Information Maximization Auto-Encoding

Dejiao Zhang
University of Michigan, Ann Arbor
dejiao@umich.edu

Tianchen Zhao
University of Michigan, Ann Arbor
ericolon@umich.edu

Laura Balzano
University of Michigan, Ann Arbor
girasole@umich.edu

Abstract

We propose the Information Maximization Autoencoder (IMAE), an information theoretic approach to simultaneously learn continuous and discrete representations in an unsupervised setting. Unlike the Variational Autoencoder framework, IMAE starts from a stochastic encoder that seeks to map each input data to a hybrid discrete and continuous representation with the objective of maximizing the mutual information between the data and their representations. A decoder is included to approximate the posterior distribution of the data given their representations, where a high fidelity approximation can be achieved by leveraging the informative representations. We show that the proposed objective is theoretically valid and provides a new perspective for understanding the tradeoffs regarding informativeness of the representation factors, disentanglement of representations, and decoding quality.

1 Introduction

A central tenet for designing and learning a model for data is that the resulting representation should be compact yet informative. Therefore, the goal of learning can be formulated as finding informative representations about the data under proper constraints. In this work, we propose an information theoretic approach to simultaneously learn continuous and discrete representations in an unsupervised setting. We start with a stochastic encoder $p_\theta(\boldsymbol{z}|\boldsymbol{x})$ and aim at maximizing the mutual information between the data $\boldsymbol{x}$ and its representation $\boldsymbol{z}$. In this setting, a reconstruction or generating phase can be obtained as the variational inference of the true posterior $p_\theta(\boldsymbol{x}|\boldsymbol{z})$. By explicitly seeking informative representations, the proposed model yields better decoding quality. Moreover, we show that the information maximization objective naturally induces a balance between the informativeness of each latent factor and the statistical independence between them, which gives a more principled way to learn semantically meaningful representations.

Another contribution of this work is proposing a framework for simultaneously learning continuous and discrete representations for categorical data. Categorical data are ubiquitous in real-world tasks, where using a hybrid discrete and continuous representation to capture both categorical information and continuous variation in data is more consistent with the natural generation process. We focus on categorical data that are similar in nature, *i.e.,* where different categories still share similar variations (features). We seek to learn semantically meaningful discrete representations while maintaining disentanglement of the continuous representations that capture the variation shared across categories.

2 Information Maximization Representation Learning

Given data $\boldsymbol{x} \in \mathbb{R}^d$, we consider learning a hybrid continuous-discrete representation, denoted respectively with variables $\boldsymbol{z} \in \mathbb{R}^{K_1}$ and $\boldsymbol{y} \in \{1, \dots, K_2\}$, using a stochastic encoder parameterized by θ, *i.e.,* $p_\theta(\boldsymbol{y}, \boldsymbol{z}|\boldsymbol{x})$. We seek to learn compact yet semantically meaningful representations in the

Third workshop on Bayesian Deep Learning (NeurIPS 2018), Montréal, Canada.

sense that they should be low dimensional but informative enough about the data. A natural approach is to maximize the mutual information [Cover and Thomas, 2012] $\mathrm{I}_\theta(\boldsymbol{x}; \boldsymbol{y}, \boldsymbol{z})$ between the data and its representation under the constraint $K_1, K_2 \ll d$. A probabilistic decoder $q_\phi(\boldsymbol{x}|\boldsymbol{y}, \boldsymbol{z})$ is adopted to approximate the true posterior $p_\theta(\boldsymbol{x}|\boldsymbol{y}, \boldsymbol{z})$, which can be hard to estimate or even intractable. The dissimilarity between them is optimized by minimizing the KL divergence $D_{\mathrm{KL}}\left(p_\theta(\boldsymbol{x}|\boldsymbol{y}, \boldsymbol{z})||q_\phi(\boldsymbol{x}|\boldsymbol{y}, \boldsymbol{z})\right)$. In summary, IMAE considers the following,

$$\text{maximize}_{\theta,\phi} \;\; \beta_0 \mathrm{I}_\theta(\boldsymbol{x}; \boldsymbol{y}, \boldsymbol{z}) - D_{\mathrm{KL}}\left(p_\theta(\boldsymbol{x}|\boldsymbol{y}, \boldsymbol{z})||q_\phi(\boldsymbol{x}|\boldsymbol{y}, \boldsymbol{z})\right) \; . \tag{1}$$

