INFORMATION MAXIMIZATION AUTO-ENCODING

Anonymous authors
Paper under double-blind review

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 (VAE) 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 the representation. A decoder is included for approximating the posterior distribution of the data given their representations, where a high fidelity approximation can be achieved by leveraging our informative learned representations. We show that our objective is theoretically valid and provides a principled framework for understanding the tradeoffs among the informativeness of each representation factor, disentanglement of representations, and the 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. Generative latent variable models are a popular approach to this problem, where a model parameterized by θ of the form $p_\theta(\boldsymbol{x}) = \int p_\theta(\boldsymbol{x}|\boldsymbol{z})p(\boldsymbol{z})d\boldsymbol{z}$ is used to represent the relationship between the data $\boldsymbol{x}$ and the low dimensional latent variable $\boldsymbol{z}$. The model is optimized by fitting the generative data distribution $p_\theta(\boldsymbol{x})$ to the training data distribution $\widehat{p}(\boldsymbol{x})$, which involves maximizing the likelihood for θ. Typically, this model is intractable even for moderately complicated functions $p_\theta(\boldsymbol{x}|\boldsymbol{z})$ and continuous latent representations $\boldsymbol{z}$. To remedy this issue, the variational autoencoder (VAE) framework (Kingma and Welling, 2013; Rezende et al., 2014) proposes maximizing the evidence lower bound (ELBO) of the marginal likelihood objective.

However, as was initially pointed out in (Hoffman and Johnson, 2016), maximizing ELBO also penalizes the mutual information between data and their representations. This in turn makes the representation learning harder and leads to poor reconstructions of the data as well. Many recent efforts have been focused on resolving this problem by revising ELBO. Generally speaking, these works fall into two categories. One line of work targets "disentangled representations" by encouraging the statistical independence between components of latent representations (Higgins et al., 2016; Kim and Mnih, 2018; Gao et al., 2018; Chen et al., 2018; Esmaeili et al., 2018), while the other line of work seeks to explicitly control or encourage the mutual information between the data and their representations (Mary Phuong, 2018; Burgess et al., 2018; Alemi et al., 2017; Dupont, 2018; Zhao et al., 2017). However, these approaches either result in an invalid lower bound for the VAE objective or cannot avoid minimizing the mutual information between data and their representations.

In this work, instead of starting with the generative latent variable model, 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 the representations $\boldsymbol{z}$. 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 targeting informative representations, the proposed model is capable of yielding better decoding quality. Moreover, as we will show in Section 3, 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 without invalidating the original objective or removing individual terms from the ELBO.

Another contribution of this work is a framework for simultaneously learning continuous and discrete representations, which yields a more flexible model for data that belong to different categories. Categorical data are ubiquitous in real-world tasks, and 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. In this work, 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 latent factors that capture the variations shared over different categories. We show that, compared to the VAE based approaches, our information maximization objective leads to a more natural yet powerful way for learning these hybrid representations.

2 Related Work

Recently, there has been a surge of interest in learning interpretable representations. Among them, β-VAE (Higgins et al., 2016) is a popular method for learning disentangled representations, which modifies ELBO by increasing the penalty on the KL divergence between the variational posterior and the factorized prior. However, by using large weight for the KL divergence term, β-VAE also penalizes the mutual information between the data and the latent representations more than a standard VAE does, resulting in more severe under utilization of the latent representation space.

Several follow up works propose different approaches to address the limitations of β-VAE. (Dupont, 2018; Alemi et al., 2017; Burgess et al., 2018; Mary Phuong, 2018) propose to constrain the mutual information between the representations and the data by pushing its upper bound, *i.e.,* the KL divergence term in ELBO, towards a progressively increased target value. However, specifying and tuning this target value can itself be very challenging, which makes this method less practical. Moreover, this extra constraint results in an invalid lower bound for the VAE objective. Alternatively, (Zhao et al., 2017) drops the mutual information term in ELBO. By pushing only the aggregated posterior towards a factorial prior, they implicitly encourage independence across the dimensions of latent representations without sacrificing the informativeness of the representations. However, simply removing the mutual information term also violates the lower bound of the VAE objective.

Another relevant line of work (Gao et al., 2018; Kim and Mnih, 2018; Chen et al., 2018; Esmaeili et al., 2018) seek to learn disentangled representations by explicitly encouraging statistical independence between latent factors. They all propose to minimize the *total correlation* term of the latent representations, either augmented as an extra term to ELBO or obtained by reinterpreting or re-weighting the terms in the VAE objective, as a way to encourage statistical independence between the representation components. In contrast, we show that our information maximization objective inherently contains the total correlation term while simultaneously seeking to maximize the informativeness of each representation factor.

In this paper, we introduce a different perspective to the growing body of the VAE based approaches for unsupervised representation learning. Starting by seeking informative representations for our data, we follow a more intuitive way to maximize the mutual information between the data and the representations. Moreover, we augment the continuous representation with a discrete one, which gives us more flexibility to model real world data that are generated from different categories. We invoke the information maximization principle (Linsker, 1988; Bell and Sejnowski, 1995) with proper constraints implied by the objective itself to avoid degenerate solutions. The proposed objective gives a more natural yet theoretically elegant way to learn semantically meaningful representations.

3 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 $\boldsymbol{y} \in \{1, \ldots, K_2\}$, using a stochastic encoder parameterized by θ, *i.e.,* $p_\theta(\boldsymbol{y}, \boldsymbol{z}|\boldsymbol{x})$. We seek semantically meaningful representations in the sense that they should be informative enough about the data. A natural approach is to maximize the mutual information (Cover and Thomas, 2012) between the data and its representations, $\mathrm{I}_\theta(\boldsymbol{x}; \boldsymbol{y}, \boldsymbol{z})$. Here the mutual information between two random variables, *e.g.,* $\boldsymbol{x}$ and $\boldsymbol{z}$, is defined as $\mathrm{I}_\theta(\boldsymbol{x}; \boldsymbol{z}) = \mathrm{H}_\theta(\boldsymbol{z}) - \mathrm{H}_\theta(\boldsymbol{z}|\boldsymbol{x})$, where $\mathrm{H}_\theta(\boldsymbol{z}) = -\mathbb{E}_{p_\theta(\boldsymbol{z})}\left[\log p_\theta(\boldsymbol{z})\right]$ is the entropy of $\boldsymbol{z}$ and $\mathrm{H}_\theta(\boldsymbol{z}|\boldsymbol{x}) = -\mathbb{E}_{p_\theta(\boldsymbol{x},\boldsymbol{z})}\left[\log p_\theta(\boldsymbol{z}|\boldsymbol{x})\right]$ is the conditional entropy of $\boldsymbol{z}$ given $\boldsymbol{x}$. The mutual information can therefore be interpreted as the

decrease in uncertainty of one random variable given another variable. In other words, it quantifies how much information one random variable has about the other.

In IMAE, 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. We want to minimize the dissimilarity between the true posterior and the approximated one, for which we consider minimizing the KL divergence between them $D_{\mathrm{KL}}\left(p_\theta(\boldsymbol{x}|\boldsymbol{y},\boldsymbol{z})||q_\phi(\boldsymbol{x}|\boldsymbol{y},\boldsymbol{z})\right)$. Therefore, the objective of IMAE is formulated as

$$\beta \mathrm{I}_\theta(\boldsymbol{x};\boldsymbol{y},\boldsymbol{z}) - D_{\mathrm{KL}}\left(p_\theta(\boldsymbol{x}|\boldsymbol{y},\boldsymbol{z})||q_\phi(\boldsymbol{x}|\boldsymbol{y},\boldsymbol{z})\right), \quad \beta > 0 \tag{1}$$

where β is introduced for a better balance between the seeking for informative latent representations and maintaining good posterior approximation quality. By leveraging the fact that $H(x)$ is independent of the optimization procedure, we can show that ptimizing (1) is equivalent to optimize the following [1].

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

where the first quantifies the informativeness of representations while the second term can be interpreted as the "reconstruction error". We set $\beta > 1$ in order to achieve a balance between the informativeness of representations and the decoding quality.

