Auto-Encoding Total Correlation Explanation

Shuyang Gao   **Rob Brekelmans**   **Greg Ver Steeg**   **Aram Galstyan**

Information Sciences Institute
University of Southern California
{gaos, brekelma}@usc.edu {gregv, galstyan}@isi.edu

Abstract

Advances in unsupervised learning enable reconstruction and generation of samples from complex distributions, but this success is marred by the inscrutability of the representations learned. We propose an information-theoretic approach to characterizing disentanglement and dependence in representation learning using multivariate mutual information, also called total correlation. The principle of Total Cor-relation Ex-planation (CorEx) has motivated successful unsupervised learning applications across a variety of domains but under some restrictive assumptions. Here we relax those restrictions by introducing a flexible variational lower bound to CorEx. Surprisingly, we find this lower bound is equivalent to the one in variational autoencoders (VAE) under certain conditions. This information-theoretic view of VAE deepens our understanding of hierarchical VAE and motivates a new algorithm, AnchorVAE, that makes latent codes more interpretable through information maximization and enables generation of richer and more realistic samples.

1 Introduction

Learning representations from data without labels has become increasingly important to solve some of the most crucial problems in machine learning including tasks in image, language, speech, etc (Bengio et al., 2013). Complex models, such as deep neural networks, have been successfully applied to generative modeling with high-dimensional data. From these methods

we can either infer hidden representations with variational autoencoders (VAE) (Kingma and Welling, 2013; Rezende et al., 2014) or generate new samples with VAE or generative adversarial networks (GAN) (Goodfellow et al., 2014).

Building on these successes, an explosive amount of recent effort has focused on *interpreting* learned representations, which could have significant implications for subsequent tasks. Methods like InfoGAN (Chen et al., 2016) and β-VAE (Higgins et al., 2017) are able to learn *disentangled* and *interpretable* representations in a completely unsupervised fashion. Information theory provides a natural framework for understanding representation learning and continues to generate new insights (Alemi et al., 2017; Shwartz-Ziv and Tishby, 2017; Achille and Soatto, 2018; Saxe et al., 2018; Kim and Mnih, 2018; Chen et al., 2018).

In this paper we motivate the problem of learning *disentangled* and *interpretable* representations in a purely information-theoretic way. Instead of making assumptions about the data generating process at the beginning, we consider the question of how informative the underlying latent variable $\mathbf{z}$ is about the original data variable $\mathbf{x}$. We would like $\mathbf{z}$ to be as informative as possible about the *relationships* in $\mathbf{x}$ while remaining as disentangled as possible in the sense of statistical independence. This principle has been previously proposed as Cor-relation Ex-planation (CorEx) (Ver Steeg and Galstyan, 2014; Ver Steeg, 2017). By optimizing appropriate information-theoretic measures, CorEx defines not only an *informative* representation but also a *disentangled* one, thus eliciting a natural comparison to the recent literature on interpretable machine learning. However, computing the CorEx objective can be challenging, and previous studies have been restricted to cases where random variables are either discrete (Ver Steeg and Galstyan, 2014), or Gaussian (Ver Steeg and Galstyan, 2017).

Our key contributions are as follows:

- We construct a variational lower bound to the CorEx objective and optimize the bound with deep

neural networks. Surprisingly, we find that under standard assumptions, the lower bound for CorEx shares the same mathematical form as the evidence lower bound (ELBO) used in VAE, suggesting that CorEx provides a dual information-theoretic perspective on representations learned by VAE.

- Going beyond the standard scenario to hierarchical VAEs or deep gaussian latent models (DLGM) (Rezende et al., 2014), we demonstrate that CorEx provides new insight into measuring how representations become progressively more disentangled at subsequent layers. In addition, the CorEx objective can be naturally decomposed into two sets of mutual information terms with an interpretation as an unsupervised information bottleneck.
- Inspired by this formulation, we propose to make some latent factors more interpretable by reweighting terms in the objective to make certain parts of the latent code uniquely informative about the inputs (instead of adding new terms to the objective, as in InfoGAN (Chen et al., 2016)).
- Finally, we show that by sampling each latent code $\mathbf{z}_i$ from the encoding distribution $p(\mathbf{z}_i) = \int_{\mathbf{x}} p(\mathbf{z}_i|\mathbf{x})p(\mathbf{x})d\mathbf{x}$ instead of the standard Gaussian prior in VAE, we can generate richer and more realistic samples than VAE even under the same network model.

2 Information Theory Background

Let $\mathbf{x} = (\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_d)$ denote a d-dimensional random variable whose probability density function is $p(\mathbf{x})$. Shannon differential entropy (Cover and Thomas, 2006) is defined in the usual way as $H(\mathbf{x}) = -\mathbb{E}_{\mathbf{x}}[\log p(\mathbf{x})]$. Let $\mathbf{z} = (\mathbf{z}_1, \mathbf{z}_2, ..., \mathbf{z}_m)$ denote an m-dimensional random variable whose probability density function is $p(\mathbf{z})$. Then mutual information between two random variables, $\mathbf{x}$ and $\mathbf{z}$, is defined as $I(\mathbf{x} : \mathbf{z}) = H(\mathbf{x}) + H(\mathbf{z}) - H(\mathbf{x}, \mathbf{z})$. Mutual information can also be viewed as the reduction in uncertainty about one variable given another variable, i.e., $I(\mathbf{x} : \mathbf{z}) = H(\mathbf{x}) - H(\mathbf{x}|\mathbf{z}) = H(\mathbf{z}) - H(\mathbf{z}|\mathbf{x})$.

A measure of multivariate mutual information called *total correlation* (Watanabe, 1960) or *multi-information* (Studený and Vejnarova, 1998) is defined as follows:

$$TC(\mathbf{x}) = \sum_{i=1}^{d} H(\mathbf{x}_i) - H(\mathbf{x}) = D_{KL}\left(p(\mathbf{x})||\prod_{i=1}^{d} p(\mathbf{x}_i)\right) \tag{1}$$

Note that $D_{KL}(\cdot)$ denotes the *Kullback-Leibler* divergence in Eq. 1. Intuitively, $TC(\mathbf{x})$ captures the total dependence across all the dimensions of $\mathbf{x}$ and is zero if and only if all $\mathbf{x}_i$ are independent. Total correlation

or statistical independence is often used to characterize *disentanglement* in recent literature on learning representations (Dinh et al., 2014; Achille and Soatto, 2017).