Given that $\mathrm{H}(x)$ is independent of the optimization procedure, we can show that optimizing (1) is equivalent to optimizing the following:

$$\text{maximize}_{\theta,\phi} \;\; \beta \mathrm{I}_\theta(\boldsymbol{x}; \boldsymbol{y}, \boldsymbol{z}) + \mathbb{E}_{p_\theta(\boldsymbol{x},\boldsymbol{y},\boldsymbol{z})}\left[\log q_\phi(\boldsymbol{x}|\boldsymbol{y}, \boldsymbol{z})\right], \quad \beta = \beta_0 - 1 > 0 \; . \tag{2}$$

We set $\beta > 0$ to balance between maximizing the informativeness of latent representations and maintaining the decoding quality. The second term is often interpreted as the "reconstruction error" which can be optimized using the reparameterization tricks proposed by [Kingma and Welling, 2013] and [Jang et al., 2016] for the continuous representation $\boldsymbol{z}$ and the discrete representation $\boldsymbol{y}$ respectively. Now we introduce our method to optimize the first term $\mathrm{I}_\theta(\boldsymbol{x}; \boldsymbol{y}, \boldsymbol{z})$ in (2).

2.1 Simultaneously seeking informativeness and disentanglement

Assuming the conditional distribution of the representation $(\boldsymbol{y}, \boldsymbol{z})$ given $\boldsymbol{x}$ is factorial, and similarly for the marginal distribution $p_\theta(\boldsymbol{y}, \boldsymbol{z}) = p_\theta(\boldsymbol{y})p_\theta(\boldsymbol{z})$, then

$$\mathrm{I}_\theta(\boldsymbol{x}; \boldsymbol{y}, \boldsymbol{z}) = \mathrm{I}_\theta(\boldsymbol{x}; \boldsymbol{y}) + \textstyle\sum_{k=1}^{K_1} \mathrm{I}_\theta(\boldsymbol{x}; \boldsymbol{z}_k) - D_{\mathrm{KL}}\left(p_\theta(\boldsymbol{z})\big|\big|\Pi_{k=1}^{K_1} p_\theta(\boldsymbol{z}_k)\right) \; . \tag{3}$$

The first two terms of the RHS quantify how much information each latent factor carries about the data. The last term is known as the *total correlation* of $\boldsymbol{z}$ [Watanabe, 1960], which quantifies the statistical independence between the continuous latent factors and achieves the minimum if and only if they are independent of each other. Eq (3) implies that maximizing $\mathrm{I}_\theta(\boldsymbol{x}; \boldsymbol{y}, \boldsymbol{z})$ can be conducted by maximizing informativeness of each latent factor while simultaneously promoting statistical independence between the continuous factors. Various Monte Carlo based sampling methods have been proposed to optimize the total correlation term [Chen et al., 2018, Esmaeili et al., 2018]; we follow this line. Next we construct tractable approximations for $\mathrm{I}_\theta(\boldsymbol{x}; \boldsymbol{z}_k)$ and $\mathrm{I}_\theta(\boldsymbol{x}; \boldsymbol{y})$ respectively.

2.2 Informative continuous representations

Without any constraints, the mutual information $\mathrm{I}_\theta(\boldsymbol{x}; \boldsymbol{z}_k)$ between the data and its continuous representation factor can be trivially maximized by severely fragmenting and exploding the latent space. Specifically, as shown in Proposition 1[1], $\boldsymbol{z}_k$ is more informative about $\boldsymbol{x}$ if it has less uncertainty given $\boldsymbol{x}$ yet captures more variance in data, *i.e.,* $\sigma_k(\boldsymbol{x})$ is small while $\mu_k(\boldsymbol{x})$ is dispersed within a large space. This can result in discontinuity of the latent representation $\boldsymbol{z}_k$, where in the extreme case each data point is associated with a delta distribution in the latent space $p_\theta(\boldsymbol{z}_k|x^{(i)}) = \delta(z_k^{(i)})$.