To optimize the first term, we follow the literature and assume that the conditional distribution of the representation $(\boldsymbol{y},\boldsymbol{z})$ given $\boldsymbol{x}$ is factorial. Given this, $\mathrm{I}_\theta(\boldsymbol{x};\boldsymbol{y},\boldsymbol{z})$ can be decomposed as[2]

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

The first two terms of the LHS quantify how much information each latent factor, *i.e.,* $\boldsymbol{y}$ or $\boldsymbol{z}_k$, carry about the data. The last term is known as *total correlation* of $(\boldsymbol{y},\boldsymbol{z})$ (Watanabe, 1960), which characterizes the dependence between the latent factors and achieves the minimum if and only if the latent factors are independent of each other. As is implied by Eq (3), the information maximization objective inherently involves a balance between maintaining the informativeness of each latent factor and the statistical independence between them. Now we proceed by first constructing tractable approximations for $\mathrm{I}_\theta(\boldsymbol{x};\boldsymbol{z}_k)$ and $\mathrm{I}_\theta(\boldsymbol{x};\boldsymbol{y})$ respectively.

3.1 Informative continuous representations

Without any constraints, the mutual information $\mathrm{I}_\theta(\boldsymbol{x};\boldsymbol{z})$ between a continuous latent factor and data can be trivially maximized by severely fragmenting the latent space. This can be justified by the fact that $\mathrm{I}_\theta(\boldsymbol{x};\boldsymbol{z}_k) = \mathrm{H}_\theta(\boldsymbol{z}_k) - \mathrm{H}_\theta(\boldsymbol{z}_k|\boldsymbol{x})$. Hence, to maximize $\mathrm{I}(\boldsymbol{x};\boldsymbol{z}_k)$, we can simply map each data $\boldsymbol{x}^{(i)}$ to a deterministic representation $\boldsymbol{z}^{(i)}$ so that the conditional entropy is minimized, while maximizing the uncertainty in $\boldsymbol{z}_k$ by always mapping to a distinct $\boldsymbol{z}_k$. To avoid this degenerate solution and motivate our objective, we call out the following proposition. While similar results have likely been established in the information theory literature, we include this proposition to give insight into our objective design.

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 $\forall k = 1,\ldots,K_1$,*

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

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

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

The lower bound in (5) implies that, in order to maximize $\mathrm{I}_\theta(\boldsymbol{x};\boldsymbol{z}_k)$, we should increase the variance of the mean $\mu_k(\boldsymbol{x})$ while decreasing the variance $\sigma_k(\boldsymbol{x})$. Note here both $\mu_k(\boldsymbol{x})$ and $\sigma_k(\boldsymbol{x})$ are random variables. This matches the discussion above, *i.e.,* $\boldsymbol{z}_k$ is more informative about $\boldsymbol{x}$ if it

[1] A detailed derivation is provided in Appendix A.1

[2] A detailed derivation is provided in Appendix A.

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.

In light of this, we can make what we described above more precise. A vanishing variance of the conditional distribution $p(\boldsymbol{z}_k|\boldsymbol{x})$ leads to a plain autoencoder that maps each data sample to a deterministic latent point, which will fragment the latent space where each data point is associated with a delta distribution in the latent space $p_\theta(\boldsymbol{z}_k|\boldsymbol{x}^i) = \delta(\boldsymbol{z}_k^i)$. On the other hand, Proposition 1 also implies that controlling the variance $\sigma_k(\boldsymbol{x})$ to be finite, $\mathrm{I}_\theta(\boldsymbol{x};\boldsymbol{z}_k)$ will be maximized by pushing $\mu_k(\boldsymbol{x})$ towards two extremes ($\pm\infty$). To avoid such a trivial solution while achieving the upper bound, a natural resolution is to squeeze $\boldsymbol{z}_k$ within the domain of a Gaussian distribution with finite mean and variance. By doing so, the representation $\boldsymbol{z}_k$ will achieve a more reasonable trade off between enlarging the spread of $\mu_k(\boldsymbol{x})$ and maintaining the continuity of $\boldsymbol{z}$.

In this paper, to do this we use a regularizing distribution $r(\boldsymbol{z})$, which is a factorial Gaussian, and push $p_\theta(\boldsymbol{z}_k)$ towards each $r(\boldsymbol{z}_k)$ by minimizing the KL divergence $D_{\mathrm{KL}}\left(p_\theta(\boldsymbol{z}_k)||r(\boldsymbol{z}_k)\right)$:

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

3.2 Informative discrete representations

Unlike the continuous representation, the mutual information $\mathrm{I}_\theta(\boldsymbol{x};\boldsymbol{y})$ between a discrete representation and data can be well approximated, given the fact that the cardinality of the space of $\boldsymbol{y}$ is typically low. To be more specific, given N i.i.d samples $\{x_n\}_{n=1}^N$ from some unknown distribution $p(\boldsymbol{x})$ of the data, the empirical estimation of $\mathrm{I}_\theta(\boldsymbol{x};\boldsymbol{y})$ under the conditional distribution $p_\theta(\boldsymbol{y}|x_n)$ follows as

$$\widehat{\mathrm{I}}_\theta(\boldsymbol{x};\boldsymbol{y}) = \widehat{\mathrm{H}}_\theta(\boldsymbol{y}) - \widehat{\mathrm{H}}_\theta(\boldsymbol{y}|\boldsymbol{x}) = \mathrm{H}\left(\frac{1}{N}\textstyle\sum_{n=1}^N p_\theta(\boldsymbol{y}|x_n)\right) - \frac{1}{N}\textstyle\sum_{n=1}^N \mathrm{H}\left(p_\theta(\boldsymbol{y}|x_n)\right)\ . \tag{7}$$

As we will now show, with a suitably large batch of samples, the empirical mutual information $\widehat{\mathrm{I}}_\theta(\boldsymbol{x};\boldsymbol{y})$ is a good approximation to the population mutual information. 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.

Proposition 2. *Let $\boldsymbol{y}$ be a discrete random variable that belongs to some categorical class $\mathcal{C}$. 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 $\boldsymbol{y} \in \mathcal{C}$ for some constant $C > 1$. Then with probability exceeding $1 - 2\delta$,*

$$\left|I_\theta(\boldsymbol{x};\boldsymbol{y}) - \widehat{I}_\theta(\boldsymbol{x};\boldsymbol{y})\right| \leq K_2\left(\max\{\log CK_2 - 1, 1\} + e\right)\sqrt{\frac{\log(2K_2/\delta)}{2N}}\ . \tag{8}$$

where N denotes the number of samples used to establish $\widehat{I}_\theta(\boldsymbol{x};\boldsymbol{y})$ according to Eq (7).

Therefore, to maximize the mutual information $\mathrm{I}_\theta(\boldsymbol{x};\boldsymbol{y})$, we consider the following:

$$\max\ \ \mathcal{L}_\theta := \widehat{\mathrm{I}}_\theta(\boldsymbol{x};\boldsymbol{y}). \tag{9}$$

Maximizing the the mutual information $\mathrm{I}_\theta(\boldsymbol{x};\boldsymbol{y})$ provides a natural way to learn discrete categorical representations. To see this, notice that $\mathrm{I}_\theta(\boldsymbol{x};\boldsymbol{y})$ contains two fundamental quantities, the category balance term $\mathrm{H}_\theta(\boldsymbol{y})$ and the category separation term $\mathrm{H}_\theta(\boldsymbol{y}|\boldsymbol{x})$. In other words, maximizing $\mathrm{I}_\theta(\boldsymbol{x};\boldsymbol{y})$ trades off uniformly assigning data over categories and seeking highly confident categorical identity for each sample $\boldsymbol{x}$. The maximum is achieved if $p_\theta(\boldsymbol{y}|\boldsymbol{x})$ is deterministic while the marginal distribution $p_\theta(\boldsymbol{y})$ is uniform, that is $\mathrm{H}_\theta(\boldsymbol{y}|\boldsymbol{x}) = 0$ and $\mathrm{H}_\theta(\boldsymbol{y}) = \log K_2$.