The conditional total correlation of $\mathbf{x}$, after observing some latent variable $\mathbf{z}$, is defined as follows,

$$TC(\mathbf{x}|\mathbf{z}) = \sum_{i=1}^{d} H(\mathbf{x}_i|\mathbf{z}) - H(\mathbf{x}|\mathbf{z})$$
$$= D_{KL}\left(p(\mathbf{x}|\mathbf{z})||\prod_{i=1}^{d} p(\mathbf{x}_i|\mathbf{z})\right) \tag{2}$$

We define a measure of *informativeness* of latent variable $\mathbf{z}$ about the dependence among the observed variables $\mathbf{x}$ by quantifying how total correlation is reduced after conditioning on some latent factor $\mathbf{z}$, i.e.,

$$TC(\mathbf{x}; \mathbf{z}) = TC(\mathbf{x}) - TC(\mathbf{x}|\mathbf{z}) \tag{3}$$

In Eq. 3, we can see that $TC(\mathbf{x}; \mathbf{z})$ is maximized if and only if the conditional distribution $p(\mathbf{x}|\mathbf{z})$ factorizes, in which case we can interpret $\mathbf{z}$ as capturing the information about *common causes* across all $\mathbf{x}_i$'s.

3 Total Correlation Explanation Representation Learning

In a typical unsupervised setting like VAE, we assume a generative model where $\mathbf{x}$ is a function of a latent variable $\mathbf{z}$ and then maximize the log likelihood of $\mathbf{x}$ under this model. From a CorEx perspective, the situation is reversed. We let $\mathbf{z}$ be some stochastic function of $\mathbf{x}$ parameterized by θ, i.e., $p_\theta(\mathbf{z}|\mathbf{x})$. Then we seek a joint distribution $p_\theta(\mathbf{x}, \mathbf{z}) = p_\theta(\mathbf{z}|\mathbf{x})p(\mathbf{x})$, where $p(\mathbf{x})$ is the underlying true data distribution, that maximizes the following objective,

$$\mathcal{L}(\theta; \mathbf{x}) = \underbrace{TC_\theta(\mathbf{x}; \mathbf{z})}_{\text{informativeness}} - \underbrace{TC_\theta(\mathbf{z})}_{\text{(dis)entanglement}} \tag{4}$$
$$= TC(\mathbf{x}) - TC_\theta(\mathbf{x}|\mathbf{z}) - TC_\theta(\mathbf{z})$$

In Eq. 4, $TC_\theta(\mathbf{x}; \mathbf{z})$ corresponds to the amount of correlation that is explained by $\mathbf{z}$ as defined in Eq. 3 and $TC_\theta(\mathbf{z})$ quantifies the dependence among the latent variables $\mathbf{z}$.

By non-negativity of total correlation, Eq. 4 naturally forms a lower bound on $TC(\mathbf{x})$, i.e., $TC(\mathbf{x}) \geq \mathcal{L}(\theta; \mathbf{x})$ for any θ. Therefore, the global maximum of Eq. 4 occurs at $TC(\mathbf{x})$, in which case $TC_{\theta^*}(\mathbf{x}|\mathbf{z}) = TC_{\theta^*}(\mathbf{z}) \equiv 0$ and $\mathbf{z}$ can be exactly interpreted as a generative model where $\mathbf{z}$ are independent random variables that generate $\mathbf{x}$, as shown in Fig. 1.

Notice that the term $TC_\theta(\mathbf{x}; \mathbf{z})$ is a bit different from the classical definition of informativeness using mutual
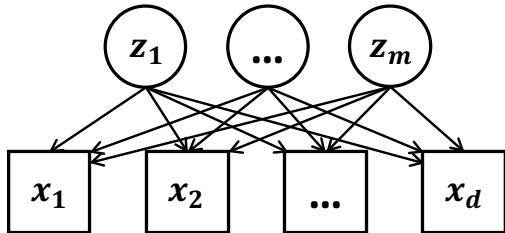
Figure 1: The graphical model for $p_{\theta*}(\mathbf{x}, \mathbf{z})$ assuming $p_{\theta*}(\mathbf{z}|\mathbf{x})$ achieves the global maximum in Eq. 4. In this model, all $\mathbf{x}_i$'s are factorized conditioned on $\mathbf{z}$, and all $\mathbf{z}_i$'s are independent.

information $I_\theta(\mathbf{x}; \mathbf{z})$ (Linsker, 1988). In fact, after combining the entropy terms in Eq. 1 and 2, the following equation holds (Ver Steeg and Galstyan, 2015),

$$TC_\theta(\mathbf{x}; \mathbf{z}) = \sum_{i=1}^{d} I_\theta(\mathbf{x}_i; \mathbf{z}) - I_\theta(\mathbf{x}; \mathbf{z}) \qquad (5)$$

The term $TC_\theta(\mathbf{x}; \mathbf{z})$ in Eq. 4 can be seen as finding a minimal latent representation $\mathbf{z}$ which, after conditioning, disentangles $\mathbf{x}$. When stacking hidden variable layers in Sec. 6, we will see that this condition can lead to interpretable features by forcing intermediate layers to be explained by higher layers under a factorized model.

Informativeness vs Disentanglement If we only consider the informativeness term $TC_\theta(\mathbf{x}; \mathbf{z})$ as in the objective, a naive solution to this problem would be just setting $\mathbf{z} = \mathbf{x}$. To avoid this, we also want the latent variables $\mathbf{z}$ to be as disentangled as possible, corresponding to the $TC(\mathbf{z})$ term encouraging independence. In other words, the objective in Eq. 4 is trying to find $\mathbf{z}$, so that $\mathbf{z}$ not only disentangles $\mathbf{x}$ as much as possible, but is itself as disentangled as possible.

4 Optimization

We first focus on optimizing the objective function defined by Eq. 4. The extension to the multi-layer (hierarchical) case is presented in the next section.

By using Eqs. 1 and 5, we expand Eq. 4 into basic information-theoretic quantities as follows:

$$\mathcal{L}(\theta; \mathbf{x}) = TC_\theta(\mathbf{x}; \mathbf{z}) - TC_\theta(\mathbf{z})$$
$$= \sum_{i=1}^{d} I_\theta(\mathbf{x}_i : \mathbf{z}) - I_\theta(\mathbf{x} : \mathbf{z}) - \sum_{i=1}^{m} H_\theta(\mathbf{z}_i) + H_\theta(\mathbf{z})$$
$$= \sum_{i=1}^{d} I_\theta(\mathbf{x}_i : \mathbf{z}) - \sum_{i=1}^{m} H_\theta(\mathbf{z}_i) + H_\theta(\mathbf{z}|\mathbf{x}) \qquad (6)$$

If we further constrain our search space $p_\theta(\mathbf{z}|\mathbf{x})$ to have

the factorized form $p_\theta(\mathbf{z}|\mathbf{x}) = \prod_{i=1}^{m} p_{\theta_i}(\mathbf{z}_i|\mathbf{x})$ [1], which is a standard assumption in most VAE models, then we get,

$$\mathcal{L}(\theta; \mathbf{x}) = TC_\theta(\mathbf{x}; \mathbf{z}) - TC_\theta(\mathbf{z})$$
$$= \sum_{i=1}^{d} I_\theta(\mathbf{x}_i : \mathbf{z}) - \sum_{i=1}^{m} I_\theta(\mathbf{z}_i : \mathbf{x}) \qquad (7)$$

We convert the two total correlation terms into two sets of mutual information terms in Eq. 7. The first term, $I_\theta(\mathbf{x}_i : \mathbf{z})$, denotes the mutual information between each input dimension $\mathbf{x}_i$ and $\mathbf{z}$, and can be broadly construed as measuring the "relevance" of the representation to each observed variable in the parlance of the information bottleneck (Tishby et al., 2000; Shwartz-Ziv and Tishby, 2017). The second term, $I_\theta(\mathbf{z}_i : \mathbf{x})$, represents the mutual information between each latent dimension $\mathbf{z}_i$ and $\mathbf{x}$ and can be viewed as the compression achieved by each latent factor. We proceed by constructing tractable bounds on these quantities.

