

Learning Representations by Maximizing Mutual Information in Variational Autoencoders

Ali Lotfi Rezaabad, and Sriram Vishwanath
The University of Texas at Austin, TX, USA
Email: alotfi@utexas.edu, sriram@austin.utexas.edu

Abstract—Variational autoencoders (VAE) have ushered in an new era of unsupervised learning methods for complex distributions. Although these techniques are elegant in their approach, they are typically not useful for representation learning. In this work, we propose a simple yet powerful class of VAEs that simultaneously result in meaningful learned representations. Our solution is to combine traditional VAEs with mutual information maximization, with the goal to enhance amortized inference in VAEs using Information Theoretic techniques. We call this approach InfoMax-VAE, and such an approach can significantly boost the quality of learned high-level representations. We realize this through explicit maximization of information measures associated with the representation. Using extensive experiments on varied datasets and setups, we show that InfoMax-VAE outperforms contemporary popular approaches, including Info-VAE and β -VAE.

I. INTRODUCTION

There is growing interest in generative models, with new powerful tools being developed for modeling complex data and reasoning under uncertainty [1]–[3]. Simultaneously, there is significant interest and progress being made in learning representations of high-dimensional data such as images [4]. Particularly, the development of variational autoencoders [5] has enabled many applications, ranging from image processing to language modeling [6]–[8]. VAEs have also found use as a representation learning model for inferring latent variables [9].

Although successful for multiple isolated applications, VAEs have not always proven to be reliable for extracting high-level representations of observations [10], [11]. As generative networks become more descriptive/richer, the extraction of meaningful representations becomes considerably more challenging. Overall, VAEs can often fail to take the advantage of its underlying mixture model, and the learned features can lose their dependencies on observations. Overall, VAEs alone may not be adequate in ensuring that the representation is accurate. In many practical applications, problem-specific solutions are presented, and our goal is to build a more general framework to build meaningful, useful representations with VAEs.

In this work, we present a simple but powerful method to train VAEs with associated guarantees towards the usefulness of learned high-level (latent) representations, by evaluating them on the basis of various metrics. Our main idea is to *induce* the maximization of mutual information (between the learned

latent representations and input) into the VAE objective. We call our resulting solution *InfoMax-VAE* as we seek to distill the information resulting from the input data into the latent codes to the highest extent possible. To this end, we formalize the development of InfoMax-VAE: As a first step, we develop a computationally tractable optimization problem. Indeed, it simultaneously acts as an autoencoder while estimating and maximizing the mutual information between input data and the resulting representation(s). We study the performance of InfoMax-VAE on different datasets across different setting to prove its advantages over other well-known approaches.

Related Work. There are several papers that discuss the issue of latent variable collapse [10], [12], [13]. In one thread of research, [6], [14] weaken and restrict the capacity of a generative network to enable higher-quality learned representations. Another proposed mechanism is to substitute simplistic priors with more sophisticated priors which encourage the model to learn features of interest. For example [15] suggests a parametric prior whose parameters are trained via a generative model. Other approaches use richer models with further modifications into the model. For example [16], [17] replace the KL divergence with the Jensen-Shannon divergence which enables the data and latent codes to be treated in a symmetric manner. In both studies adversarial training is leveraged to estimate the Jensen-Shannon divergence.

There is also a recent body of work that works towards enabling the maximization of the mutual information in VAEs. For example, [11] suggests several skip connections from the latent codes to the output of VAE to implicitly force higher dependency between the latent codes and observations. In [18], the authors propose to maximize the mutual information between learned latent representations and input data which is realized by adding the mutual information to the VAE objective. Their approach is to estimate $q_\phi(z)$ using Monte-Carlo and then calculate the mutual information. However, such an estimation is computationally expensive and also limits the performance benefits [19]. With a similar goal, Info-VAE [20] proposed to evade the calculation of mutual information by recasting the objective. In so doing, Info-VAE minimizes the maximum-mean discrepancy (MMD) distance (or the KL divergence) between the marginal of the inference network and the prior to implicitly increase the corresponding mutual information contained in the model. Thus, Info-VAE effectively becomes a mixture of AAE and β -VAE. Using Info-VAE, the best results are achieved once the adversarial learning in AAE is replaced

This research was supported by Army under grant W911NF-17-S-0002 and by ONR under grant N000141912590. We have made our code publicly available; see <https://github.com/AliLotfi92/InfoMaxVAE>

by the MMD distance. Also, Info-VAE is limited in the choice of the coefficient for the mutual information which determines the *information preference*. We explain this with more details in the long version of this paper [21]. In our work, however, we *explicitly estimate* and *maximize* the mutual information with means of another deep neural network, and it offers much greater flexibility in the selection of the mutual information coefficient. Also, we believe this flexibility is the primary reason why InfoMax-VAE outperforms InfoVAE in all models and datasets, as we enable the resulting VAE to uncover more information-rich latent codes compared to InfoVAE.