Overall Objective Our overall objective can therefore be summarized as

$$\beta\left(\max_{\theta,\phi}\ \ \mathcal{L}_\theta(\boldsymbol{z}) + \mathcal{L}_\theta(\boldsymbol{y}) - D_{\mathrm{KL}}\left[p(\boldsymbol{y},\boldsymbol{z})||p(\boldsymbol{y})\Pi_{k=1}^{K_1}p(\boldsymbol{z}_k)\right]\right) + \mathbb{E}_{p_\theta(\boldsymbol{x},\boldsymbol{y},\boldsymbol{z})}\left[\log q_\phi(\boldsymbol{x}|\boldsymbol{y},\boldsymbol{z})\right]\ . \tag{10}$$

where $\mathcal{L}_\theta(\boldsymbol{z})$ and $\mathcal{L}_\theta(\boldsymbol{y})$ are defined in Eqs (6) and (12), respectively. As we discussed before, the first three terms associate 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. One the other hand, 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 factors can be induced by minimizing the total correlation term $D_{\mathrm{KL}}\left[p(\boldsymbol{y},\boldsymbol{z})||p(\boldsymbol{y})\Pi_{k=1}^{K_1}p(\boldsymbol{z}_k)\right]$. Therefore, trade-offs can be formalized among different terms so as to achieve a better balance regarding the informativeness of each latent factor, disentanglement of the representation, and better decoding quality. This motivates us to consider the following objective, with tradeoff parameters $\beta,\gamma>0$:

$$\max_{\theta,\phi}\ \beta\left(\mathcal{L}_\theta(\boldsymbol{z})+\mathcal{L}_D(\boldsymbol{y})-\gamma D_{\mathrm{KL}}\left[p(\boldsymbol{y},\boldsymbol{z})||p(\boldsymbol{y})\Pi_{k=1}^{K_1}p(\boldsymbol{z}_k)\right]\right)+\mathbb{E}_{p_\theta(\boldsymbol{x},\boldsymbol{y},\boldsymbol{z})}\left[\log q_\phi(\boldsymbol{x}|\boldsymbol{y},\boldsymbol{z})\right]\ . \tag{11}$$

Similar to VAE, the reconstruction term $\mathbb{E}_{p_\theta(\boldsymbol{x},\boldsymbol{y},\boldsymbol{z})}\left[\log q_\phi(\boldsymbol{x}|\boldsymbol{y},\boldsymbol{z})\right]$ is optimized using reparameterization tricks proposed by (Kingma and Welling, 2013) and (Jang et al., 2016) for continuous representations $\boldsymbol{z}$ and discrete representation $\boldsymbol{y}$ respectively. As for the marginal distributions, we approximate them using Monte Carlo sampling methods. Please refer to B for details.

4 Experimental Results

We compare IMAE against a number of VAE based approaches summarized in Figure 1. We would like to demonstrate that IMAE can (i) successfully learn a hybrid of continuous ($\boldsymbol{z}$) and discrete ($\boldsymbol{y}$) representations, with $\boldsymbol{y}$ matching the nature categorical information $\boldsymbol{y}_{\textbf{true}}$ well and $\boldsymbol{z}$ capturing the disentangled feature information shared across categories; (ii) outperform the VAE based models on achieving a better trade-off between the interpretability of the representations and the decoding quality. We evaluate all models on three different datasets: MNIST, Fashion MNIST (F-MNIST) and CelebA. We choose the priors $r(\boldsymbol{y})$ and $r(\boldsymbol{z})$ as isotropic Gaussian distribution and uniform distribution over categories correspondingly. Detailed experiment settings is provided in Appendix E. For notational convenience, we drop the subscripts θ and ϕ in this section.

$$\begin{aligned}\mathcal{L}_{\mathrm{VAE}} &= \boxed{\mathbb{E}_{p(\boldsymbol{y},\boldsymbol{z}|\boldsymbol{x})}\left[q(\boldsymbol{x}|\boldsymbol{y},\boldsymbol{z})\right]-D_{\mathrm{KL}}\left(p(\boldsymbol{z}|\boldsymbol{x})||r(\boldsymbol{z})\right)-D_{\mathrm{KL}}\left(p(\boldsymbol{y}|\boldsymbol{x})||r(\boldsymbol{y})\right)} \leftarrow \text{ELBO}\\ &= \underbrace{\mathbb{E}_{p(\boldsymbol{y},\boldsymbol{z}|\boldsymbol{x})}\left[q(\boldsymbol{x}|\boldsymbol{y},\boldsymbol{z})\right]}_{①}-\underbrace{I(\boldsymbol{x};\boldsymbol{y})}_{②}-\underbrace{D_{\mathrm{KL}}\left(p(\boldsymbol{y})||r(\boldsymbol{y})\right)}_{③}-\underbrace{I(\boldsymbol{x};\boldsymbol{z})}_{④}-\underbrace{D_{\mathrm{KL}}\left(p(\boldsymbol{z})||r(\boldsymbol{z})\right)}_{⑤}\end{aligned}$$

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

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

Figure 1: 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 $C_{\boldsymbol{z}}$.

To evaluate the ability of IMAE on jointly learning the continuous and discrete representations, we conduct both qualitative and quantity experiments on MNIST and FMNIST.

4.1 Informative continuous latent representations yield interpretability

In this section, we evaluate how the informative representations can yield better interpretability. We start by qualitatively examining Proposition 1. As we can see, Figure 2 validates Proposition 1 in a way that, with roughly same amount of variance for each latent variable $\boldsymbol{z}_k$, those achieving high mutual information between the data and themselves have the means $\mu_k(\boldsymbol{x})$ of the conditional probability $p(\boldsymbol{z}_k|\boldsymbol{x})$ disperse across data samples and the variance $\sigma_k(\boldsymbol{x})$ decrease to small values for all data samples. As a qualitative assessment of how the informative representations lead to better interpretability for continuous latent factors, we traverse the latent dimensions corresponding with the largest, medium, and smallest mutual information between themselves and the data. As shown in Figure 2, informative representations is capable of uncovering intuitive factors of variation in the data, while the non-informative one cannot. This matches the insight from recently developed

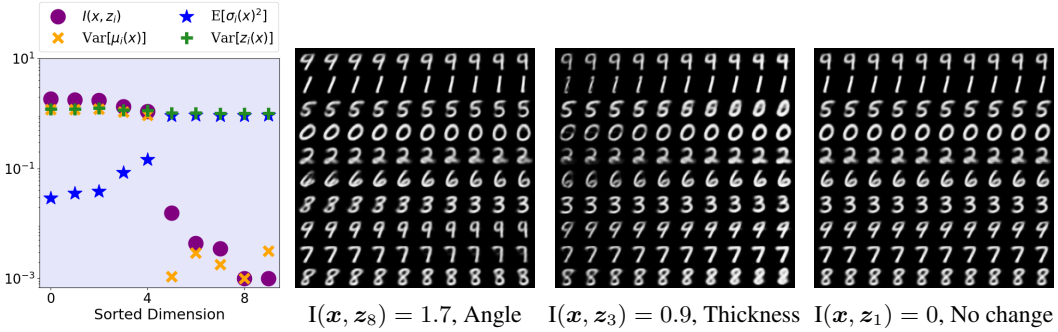


Figure 2: **From left to right:** (1) Illustration of Proposition 1. (2) - (5) Latent traverse for IMAE on MNIST. We get the initial latent representations by feeding the encoder with randomly selected images. For each image, we then traverse the latent dimensions corresponding with the largest, medium, and smallest mutual information between themselves and the data by systematically manipulating each of them within the range $[-2, 2]$ while keeping all other dimensions fixed.

methods (Chen et al., 2016; Gao et al., 2018) that interpretable representations can be attained by maximizing the mutual information between themselves and the data.

4.2 Unsupervised learning of discrete latent factor

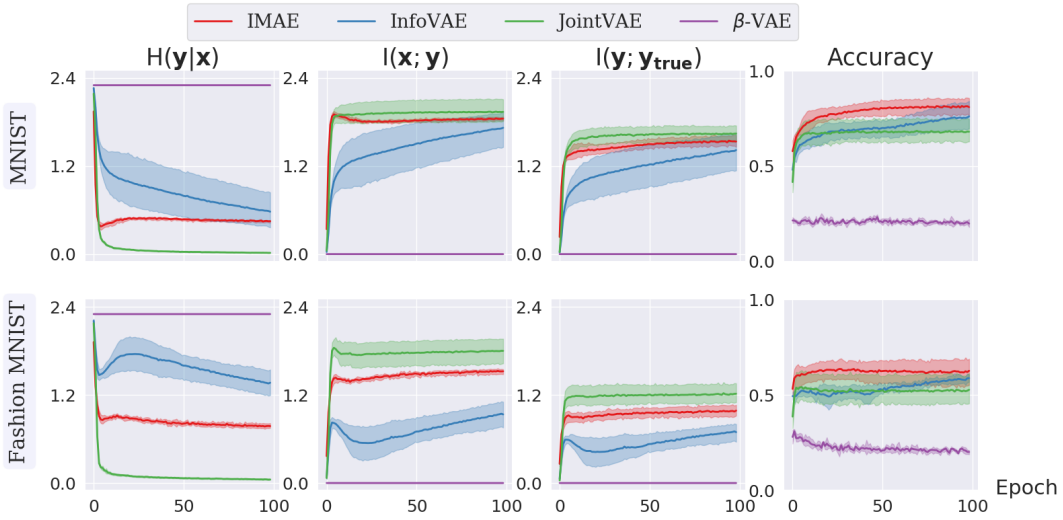
Before we proceed on the numerical results, we introduce an assumption on the conditional distribution $p(\boldsymbol{y}|\boldsymbol{x})$, which is crucial for learning categorical representations. A reasonable assumption for categorical representation is that the resulting distribution $p_\theta(\boldsymbol{y}|\boldsymbol{x})$ should be locally smooth such that the data that are close on their manifold should have high probability of being assigned to the same category (Agakov, 2005). This assumption is especially crucial for using deep neural network to learn discrete representations, given that the high capacity of neural network allows it to easily learn a non-smooth function $p_\theta(\boldsymbol{y}|\boldsymbol{x})$ which can abruptly change its predictions without guaranteeing similar data samples have the same categorical representation. To remedy this issue, we adopt the virtual adversarial training technique (VAT) proposed by (Miyato et al., 2016) to encourage the local smoothness of $p_\theta(\boldsymbol{y}|\boldsymbol{x})$ and consider the following [3]