4.1 Variational Lower Bound for $I_\theta(\mathbf{x}_i : \mathbf{z})$

Barber and Agakov (2003) derived the following lower bound for mutual information by using the non-negativity of KL-divergence, i.e., $\Sigma_{\mathbf{x}_i} p(\mathbf{x}_i|\mathbf{z}) \log \frac{p(\mathbf{x}_i|\mathbf{z})}{q(\mathbf{x}_i|\mathbf{z})} \geq 0$ gives,

$$I_\theta(\mathbf{x}_i : \mathbf{z}) \geq H(\mathbf{x}_i) + \langle \ln q_\phi(\mathbf{x}_i|\mathbf{z}) \rangle_{p_\theta(\mathbf{x}, \mathbf{z})} \qquad (8)$$

where the angled brackets represent expectations and $q_\phi(\mathbf{x}_i|\mathbf{z})$ is any arbitrary distribution parametrized by ϕ. We need a variational distribution $q_\phi(\mathbf{x}_i|\mathbf{z})$ because the posterior distribution $p_\theta(\mathbf{x}|\mathbf{z}) = p_\theta(\mathbf{z}|\mathbf{x})p(\mathbf{x})/p_\theta(\mathbf{z})$ is hard to calculate due to fact that the true data distribution $p(\mathbf{x})$ is unknown, although approximating the normalization factor $p_\theta(\mathbf{z})$ can be tractable compared to VAE. A detailed comparison with VAE will be made in Sec. 5.

4.2 Variational Upper Bound for $I_\theta(\mathbf{z}_i : \mathbf{x})$

We again use the non-negativity of KL-divergence, i.e., $\Sigma_{\mathbf{z}_i} p(\mathbf{z}_i) \log \frac{p(\mathbf{z}_i)}{r(\mathbf{z}_i)} \geq 0$, to obtain

$$I_\theta(\mathbf{x} : \mathbf{z}_i)$$
$$= \int d\mathbf{x} d\mathbf{z}_i p_\theta(\mathbf{z}_i, \mathbf{x}) \log p_\theta(\mathbf{z}_i|\mathbf{x}) - \int d\mathbf{z}_i p_\theta(\mathbf{z}_i) \log p_\theta(\mathbf{z}_i)$$
$$\leq \int d\mathbf{x} d\mathbf{z}_i p_\theta(\mathbf{z}_i, \mathbf{x}) \log p_\theta(\mathbf{z}_i|\mathbf{x}) - \int d\mathbf{z}_i p_\theta(\mathbf{z}_i) \log r_\alpha(\mathbf{z}_i)$$
$$= \int d\mathbf{x} d\mathbf{z}_i p_\theta(\mathbf{x}) p_\theta(\mathbf{z}_i|\mathbf{x}) \log \frac{p_\theta(\mathbf{z}_i|\mathbf{x})}{r_\alpha(\mathbf{z}_i)}$$
$$= D_{KL}(p_\theta(\mathbf{z}_i|\mathbf{x}) \| r_\alpha(\mathbf{z}_i)) \qquad (9)$$

---

[1] Each marginal distribution $p_{\theta_i}(\mathbf{z}_i|\mathbf{x})$ is parametrized by a different θ_i. But we will omit the subscript i under θ for simplicity, as well as ϕ, α in the following context.

where $r_\alpha(\mathbf{z}_i)$ represents an arbitrary distribution parametrized by α.

Combining bounds in Eqs. 8 and 9 into Eq. 7, we get,

$$
\begin{aligned}
\mathcal{L}(\theta; \mathbf{x}) &= \sum_{i=1}^{d} I_\theta(\mathbf{x}_i : \mathbf{z}) - \sum_{i=1}^{m} I_\theta(\mathbf{z}_i : \mathbf{x}) \\
&\geq \sum_{i=1}^{d} H(\mathbf{x}_i) + \langle \ln q_\phi(\mathbf{x}_i|\mathbf{z}) \rangle_{p_\theta(\mathbf{x},\mathbf{z})} \\
&\quad - \sum_{i=1}^{m} D_{KL}(p_\theta(\mathbf{z}_i|\mathbf{x}) || r_\alpha(\mathbf{z}_i))
\end{aligned}
\tag{10}
$$

We then can jointly optimize the lower bound in Eq. 10 w.r.t. both the stochastic parameter θ and the variational parameters ϕ and α.

5 Connection to Variational Autoencoders

Remarkably, Eq. 10 has a very similar form with the lower bound introduced in variational autoencoders, except it is decomposed into each dimension $\mathbf{x}_i$ and $\mathbf{z}_i$. To pursue this similarity further, we denote

$$
q_\phi(\mathbf{x}|\mathbf{z}) = \prod_{i=1}^{d} q_\phi(\mathbf{x}_i|\mathbf{z}), \quad r_\alpha(\mathbf{z}) = \prod_{i=1}^{m} r_\alpha(\mathbf{z}_i)
\tag{11}
$$

Then, by rearranging the terms in Eq. 10, we obtain,

$$
\begin{aligned}
\mathcal{L}(\theta; \mathbf{x}) &= \sum_{i=1}^{d} I_\theta(\mathbf{x}_i : \mathbf{z}) - \sum_{i=1}^{m} I_\theta(\mathbf{z}_i : \mathbf{x}) \\
&\geq \left(\sum_{i=1}^{d} H(\mathbf{x}_i) \right) + \left\langle \ln \underbrace{q_\phi(\mathbf{x}|\mathbf{z})}_{\text{decoder}} \right\rangle_{p_\theta(\mathbf{x},\mathbf{z})} \\
&\quad - D_{KL}(\underbrace{p_\theta(\mathbf{z}|\mathbf{x})}_{\text{encoder}} || r_\alpha(\mathbf{z}))
\end{aligned}
\tag{12}
$$

The first term in the bound, $\sum_{i=1}^{d} H(\mathbf{x}_i)$, is a constant and has no effect on the optimization. The remaining expression coincides with the VAE objective as long as $r_\alpha(\mathbf{z})$ is a standard Gaussian: the second term corresponds to the reconstruction error, and the third term is the KL-divergence term in VAE.

