We use $\beta = 0.1$ and $\gamma = 4.0$. This means that we *relax* the constraint on the TC at the group level. In other words, the HFVAE model should learn two groups of 25 topics in a manner that *allows* correlations between topics in *different* groups, and *prevents* correlations within a group.

The HFVAE learns correlations between topics that are distinct yet not unrelated, e.g. religion and politics in the middle east, whereas the VAE does not uncover any significant correlations between topics. As in the MNIST and F-MNIST examples, there is also a significant degree of pruning. The HFVAE learns 11 topics with a significant mutual information $I(\boldsymbol{x}; \boldsymbol{z}_d)$, whereas the VAE has a nonzero mutual information for all 50 topics. See Appendix Section D for the full results.

5.3 Unsupervised Learning of Discrete Labels

An interesting question relating to the use of discrete latent variables in our framework is how effective these variables are at improving disentanglement. In general, a discrete variable over K dimensions expresses a sparsity constraint over those dimensions, as any choice made from that variable is constrained to lie on the vertices of the K-dimensional simplex it represents. This interpretation broadly carries over even in the case of continuous relaxations such as the Concrete distribution that we employ to enable reparameterization for gradient-based methods.

The ability of our framework to disentangle discrete factors of variation *using* the discrete variables depends then, on how separable the underlying classes or identities actually are. For example, in the case of F-MNIST, a number of classes (i.e., shoes, trousers, dresses, etc.) are visually distinctive enough that they can be captured faithfully by the discrete latent variable. However, in the case of dSprites, while the shapes (squares, ellipses, and hearts) are conceptually distinct, visually, they can often be difficult to distinguish, with actual differences in the scale of a few pixels. Here, our approach is less effective at disentangling the relevant factors with the discrete latent. The MNIST dataset lies in the middle of these two extremes, allowing for clear separation in terms of the digits themselves, but also blurring the lines partially with how similar some of them may appear—a 9 can appear quite similar to a 4, for example, under some small perturbation.

Figure 5 showcases the ability of the HFVAE to disentangle discrete latent variables from continuous factors of variation in an unsupervised manner. Here, we have sampled a single data point for each digit and vary the "thickness"-encoding dimension for each of the β-VAE and HFVAE models. Clearly, HFVAE does a better job at disentangling digit vs. thickness. To quantify this ability, as before, we measure the mutual information $I(\boldsymbol{x}; \boldsymbol{z}_d)$ between the label

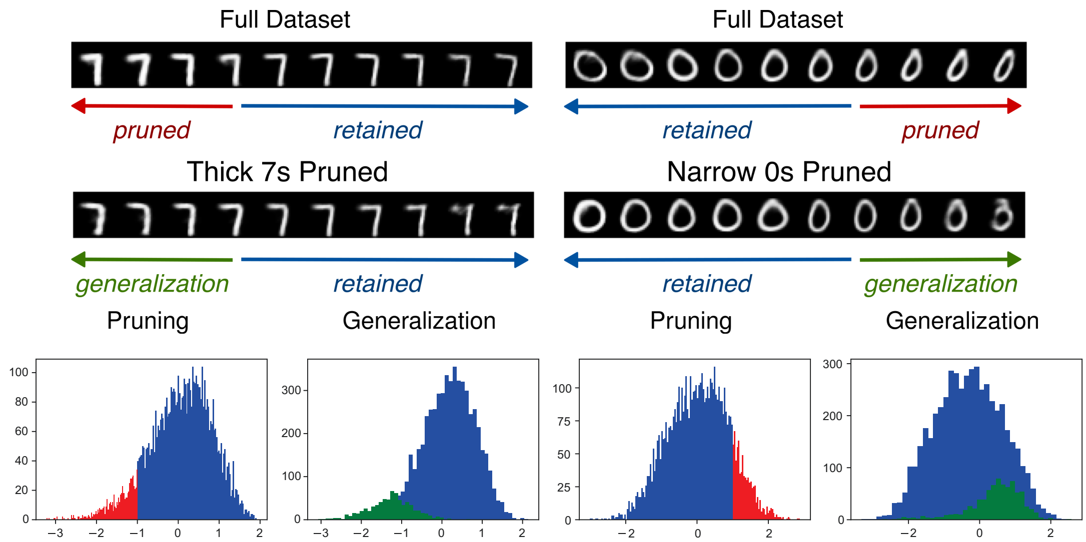

Figure 7: Generalization to unseen combinations of factors. A HFVAE is trained on the full dataset, and then retrained after a subset of the data is pruned. We then test generalization on the removed portion of the data.

and each latent dimension, and confirm that the HFVAE learns to encode the information on digits in the discrete variable. In the case of the β-VAE, this information is less clearly captured, spread out across the available dimensions.