$$\max \quad \mathcal{L}_D := \mathrm{I}_\theta(\boldsymbol{x}; \boldsymbol{y}) - \lambda \mathbb{E}_{\widehat{p}(\boldsymbol{x})} \left[\max_{\|\eta\| \leq \epsilon} \mathrm{H}\left(p_\theta(\boldsymbol{y}|\boldsymbol{x}); p_\theta(\boldsymbol{y}|\boldsymbol{x}+\eta)\right) \right] \ , \lambda > 0 \ . \tag{12}$$

Here the second term of RHS explicitly enforces $p_\theta(\boldsymbol{y}|\boldsymbol{x})$ being consistent within the ϵ norm ball of each data sample so as to maintain the local smoothness of the prediction model. We augment all models explored in this section with this VAT regularization, which is very important for all of them except β-VAE to learn semantically meaningful discrete representation.

Figure (3a) shows that, by simply pushing the prediction $p(\boldsymbol{y}|\boldsymbol{x})$ towards a uniform distribution $r(\boldsymbol{y})$, β-VAE sacrifices the mutual information $\mathrm{I}(\boldsymbol{x}; \boldsymbol{y})$ and hence struggles in discovering semantically meaningful discrete latent factor even with the VAT. As a comparison, InfoVAE achieves much better performance by dropping $\mathrm{I}(\boldsymbol{x}; \boldsymbol{y})$ from ELBO. By minimizing the KL divergence $D_{\mathrm{KL}}(p(\boldsymbol{y})||r(\boldsymbol{y}))$ minimizing $D_{\mathrm{KL}}(p(\boldsymbol{y})||r(\boldsymbol{y}))$, InfoVAE also pushes the whole dataset to uniformly distributed over categories. Therefore, for data that is distinctive enough between categories, *e.g.,* MNIST, minimizing this KL divergence together with the local smoothness encouraged regularization VAT can yield good performance. However, InfoVAE struggles as we move to less distinctive data, *e.g.,* F-MNIST, for which it cannot give fairly confident category separation by only requiring local smoothness. This is demonstrated by high value of the conditional entropy $H(\boldsymbol{y}|\boldsymbol{x})$ and low values of $\mathrm{I}(\boldsymbol{x}; \boldsymbol{y})$ and $\mathrm{I}(\boldsymbol{y}; \boldsymbol{y}_{\text{true}})$. In contrast, IMAE achieves much better performance by explicitly encouraging confident category separation via minimizing $\mathrm{H}(\boldsymbol{y}|\boldsymbol{x})$, while using VAT to maintain local smoothness so as to prevent overfitting of neural network.

[3] 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).



(a) Evolution of the key quantities for discrete representations.

Figure 3: (a) Tracking the key quantities for categorical representations.

By pushing the upper bound $D_{\mathrm{KL}}\left(p(\boldsymbol{y}|\boldsymbol{x})||r(\boldsymbol{y})\right)$ of $\mathrm{I}(\boldsymbol{x};\boldsymbol{y})$ towards a progressively increasing target value $C_{\boldsymbol{y}}$, JointVAE performs well on both MNIST and F-MNIST. However, we found it can easily get stuck at some bad local optima where $\mathrm{I}(\boldsymbol{x};\boldsymbol{y})$ is close to the maximum while the accuracy are lower than IMAE. This is an implication of overfitting, *i.e.,* the prediction $p(\boldsymbol{y}|)$ are over confident about it's prediction. A possible explanation is that once JointVAE enter the local region of a local optima, progressively increasing the target value can only lead to oscillation within that region. Although, it might be possible to escape such local region by increasing the target value more progressively at the corresponding training steps, specifying this increasing rate can be very challenging across datasets.

4.3 Better trade-offs among interpretability and decoding quality

In this section, we examine the trade-offs between the interpretability of latent representations and the decoding quality induced by different models. Table (4a) shows that IMAE achieves better overall trade-off with respect to the decoding quality, informativeness of latent representations and the statistical independence between latent factors. This is further demonstrated by Figure (4b) to (4e), where we traverse the encoding "angle" factor conditioned on the discrete representations learned by different models. As we can see, different models learn the digit type information to various degree. Being able to learn meaningful discrete representations encourages more separated mainfold structures between categories, *i.e.,* a more interpretable embedding space, which in turn allows more flexibilities to represent the data. On the other hand, compared to JointVAE and InfoVAE, IMAE attains less overlap between the mainfolds associated with each category. We attribute this to the effect of explicitly seeking for statistical independence between the latent factors. This can serves as an evidence of statistical independence induce disentanglement, though it's not sufficient.

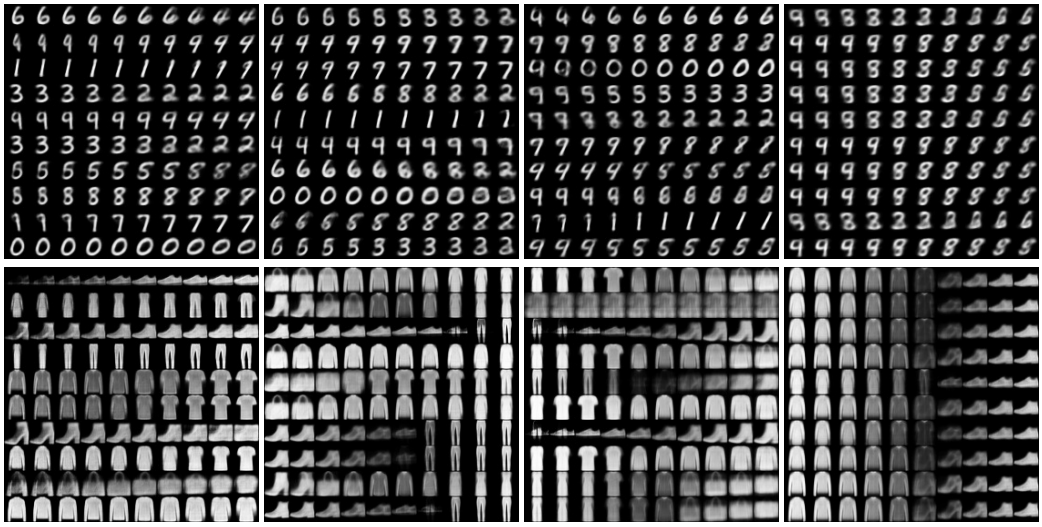
On the other hand, an important insight from recent work, *e.g.,* (Zhao et al., 2017; Tolstikhin et al., 2017; Makhzani and Frey, 2017), is that informative representations can give better decoding quality. This is demonstrated in Table (4a), where IMAE gives better reconstruction error as a result of being able to learn more informative and interpretable representations.

5 Conclusion

We have proposed IMAE, a novel approach for simultaneously learning the categorical information of data while uncovering latent continuous features shared across categories. Different from VAE,

(a) Learning statistics on MNIST/F-MNIST.