Comparison The CorEx objective starts with a defined encoder $p_\theta(\mathbf{z}|\mathbf{x})$ and seeks a decoder $q_\phi(\mathbf{x}|\mathbf{z})$ via variational approximation to the true posterior, while VAE is exactly the opposite. Moreover, in VAE, we need a variational approximation to the posterior because the normalization constant is intractable, while in CorEx, the variational distribution is needed because we do not know the true data distribution $p(\mathbf{x})$. It is also worth mentioning that the lower bound in Eq. 12

requires a fully factorized form of the decoder $q_\phi(\mathbf{x}|\mathbf{z})$, unlike VAE where $q_\phi(\mathbf{x}|\mathbf{z})$ can be flexible [2].

As pointed out by Zhao et al. (2017), if we choose to use a more expressive distribution family, such as PixelRNN/PixelCNN (Van Oord et al., 2016; Gulrajani et al., 2017) for the decoder in a VAE, the model tends to neglect the latent codes all together, i.e., $I(\mathbf{x} : \mathbf{z}) = 0$. This problem, however, does not exist in CorEx, since it explicitly requires $\mathbf{z}$ to be *informative* about $\mathbf{x}$ in the objective function. It is this *informativeness* term that leads the CorEx objective to a factorized decoder family $q_\phi(\mathbf{x}|\mathbf{z})$. In fact, if we assume $I_\theta(\mathbf{x} : \mathbf{z}) = 0$, then we will get $TC(\mathbf{x}) = TC_\theta(\mathbf{x}|\mathbf{z})$ and an *informativeness* term $TC_\theta(\mathbf{x}; \mathbf{z})$ of zero, meaning CorEx will avoid such undesirable solutions.

Stacking CorEx and Hierarchical VAE Notice that if Eq. 4 does not achieve the global maximum, it might be the case that the latent variable $\mathbf{z}$ is still not disentangled enough, i.e., $TC_\theta(\mathbf{z}) > 0$. If this is true, we can re-apply the CorEx principle (Ver Steeg and Galstyan, 2015) and learn another layer of latent variables $\mathbf{z}^{(2)}$ on top of $\mathbf{z}$ and redo the optimization on $\theta^{(2)}$ w.r.t. the following equation, i.e.,

$$
\begin{aligned}
\mathcal{L}(\theta^{(2)}; \mathbf{z}) &= TC_{\theta^{(2)}}(\mathbf{z}; \mathbf{z}^{(2)}) - TC_{\theta^{(2)}}(\mathbf{z}^{(2)}) \\
&= TC_\theta(\mathbf{z}) - TC_{\theta^{(2)}}(\mathbf{z}|\mathbf{z}^{(2)}) - TC_{\theta^{(2)}}(\mathbf{z}^{(2)})
\end{aligned}
\tag{13}
$$

To generalize, suppose there are L layers of latent variables, $\mathbf{z}^{(1)}, \mathbf{z}^{(2)}, ..., \mathbf{z}^{(L)}$ and we further denote the observed variable $\mathbf{x} \equiv \mathbf{z}^{(0)}$, then one can stack each latent variable $\mathbf{z}^{(l)}$ on top of $\mathbf{z}^{(l-1)}$, and jointly optimize the summation of the corresponding objectives as shown in Eqs. 4 and 13, i.e.,

$$
\mathcal{L}(\theta^{(1,2,...,L)}; \mathbf{x}) = \sum_{l=1}^{L} \mathcal{L}(\theta^{(l)}; \mathbf{z}^{(l-1)})
\tag{14}
$$

By simple expansion of Eq. 14 and cancellation of intermediate TC terms, we get

$$
\begin{aligned}
&\mathcal{L}(\theta^{(1,2,...,L)}; \mathbf{x}) \\
&= \mathcal{L}(\theta^{(1)}; \mathbf{z}^{(0)}) + \mathcal{L}(\theta^{(2)}; \mathbf{z}^{(1)}) + ... + \mathcal{L}(\theta^{(L)}; \mathbf{z}^{(L-1)}) \\
&= TC(\mathbf{x}) - \sum_{l=1}^{L} TC_{\theta^{(l)}}(\mathbf{z}^{(l-1)}|\mathbf{z}^{(l)}) - TC_{\theta^{(L)}}(\mathbf{z}^{(L)}) \\
&\leq TC(\mathbf{x})
\end{aligned}
\tag{15}
$$

Furthermore, if we have $\mathcal{L}(\theta^{(l)}; \mathbf{z}^{(l-1)}) > 0$ for all l, then we get,

$$
\begin{aligned}
\mathcal{L}(\theta^{(1)}; \mathbf{x}) \leq \mathcal{L}(\theta^{(1,2)}; \mathbf{x}) \leq ... &\leq \mathcal{L}(\theta^{(1,...,L)}; \mathbf{x}) \\
&\leq TC(\mathbf{x})
\end{aligned}
\tag{16}
$$

---

[2] In this paper, we also restrict the encoder distribution $p_\theta(\mathbf{z}|\mathbf{x})$ to have a factorized form which follows the standard network structures in VAE, but it is not a necessary condition to achieve the lower bound shown in Eq. 12.

Eq. 16 shows that stacking latent factor representations results in progressively better lower bounds for $TC(\mathbf{x})$.

To optimize Eq. 14, we reuse Eqs. 7, 8 and 9 and get,

$$
\begin{aligned}
\mathcal{L}(\theta^{(1,2,..,L)}; \mathbf{x}) \geq \quad & \sum_i H(\mathbf{z}_i^{(0)}) \\
+ \quad & \sum_{l=1}^{L} \sum_i \left\langle \ln q_{\phi^{(l)}}(\mathbf{z}_i^{(l-1)}|\mathbf{z}^{(l)}) \right\rangle_{p_\theta(\mathbf{z})} \\
- \quad & \sum_{l=1}^{L} \sum_i \left\langle \ln p_{\theta^{(l)}}(\mathbf{z}_i^{(l)}|\mathbf{z}^{(l-1)}) \right\rangle_{p_\theta(\mathbf{z})} \\
+ \quad & \sum_i \left\langle \ln r_\alpha(\mathbf{z}_i^{(L)}) \right\rangle_{p_\theta(\mathbf{z})} \qquad (17)
\end{aligned}
$$