II. BACKGROUND AND NOTATION

Following the same lines as the literature on variational inference, we assume that we have a set of observed data $\mathbf{X} = \{\mathbf{x}^{(i)}\}_{i=1}^N$, consisting of N i.i.d. samples \mathbf{x} . Indeed, these samples are assumed to come from a distribution $q(\mathbf{x})$, where we have access to its empirical distribution rather than its explicit form. We assume that samples are being drawn from the posterior $p_\theta(\mathbf{x}|\mathbf{z})$, with θ as the generative model parameters and \mathbf{z} is the hidden latent variable (latent features, representations, or high-level abstraction). The prior distribution is denoted by $p(\mathbf{z})$ and the amortized inference distribution by $q_\phi(\mathbf{z}|\mathbf{x})$ (it is also called variational posterior distribution), which is used to map the data to latent variable space, and $p_\theta(\mathbf{x}|\mathbf{z})$ enables us to return to the input data space. Naturally, $q_\phi(\mathbf{z}|\mathbf{x})$ is called the *encoder* or *inference* model, while $p_\theta(\mathbf{x}|\mathbf{z})$ refers to the *decoder* or *generative* model. Importantly, the variational posterior distributions are typically designed for easy sampling, and are often modeled using deep neural networks. VAEs seek to maximize the variational expression for maximum likelihood: $\mathcal{L}_{\phi,\theta}$ with respect to parameters ϕ and θ , where we have:

$$\begin{aligned} \mathcal{L}_{\text{VAE}}(\phi, \theta) &= \\ \mathbb{E}_{q(\mathbf{x})}[\mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}[\log p_\theta(\mathbf{x}|\mathbf{z})] - \text{KL}(q_\phi(\mathbf{z}|\mathbf{x})||p(\mathbf{z}))] & \quad (1) \\ = -\text{KL}(q_\phi(\mathbf{x}, \mathbf{z})||p_\theta(\mathbf{x}, \mathbf{z})) + \text{const.} \end{aligned}$$

The expectations $\mathbb{E}_{q(\mathbf{x})}$ and $\mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}$ are empirically approximated via sampling, where samples are drawn based on $\mathbf{x}^{(i)} \sim q(\mathbf{x})$ and $\mathbf{z}^{(i)} \sim q_\phi(\mathbf{z}|\mathbf{x})$, and the latter is realized via the reparameterization trick [5]. The associated KL divergence can be computed both analytically or using an approach similar to the one above. Likewise [5], we call the first term in optimization as *reconstruction error*, while the KL divergence is interpreted as a *regularizer*.

III. CHALLENGES IN MEANINGFUL/USEFUL REPRESENTATIONS USING VAEs

Although VAEs remain very popular for numerous applications ranging from image processing to language modeling, they typically suffer from challenges in enabling meaningful and useful representations \mathbf{z} . Indeed, under appropriate situations (where the sets θ and ϕ are defined appropriately) both inference and generative models collaborate in producing an acceptable $p_\theta(\mathbf{x}|\mathbf{z})$ and an accurate amortized inference. However, finding suitable models for inference and generative networks across

different tasks and datasets is challenging - when the generative model is expressive, a vanilla VAE sacrifices log-likelihood in favor of amortized inference [12]. As a consequence, we obtain latent variables which are independent from the observed data, in fact, $q_\phi(\mathbf{z}|\mathbf{x}) = q_\phi(\mathbf{z})$.

To understand the origin of this discrepancy, we must return to the original problem. Particularly, a maximum likelihood technique is leveraged to minimize the bound on the KL divergence between the true data distribution $q(\mathbf{x})$ and the model's marginal distribution $p_\theta(\mathbf{x})$, $\text{KL}(q(\mathbf{x})||p_\theta(\mathbf{x}))$; whereas the quality of the latent variables only depends on $q_\phi(\mathbf{z}|\mathbf{x})$. Thus, myopic maximum likelihood without additional constraints on the posterior is insufficient when aiming to uncover relevant and information-rich latent variables.