	$\mathrm{KL}(\boldsymbol{x};\widetilde{\boldsymbol{x}})$	$\mathrm{I}(\boldsymbol{x};\boldsymbol{y})$	$\mathrm{I}(\boldsymbol{x};\boldsymbol{z})$	$\mathrm{TC}(\boldsymbol{y};\boldsymbol{z})$	$\mathrm{KL}(\boldsymbol{x};\widetilde{\boldsymbol{x}})$	$\mathrm{I}(\boldsymbol{x};\boldsymbol{y})$	$\mathrm{I}(\boldsymbol{x};\boldsymbol{z})$	$\mathrm{TC}(\boldsymbol{y};\boldsymbol{z})$
β-VAE	141.5	0.0	4.4	**0.0**	252.5	0.0	4.6	**0.0**
J-VAE	127.3	**1.9**	5.6	2.3	256.0	**1.8**	4.7	2.6
I-VAE	121.5	1.7	6.8	1.5	247.1	0.9	6.0	0.7
IMAE	**117.9**	1.8	**7.2**	0.8	**234.9**	1.7	**8.5**	**0.0**



(b) IMAE (c) InfoVAE (d) JointVAE (e) β-VAE

Figure 4: IMAE achieves better disentanglement between discrete and continuous representations. Here we set $\beta, \lambda = 5$ for all models and choose $\gamma = 2$ for IMAE. In Table 4a, $\mathrm{TC}(\boldsymbol{y};\boldsymbol{z})$ is referred to the total correlation of $(\boldsymbol{y}, \boldsymbol{z})$ which is introduced in Section 3. We traverse the continuous latent variable conditioned on the learnt discrete representation. The initial values of $\boldsymbol{z}$ for each category learnt by a model is selected as the sample average of the representations associated with that category. Then we traverse encoding "angle" latent factor within [-2, 2] over categories. The results for (a) - (e) are reported on the 10000 testing images.

IMAE starts with a stochastic encoder that seeks to maximize the mutual information between data and their representations, where a decoder is used to approximate the true posterior distribution of the data given the representations. This model targets at informative representations directly, which in turn naturally yields an objective that is capable of simultaneously inducing semantically meaningful representations and maintaining good decoding quality, which is further demonstrated by the numerical results.

Unsupervised joint learning of disentangled continuous and discrete representations is a challenging problem due to the lack of 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 the 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.

REFERENCES

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

A. A. Alemi, B. Poole, I. Fischer, J. V. Dillon, R. A. Saurous, and K. Murphy. An information-theoretic analysis of deep latent-variable models. *arXiv preprint arXiv:1711.00464*, 2017.

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

C. P. Burgess, I. Higgins, A. Pal, L. Matthey, N. Watters, G. Desjardins, and A. Lerchner. Understanding disentangling in β-vae. *arXiv preprint arXiv:1804.03599*, 2018.

T. Q. Chen, X. Li, R. Grosse, and D. Duvenaud. Isolating sources of disentanglement in variational autoencoders. *arXiv preprint arXiv:1802.04942*, 2018.

X. Chen, Y. Duan, R. Houthooft, J. Schulman, I. Sutskever, and P. Abbeel. Infogan: Interpretable representation learning by information maximizing generative adversarial nets. In *Advances in Neural Information Processing Systems*, pages 2172–2180, 2016.

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

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

B. Esmaeili, H. Wu, S. Jain, S. Narayanaswamy, B. Paige, and J.-W. van de Meent. Hierarchical disentangled representations. *arXiv preprint arXiv:1804.02086*, 2018.

S. Gao, R. Brekelmans, G. V. Steeg, and A. Galstyan. Auto-encoding total correlation explanation. *arXiv preprint arXiv:1802.05822*, 2018.

I. Higgins, L. Matthey, A. Pal, C. Burgess, X. Glorot, M. Botvinick, S. Mohamed, and A. Lerchner. beta-vae: Learning basic visual concepts with a constrained variational framework. 2016.

M. D. Hoffman and M. 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.

E. Jang, S. Gu, and B. Poole. Categorical reparameterization with gumbel-softmax. *arXiv preprint arXiv:1611.01144*, 2016.

H. Kim and A. Mnih. Disentangling by factorising. *arXiv preprint arXiv:1802.05983*, 2018.

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

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

A. Makhzani and B. J. Frey. Pixelgan autoencoders. In *Advances in Neural Information Processing Systems*, pages 1972–1982, 2017.

N. K. R. T. S. N. Mary Phuong, Max Welling. The mutual autoencoder: Controlling information in latent code representations. 2018. URL `https://openreview.net/forum?id=HkbmWqxCZ`.

T. Miyato, A. M. Dai, and I. Goodfellow. Virtual adversarial training for semi-supervised text classification. 2016.

D. J. Rezende, S. Mohamed, and D. Wierstra. Stochastic backpropagation and approximate inference in deep generative models. *arXiv preprint arXiv:1401.4082*, 2014.

I. Tolstikhin, O. Bousquet, S. Gelly, and B. Schoelkopf. Wasserstein auto-encoders. *arXiv preprint arXiv:1711.01558*, 2017.

R. Vershynin. *High-Dimensional Probability: An Introduction with Applications in Data Science*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 2018.

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

S. Zhao, J. Song, and S. Ermon. Infovae: Information maximizing variational autoencoders. *arXiv preprint arXiv:1706.02262*, 2017.

A PROOF OF SECTION 3

Decomposition of $I_\theta(\boldsymbol{x};\boldsymbol{y},\boldsymbol{z})$ Let $\boldsymbol{b}=(\boldsymbol{z},\boldsymbol{y})$ denote the joint random variable consisting of the continuous random variable $\boldsymbol{b}$ and discrete random variable $\boldsymbol{y}$.

Note that $I_\theta(\boldsymbol{x};\boldsymbol{y},\boldsymbol{z})=I_\theta(\boldsymbol{x};\boldsymbol{b})$ can be written as:

$$\begin{aligned} I_\theta(\boldsymbol{x};\boldsymbol{b}) &= -\int_{\mathcal{X}} p(\boldsymbol{x}) \int_{\mathcal{Z}} p_\theta(\boldsymbol{b}|\boldsymbol{x}) \log p_\theta(\boldsymbol{b}) d\boldsymbol{b} d\boldsymbol{x} + \int_{\mathcal{X}} p(\boldsymbol{x}) \int_{\mathcal{Z}} p_\theta(\boldsymbol{b}|\boldsymbol{x}) \log p_\theta(\boldsymbol{b}|\boldsymbol{x}) d\boldsymbol{b} d\boldsymbol{x} \\ &= -\int_{\mathcal{Z}} p_\theta(\boldsymbol{b}) \log p_\theta(\boldsymbol{b}) d\boldsymbol{b} + \int_{\mathcal{X}} p(\boldsymbol{x}) \int_{\mathcal{Z}} p_\theta(\boldsymbol{b}|\boldsymbol{x}) \log p_\theta(\boldsymbol{b}|\boldsymbol{x}) d\boldsymbol{b} d\boldsymbol{x} \ . \end{aligned} \tag{13}$$

The second term in equation 13 has the form:

$$\begin{aligned} \int_{\mathcal{X}} p(\boldsymbol{x}) \int_{\mathcal{Z}} p_\theta(\boldsymbol{b}|\boldsymbol{x}) \log p_\theta(\boldsymbol{b}|\boldsymbol{x}) d\boldsymbol{b} d\boldsymbol{x} &\overset{\vartheta_1}{=} \sum_{k=1}^{K_1+1} \int_{\mathcal{X}} p(\boldsymbol{x}) \int_{\mathcal{Z}} p_\theta(\boldsymbol{b}|\boldsymbol{x}) \log p_\theta(\boldsymbol{b}_k|\boldsymbol{x}) d\boldsymbol{b} d\boldsymbol{x} \\ &= \sum_{k=1}^{K_1+1} H_\theta(\boldsymbol{b}_k|\boldsymbol{x}) \ , \end{aligned} \tag{14}$$

where ϑ_1 follows by the assumption that $p_\theta(\boldsymbol{b}|\boldsymbol{x})$ is factorial.