Enforcing independence relations at each layer, we denote

$$
\begin{aligned}
q_\phi(\mathbf{x}, \mathbf{z}) &= \prod_i r_\alpha(\mathbf{z}_i^{(L)}) \times \prod_{l=1}^{L} \prod_i q_{\theta^{(l)}}(\mathbf{z}_i^{(l-1)}|\mathbf{z}^{(l)}) \\
p_\theta(\mathbf{z}|\mathbf{x}) &= \prod_{l=1}^{L} p_{\theta^{(l)}}(\mathbf{z}^{(l)}|\mathbf{z}^{(l-1)})
\end{aligned} \qquad (18)
$$

and obtain

$$
\begin{aligned}
\mathcal{L}(\theta^{(1,2,..,L)}; \mathbf{x}) \geq \quad & \sum_i H(\mathbf{z}_i^{(0)}) \\
+ \quad & \left\langle \ln \frac{q_\phi(\mathbf{x}, \mathbf{z})}{p_\theta(\mathbf{z}|\mathbf{x})} \right\rangle_{p_\theta(\mathbf{z}|\mathbf{x})p(\mathbf{x})} \qquad (19)
\end{aligned}
$$

One can now see that the second term of the RHS in Eq. 19 has the same form as deep latent Gaussian models (Rezende et al., 2014) (also known as hierarchical VAE) as long as the latent code distribution $r_\alpha(\mathbf{z}^{(L)})$ on the top layer follows standard normal and $q_{\theta^{(l)}}(\mathbf{z}^{(l-1)}|\mathbf{z}^{(l)})$ on each layer is parametrized by Gaussian distributions.

One immediate insight from this connection is that, as long as each $\mathcal{L}(\theta^{(l)}; \mathbf{z}^{(l-1)})$ is greater than zero in Eq. 14, then by expanding the definition of each term we can easily see that $\mathbf{z}^{(l)}$ is more disentangled than $\mathbf{z}^{(l-1)}$, i.e., $TC(\mathbf{z}^{(l-1)}) > TC(\mathbf{z}^{(l)})$ if $TC(\mathbf{z}^{(l-1)}|\mathbf{z}^{(l)}) - TC(\mathbf{z}^{(l)}) > 0$. Therefore, each latent layer of hierarchical VAE will be more and more disentangled if $\mathcal{L}(\theta^{(l)}; \mathbf{z}^{(l-1)}) > 0$ for each l. This interpretation also provides a criterion for determining the depth of a hierarchical representation; we can add layers as long as the corresponding term in the objective is positive so that the overall lower bound on $TC(\mathbf{x})$ is increasing.

Despite reaching the same final expression, approaching this result from an information-theoretic optimization rather than generative modeling perspective offers some advantages. First of all, we have much more flexibility in specifying the distribution of latent factors, as we can directly sample from this distribution using our encoder. Secondly, the connection with mutual information suggests intuitive modifications of our objective that increase the interpretability of results. These advantages will be explored in more depth in Sec. 6.

6 Applications

6.1 Disentangling Latent Codes via Hierarchical VAE / Stacking CorEx on MNIST

We train a simple hierarchical VAE/stacking CorEx model with two stochastic layers on the MNIST dataset. The graphical model is shown in Fig. 2. For each stochastic layer, we use a neural network to parametrize the distribution p_θ and q_ϕ, and we set r_α to be a fixed standard Gaussian. We use a 784-512-512-64 fully
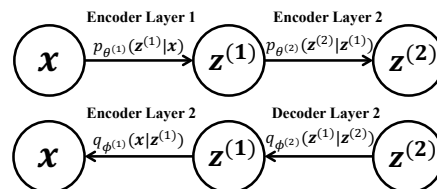
Figure 2: Encoder and decoder models for MNIST, where $\mathbf{z}^{(1)}$ is 64 dimensional continuous variable and $\mathbf{z}^{(2)}$ is a discrete variable (one hot vector with length ten).

connected network between $\mathbf{x}$ and $\mathbf{z}^{(1)}$ and a 64-32-32-16-16-10 dense network between $\mathbf{z}^{(1)}$ and $\mathbf{z}^{(2)}$, with ReLU activations in both. The output of $\mathbf{z}^{(2)}$ is a ten-dimensional one hot vector, where we decode based on each one-hot representation and weight the result according to their softmax probabilities.

After training the model, we find the learned discrete variable $\mathbf{z}^{(2)}$ on the top layer gives us an unsupervised classification accuracy of 85%, competitive with the more complex method shown in Dilokthanakul et al. (2016).

To verify that the top layer $\mathbf{z}^{(2)}$ helps disentangle the middle layer $\mathbf{z}^{(1)}$ by encouraging conditional independence of $\mathbf{z}^{(1)}$ given $\mathbf{z}^{(2)}$, we calculate the mutual information $I_\theta(\mathbf{x} : \mathbf{z}_i^{(1)})$ between input $\mathbf{x}$ and each dimension $\mathbf{z}_i^{(1)}$. We then select the top two dimensions with the most mutual information, denote these two dimensions as $\mathbf{z}_a^{(1)}$, $\mathbf{z}_b^{(1)}$. We find $I_\theta(\mathbf{x} : \mathbf{z}_a^{(1)}) = 2.71$ and $I_\theta(\mathbf{x} : \mathbf{z}_b^{(1)}) = 2.56$. We then generate new digits by first fixing the discrete latent variable $\mathbf{z}^{(2)}$ on the top layer, and sampling latent codes $\mathbf{z}^{(1)}$ from $q_\phi(\mathbf{z}^{(1)}|\mathbf{z}^{(2)})$. We systematically vary the noise from -2 to 2 through

$q_\phi(\mathbf{z}_a^{(1)}|\mathbf{z}^{(2)})$ and $q_\phi(\mathbf{z}_b^{(1)}|\mathbf{z}^{(2)})$ while keeping the other dimensions of $\mathbf{z}^{(1)}$ fixed, and visualize the results in Fig. 3. We can see that this simple two-layer structure
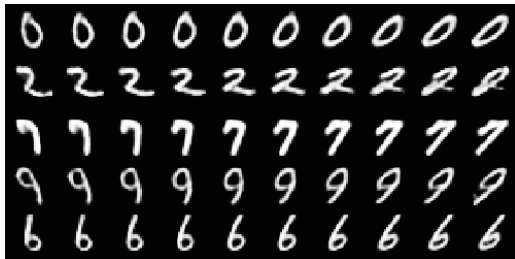
(a) Manipulating $\mathbf{z}_a^{(1)}$ with MNIST. (Azimuth)
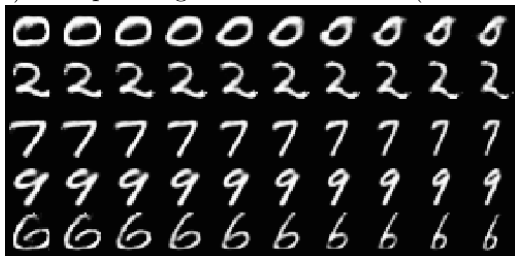
(b) Manipulating $\mathbf{z}_b^{(1)}$ with MNIST. (Width)

Figure 3: **Varying the latent codes of $\mathbf{z}^{(1)}$ on MNIST:** In both figures, each row corresponds to a fixed discrete number in layer $\mathbf{z}^{(2)}$. Different columns correspond to the varying noise from the selected latent node in layer $\mathbf{z}^{(1)}$ from left to right, while keeping other latent codes fixed. In (a), varying the noise results in different rotations of the digit; In (b), a small (large) value of the latent code corresponds to wider (narrower) digit.

automatically disentangles and learns the interpretable factors on MNIST (width and rotation). We attribute this behavior to stacking, where the top layer disentangles the middle layer and makes the latent codes more interpretable through samples from $q_\phi(\mathbf{z}^{(1)}|\mathbf{z}^{(2)})$.