In addition, evidence lower bound (ELBO) imposes a regularizer over latent codes, $\text{KL}(q_\phi(\mathbf{z}|\mathbf{x})||p(\mathbf{z}))$, where it seeks in the family set of ϕ for those solutions that minimize this KL divergence. As a result, it also reduces the usefulness of latent codes by encouraging $q_\phi(\mathbf{z}|\mathbf{x})$ to be matched to $p(\mathbf{z})$, which bears no relationship with observed data. Such an approach minimizes the upper bound of the mutual information between the representations and input data. To observe this, note that

$$\begin{aligned} \mathbb{E}_{q(\mathbf{x})}[\text{KL}(q_\phi(\mathbf{z}|\mathbf{x})||p(\mathbf{z}))] &= \int q_\phi(\mathbf{x}, \mathbf{z}) \log \frac{q_\phi(\mathbf{z}|\mathbf{x})}{p(\mathbf{z})} d\mathbf{x}d\mathbf{z} \\ &\geq \int q_\phi(\mathbf{x}, \mathbf{z}) \log \frac{q_\phi(\mathbf{z}|\mathbf{x})}{p(\mathbf{z})} d\mathbf{x}d\mathbf{z} - \text{KL}(q_\phi(\mathbf{z})||p(\mathbf{z})) \\ &= \int q_\phi(\mathbf{x}, \mathbf{z}) [\log \frac{q_\phi(\mathbf{z}|\mathbf{x})}{p(\mathbf{z})} - \log \frac{q_\phi(\mathbf{z})}{p(\mathbf{z})}] d\mathbf{x}d\mathbf{z} \\ &= \int q_\phi(\mathbf{x}, \mathbf{z}) \log \frac{q_\phi(\mathbf{z}|\mathbf{x})}{q_\phi(\mathbf{z})} d\mathbf{x}d\mathbf{z} \\ &= I_{q_\phi}(\mathbf{x}; \mathbf{z}). \end{aligned} \quad (2)$$

The inequity arises from the fact that the KL divergence does not take negative values. Hence, as vanilla VAEs push the model to minimize the KL divergence between the variational posterior $q_\phi(\mathbf{z}|\mathbf{x})$ and prior $p(\mathbf{z})$, they also force the representations to carry less information from input data. Actually, this may potentially result in very poor learned representations. In practice, by employing expressive generative networks, the problem is exacerbated as the model sacrifices the inference in favor of the the likelihood. Indeed, the model becomes capable of recovering data from noise, regardless of latent codes. Therefore, a vanilla VAE may not be enough to discover accurate high-level abstractions of input data.

IV. REPRESENTATION LEARNING USING VAE

A. InfoMax Variational Autoencoders

As discussed earlier, VAEs without additional constraints can prove to be unreliable for representation learning. One reason, as showed, is that the mutual information is not regarded in their objective appropriately. This bring us to the point to begin a new family of VAEs, so called *InfoMax-VAE* that effectively mitigates this issue by putting forth the *explicit maximization* of the mutual information between representations and data

into VAEs. Therefore, we have an optimization problem of form,

$$\max_{\phi, \theta} \mathbb{E}_{q(\mathbf{x})} [\mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})} [\log p_\theta(\mathbf{x}|\mathbf{z})] - \beta \text{KL}(q_\phi(\mathbf{z}|\mathbf{x})||p(\mathbf{z}))] + \alpha I_{q_\phi}(\mathbf{x}; \mathbf{z}), \quad (3)$$

where $\beta, \alpha \geq 0$ are defined to be regularization coefficients for the KL divergence and mutual information. Varying α changes the amount of information in inferring representations (or so called *information preference*). See [21] for further explanation on the interpretation of the objective. Now, evaluating $I_{q_\phi}(\mathbf{x}; \mathbf{z})$ is, in general, computationally challenging and intractable since it involves mixtures of a large number of components $q_\phi(\mathbf{z}) = \int q_\phi(\mathbf{x}, \mathbf{z}) d\mathbf{x}$. Another key question is how to effectively estimate the mutual information by drawing samples from the joint and marginals, which will be addressed in the following subsection.