For the first term in equation 13, we have:

$$\begin{aligned} \int_{\mathcal{Z}} p_\theta(\boldsymbol{b}) \log p_\theta(\boldsymbol{b}) d\boldsymbol{b} &= \int_{\mathcal{Z}} p_\theta(\boldsymbol{b}) \log \frac{p_\theta(\boldsymbol{b})}{\Pi_{k=1}^{K_1+1} p_\theta(\boldsymbol{b}_k)} d\boldsymbol{b} + \sum_{k=1}^{K_1+1} \int_{\mathcal{Z}} p_\theta(\boldsymbol{b}) \log p_\theta(\boldsymbol{b}_k) d\boldsymbol{b} \\ &= D_{\mathrm{KL}}\left(p_\theta(\boldsymbol{b}) || \Pi_{k=1}^{K_1+1} p_\theta(\boldsymbol{b}_k)\right) - \sum_{k=1}^{K_1+1} H_\theta(\boldsymbol{b}_k) \ . \end{aligned} \tag{15}$$

Substituting equation 14 and equation 15 into equation 13 yields the result:

$$\begin{aligned} I_\theta(\boldsymbol{x};\boldsymbol{y},\boldsymbol{z}) = I_\theta(\boldsymbol{x};\boldsymbol{b}) &= H_\theta(\boldsymbol{b}_k) - D_{\mathrm{KL}}\left(p_\theta(\boldsymbol{b}) || \Pi_{k=1}^{K_1+1} p_\theta(\boldsymbol{b}_k)\right) - \sum_{k=1}^{K_1+1} H_\theta(\boldsymbol{b}_k|\boldsymbol{x}) \\ &= \sum_{k=1}^{K_1+1} I_\theta(\boldsymbol{x};\boldsymbol{b}_k) - D_{\mathrm{KL}}\left(p_\theta(\boldsymbol{b}) || \Pi_{k=1}^{K_1+1} p_\theta(\boldsymbol{b}_k)\right) \\ &= I_\theta(\boldsymbol{x};\boldsymbol{y}) + \sum_{k=1}^{K_1} I_\theta(\boldsymbol{x};\boldsymbol{z}_k) - D_{\mathrm{KL}}\left(p_\theta(\boldsymbol{y},\boldsymbol{z}) || p_\theta(\boldsymbol{y}) \Pi_{k=1}^{K_1} p_\theta(\boldsymbol{z}_k)\right) \ . \end{aligned} \tag{16}$$

proof of proposition 1

Proof. We start with computing the expectation of $\boldsymbol{z}_k$:

$$\begin{aligned} \mathbb{E}_\theta\left[\boldsymbol{z}_k\right] = \int_{\mathcal{Z}_k} \boldsymbol{z}_k \int_{\mathcal{X}} p_\theta(\boldsymbol{z}_k|\boldsymbol{x}) p(\boldsymbol{x}) d\boldsymbol{x} d\boldsymbol{z}_k &= \int_{\mathcal{X}} p(\boldsymbol{x}) \int_{\mathcal{Z}_k} \boldsymbol{z}_k p_\theta(\boldsymbol{z}_k|\boldsymbol{x}) d\boldsymbol{z}_k d\boldsymbol{x} \\ &= \int_{\mathcal{X}} p(\boldsymbol{x}) \mu_k(\boldsymbol{x}) d\boldsymbol{x} = \mathbb{E}_{\boldsymbol{x}}\left[\mu_k(\boldsymbol{x})\right] \ . \end{aligned} \tag{17}$$

Then the variance of $\boldsymbol{z}_k$ followed as:

$$\begin{aligned} \mathrm{Var}_\theta\left[\boldsymbol{z}_k\right] &= \int_{\mathcal{Z}_k} \boldsymbol{z}_k^2 \int_{\mathcal{X}} p_\theta(\boldsymbol{z}_k|\boldsymbol{x}) p(\boldsymbol{x}) d\boldsymbol{x} d\boldsymbol{z}_k - \mathbb{E}_{\boldsymbol{x}}\left[\mu_k(\boldsymbol{x})\right]^2 \\ &= \int_{\mathcal{X}} p(\boldsymbol{x}) \int_{\mathcal{Z}_k} \boldsymbol{z}_k^2 p_\theta(\boldsymbol{z}_k|\boldsymbol{x}) d\boldsymbol{z}_k d\boldsymbol{x} - \mathbb{E}_{\boldsymbol{x}}\left[\mu_k(\boldsymbol{x})\right]^2 \\ &= \int_{\mathcal{X}} p(\boldsymbol{x}) \left[\sigma_k^2(\boldsymbol{x}) + \mu_k(\boldsymbol{x})^2\right] d\boldsymbol{x} - \mathbb{E}_{\boldsymbol{x}}\left[\mu_k(\boldsymbol{x})\right]^2 \\ &= \mathbb{E}_{\boldsymbol{x}}\left[\sigma_k^2(\boldsymbol{x})\right] + \mathrm{Var}_{\boldsymbol{x}}\left[\mu_k(\boldsymbol{x})\right] \ . \end{aligned} \tag{18}$$

Note that

$$I_\theta(\boldsymbol{x};\boldsymbol{z}_k) = H_\theta(\boldsymbol{z}_k) - H_\theta(\boldsymbol{z}_k|\boldsymbol{x}) \ , \tag{19}$$

for which we have the following,

$$\begin{aligned} H_\theta(\boldsymbol{z}_k|\boldsymbol{x}) &= -\int_{\mathcal{X}} p(\boldsymbol{x}) \int_{\mathcal{Z}_k} p_\theta(\boldsymbol{z}_k|\boldsymbol{x}) \log p_\theta(\boldsymbol{z}_k|\boldsymbol{x}) d\boldsymbol{z} d\boldsymbol{x} \\ &= \frac{1}{2}\int_{\mathcal{X}} p(\boldsymbol{x}) \log\left(2\pi e \sigma_k^2(\boldsymbol{x})\right) d\boldsymbol{x} \\ &= \frac{1}{2}\left(\log(2\pi e) + \mathbb{E}_{\boldsymbol{x}}\left[\log \sigma_k^2(\boldsymbol{x})\right]\right) . \end{aligned} \tag{20}$$

For the entropy of $\boldsymbol{z}_k$, we leverage the fact that $H_\theta(\boldsymbol{z}_k)$ is upper bounded by the entropy of a Gaussian distributed random variable with the same mean and variance, that is

$$H_\theta(\boldsymbol{z}_k) \leq \frac{1}{2}\left(\log 2\pi e + \log\left(\mathbb{E}_{\boldsymbol{x}}\left[\sigma_k^2(\boldsymbol{x})\right] + \mathrm{Var}_{\boldsymbol{x}}\left[\mu_k(\boldsymbol{x})\right]\right)\right) \tag{21}$$

Substituting equation 20 and equation 21 into equation 19 completes the proof. □

proof of proposition 2

Proof. Let $\widehat{p}_\theta(\boldsymbol{y}) = \frac{1}{N}\sum_{n=1}^N p_\theta(\boldsymbol{y}|\boldsymbol{x}_n)$ denote the Monte Carlo estimator of the true probability $p_\theta(\boldsymbol{y}) = \int_{\mathcal{X}} p(\boldsymbol{x}) p_\theta(\boldsymbol{y}|\boldsymbol{x}) d\boldsymbol{x} = \mathbb{E}_{\boldsymbol{x}}\left[p_\theta(\boldsymbol{y}|\boldsymbol{x})\right]$. Note that $p_\theta(\boldsymbol{y}|\boldsymbol{x}) \in [0,1]$ for all $\boldsymbol{x} \in \mathcal{X}$, then applying the Hoeffding's inequality for bounded random variables [Theorem 2.2.6, (Vershynin, 2018)] yields,

$$\mathbb{P}\left(|\widehat{p}_\theta(\boldsymbol{y}) - p_\theta(\boldsymbol{y})| \geq t\right) = \mathbb{P}\left(\left|\frac{1}{N}\sum_{n=1}^N p_\theta(\boldsymbol{y}|\boldsymbol{x}_n) - \mathbb{E}_{\boldsymbol{x}}\left[p_\theta(\boldsymbol{y}|\boldsymbol{x})\right]\right| \geq t\right) \leq 2\exp\left(-2Nt^2\right) \tag{22}$$

Let $\delta' = 2\exp\left(-2Nt^2\right)$, it then follows,

$$\mathbb{P}\left(|\widehat{p}_\theta(\boldsymbol{y}) - p_\theta(\boldsymbol{y})| < \sqrt{\frac{\log(2/\delta')}{2N}}\right) \geq 1 - \delta' \tag{23}$$

Given equation 23, we first establish the concentration results of the entropy $H_{\widehat{p}_\theta}(\boldsymbol{y})$ with respect to the empirical distribution $\widehat{p}_\theta(\boldsymbol{y})$. Assume For all $y \in \mathcal{C}$, we have $p_\theta(y), \widehat{p}_\theta(y)$ bounded below by $1/(CK_2)$ for some fixed constant $C > 1$. This assumption is practical since the distributions of true data and predicted data are approximately uniform and therefore $p_\theta(y), \widehat{p}_\theta(y) \approx 1/K_2$ for all $y \in \mathcal{C}$. Consider the function $t\log t$, with derivative $1 + \log t \in [1 - \log CK_2, 1]$ for $t \in [1/(CK_2), 1]$,

$$\begin{aligned} &|\widehat{p}_\theta(y)\log\widehat{p}_\theta(y) - p_\theta(y)\log p_\theta(y)| = \left|\int_{p_\theta(y)}^{\widehat{p}_\theta(y)} (1+\log t) dt\right| \\ &\leq \left|\int_{p_\theta(y)}^{\widehat{p}_\theta(y)} |1+\log t| dt\right| \leq \left|\int_{p_\theta(y)}^{\widehat{p}_\theta(y)} \max\{\log CK_2 - 1, 1\} dt\right| \\ &\leq \max\{\log CK_2 - 1, 1\}|\widehat{p}_\theta(y) - p_\theta(y)| \end{aligned} \tag{24}$$