Proposition 1 *Suppose the conditional distribution $p_\theta(\boldsymbol{z}|\boldsymbol{x})$ is a factorial Gaussian distribution with mean $\mu(\boldsymbol{x})$ and covariance $\Sigma(\boldsymbol{x})$. Let $\sigma(\boldsymbol{x}) \in \mathbb{R}^{K_1}$ denote the diagonal entries of $\Sigma(\boldsymbol{x})$, then*

$$I_\theta(\boldsymbol{x}; \boldsymbol{z}_k) \leq \frac{1}{2} \log\left[\left(\mathbb{E}_{\boldsymbol{x}}\left[\sigma_k^2(\boldsymbol{x})\right] + \mathrm{Var}_{\boldsymbol{x}}\left[\mu_k(\boldsymbol{x})\right]\right)\right] - \frac{1}{2}\mathbb{E}_{\boldsymbol{x}}\left[\log \sigma_k^2(\boldsymbol{x})\right] \; , \quad k = 1, \ldots, K_1 \; . \tag{4}$$

The equality in (4) is attained if and only if $\boldsymbol{z}_k$ is Gaussian distributed, given which we have

$$I_\theta(\boldsymbol{x}; \boldsymbol{z}_k) \geq \frac{1}{2} \log\left(1 + \mathrm{Var}_{\boldsymbol{x}}\left[\mu_k(\boldsymbol{x})\right] / \mathbb{E}_{\boldsymbol{x}}\left[\sigma_k^2(\boldsymbol{x})\right]\right) \; , \quad k = 1, \ldots, K_1 \; . \tag{5}$$

To remedy this issue while achieving the upper bound in Proposition 1, a natural resolution is to squeeze $\boldsymbol{z}_k$ within the domain of a Gaussian distribution with finite variance. By doing so, we can get a more reasonable trade-off between enlarging the spread of $\mu_k(\boldsymbol{x})$ and maintaining the continuity of $\boldsymbol{z}_k$, while achieving the maximal $\mathrm{I}(\boldsymbol{x}; \boldsymbol{z})$ among all possible solutions with the same variance of $\boldsymbol{z}_k$. Therefore, we consider the following as the surrogate for maximizing $\mathrm{I}_\theta(\boldsymbol{x}; \boldsymbol{z}_k)$,

$$\max_\theta \; \mathcal{L}_\theta(\boldsymbol{z}) := -\textstyle\sum_{k=1}^{K_1} D_{\mathrm{KL}}\left(p_\theta(\boldsymbol{z}_k)||r(\boldsymbol{z}_k)\right) \; . \tag{6}$$

Here $r(\boldsymbol{z}_k)$ is an i.i.d scaled normal distribution with variance being some prefixed finite value.

[1]While similar results have likely been established, we include Proposition 1 to motivate our objective design.

2.3 Informative discrete representations

The mutual information $\mathrm{I}_\theta(\boldsymbol{x};\boldsymbol{y})$ between the data and its discrete representation can be well approximated, given the fact that the cardinality of the space of $\boldsymbol{y}$ is typically low. To be more precise, Proposition 2 shows that, with a suitably large batch of samples, the empirical mutual information $\widehat{\mathrm{I}}_\theta(\boldsymbol{x};\boldsymbol{y})$ is a good approximation to $\mathrm{I}_\theta(\boldsymbol{x};\boldsymbol{y})$. This enables us to optimize $\mathrm{I}_\theta(\boldsymbol{x};\boldsymbol{y})$ in a theoretically justifiable way that is amenable to stochastic gradient descent with minibatches of data. Hence, to maximize $\mathrm{I}_\theta(\boldsymbol{x};\boldsymbol{y})$ we consider the following,

$$\max_\theta \ \mathcal{L}_\theta(\boldsymbol{y}) := \widehat{\mathrm{I}}_\theta(\boldsymbol{x};\boldsymbol{y}) = \mathrm{H}\left(\tfrac{1}{M}\textstyle\sum_{m=1}^{M} p_\theta(\boldsymbol{y}|x_m)\right) - \tfrac{1}{M}\textstyle\sum_{m=1}^{M}\mathrm{H}\left(p_\theta(\boldsymbol{y}|x_m)\right). \tag{7}$$

Proposition 2 *Assume the marginal probabilities of the true and predicted labels are bounded below, i.e.* $p_\theta(\boldsymbol{y}), \widehat{p}_\theta(\boldsymbol{y}) \in [1/(CK_2), 1]$ *for all* $\boldsymbol{y} \in \{1,\dots,K_2\}$ *with some constant* $C > 1$. *Then* $\mathbb{P}\left(\left|I_\theta(\boldsymbol{x};\boldsymbol{y}) - \widehat{I}_\theta(\boldsymbol{x};\boldsymbol{y})\right| \le K_2\left(\max\{\log CK_2 - 1, 1\} + e\right)\sqrt{\frac{\log(2K_2/\delta)}{2M}}\right) \ge 1 - 2\delta$, $\delta \in (0,1)$.