5.4 Hyperparameter Analysis

Here, we analyze the influence of γ and β on the trade-off between total correlation and mutual information. We also show why setting $\gamma \neq \beta$ is necessary for better disentanglement, in comparison to previous objectives [Chen et al., 2018, Kim and Mnih, 2018] where $\gamma = \beta$. We investigate both these facets on the MNIST dataset. Figure 6 (left) shows the results of running our model for 50 restarts, for each of seven different values of $\gamma = \beta$, indicating positive correlation between mutual information $I(x; z, c)$ and total correlation $TC(z, c)$. In this figure, the ideal is the lower-right corner, corresponding to high mutual information and low total correlation. Figure 6 (right) shows how the mutual information gap (MIG), defined by Chen et al. [2018] to be $I(\boldsymbol{y}; c) - max_i I(\boldsymbol{y}; \boldsymbol{z}_i)$, varies as a function of β, for a fixed $\gamma = 3$. We observe that higher values of β result in the concrete variable better capturing the label information.

5.5 Zero-shot Generalization

A particular feature of disentangled representations is their *utility*, with evidence from human cognition Lake et al. [2017], suggesting that learning independent factors can aid in generalization to previously unseen combinations of factors. For example, one can imagine a pink elephant even if one has (sadly) not encountered such an entity previously.

To evaluate if the representations learned using HFVAEs exhibit such properties, we introduce a novel measure of disentanglement quality. Having first trained a model with the chosen data and objective, here MNIST and the HFVAE, we prune the dataset, removing data containing some particular combinations of factors, say images depicting a thick number 7, or a narrow 0. We then re-train with the modified dataset, using the pruned data as unseen test data.

Figure 7 shows the results of this experiment. As can be seen, the model trained on pruned data is able to successfully reconstruct digits with values for the stroke and character width that were never encountered during training. The histograms for the feature values show the ability of HFVAE to correctly encode features from previously unseen examples.

6 Discussion

Much of the work on learning disentangled representations thus far has focused on cases where the factors of variation are uncorrelated scalar variables. As we begin to apply these techniques to real world datasets, we are likely to encounter correlations between latent variables, particularly when there are causal dependencies between them. This work is a first step towards learning of more structured disentangled representations. By enforcing statistical independence between groups of variables, or relaxing this constraint, we now have the capability to disentangle variables that have higher-dimensional representations. An avenue of future work is to develop datasets that allow us to more rigorously test our ability to characterize correlations between higher-dimensional variables.

Acknowledgements

We would like to thank our reviewers and the area chair for their thoughtful comments. This work was supported by NSF award 1835309, NIH grant R01CA199673 from NCI, and MSKCC's Cancer Center core support NIH grant P30CA008748 from NCI. NS was supported by ERC grant ERC-2012-AdG 321162-HELIOS, EPSRC grant Seebibyte EP/M013774/1 and EPSRC/MURI grant EP/N019474/1. BP was supported by The Alan Turing Institute under the EPSRC grant EP/N510129/1. JWM would additionally like to acknowledge generous support from the Intel Corporation, the 3M Corporation, and startup funds from Northeastern University.

References

A. Achille and S. Soatto. Information Dropout: Learning Optimal Representations Through Noisy Computation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PP(99):1–1, 2018.

Alexander Alemi, Ben Poole, Ian Fischer, Joshua Dillon, Rif A Saurous, and Kevin Murphy. Fixing a broken elbo. In *International Conference on Machine Learning*, pages 159–168, 2018.

Alexander A. Alemi, Ian Fischer, Joshua V. Dillon, and Kevin Murphy. Deep Variational Information Bottleneck. *arXiv:1612.00410 [cs, math]*, December 2016.