Summing over $\mathcal{C}$ gives

$$\left|\widehat{H}_\theta(\boldsymbol{y}) - H_\theta(\boldsymbol{y})\right| \leq K_2 \max\{\log CK_2 - 1, 1\}|\widehat{p}_\theta(y) - p_\theta(y)| \ . \tag{25}$$

Let $\delta = K_2\delta'$, then equation 23 together with equation 25 yields

$$\mathbb{P}\left(\left|\widehat{H}_\theta(\boldsymbol{y}) - H_\theta(\boldsymbol{y})\right| < K_2 \max\{\log CK_2 - 1, 1\}\sqrt{\frac{\log(2K_2/\delta)}{2N}}\right) \geq 1 - \delta \tag{26}$$

Next we are going to bound the divergence between $\widehat{H}_\theta(\boldsymbol{y}|\boldsymbol{x})$ and $H_\theta(\boldsymbol{y}|\boldsymbol{x})$ which are defined as,

$$\widehat{H}_\theta(\boldsymbol{y}|\boldsymbol{x}) = -\frac{1}{N}\sum_{n=1}^{N}\sum_{\boldsymbol{y}} p_\theta(\boldsymbol{y}|\boldsymbol{x}_n)\log p_\theta(\boldsymbol{y}|\boldsymbol{x}_n),$$
$$H_\theta(\boldsymbol{y}|\boldsymbol{x}) = -\int_{\boldsymbol{x}\in\mathcal{X}}\sum_{\boldsymbol{y}} p_\theta(\boldsymbol{y}|\boldsymbol{x})\log p_\theta(\boldsymbol{y}|\boldsymbol{x})\ .$$

Note that $h\log h \in [-1/e, 0]$ for all $h \in [0,1]$, then again applying [Theorem 2.2.6, (Vershynin, 2018)] yields,

$$\mathbb{P}\left(\left|\frac{1}{N}\sum_{n=1}^{N} p_\theta(\boldsymbol{y}|\boldsymbol{x}_n)\log p_\theta(\boldsymbol{y}|\boldsymbol{x}_n) - \mathbb{E}_{p(\boldsymbol{x})}\left[p_\theta(\boldsymbol{y}|\boldsymbol{x})\log p_\theta(\boldsymbol{y}|\boldsymbol{x})\right]\right| < t\right) \leq 2\exp\left(-2t^2e^2N\right) \tag{27}$$

Following the similar arguments as before, let $\delta' = 2\exp\left(-2t^2e^2N\right)$, then

$$\mathbb{P}\left(\left|\frac{1}{N}\sum_{n=1}^{N} p_\theta(\boldsymbol{y}|\boldsymbol{x}_n)\log p_\theta(\boldsymbol{y}|\boldsymbol{x}_n) - \mathbb{E}_{p(\boldsymbol{x})}\left[p_\theta(\boldsymbol{y}|\boldsymbol{x})\log p_\theta(\boldsymbol{y}|\boldsymbol{x})\right]\right| < \sqrt{\frac{e^2\log(2/\delta')}{2N}}\right) \leq \delta' \tag{28}$$

Now let $\delta = K_2\delta'$, then applying the union bound we have

$$\begin{aligned}|\widehat{H}_\theta(\boldsymbol{y}|\boldsymbol{x}) - H_\theta(\boldsymbol{y}|\boldsymbol{x})| &\leq \sum_{\boldsymbol{y}\in\mathcal{C}}\left|\frac{1}{N}\sum_{n=1}^{N} p_\theta(\boldsymbol{y}|\boldsymbol{x}_n)\log p_\theta(\boldsymbol{y}|\boldsymbol{x}_n) - \mathbb{E}_{p(\boldsymbol{x})}\left[p_\theta(\boldsymbol{y}|\boldsymbol{x})\log p_\theta(\boldsymbol{y}|\boldsymbol{x})\right]\right| \\ &\leq K_2\sqrt{\frac{e^2\log(2K_2/\delta)}{2N}}\end{aligned} \tag{29}$$

hold with probability $1-\delta$.

Conclude from equation 26 and equation 29, we have

$$\begin{aligned}\left|I_\theta(\boldsymbol{x};\boldsymbol{y}) - \widehat{I}_\theta(\boldsymbol{x};\boldsymbol{y})\right| &\leq \left|H_\theta(\boldsymbol{y}) - \widehat{H}_\theta(\boldsymbol{y})\right| + \left|H_\theta(\boldsymbol{y}|\boldsymbol{x}) - \widehat{H}_\theta(\boldsymbol{y}|\boldsymbol{x})\right| \\ &= K_2\left(\max\{\log CK_2 - 1, 1\} + e\right)\sqrt{\frac{\log(2K_2/\delta)}{N}}\ .\end{aligned} \tag{30}$$

hold with probability at least $1-2\delta$. □

A.1 Balance between posterior inference fidelity and information maximization objective

Notice that

$$I_\theta(\boldsymbol{x};\boldsymbol{y},\boldsymbol{z}) = \mathrm{H}(\boldsymbol{x}) + \mathbb{E}_{p_\theta(\boldsymbol{x},\boldsymbol{y},\boldsymbol{z})}\left[\log q_\phi(\boldsymbol{x}|\boldsymbol{y},\boldsymbol{z})\right] + D_{\mathrm{KL}}\left(p_\theta(\boldsymbol{x}|\boldsymbol{y},\boldsymbol{z})||q_\phi(\boldsymbol{x}|\boldsymbol{y},\boldsymbol{z})\right) \tag{31}$$

Therefore,

$$I_\theta(\boldsymbol{x};\boldsymbol{y},\boldsymbol{z}) - D_{\mathrm{KL}}\left(p_\theta(\boldsymbol{x}|\boldsymbol{y},\boldsymbol{z})||q_\phi(\boldsymbol{x}|\boldsymbol{y},\boldsymbol{z})\right) = \mathrm{H}(\boldsymbol{x}) + \mathbb{E}_{p_\theta(\boldsymbol{x},\boldsymbol{y},\boldsymbol{z})}\left[\log q_\phi(\boldsymbol{x}|\boldsymbol{y},\boldsymbol{z})\right] \tag{32}$$

Therefore, our objective can also be motivated via

$$\begin{aligned}&\max\ \ \beta I_\theta(\boldsymbol{x};\boldsymbol{y},\boldsymbol{z}) - D_{\mathrm{KL}}\left(p_\theta(\boldsymbol{x}|\boldsymbol{y},\boldsymbol{z})||q_\phi(\boldsymbol{x}|\boldsymbol{y},\boldsymbol{z})\right), \qquad \beta > 1 \\ &\Rightarrow \max\ \ (\beta-1)I_\theta(\boldsymbol{x};\boldsymbol{y},\boldsymbol{z}) + \mathbb{E}_{p_\theta(\boldsymbol{x},\boldsymbol{y},\boldsymbol{z})}\left[\log q_\phi(\boldsymbol{x}|\boldsymbol{y},\boldsymbol{z})\right]\end{aligned} \tag{33}$$

where β trade-off the informativeness of the latent representation and generation fidelity.

B APPROXIMATION OF THE MARGINAL DISTRIBUTION

In order to optimize the objective function (11), we need to compute the marginal probabilities $p_\theta(\boldsymbol{y})$, $p_\theta(\boldsymbol{z})$ and $p_\theta(\boldsymbol{z}_i)$. Computing these quantities requires the entire dataset, *e.g.,* $p_\theta(\boldsymbol{z}) = \int_{\mathcal{X}} p_\theta(\boldsymbol{z}, \boldsymbol{x}) d\boldsymbol{x} = \mathbb{E}_{\boldsymbol{x}}\left[p_\theta(\boldsymbol{z}|\boldsymbol{x})\right]$. To scale up our method to large datasets, we would like to estimate based on the minibatch data, e.g., $p_\theta(\boldsymbol{z}) \approx \frac{1}{B}\sum_{i=1}^{B} p_\theta(\boldsymbol{z}|\boldsymbol{x}_i)$.