B. Dual Form of Mutual Information

We start the discussion with noting that mutual information is the KL divergence between the joint and associated marginals: $I_{q_\phi}(\mathbf{x}; \mathbf{z}) = \text{KL}(q_\phi(\mathbf{x}, \mathbf{z})||q(\mathbf{x})q_\phi(\mathbf{z}))$. Interestingly, we can replace this KL divergence with any other strict divergences¹, D , which might prove to be better suited from an algorithmic perspective. In doing so, we can maximize other distances between the joint and marginals.

$$\max_{\phi, \theta} \mathbb{E}_{q(\mathbf{x})} [\mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})} [\log p_\theta(\mathbf{x}|\mathbf{z})] - \beta \text{KL}(q_\phi(\mathbf{z}|\mathbf{x})||p(\mathbf{z}))] + \alpha D(q_\phi(\mathbf{x}, \mathbf{z})||q(\mathbf{x})q_\phi(\mathbf{z})). \quad (4)$$

For instance, if we choose f -divergence, a large class of different divergences which includes the KL divergence, we get an alternate optimization problem, and by substituting the variational f -divergence we will have an objective of form:

$$\begin{aligned} & \max_{\phi, \theta} \mathbb{E}_{q(\mathbf{x})} [\mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})} [\log p_\theta(\mathbf{x}|\mathbf{z})] - \beta \text{KL}(q_\phi(\mathbf{z}|\mathbf{x})||p(\mathbf{z}))] \\ & \quad + \alpha D_f(q_\phi(\mathbf{x}, \mathbf{z})||q(\mathbf{x})q_\phi(\mathbf{z})), \\ & = \max_{\phi, \theta} \mathbb{E}_{q(\mathbf{x})} [\mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})} [\log p_\theta(\mathbf{x}|\mathbf{z})] - \beta \text{KL}(q_\phi(\mathbf{z}|\mathbf{x})||p(\mathbf{z}))] \\ & \quad + \alpha \mathbb{E}_{q(\mathbf{x})q_\phi(\mathbf{z})} [f(\frac{q_\phi(\mathbf{x}, \mathbf{z})}{q(\mathbf{x})q_\phi(\mathbf{z})})], \\ & \geq \max_{\phi, \theta, t} \mathbb{E}_{q(\mathbf{x})} [\mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})} [\log p_\theta(\mathbf{x}|\mathbf{z})] - \beta \text{KL}(q_\phi(\mathbf{z}|\mathbf{x})||p(\mathbf{z}))] \\ & \quad + \alpha (\mathbb{E}_{q_\phi(\mathbf{x}, \mathbf{z})} [t(\mathbf{x}, \mathbf{z})] - \mathbb{E}_{q(\mathbf{x})q_\phi(\mathbf{z})} [f^*(t(\mathbf{x}, \mathbf{z}))]), \end{aligned} \quad (5)$$

where f^* is the convex conjugate function of f , and t represents all possible functions. Such an inequality is imposed both due to Jensen's inequality and due to the restriction on exploring all possible functions t . As a special case, if we take $f(t)$ to be $t \log t$, which corresponds to the KL divergence (or the mutual information between \mathbf{x} and \mathbf{z}), we get the following dual representation for InfoMax-VAE,

$$\max_{\phi, \theta, t} \mathbb{E}_{q(\mathbf{x})} [\mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})} [\log p_\theta(\mathbf{x}|\mathbf{z})] - \beta \text{KL}(q_\phi(\mathbf{z}|\mathbf{x})||p(\mathbf{z}))] + \alpha (\mathbb{E}_{q_\phi(\mathbf{x}, \mathbf{z})} [t(\mathbf{x}, \mathbf{z})] - \mathbb{E}_{q(\mathbf{x})q_\phi(\mathbf{z})} [\exp(t(\mathbf{x}, \mathbf{z}) - 1)]). \quad (6)$$

¹strict in the sense that $D(q(\cdot)||p(\cdot)) = 0 \iff q(\cdot) = p(\cdot)$

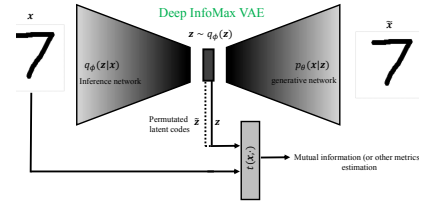


Fig. 1: Architecture of InfoMax-VAE, a family of VAEs which encourages VAEs to learn useful high-level representations of data, the top networks can be convolutional or FC NNs. The bottom network ($t(\mathbf{x}, \cdot)$) is an MLP which estimates the mutual information between inferred latent representations \mathbf{z} and input data \mathbf{x} . Or, it can estimate any other strict divergence/distance between joint $q_\phi(\mathbf{x}, \mathbf{z})$ and the marginals $q(\mathbf{x}), q_\phi(\mathbf{z})$.