Overall Objective As a summary of (3) (6) and (7), so far we have

$$\max_{\theta,\phi} \ \beta\left(\mathcal{L}_\theta(\boldsymbol{z}) + \mathcal{L}_\theta(\boldsymbol{y}) - D_{\mathrm{KL}}\left[p(\boldsymbol{z})||\Pi_{k=1}^{K_1} p(\boldsymbol{z}_k)\right]\right) + \mathbb{E}_{p_\theta(\boldsymbol{x},\boldsymbol{y},\boldsymbol{z})}\left[\log q_\phi(\boldsymbol{x}|\boldsymbol{y},\boldsymbol{z})\right] .$$

The first three terms are associated with our information maximization objective, while the last one aims at better approximation of the posterior $p_\theta(\boldsymbol{x}|\boldsymbol{y},\boldsymbol{z})$. A better balance between these two targets can be achieved by weighting them differently. The informativeness of each latent factor can be optimized through $\mathcal{L}_\theta(\boldsymbol{z})$ and $\mathcal{L}_\theta(\boldsymbol{y})$, while statistically independent latent continuous factors can be promoted by minimizing the total correlation term. Therefore, trade-offs can be formalized regarding the informativeness of each latent factor, disentanglement of the representation, and better decoding quality. With this final adjustment, we settle on the following overall objective:

$$\begin{aligned}\max_{\theta,\phi} \ \mathcal{L}_{\text{IMAE}} &:= \mathbb{E}_{p_\theta(\boldsymbol{x},\boldsymbol{y},\boldsymbol{z})}\left[\log q_\phi(\boldsymbol{x}|\boldsymbol{y},\boldsymbol{z})\right] \\ &\quad + \beta\mathcal{L}_\theta(\boldsymbol{y}) + \beta\mathcal{L}_\theta(\boldsymbol{z}) - \gamma D_{\mathrm{KL}}\left[p_\theta(\boldsymbol{z})||\Pi_{k=1}^{K_1} p_\theta(\boldsymbol{z}_k)\right] , \quad \beta,\gamma > 0 .\end{aligned} \tag{8}$$

3 Numerical Results

We compare IMAE against a number of VAE based approaches, including β-VAE [Higgins et al., 2016], InfoVAE [Zhao et al., 2017] and JointVAE [Dupont, 2018]. A summarization of these models is provided in Appendix A. We would like to show that IMAE can (i) successfully learn a hybrid continuous and discrete representation, with $\boldsymbol{y}$ matching the natural categorical information well and $\boldsymbol{z}$ capturing the disentangled feature information shared across categories; (ii) outperform the other models on achieving a better trade-off between the interpretability of the representation and the decoding quality. We choose the priors $r(\boldsymbol{z})$ and $r(\boldsymbol{y})$ as isotropic Gaussian distribution and uniform distribution over categories respectively.

3.1 Informative representations yield better interpretability

Figure 1 validates Proposition 1 by showing that, with roughly same amount of variance for each latent variable $\boldsymbol{z}_k$, those achieving high mutual information with the data have mean values $\mu_k(\boldsymbol{x})$ of the conditional probability $p(\boldsymbol{z}_k|\boldsymbol{x})$ disperse across data samples and variances $\sigma_k(\boldsymbol{x})$ decrease to small values for all data samples. As seen in Figure 1(b)-(d), informative variables in the continuous representation have uncovered intuitive latent factors of the variation in the data, while the factor $\boldsymbol{z}_8$ has no mutual information with the data and shows no variation. We observe the same phenomenon for the discrete representation $\boldsymbol{y}$ in Figure 1(e)&(f), which were obtained with two different values of β and γ, where the more informative one matches the natural labels better.