6.2 Learning Interpretable Representations through Information Maximizing VAE / CorEx on CelebA

One important insight from recently developed methods, like InfoGAN, is that we can maximize the mutual information between a latent code and the observations to make the latent code more interpretable.

While it seems ad-hoc to add an additional mutual information term in the original VAE objective, a more natural analogue arises in the CorEx setting. Looking at the formulation in Eq. 7, it already contains two sets of mutual information terms. If one would like to *anchor* a latent variable, say $\mathbf{z}_a$, to have higher mutual information with the observation $\mathbf{x}$, then one can simply modify the objective by replacing the unweighted sum

with a weighted one,

$$\mathcal{L}_{anchor}(\theta; \mathbf{x}) \qquad (20)$$
$$= TC_\theta(\mathbf{x}; \mathbf{z}) - TC_\theta(\mathbf{z}) + \lambda I_\theta(\mathbf{z}_a : \mathbf{x})$$
$$= \sum_{i=1}^{d} I_\theta(\mathbf{x}_i : \mathbf{z}) - \sum_{i=1, i\neq a}^{m} I_\theta(\mathbf{z}_i : \mathbf{x}) - (1-\lambda)I_\theta(\mathbf{z}_a : \mathbf{x})$$

Eq. 20 suggests that mutual information maximization in CorEx is achieved by modifying the corresponding weights of the second term $I_\theta(\mathbf{z}_i : \mathbf{x})$ in Eq. 7. We then use the lower bound in Eq. 10 to obtain

$$
\begin{aligned}
\mathcal{L}_{anchor}(\theta; \mathbf{x}) \;\geq\; & \sum_{i=1}^{d} H(\mathbf{x}_i) \qquad (21)\\
+ & \; \langle \ln q_\phi(\mathbf{x}_i|\mathbf{z})\rangle_{p_\theta(\mathbf{x},\mathbf{z})} \\
- & \sum_{i=1, i\neq a}^{m} D_{KL}(p_\theta(\mathbf{z}_i|\mathbf{x})||r_\alpha(\mathbf{z}_i)) \\
- & \; (1-\lambda)D_{KL}(p_\theta(\mathbf{z}_a|\mathbf{x})||r_\alpha(\mathbf{z}_a))
\end{aligned}
$$

Eq. 21 shows that in VAE, we can decrease the weight of KL-divergence for particular latent codes to achieve mutual information maximization. We call this new approach **AnchorVAE** in Eq. 21. Notice that there is a subtle difference between AnchorVAE and β-VAE (Higgins et al., 2017). In β-VAE, the weights of KL-divergence term for all latent codes are the same, while in AnchorVAE, only the weights of specified factors have been changed to encourage high mutual information. With some prior knowledge of the underlying factors of variation, AnchorVAE encourages the model to concentrate this explanatory power in a limited number of variables.
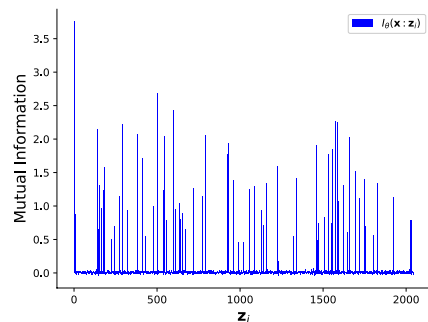

Figure 4: Mutual information between input data $\mathbf{x}$ and each latent variable $\mathbf{z}_i$ in CelebA with AnchorVAE. It is clear that the anchored first five dimensions have the highest mutual information with $\mathbf{x}$.

We trained AnchorVAE on the CelebA dataset with 2048 latent factors, with mean square error for reconstruction loss. We adopted a three layer convolu-

tional neural network structure. The weights of KL-divergence of first five latent variables are set to be 0.5 to let them have higher mutual information than other latent variables. The mutual information is plotted in Fig. 4 after training. We find these five latent variables have the highest mutual information of around 3.5, demonstrating the mutual information maximization effect in AnchorVAE.

To evaluate the interpretability of those anchored variables for generating new samples, we manipulate the first five latent variables while keep other dimensions fixed. Fig. 5 summarizes the result. We observe that all five anchored latent variables learn intuitive factors of variation in the data. It is interesting to see that latent variable $\mathbf{z}_0$ and $\mathbf{z}_4$ are very similar, both vary the generated images from white to black in some sense. However, these two latent factors are actually very different, $\mathbf{z}_0$ emphasizes skin color variation while $\mathbf{z}_4$ controls the position of the light source.
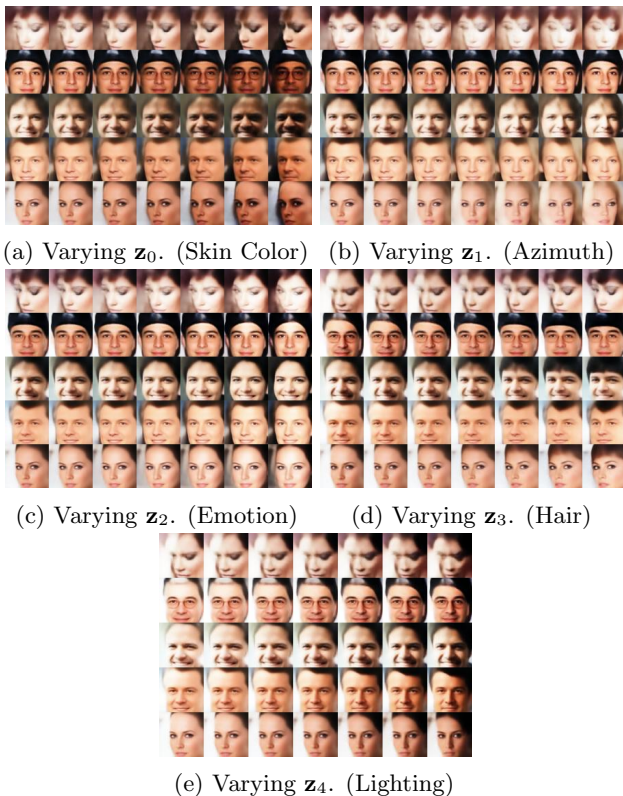