To evaluate $\mathbb{E}_{q_\phi(\mathbf{x}, \mathbf{z})}$ and $\mathbb{E}_{q(\mathbf{x})q_\phi(\mathbf{z})}$ in a tractable manner, we take an alternative approach. First, we observe that we can simply draw samples from $(\mathbf{x}^{(i)}, \mathbf{z}^{(i)}) \sim q_\phi(\mathbf{x}, \mathbf{z}) = q_\phi(\mathbf{z}|\mathbf{x})q(\mathbf{x})$ thanks to the reparameterization trick and having access to the empirical distribution of input data $q(\mathbf{x}) = \frac{1}{N} \sum \delta_{\mathbf{x}^{(i)}}(\mathbf{x})$. Also to get samples from the marginal $q_\phi(\mathbf{z})$ we can randomly choose a datapoint $\mathbf{x}^{(j)}$ then sample from $\mathbf{z} \sim q_\phi(\mathbf{z}|\mathbf{x}^{(j)})$. In practice, however, we can effectively get samples from a batch and then permute representations \mathbf{z} across the batch. This trick is first used in [22], and proved to be sufficiently accurate as long as the batch size is large enough, *i.e.*, 64.

Finally, while f -divergence family offers a large class of different divergences, our proposed method is also capable of taking other divergence families or other dual representations which might enable tighter bounds. As an example, we can also use Donsker-Varadhan dual representation for the KL divergence. In doing so, we obtain a tighter lower bound than f -dual representation for the KL divergence. See Appendix C of our companion paper [21] for more details. A detailed description of InfoMax-VAE approach is presented in Algorithm 1 and Figure 1.

V. RESULTS

In this section, we evaluate our proposed InfoMax-VAE, and demonstrate that it consistently discovers more efficient high-level representations compared to other well-known approaches. To this end, we conduct experiments across the following datasets to compare the behavior of InfoMax-VAE against vanilla VAEs and its variant frameworks: 1) **MNIST**: 60,000 gray scale 28x28 images, 2) **Binarized MNIST**: the binary version of MNIST (with autoregressive decoder; results are provided in our companion paper [21]), 3) **Fashion MNIST**: 60,000 gray scale 28x28 images, 4) **CIFAR-10,100**: 60,000 RGB 32x32x3 images in 10 and 100 classes, 5) **CelebA**(shrunk and cropped version, see [21]): 12,000 RGB 64x64x3 images of celebrities. Other details of experiments such as hyperparameter setting, optimization, and the architectures of inference and generative networks are provided in [21]. In all experiments we get the best results by the choice of $f(t) = t \log t$, see Appendix C of [21]. Also,

Algorithm 1 InfoMax-VAE

Input: \mathcal{B} as a batch size of b , latent variable dimension \mathbf{z}_{dim} , α , observations $\{\mathbf{x}\}_{i=1}^N$, VAE/Mutual information optimizers: G , G_t

- 1: Initialize ϕ, θ, t
- 2: **repeat**
- 3: Randomly select b observed datapoints from $\{\mathbf{x}\}_{i=1}^b$
- 4: Get samples of $\mathbf{z}^{(i)} \sim q_\phi(\mathbf{z}|\mathbf{x}^{(i)})$, form $\{(\mathbf{x}, \mathbf{z})\}_{i=1}^b$,
- 5: permute latent codes $\{\mathbf{z}\}_{i=1}^b$ to get $\{(\mathbf{x}, \tilde{\mathbf{z}})\}_{i=1}^b$
- 6: $\theta, \phi \leftarrow G(\nabla_{\theta, \phi} [\frac{1}{b} \sum_{i=1}^b \log \frac{p_\theta(\mathbf{x}^{(i)}|\mathbf{z}^{(i)})p(\mathbf{z})^\beta}{q_\phi(\mathbf{z}^{(i)}|\mathbf{x}^{(i)})^\beta} + \frac{\alpha}{b} (\sum_{i=1}^b t(\mathbf{x}^{(i)}, \mathbf{z}^{(i)}) - \sum_{i=1}^b f^*(t(\mathbf{x}^{(i)}, \tilde{\mathbf{z}}^{(i)})))]])$
- 7: $t \leftarrow G_t(\nabla_t [\frac{1}{b} \sum_{i=1}^b t(\mathbf{x}^{(i)}, \mathbf{z}^{(i)}) - \frac{1}{b} \sum_{i=1}^b f^*(t(\mathbf{x}^{(i)}, \tilde{\mathbf{z}}^{(i)})])]$
- 8: **until** convergence

TABLE I: Performance of InfoMax-VAE vs vanilla, β -, and Info VAEs on MNIST w.r.t log-likelihood-, mutual information, KL divergence on MNIST dataset with FC and convolutional NNs.