3.2 Quantitative comparisons

In this section, we perform quantitative evaluations on MNIST and dSprites. Before we present our main results, we first describe an assumption that we make on the discrete representations. A

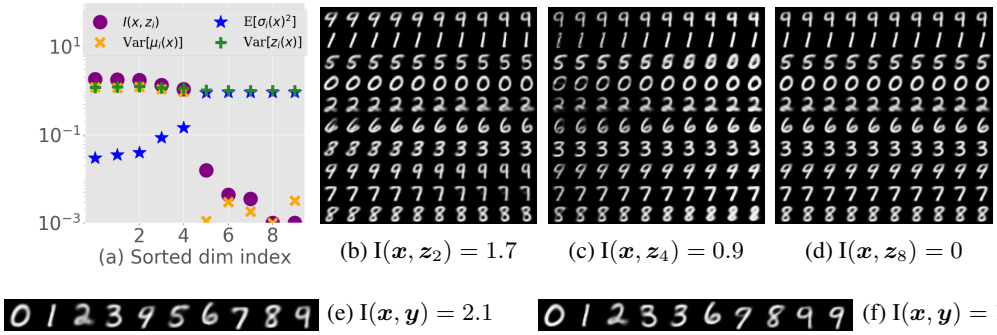

Figure 1: **IMAE on MNIST**. **(a)** Illustration of Proposition 1. **(b)-(d)** Latent traverse on the continuous representations $\boldsymbol{z}$. The rows are conditioned on the discrete representations $\boldsymbol{y}$ learnt by IMAE, where for each row $\boldsymbol{z}$ is initialized by feeding the encoder with randomly selected data whose labels are predicted as $\boldsymbol{y}$. We then manipulate each selected $\boldsymbol{z}_k$ within $[-2, 2]$ while keeping all other dimensions fixed. **(e) & (f)** Discrete representations learnt by IMAE with different β values.

reasonable assumption is that the conditional distribution $p(\boldsymbol{y}|\boldsymbol{x})$ should be locally smooth so that the data samples that are close on their manifold should have high probability of being assigned to the same category [Agakov, 2005]. This assumption is crucial for using neural networks to learn discrete representations, since it's easy for a high capacity model to learn a non-smooth function $p(\boldsymbol{y}|\boldsymbol{x})$ that can abruptly change its predictions without guaranteeing similar data samples will be mapped to similar $\boldsymbol{y}$. To remedy this issue, we adopt the virtual adversarial training (VAT) trick proposed by [Miyato et al., 2016] and augment $\mathcal{L}_\theta(\boldsymbol{y})$ as follows:[2]

$$\max \;\; \mathcal{L}_\theta(\boldsymbol{y}) := \widehat{\mathrm{I}}_\theta(\boldsymbol{x}; \boldsymbol{y}) - \mathbb{E}_{\widehat{p}(\boldsymbol{x})} \left[\max_{\|\eta\| \le \epsilon} \mathrm{H}\left(p_\theta(\boldsymbol{y}|\boldsymbol{x}); p_\theta(\boldsymbol{y}|\boldsymbol{x} + \eta) \right) \right] \; . \tag{9}$$

The second term of RHS regularizes $p_\theta(\boldsymbol{y}|\boldsymbol{x})$ to be consistent within the ϵ norm ball of each data sample so as to maintain local smoothness of the prediction model. *For fair comparison, we augment all four methods with VAT. We found that using VAT is essential for all of them except β-VAE to learn interpretable discrete representations.*

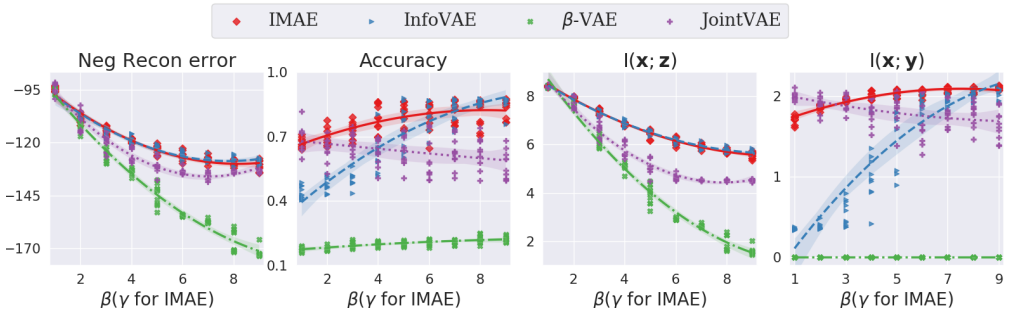

Figure 2: **IMAE on MNIST**. We track the key quantities for different models by sweeping β. We set $\gamma = 2\beta$ for IMAE. For each β, we run each method over 10 random initializations.