(a) Varying $\mathbf{z}_0$. (Skin Color)   (b) Varying $\mathbf{z}_1$. (Azimuth)

(c) Varying $\mathbf{z}_2$. (Emotion)   (d) Varying $\mathbf{z}_3$. (Hair)

(e) Varying $\mathbf{z}_4$. (Lighting)

Figure 5: **Manipulating latent codes $\mathbf{z}_0, \mathbf{z}_1, \mathbf{z}_2, \mathbf{z}_3, \mathbf{z}_4$ on CelebA using AnchorVAE**: We show the effect of the anchored latent variables on the outputs while traversing their values from [-3,3]. Each row of represents a different seed image to encode latent codes. Each anchored latent code represents a different factor on interpretablility. (a) Skin Color (b) Azimuth (c) Emotion (Smile) (d) Hair (less or more) (e) Lighting.

We also trained the original VAE objective with same network structure, and examine the top five latent codes with highest mutual information. Fig. 6 shows the results of manipulating the top two latent codes $\mathbf{z}_{130}$, $\mathbf{z}_{610}$, with mutual information $I(\mathbf{z}_{130} : \mathbf{x}) = 3.1$ and $I(\mathbf{z}_{610} : \mathbf{x}) = 2.8$ respectively. We can see that they reflect an entangled representation. The other three latent codes demonstrate similar entanglement which are omitted here.
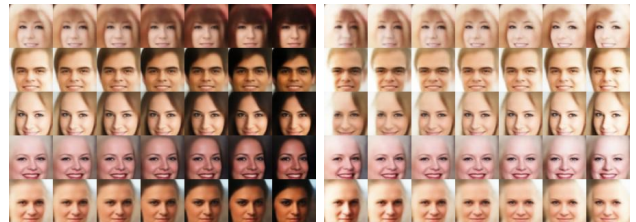
(a) $\mathbf{z}_{130}$ entangles skin color with hair   (b) $\mathbf{z}_{610}$ entangles emotion with azimuth

Figure 6: **Manipulating top two latent codes with most mutual information on CelebA using original VAE**. We observe that both latent codes learned entangled representations. (a) $\mathbf{z}_{130}$ entangles skin color with hair (b) $\mathbf{z}_{610}$ entangles emotion with azimuth.

6.3 Generating Richer and More Realistic Images via CorEx

Let us revisit the variational upper bound on $I_\theta(\mathbf{x} : \mathbf{z}_i)$ in Eq. 9. In this upper bound, VAE chooses $r_\alpha(\mathbf{z}_i)$ to be a standard normal distribution. But notice that this upper bound becomes tight when $r_\alpha(\mathbf{z}_i) = p_\theta(\mathbf{z}_i)$, i.e.,

$$I_\theta(\mathbf{x} : \mathbf{z}_i) \equiv D_{KL}(p_\theta(\mathbf{z}_i|\mathbf{x})\,||p_\theta(\mathbf{z}_i))$$
$$\leq D_{KL}(p_\theta(\mathbf{z}_i|\mathbf{x})\,||r_\alpha(\mathbf{z}_i))$$

where $p_\theta(\mathbf{z}_i) = \int_\mathbf{x} p_\theta(\mathbf{z}_i|\mathbf{x})p(\mathbf{x})d\mathbf{x}$. Therefore, after training the model, we can approximate the true distribution $p_\theta(\mathbf{z}_i) \approx \frac{1}{N}\sum_{i=1}^{N} p_\theta(\mathbf{z}_i|\mathbf{x}^{[i]})$ by first sampling a data point $\mathbf{x}^{[i]}$, and then sampling from the conditional $p_\theta(\mathbf{z}_i|\mathbf{x}^{[i]})$. Repeating this process across latent dimensions, we can use the factorized distribution $\prod_{i=1}^{m} p_\theta(\mathbf{z}_i)$ to generate new data instead of sampling from a standard normal. In this way, we obtain more realistic images since we are sampling from a tighter lower bound to the CorEx objective.

We ran a traditional VAE on the celebA dataset with the log-normal loss as the reconstruction error and 128 latent codes. We calculate the variance of each $p_\theta(\mathbf{z}_i)$, and plot the cumulative distribution of these variances in Fig. 7a. One can see that around 20% of the latent variables actually have a variance greater than two. We
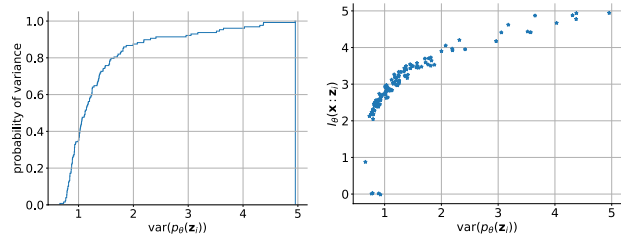
(a) Cumulative distribution of variance for each $p_\theta(\mathbf{z}_i)$) (b) Variance of $p_\theta(\mathbf{z}_i)$ versus mutual information $I_\theta(\mathbf{x} : \mathbf{z}_i)$

Figure 7: Variance statistics for $p_\theta(\mathbf{z})$ on celebA after training a standard VAE with 128 latent codes.

have plotted variance versus the mutual information in Fig. 7b, in which we can see that higher variance in $\mathbf{z}_i$ corresponds to higher mutual information $I(\mathbf{x} : \mathbf{z}_i)$. In this case, using a standard normal distribution with variance 1 for all $\mathbf{z}_i$'s would be far from optimal for generating the data.

(a) Latent codes are generated from standard normal (b) Latent codes are generated from $\prod_{i=1}^m p_\theta(\mathbf{z}_i)$

Figure 8: **Different sampling strategies of latent codes for CelebA dataset on VAE / CorEx.** Sampling latent codes from $\prod_{i=1}^m p_\theta(\mathbf{z}_i)$ in (b) yields better quality images than sampling from a standard normal distribution in (a).

Fig. 8 shows the generated images by either sampling the latent code from a standard normal distribution or the factorized distribution $\prod_{i=1}^m p_\theta(\mathbf{z}_i)$. We can see that Fig. 8b not only tends to generate more realistic images than Fig. 8a but also, it also exhibits more diversity than Fig. 8a. We attribute this improvement to the more flexible nature of our latent code distribution.