Now consider the first term on RHS of equation 6:

$$\mathrm{KL}\left(p_\theta(\boldsymbol{z})||r(\boldsymbol{z})\right) = \int_{\mathcal{Z}} p_\theta(\boldsymbol{z}) \log \frac{p_\theta(\boldsymbol{z})}{r(\boldsymbol{z})} d\boldsymbol{z} \tag{34}$$

We will only present our approximation for $\int_{\mathcal{Z}} p_\theta(\boldsymbol{z}) \log p_\theta(\boldsymbol{z}) d\boldsymbol{z}$ and $\int_{\mathcal{Z}} p_\theta(\boldsymbol{z}) \log r(\boldsymbol{z}) d\boldsymbol{z}$ can be approximated in a similar fashion.

$$\begin{aligned}\int_{\mathcal{Z}} p_\theta(\boldsymbol{z}) \log p_\theta(\boldsymbol{z}) d\boldsymbol{z} &= \int_{\mathcal{X}} d\boldsymbol{x} p(\boldsymbol{x}) \int_{\mathcal{Z}} p_\theta(\boldsymbol{z}|\boldsymbol{x}) \log p_\theta(\boldsymbol{z}) d\boldsymbol{z} = \frac{1}{B}\sum_{i=1}^{B} \int_{\mathcal{Z}} p_\theta(\boldsymbol{z}|\boldsymbol{x}_i) \log p_\theta(\boldsymbol{z}) d\boldsymbol{z} \\ &= \frac{1}{B}\sum_{i=1}^{B} \log p_\theta(\boldsymbol{z}_i) = \frac{1}{B}\sum_{i=1}^{B} \log\left(\int_{\mathcal{X}} d\boldsymbol{x} p(\boldsymbol{x}) p_\theta(\boldsymbol{z}_i|\boldsymbol{x})\right) = \frac{1}{B}\sum_{i=1}^{B} \log \frac{1}{B}\sum_{j=1}^{B} p_\theta(\boldsymbol{z}_i|\boldsymbol{x}_j)\ ,\end{aligned} \tag{35}$$

where B denotes the batch size, $\boldsymbol{z}_i$ is sampled from $p_\theta(\boldsymbol{z}|\boldsymbol{x}_i)$ and different batches are iterated by i, j respectively.

We estimate the integral of $\boldsymbol{z}$ by sampling $\boldsymbol{z} \sim p_\theta(\boldsymbol{z}|\boldsymbol{x}_i)$ and perform the Monte Carlo approximation. Although we minimize the unbiased estimator of the lower bound of the KL divergence, the term inside the logarithm is a summation of probability densities of Gaussians. In particular, we record the distribution of the variances output by our encoder and observe that the mean of the variances of the Gaussians is bounded between 0.2 and 2, which implies that the values of probability densities do not range in a large scale. Since logarithm is locally affine, we argue that our bound in (35) is tight. Other quantities involved in our objective function (11) are estimated in a similar fashion.

C CONNECTIONS TO VAE

In VAE, they assume a generative model specified by a stochastic decoder $p_\theta(\boldsymbol{x}|\boldsymbol{z})$, taking the continuous representation as an example, and seek an encoder $q_\phi(\boldsymbol{z}|\boldsymbol{x})$ as a variational approximation of the true posterior $p_\theta(\boldsymbol{z}|\boldsymbol{x})$. The model is fitted by maximizing the evidence lower bound (ELBO) of the marginal likelihood,

$$\mathbb{E}_{\boldsymbol{x}}\left[\log p_\theta(\boldsymbol{x})\right] \geq \mathcal{L}(\boldsymbol{x}, \theta, \phi) = \mathbb{E}_{q_\phi(\boldsymbol{z}|\boldsymbol{x})}\left[\log p_\theta(\boldsymbol{x}|\boldsymbol{z})\right] - \mathbb{E}_{\boldsymbol{x}}\left[D_{\mathrm{KL}}\left(q_\phi(\boldsymbol{z}|\boldsymbol{x})||r(\boldsymbol{z})\right)\right]\ . \tag{36}$$

Here the KL divergence term can be further decomposed as (Hoffman and Johnson, 2016),

$$\mathbb{E}_{\boldsymbol{x}}\left[D_{\mathrm{KL}}\left(q_\phi(\boldsymbol{z}|\boldsymbol{x})||r(\boldsymbol{z})\right)\right] = I_\theta(\boldsymbol{x}; \boldsymbol{z}) + \mathbb{E}_{\boldsymbol{x}}\left[D_{\mathrm{KL}}\left(q_\phi(\boldsymbol{z})||r(\boldsymbol{z})\right)\right]\ . \tag{37}$$

That is, minimizing the KL divergence also penalizes the mutual information $I_\theta(\boldsymbol{x}; \boldsymbol{z})$, thus reduces the amount of information $\boldsymbol{z}$ has about $\boldsymbol{x}$. This can make the inference task $q_\phi(\boldsymbol{z}|\boldsymbol{x})$ hard and lead to poor reconstructions of $\boldsymbol{x}$ as well. Many recent efforts have been focused on resolving this problem by revising ELBO. Although approaches differ, it can be summarized as either dropping the mutual information term in Eq (37), or encouraging statistical independence across the dimensions of $\boldsymbol{z}$ by increasing the penalty on the total correlation term extracted from the KL divergence $D_{\mathrm{KL}}\left(q_\phi(\boldsymbol{z})||r(\boldsymbol{z})\right)$ with respect to $q_\phi(\boldsymbol{z})$. However, these approaches either result in an invalid lower bound for the VAE objective, or cannot avoid minimizing the mutual information $I_\theta(\boldsymbol{x}; \boldsymbol{z})$ between the representation and the data.

In contrast, IMAE starts with a stochastic encoder $p_\theta(\boldsymbol{z}|\boldsymbol{x})$ and aims at maximizing the mutual information between the data $\boldsymbol{x}$ and the representations $\boldsymbol{z}$ from the very beginning. By following the constraints which are naturally implied by the objective in order to avoid degenerated solutions,

IMAE targets at both informative and statistical independent representations. On the other hand, in IMAE the decoder $q_\phi(\boldsymbol{x}|\boldsymbol{z})$ serves as a variational approximation to the true posterior $p_\theta(\boldsymbol{x}|\boldsymbol{z})$. As we will show in Section 4, being able to learn more interpretable representations allows IMAE to reconstruct and generate data with better quality.

D VAT STABILIZES THE LEARNING OF CATEGORICAL REPRESENTATIONS

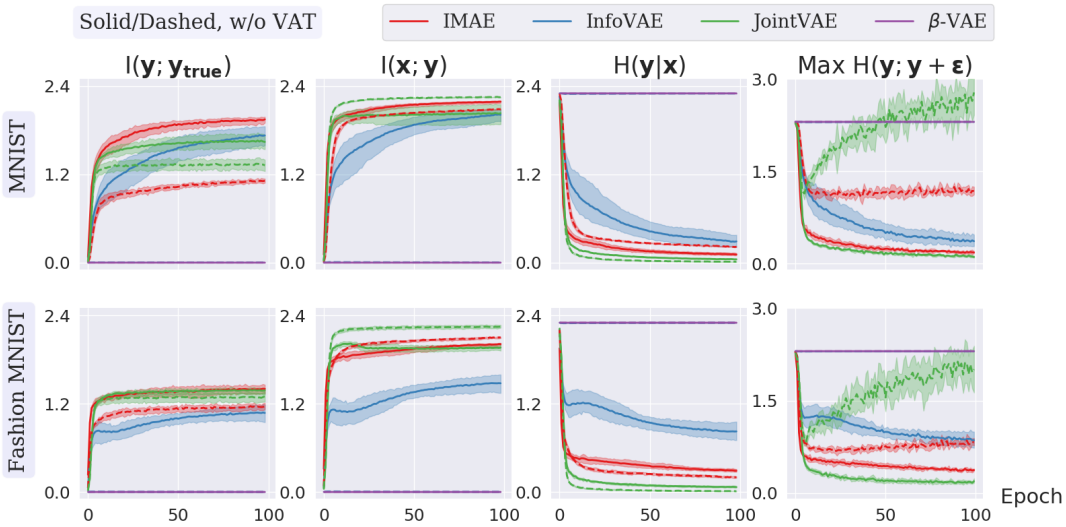


Figure 5: Prevent over confidence predictions by encouraging local smoothness

E EXPERIMENTAL SETTINGS

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

We train all models using momentum, with initial learning rate being $1e$-3. We decay the learning rate by 0.98 every epoch.