Method	\mathbf{z}_{dim}	Architecture	log-likelihood	MI	KL	AU
VAE	2	FC	-131.36	2.8882	6.8113	2
β -VAE	2	FC	-140.26	1.8139	4.4933	2
Info-VAE	2	FC	-197.60	0.3199	9.8007	2
InfoMax-VAE	2	FC	-131.50	3.6180	24.186	2
VAE	20	CNN	-82.96	2.6255	18.3929	10
β -VAE	20	CNN	-123.34	1.7995	5.2494	13
Info-VAE	20	CNN	-85.12	2.1803	132.5293	20
InfoMax-VAE	20	CNN	-83.36	4.1612	23.4517	20

TABLE II: Performance of InfoMax-VAE vs Vanilla, β -, and Info- VAEs on MNIST (Top) and Fashion MNIST (Bottom) w.r.t log-likelihood and AU. The model is fixed at having 20 latent variables. InfoMax-VAE outperforms the others on the aforementioned metrics as varying the complexity of the generative network.

Dataset	Layers	log-likelihood				AU			
		VAE	β -VAE	Info-VAE	InfoMax-VAE	VAE	β -VAE	Info-VAE	InfoMax-VAE
MNIST	2	-86.60	-154.47	-120.61	-103.14	20	19	20	20
	4	-115.74	-164.64	-130.76	-99.68	20	18	20	20
	10	-153.35	-175.33	-159.23	-146.10	11	14	19	20
Fashion MNIST	2	-253.87	-265.00	-253.37	-252.17	20	20	20	20
	4	-254.44	-291.25	-245.07	-243.46	18	13	20	20
	10	-284.82	-277.54	-266.12	-262.61	9	17	12	20

we examine the effects of α and β in our model, please see Appendix E of [21].

Quantitative Evaluation. To conduct a thorough evaluation of the capability of InfoMax-VAEs in learning representations we employ three different metrics: mutual information, KL divergence, and active units (7). First, to measure the amount of information carried through to latent factors, we utilize the method proposed in [23]. We first train autoencoders on a training dataset, and then we provide the observed data and achieved latent representations into another network to estimate the resulting mutual information. A detailed description of this technique can be found in [23]. In doing so, we can estimate the mutual information between the input data and latent codes. The KL divergence as in (1) is another metric which represents the divergence between the variational posterior and prior, $\text{KL}(q_\phi(\mathbf{z}|\mathbf{x})||p(\mathbf{z}))$.

The active units (AU) metric is another measure to assess

latent variable collapse defined in [24]. With this metric, we examine each dimension of the latent code independently from the others, and the believe is that the distribution of a latent dimension changes based on the input data if it keeps useful information. Therefore, it can be expressed as:

$$\text{AU} = \sum_{d=1}^D \mathbb{I}\{\text{Cov}_x(\mathbb{E}_{q_\phi(z_d|\mathbf{x})}[z_d]) \geq \epsilon\}, \quad (7)$$

where z_d corresponds to the d -th dimension of the latent code \mathbf{z} , and $\epsilon = 0.05$ is a threshold. Also \mathbb{I} is an indicator which is 1 if its argument is true and 0 otherwise. In reliance on these metrics, we compare InfoMax-VAE to other well-known models of VAEs: β -VAE and Info-VAE. Moreover, we reported the log-likelihood to examine how much the obtained latent codes contribute to retrieve the input.

We summarize the results on MNIST test dataset for networks with FC and Convolutional layers with different size of

TABLE III: Performance of InfoMax-VAE vs Vanilla, β , and Info- VAEs on MNIST (Top) and Fashion MNIST (Bottom) w.r.t mutual information and KL divergence. The model is fixed at having 20 latent variables. InfoMax-VAE outperforms the other frameworks on the mentioned metrics as varying the complexity of the generative network.

Dataset	Layers	Mutual Information				KL Divergence			
		VAE	β -VAE	Info-VAE	InfoMax-VAE	VAE	β -VAE	Info-VAE	InfoMax-VAE
MNIST	2	4.32	2.26	3.42	5.03	23.50	5.17	43.25	26.65
	4	4.37	2.02	2.92	4.60	19.18	3.80	22.59	26.87
	10	3.24	1.35	2.54	4.02	8.25	1.34	18.33	12.99
Fashion MNIST	2	3.55	2.04	3.36	3.58	16.96	5.57	84.87	18.38
	4	3.12	2.12	3.56	3.85	14.28	5.5	82.09	16.90
	10	3.04	2.53	3.57	3.87	9.64	4.82	43.42	15.72

TABLE IV: Performance of InfoMax-VAE vs Vanilla, β -, and Info- VAEs on CIFAR-10 and CIFAR-100. The inference and generative networks have the same architecture in all scenarios. InfoMax-VAE outperforms other techniques in both classification and the activation of latent codes.