Diane Bouchacourt, Ryota Tomioka, and Sebastian Nowozin. Multi-level variational autoencoder: Learning disentangled representations from grouped observations. *arXiv preprint arXiv:1705.08841*, 2017.

Samuel R Bowman, Luke Vilnis, Oriol Vinyals, Andrew M Dai, Rafal Jozefowicz, and Samy Bengio. Generating sentences from a continuous space. *CoNLL 2016*, page 10, 2016.

Christopher P Burgess, Irina Higgins, Arka Pal, Loic Matthey, Nick Watters, Guillaume Desjardins, and Alexander Lerchner. Understanding disentangling in β-vae. *arXiv preprint arXiv:1804.03599*, 2018.

Tian Qi Chen, Xuechen Li, Roger Grosse, and David Duvenaud. Isolating sources of disentanglement in variational autoencoders. *arXiv preprint arXiv:1802.04942*, 2018.

Xi Chen, Yan Duan, Rein Houthooft, John Schulman, Ilya Sutskever, and Pieter Abbeel. Infogan: Interpretable representation learning by information maximizing generative adversarial nets. In *Advances in Neural Information Processing Systems*, pages 2172–2180, 2016a.

Xi Chen, Diederik P. Kingma, Tim Salimans, Yan Duan, Prafulla Dhariwal, John Schulman, Ilya Sutskever, and Pieter Abbeel. Variational lossy autoencoder. *arXiv preprint arXiv:1611.02731*, 2016b.

Jeff Donahue, Philipp Krähenbühl, and Trevor Darrell. Adversarial feature learning. *arXiv preprint arXiv:1605.09782*, 2016.

Vincent Dumoulin, Ishmael Belghazi, Ben Poole, Olivier Mastropietro, Alex Lamb, Martin Arjovsky, and Aaron Courville. Adversarially learned inference. *arXiv preprint arXiv:1606.00704*, 2016.

Emilien Dupont. Joint-vae: Learning disentangled joint continuous and discrete representations. *arXiv preprint arXiv:1804.00104*, 2018.

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

Jesse Engel, Matthew Hoffman, and Adam Roberts. Latent Constraints: Learning to Generate Conditionally from Unconditional Generative Models. *arXiv:1711.05772 [cs, stat]*, November 2017.

Shuyang Gao, Rob Brekelmans, Greg Ver Steeg, and Aram Galstyan. Auto-encoding total correlation explanation. *arXiv preprint arXiv:1802.05822*, 2018.

Leon A Gatys, Alexander S Ecker, and Matthias Bethge. A neural algorithm of artistic style. *arXiv preprint arXiv:1508.06576*, 2015.

Rafael Gómez-Bombarelli, Jennifer N Wei, David Duvenaud, José Miguel Hernández-Lobato, Benjamín Sánchez-Lengeling, Dennis Sheberla, Jorge Aguilera-Iparraguirre, Timothy D Hirzel, Ryan P Adams, and Alán Aspuru-Guzik. Automatic chemical design using a data-driven continuous representation of molecules. *ACS Central Science*, 2018.

Ishaan Gulrajani, Kundan Kumar, Faruk Ahmed, Adrien Ali Taiga, Francesco Visin, David Vazquez, and Aaron Courville. PixelVAE: A latent variable model for natural images. In *International Conference on Machine Learning*, 2017.

Irina Higgins, Loic Matthey, Arka Pal, Christopher Burgess, Xavier Glorot, Matthew Botvinick, Shakir Mohamed, and Alexander Lerchner. beta-VAE: Learning basic visual concepts with a constrained variational framework. In *International Conference on Learning Representations*, 2016.

Matthew D. Hoffman and Matthew J. Johnson. Elbo surgery: Yet another way to carve up the variational evidence lower bound. In *Workshop in Advances in Approximate Bayesian Inference, NIPS*, 2016.

Sarthak Jain, Edward Banner, Jan-Willem van de Meent, Iain J. Marshall, and Byron C. Wallace. Learning Disentangled Representations of Texts with Application to Biomedical Abstracts. *arXiv:1804.07212 [cs]*, April 2018.

Eric Jang, Shixiang Gu, and Ben Poole. Categorical reparameterization with Gumbel-softmax. In *International Conference on Learning Representations*, 2017.