MNIST Figure 4 shows that, by simply pushing the conditional distribution $p(\boldsymbol{y}|\boldsymbol{x})$ towards the uniform distribution $r(\boldsymbol{y})$, β-VAE sacrifices the mutual information $\mathrm{I}(\boldsymbol{x}; \boldsymbol{y})$ and hence struggles in learning interpretable discrete representation even with VAT. On the other hand, large β values drive β-VAE to sacrifice more mutual information $\mathrm{I}(\boldsymbol{x}; \boldsymbol{z})$ between the data and its continuous representations, which together with the less informative discrete representation result in poor decoding quality. In contrast, the other three methods can remedy this issue to different degrees, and hence attain better trade-off regarding informativeness of latent representations and decoding quality. Compared to JointVAE and InfoVAE, IMAE is more capable of learning discrete representations over a wide range of β, γ values, which implies that less overlap between the manifolds of different categories

[2] In this paper, we set $\epsilon = 1$ across datasets. VAT can be effectively approximated by a pair of forward and backward passes [Miyato et al., 2016].

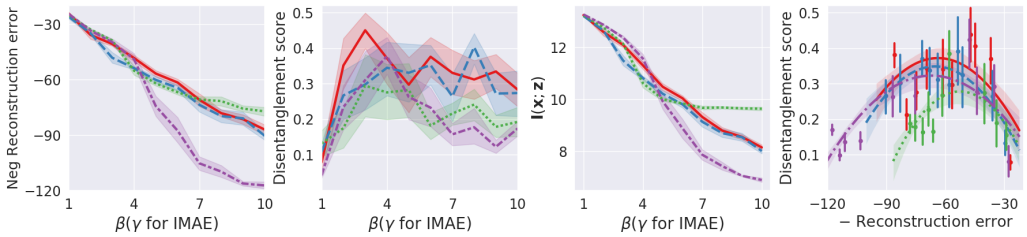

Figure 3: **Disentanglement comparison on dSprites.** IMAE performs well regarding the disentanglement score vs. decoding quality trade-off, especially in the region of interest where both decoding quality and informativeness of representations are fairly good. The results are reported by training each method with $\beta \in [1, 10]$ (we set $\beta = \gamma/2$ with $\gamma \in [1, 10]$ for IMAE). For each β value, every method is trained over 10 random initializations. Shade regions indicate the 80% confidence intervals.

is induced. As a result, IMAE is expected to yield better decoding quality for each category, seen for MNIST in Figure 2. Although InfoVAE and JointVAE also learn comparatively good discrete representations with large and small β values respectively, the corresponding results of these two regions have poor decoding quality or a much lower disentanglement score (see Figure 3). In contrast, IMAE consistently performs well with different hyperparameter values, especially in the region of interest where the decoding quality and the informativeness of latent representations are good enough.

2D Shapes We now evaluate IMAE on dSprites where the ground truth of both continuous and discrete representaion factors are available. We use the disentanglement metric proposed by [Chen et al., 2018]. The higher the disentanglement score is, the more disentangled the representation factors are. [3]

As shown in Figure 3, with large β values, β-VAE penalizes the mutual information too much and this degrades the usefulness of representations, while all other three methods achieve higher disentanglement score with better decoding quality. For JointVAE, higher β values push the upper bound of the mutual information to converge to the prefixed target value; it therefore can maintain more mutual information between the data and the overall latent representations and give better decoding quality. However, the associated disentanglement quality is poor. This implies that simply restricting the overall capacity of the latent representations is not enough for learning disentangled representations. While InfoVAE yields a comparatively better disentanglement score by pushing the marginal joint distribution of the representations towards a factorial distribution more aggressively with large values of β, the associated decoding quality and informativeness of latent representations are both poor. In contrast, IMAE is capable of achieving a better trade-off between the disentanglement score and the decoding quality in the region of interest where the decoding quality as well as the informativeness are fairly good. We attribute this to the effect of explicitly seeking statistically independent latent factors by minimizing the total correlation term in our objective.

4 Conclusion

Unsupervised joint learning of disentangled continuous and discrete representations is a challenging problem due to the lack of a prior for semantic awareness and other inherent difficulties that arise in learning discrete representations. This work takes a step towards achieving this goal. A limitation of our model is that it pursues disentanglement by assuming or trying to encourage independent scalar latent factors, which may not always be sufficient for representing real data. For example, data may exhibit category specific variation, or a subset of latent factors might be correlated. This motivates us to explore more structured disentangled representations; one possible direction is to encourage group independence. We leave this for future work.

5 Acknowledgement

Dejiao Zhang and Laura Balzano were supported by DARPA grant 16-43-D3M-FP-037. The authors also would like to thank Zeyu Sun for the initial participation in this project.