Dataset	z_{dim}	Accuracy %				Active Units			
		VAE	β -VAE	Info-VAE	InfoMax-VAE	VAE	β -VAE	Info-VAE	InfoMax-VAE
CIFAR-10	100	27.52	24.12	31.45	32.55	99	90	93	100
	200	32.83	25.82	39.05	41.75	187	182	193	200
	500	31.61	24.59	32.74	40.36	490	498	499	500
CIFAR-100	100	15.82	11.74	16.57	18.26	99	90	93	100
	200	14.49	11.46	16.36	17.24	187	182	193	200
	500	10.21	10.59	10.81	16.46	490	498	499	500

the latent dimension in Table I. As the results in Table I shows, flexible networks typically result in poor amortized inference. Indeed, VAEs attempt to compress the aggregated posterior on the center for both FC and convolutional NNs which guarantee the minimization of $KL(q_\phi(z|x)||p(z))$. This process severely merges the latent codes regardless of their categories. InfoMax-VAEs mitigate this by assuring maximization of the mutual information between the input and latent codes. Further, InfoMax-VAEs achieve higher mutual information, KL divergence, AUs compared to conventional VAEs and β -VAEs. For CNNs with 20 latent dimensions, Info-VAEs end up with a large KL divergence which can be observed as its very poor performance in matching the marginalized posterior and prior, see Appendix B of [21] for further details. In InfoMax-VAEs, we require that the model encodes sufficient information about observations, which also diligently preserve the KL divergence between $q_\phi(z)$ and $p(z)$ from blowing up.

Table II & III show the studies on MNIST and Fashion MNIST datasets as the generative networks become richer. In all experiments in this segment, we fix the inference network and set the dimension of latent codes to be 20. Table II & III suggest that as the generative network becomes more expressive, the latent codes become less reliant on the observations. We can infer this from the evaluated metrics. The InfoMax-VAE, however, helps to prevent from mode collapse and it performs better on all metrics. These results indicate that InfoMax-VAE has a strong inductive bias to keep more information in latent codes even in the presence of complex generative networks. Further, the high KL distance achieved by Info-VAE shows its poor performance in matching $q_\phi(z)$ and $p(z)$, please see

Appendix B of [21].

Generalization. Another evaluation that we performed is the classification task directly on the learned features of the data. Both the inference and generative networks are fixed after training. For this part, we performed evaluation on CIFAR-10 and CIFAR-100 datasets with different dimensions of latent codes. Table IV shows the results of InfoMax-VAE against vanilla, β -, and Info- VAEs performed on test dataset. Further, the number of active units are reported for each scenarios. We observe that InfoMax-VAE outperforms other models both in classification and activating all available latent codes. These results suggest that not only the proposed InfoMax-VAE is capable of learning useful and meaningful representations, but it also reveals that the learned features are generalized better than the aforementioned frameworks.

VI. CONCLUSION

In this paper we find that the conventional objective of VAEs is insufficient towards obtaining general, useful representations. We also determine that rich generative networks discourage the model from learning constructive representations. We propose a new information-based VAE that constrain latent representations so that the amount of information that they store from the observations is maximized, even in the presence of very rich networks. We perform extensive computational experiments and compare our work to other well-known approaches, where the proposed InfoMax-VAE outperforms them based on different metrics. We perform extensive computational experiments and compare our work to other well-known approaches, where the

proposed InfoMax-VAE outperforms them based on different metrics.