Theofanis Karaletsos, Serge Belongie, and Gunnar Rätsch. Bayesian representation learning with oracle constraints. *arXiv preprint arXiv:1506.05011*, 2015.

Hyunjik Kim and Andriy Mnih. Disentangling by factorising. *arXiv preprint arXiv:1802.05983*, 2018.

Diederik P. Kingma and Max Welling. Auto-encoding variational bayes. In *International Conference on Learning Representations*, 2013.

Diederik P Kingma, Shakir Mohamed, Danilo Jimenez Rezende, and Max Welling. Semi-supervised learning with deep generative models. In *Advances in Neural Information Processing Systems*, pages 3581–3589, 2014.

Tejas D Kulkarni, William F Whitney, Pushmeet Kohli, and Josh Tenenbaum. Deep convolutional inverse graphics network. In *Advances in Neural Information Processing Systems*, pages 2539–2547, 2015.

Abhishek Kumar, Prasanna Sattigeri, and Avinash Balakrishnan. Variational inference of disentangled latent concepts from unlabeled observations. *arXiv preprint arXiv:1711.00848*, 2017.

Matt J Kusner, Brooks Paige, and José Miguel Hernández-Lobato. Grammar variational autoencoder. In *International Conference on Machine Learning*, 2017.

Brenden M Lake, Tomer D Ullman, Joshua B Tenenbaum, and Samuel J Gershman. Building machines that learn and think like people. *Behavioral and Brain Sciences*, 40, 2017.

Ken Lang. 20 newsgroups data set, 2007. URL `http://www.ai.mit.edu/people/jrennie/20Newsgroups/`. [Online; accessed 18-May-2018].

Yann LeCun, Corinna Cortes, and CJ Burges. Mnist handwritten digit database. *AT&T Labs [Online]. Available: http://yann.lecun. com/exdb/mnist*, 2, 2010.

Dawen Liang, Rahul G. Krishnan, Matthew D. Hoffman, and Tony Jebara. Variational Autoencoders for Collaborative Filtering. *arXiv:1802.05814 [cs, stat]*, February 2018.

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

Chris J Maddison, Andriy Mnih, and Yee Whye Teh. The concrete distribution: A continuous relaxation of discrete random variables. In *International Conference on Learning Representations*, 2017.

Alireza Makhzani, Jonathon Shlens, Navdeep Jaitly, Ian Goodfellow, and Brendan Frey. Adversarial autoencoders. *arXiv preprint arXiv:1511.05644*, 2015.

Yishu Miao, Lei Yu, and Phil Blunsom. Neural variational inference for text processing. In *International Conference on Machine Learning*, pages 1727–1736, 2016.

Aaron van den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew Senior, and Koray Kavukcuoglu. Wavenet: A generative model for raw audio. *arXiv preprint arXiv:1609.03499*, 2016.

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

N Siddharth, Brooks Paige, Jan-Willem Van de Meent, Alban Desmaison, Frank Wood, Noah D Goodman, Pushmeet Kohli, and Philip HS Torr. Learning disentangled representations with semi-supervised deep generative models. In *Advances in Neural Information Processing Systems*, 2017.

Akash Srivastava and Charles Sutton. Autoencoding variational inference for topic models. *arXiv preprint arXiv:1703.01488*, 2017.

Lucas Theis, Wenzhe Shi, Andrew Cunningham, and Ferenc Huszár. Lossy image compression with compressive autoencoders. *arXiv preprint arXiv:1703.00395*, 2017.

Andreas Veit, Serge Belongie, and Theofanis Karaletsos. Disentangling Nonlinear Perceptual Embeddings With Multi-Query Triplet Networks. *arXiv preprint arXiv:1603.07810*, 2016.

Han Xiao, Kashif Rasul, and Roland Vollgraf. Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms, 2017.

Weidi Xu, Haoze Sun, Chao Deng, and Ying Tan. Variational autoencoder for semi-supervised text classification. In *AAAI*, pages 3358–3364, 2017.

Shengjia Zhao, Jiaming Song, and Stefano Ermon. InfoVAE: Information maximizing variational autoencoders. *arXiv preprint arXiv:1706.02262*, 2017.