[3] Although the truth discrete factor is provided, we evaluate the disentanglement quality only in terms of the continuous representations since the pixel-wise difference between different categories are very small.

References

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

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

Eric Jang, Shixiang Gu, and Ben Poole. Categorical reparameterization with gumbel-softmax. *arXiv preprint arXiv:1611.01144*, 2016.

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

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

Babak Esmaeili, Hao Wu, Sarthak Jain, Siddharth Narayanaswamy, Brooks Paige, and Jan-Willem van de Meent. Hierarchical disentangled representations. *arXiv preprint arXiv:1804.02086*, 2018.

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. 2016.

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

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

Felix Vsevolodovich Agakov. *Variational Information Maximization in Stochastic Environments*. PhD thesis, University of Edinburgh, 2005.

Takeru Miyato, Andrew M Dai, and Ian Goodfellow. Virtual adversarial training for semi-supervised text classification. 2016.

A Supplement for the numerical results

$$\mathcal{L}_{\text{VAE}} = \boxed{\mathbb{E}_{p(\boldsymbol{y},\boldsymbol{z}|\boldsymbol{x})}\left[q(\boldsymbol{x}|\boldsymbol{y},\boldsymbol{z})\right] - D_{\text{KL}}\left(p(\boldsymbol{z}|\boldsymbol{x})||r(\boldsymbol{z})\right) - D_{\text{KL}}\left(p(\boldsymbol{y}|\boldsymbol{x})||r(\boldsymbol{y})\right)} \leftarrow \text{ELBO}$$

$$= \underbrace{\mathbb{E}_{p(\boldsymbol{y},\boldsymbol{z}|\boldsymbol{x})}\left[q(\boldsymbol{x}|\boldsymbol{y},\boldsymbol{z})\right]}_{①} - \underbrace{I(\boldsymbol{x};\boldsymbol{y})}_{②} - \underbrace{D_{\text{KL}}\left(p(\boldsymbol{y})||r(\boldsymbol{y})\right)}_{③} - \underbrace{I(\boldsymbol{x};\boldsymbol{z})}_{④} - \underbrace{D_{\text{KL}}\left(p(\boldsymbol{z})||r(\boldsymbol{z})\right)}_{⑤}$$

β-VAE: $① - \beta\left(② + ③\right) - \beta\left(④ + ⑤\right)$ InfoVAE: $① - \beta\, ③ - \beta\, ⑤$

Joint-VAE: $① - \beta\left|② + ③ - C_{\boldsymbol{y}}\right| - \beta\left|④ + ⑤ - C_{\boldsymbol{z}}\right|$

Figure 4: Summarization of relevant works. β-VAE modifies ELBO by increasing the penalty on the KL divergence terms. Both InfoVAE and WAE drop the mutual information terms from ELBO. JointVAE seeks to control the mutual information by pushing the their upper bounds, *i.e.,* the associated KL divergence terms, towards progressively increased target values, $C_{\boldsymbol{y}}$ and $C_{\boldsymbol{z}}$.

Table 1: Encoder and Decoder architecture for MNIST and Fashion MNIST.

Encoder	**Decoder**
Input vectorized 28×28 grayscale image	Input $\boldsymbol{y} \in \mathbb{R}^{10}$ and $\boldsymbol{z} \in \mathbb{R}^{10}$
FC. 500 BatchNorm ReLU	FC. 500 ReLU
FC. 2×500 BatchNorm ReLU	FC. 500 ReLU
FC. 20 $(\mu_{\boldsymbol{z}}, \log \sigma_{\boldsymbol{z}})$ + 10 $(p_{\boldsymbol{y}})$	FC. 28×28 Sigmoid

Table 2: Encoder and Decoder architecture for dSprites.

Encoder	**Decoder**
Input vectorized 64×64 grayscale image	Input $\boldsymbol{y} \in \mathbb{R}^{3}$ and $\boldsymbol{z} \in \mathbb{R}^{10}$
FC. 1200 ReLU	FC. 1200 ReLU
FC. 1200 ReLU	FC. 1200 ReLU
FC. 2×1200 ReLU	FC. 1200 ReLU
FC. 20 $(\mu_{\boldsymbol{z}}, \log \sigma_{\boldsymbol{z}})$ + 3 $(p_{\boldsymbol{y}})$	FC. 28×28 Sigmoid