REFERENCES

- [1] Yunchen Pu, Win Yuan, Andrew Stevens, Chunyuan Li, and Lawrence Carin. A deep generative deconvolutional image model. In *Artificial Intelligence and Statistics*, pages 741–750, 2016.
- [2] Alec Radford, Luke Metz, and Soumith Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv preprint arXiv:1511.06434*, 2015.
- [3] Sebastian Nowozin, Botond Cseke, and Ryota Tomioka. f-gan: Training generative neural samplers using variational divergence minimization. In *Advances in Neural Information Processing Systems 29*, pages 271–279, 2016.
- [4] R Devon Hjelm, Alex Fedorov, Samuel Lavoie-Marchildon, Karan Grewal, Adam Trischler, and Yoshua Bengio. Learning deep representations by mutual information estimation and maximization. *arXiv preprint arXiv:1808.06670*, 2018.
- [5] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.
- [6] Samuel R Bowman, Luke Vilnis, Oriol Vinyals, Andrew M Dai, Rafal Jozefowicz, and Samy Bengio. Generating sentences from a continuous space. *arXiv preprint arXiv:1511.06349*, 2015.
- [7] Yunchen Pu, Zhe Gan, Ricardo Henao, Xin Yuan, Chunyuan Li, Andrew Stevens, and Lawrence Carin. Variational autoencoder for deep learning of images, labels and captions. In *Advances in neural information processing systems*, pages 2352–2360, 2016.
- [8] Jiacheng Xu and Greg Durrett. Spherical latent spaces for stable variational autoencoders. *arXiv preprint arXiv:1808.10805*, 2018.
- [9] Michael Tschannen, Olivier Bachem, and Mario Lucic. Recent advances in autoencoder-based representation learning. *arXiv preprint arXiv:1812.05069*, 2018.
- [10] Alexander A Alemi, Ben Poole, Ian Fischer, Joshua V Dillon, Rif A Saurous, and Kevin Murphy. Fixing a broken elbo. *arXiv preprint arXiv:1711.00464*, 2017.
- [11] Adji B. Dieng, Yoon Kim, Alexander M. Rush, and David M. Blei. Avoiding latent variable collapse with generative skip models. In *Proceedings of Machine Learning Research*, volume 89, pages 2397–2405. PMLR, 16–18 Apr 2019.
- [12] Xi Chen, Diederik P Kingma, Tim Salimans, Yan Duan, Prafulla Dhariwal, John Schulman, Ilya Sutskever, and Pieter Abbeel. Variational lossy autoencoder. *arXiv preprint arXiv:1611.02731*, 2016.
- [13] Shengjia Zhao, Jiaming Song, and Stefano Ermon. Towards deeper understanding of variational autoencoding models. *arXiv preprint arXiv:1702.08658*, 2017.
- [14] Zichao Yang, Zhiting Hu, Ruslan Salakhutdinov, and Taylor Berg-Kirkpatrick. Improved variational autoencoders for text modeling using dilated convolutions. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70, pages 3881–3890. PMLR, 06–11 Aug 2017.
- [15] Aaron van den Oord, Oriol Vinyals, and koray kavukcuoglu. Neural discrete representation learning. In *Advances in Neural Information Processing Systems 30*, pages 6306–6315, 2017.
- [16] Chunyuan Li, Hao Liu, Changyou Chen, Yuchen Pu, Liqun Chen, Ricardo Henao, and Lawrence Carin. Alice: Towards understanding adversarial learning for joint distribution matching. In *Advances in Neural Information Processing Systems 30*, pages 5495–5503, 2017.
- [17] Yuchen Pu, Weiyao Wang, Ricardo Henao, Liqun Chen, Zhe Gan, Chunyuan Li, and Lawrence Carin. Adversarial symmetric variational autoencoder. In *Advances in Neural Information Processing Systems 30*, pages 4330–4339. Curran Associates, Inc., 2017.
- [18] Matthew D Hoffman and Matthew J Johnson. Elbo surgery: yet another way to carve up the variational evidence lower bound. In *Workshop in Advances in Approximate Bayesian Inference, NIPS*, volume 1, 2016.
- [19] Hyunjik Kim and Andriy Mnih. Disentangling by factorising. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 2649–2658, Stockholmsmässan, Stockholm Sweden, 10–15 Jul 2018. PMLR.
- [20] Shengjia Zhao, Jiaming Song, and Stefano Ermon. Infovae: Balancing learning and inference in variational autoencoders. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 5885–5892, 2019.
- [21] Ali Lotfi Rezaabad and Sriram Vishwanath. Learning representations by maximizing mutual information in variational autoencoders. *arXiv preprint arXiv:1912.13361*, 2019.
- [22] Miguel A Arcones and Evarist Gine. On the bootstrap of u and v statistics. *The Annals of Statistics*, pages 655–674, 1992.
- [23] Mohamed Ishmael Belghazi, Aristide Baratin, Sai Rajeswar, Sherjil Ozair, Yoshua Bengio, Aaron Courville, and R Devon Hjelm. Mine: mutual information neural estimation. *arXiv preprint arXiv:1801.04062*, 2018.
- [24] Yuri Burda, Roger Grosse, and Ruslan Salakhutdinov. Importance weighted autoencoders. *arXiv preprint arXiv:1509.00519*, 2015.