7 Related Works

The notion of disentanglement in representation learning lacks a unique characterization, but generally refers to latent factors which are individually interpretable, amenable to simple downstream modeling or transfer learning, and invariant to nuisance variation in the data

(Bengio et al., 2013). We adopt the common definition of statistical independence (Achille and Soatto, 2017; Dinh et al., 2014) by minimizing total correlation, an idea with a rich history (Barlow, 1989; Comon, 1994; Schmidhuber, 1992). However, there are numerous alternatives not rooted in independence. Higgins et al. (2017) measures disentanglement by the identifiability of changes in a single latent dimension. The work of (Thomas et al., 2017; Bengio et al., 2017) is similar in spirit, identifying disentangled factors as changes in a latent embedding that can be controlled via reinforcement learning.

Our work also provides a complementary perspective to a growing body of research connecting information theory and variational inference (Achille and Soatto, 2017, 2018; Alemi et al., 2017), much of which is motivated by the Information Bottleneck (IB) method (Tishby et al., 2000).

Our work is based on the unsupervised principle of Correlation Explanation, but other variants of CorEx also exist, such as Anchored CorEx (Gallagher et al., 2017), in this case, the anchoring worked on a different term and were used in semi-supervised settings.

8 Conclusion

Deep learning enables us to construct latent representations that reconstruct or generate samples from complex, high-dimensional distributions. Unfortunately, these powerful models do not necessarily produce representations with structures that match human intuition or goals. Subtle changes to training objectives lead to qualitatively different representations, but our understanding of this dependence remains tenuous.

Information theory has proven fruitful for understanding the competition between compression and relevance preservation in supervised learning (Shwartz-Ziv and Tishby, 2017). We explored a similar trade-off in unsupervised learning, between information maximization and disentanglement of the learned factors. Writing this objective in terms of mutual information led to two surprising connections. First, we came to an unsupervised information formulation that trades off compression and reconstruction relevance. Second, we found that by making appropriate variational approximations, we could reproduce the venerable VAE objective. This new perspective on VAE enabled more flexible distributions for latent codes and motivated new generalizations of the objective to localize interpretable information in latent codes. Ultimately, this led us to a novel learning objective that generated latent factors capturing intuitive structures in image data. We hope this alternative formulation of unsupervised learning continues to provide useful insights into this challenging problem.

References

Alessandro Achille and Stefano Soatto. On the emergence of invariance and disentangling in deep representations. *arXiv preprint arXiv:1706.01350*, 2017.

Alessandro Achille and Stefano Soatto. Information dropout: Learning optimal representations through noisy computation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2018.

Alexander A Alemi, Ian Fischer, Joshua V Dillon, and Kevin Murphy. Deep variational information bottleneck. *International Conference on Learning Representations*, 2017.

David Barber and Felix Agakov. The im algorithm: a variational approach to information maximization. In *Proceedings of the 16th International Conference on Neural Information Processing Systems*, pages 201–208. MIT Press, 2003.

Horace Barlow. Unsupervised learning. *Neural computation*, 1(3):295–311, 1989.

Emmanuel Bengio, Valentin Thomas, Joelle Pineau, Doina Precup, and Yoshua Bengio. Independently controllable features. *arXiv preprint arXiv:1703.07718*, 2017.

Yoshua Bengio, Aaron Courville, and Pascal Vincent. Representation learning: A review and new perspectives. *IEEE transactions on pattern analysis and machine intelligence*, 35(8):1798–1828, 2013.

Tian Qi Chen, Xuechen Li, Roger B Grosse, and David K Duvenaud. Isolating sources of disentanglement in variational autoencoders. In *Advances in Neural Information Processing Systems 31*, pages 2610–2620. 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, 2016.

Pierre Comon. Independent component analysis, a new concept? *Signal processing*, 36(3):287–314, 1994.

Thomas M Cover and Joy A Thomas. *Elements of information theory*. Wiley-Interscience, 2006.

Nat Dilokthanakul, Pedro AM Mediano, Marta Garnelo, Matthew CH Lee, Hugh Salimbeni, Kai Arulkumaran, and Murray Shanahan. Deep unsupervised clustering with gaussian mixture variational autoencoders. *arXiv preprint arXiv:1611.02648*, 2016.

Laurent Dinh, David Krueger, and Yoshua Bengio. Nice: Non-linear independent components estimation. *arXiv preprint arXiv:1410.8516*, 2014.

Ryan J Gallagher, Kyle Reing, David Kale, and Greg Ver Steeg. Anchored correlation explanation: Topic modeling with minimal domain knowledge. *Transactions of the Association for Computational Linguistics*, 5:529–542, 2017.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680, 2014.

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

Hyunjik Kim and Andriy Mnih. Disentangling by factorising. In *Proceedings of the 35th International Conference on Machine Learning*, 2018.

Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

Ralph Linsker. Self-organization in a perceptual network. *Computer*, 21(3):105–117, 1988.

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

Michael A Saxe, Yamini Bansal, Joel Dapello, Madhu Advani, Artemy Kolchinsky, Brendan T Daniel, and David D Cox. On the information bottleneck theory of deep learning. *International Conference on Learning Representations*, 2018.

Jürgen Schmidhuber. Learning factorial codes by predictability minimization. *Neural Computation*, 4(6):863–879, 1992.

Ravid Shwartz-Ziv and Naftali Tishby. Opening the black box of deep neural networks via information. *arXiv preprint arXiv:1703.00810*, 2017.

M Studenỳ and J Vejnarova. The multiinformation function as a tool for measuring stochastic dependence. In *Learning in graphical models*, pages 261–297. Springer, 1998.

Valentin Thomas, Jules Pondard, Emmanuel Bengio, Marc Sarfati, Philippe Beaudoin, Marie-Jean Meurs, Joelle Pineau, Doina Precup, and Yoshua Bengio. Independently controllable features. *arXiv preprint arXiv:1708.01289*, 2017.

Naftali Tishby, Fernando C Pereira, and William Bialek. The information bottleneck method. *arXiv preprint physics/0004057*, 2000.

Aaron Van Oord, Nal Kalchbrenner, and Koray Kavukcuoglu. Pixel recurrent neural networks. In *International Conference on Machine Learning*, pages 1747–1756, 2016.

Greg Ver Steeg. Unsupervised learning via total correlation explanation. *IJCAI*, 2017.

Greg Ver Steeg and Aram Galstyan. Discovering structure in high-dimensional data through correlation explanation. In *Advances in Neural Information Processing Systems*, pages 577–585, 2014.

Greg Ver Steeg and Aram Galstyan. Maximally informative hierarchical representations of high-dimensional data. In *Artificial Intelligence and Statistics*, pages 1004–1012, 2015.

Greg Ver Steeg and Aram Galstyan. Low complexity gaussian latent factor models and a blessing of dimensionality. *arXiv preprint arXiv:1706.03353*, 2017.

Satosi Watanabe. Information theoretical analysis of multivariate correlation. *IBM Journal of research and development*, 4(1):66–82, 1960.

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