THE VARIATIONAL INFOMAX AUTOENCODER

Anonymous authors
Paper under double-blind review

ABSTRACT

We propose the Variational InfoMax AutoEncoder (VIMAE), an autoencoder based on a new learning principle for unsupervised models: the Capacity-Constrained InfoMax, which allows the learning of a disentangled representation while maintaining optimal generative performance. The variational capacity of an autoencoder is defined and we investigate its role. We associate the two main properties of a Variational AutoEncoder (VAE), generation quality and disentangled representation, to two different information concepts, respectively Mutual Information and network capacity. We deduce that a small capacity autoencoder tends to learn a more robust and disentangled representation than a high capacity one. This observation is confirmed by the computational experiments.

1 INTRODUCTION

A common assumption in machine learning is that any visible data $x \in \mathcal{X}$ is completely described by some generative factor o, living in a smaller hidden space $\mathcal{O}$, i.e. $x = g(o)$ with g a (possibly stochastic) generative function. The aim of unsupervised representation learning research is to find a *representation* z of the generative factor o living in a known space $\mathcal{Z}$ describing, as well as o, the visible data x. This is particularly relevant because the learnt small representation z is task agnostic and, in principle, can be used as input for networks performing different tasks, leading to faster and more robust learning (*generalisation property*), (Rifai et al., 2011).

Many models $f_\phi : \mathcal{X} \to \mathcal{Z}$ trying to learn such representations have been proposed (Dinh et al., 2016; Hinton et al., 2006; Maddison et al., 2017; Radford et al., 2015), but recently in order to solve this problem it was proposed to consider a dual problem: define a priori z and find a generator map g_θ, such that for any z, $g_\theta(z)$ is an element of $\mathcal{X}$. In particular, two families of probabilistic generative models have become dominant: Variational AutoEncoder (VAE) (Kingma & Welling, 2013; Rezende et al., 2014) and Generative Adversarial Network (GAN) (Goodfellow et al., 2014). The common idea of the two approaches is that a good generator $p_\theta(x|z)$ is the one able to generate the data that is as close as possible to the visible one, i.e. that with respect a certain metric D, the distance between the marginal $p_\theta(x) = \mathbb{E}_{p(z)}[p_\theta(x|z)]$ and the visible distribution $p_D(x)$ is minimal.

In this manuscript we restrict our attention to the VAE model, since by its architecture, it is the only one where the learnt representation can be used as input for networks performing different tasks. Although VAE, by its training robustness and general good generative performance is the most popular model for representation learning, in particular cases it suffers from the *uninformative representation* issue: the representation is entangled and the generative model tends to be independent of z, i.e. $p_\theta(x|z) \approx p_\theta(x)$. As highlighted in the next section such behaviour is intrinsic in the variational loss, the Evidence Lower BOund (ELBO), encouraging a less informative representation.

We propose a method that learns the most informative generative model between the visible and hidden representations, while maintaining a bounded capacity of the encoding network. This is in accordance with the many attempts to solve such issues, by reducing the encoding information to have a more disentangled representation (Higgins et al., 2017; Burgess et al., 2018) or, oppositely, removing the encoding information penalty (Zhao et al., 2017) to improve the generative quality performance.

We present a new information theoretic derivation of a Variational InfoMax (VIM) objective, which turns out to be the Variational Wasserstein Distance suggested to optimise the Wasserstein AutoEncoder (WAE) (Tolstikhin et al., 2017). Our derivation allows us to define the capacity of the variational network and individuate the parameters to optimise in order to have both good generative performance and a disentangled representation. Indeed, we argue that in order to learn good representations it is sufficient to bound the capacity of the network and not the encoding information. That means, from a pure information perspective, that an unsupervised network should optimise a Capacity-Constrained InfoMax objective, a principle slightly different from the Information Bottleneck (Tishby et al., 2000), where the capacity of the encoding network is bounded instead of its information.

The theoretical arguments are confirmed by the performed experiments where we observe that, differently from what was argued in previous works (Higgins et al., 2017; Burgess et al., 2018), it is possible to train a model that is able to learn *good* (able to generalise) representations while maintaining optimal generative performance. The main contributions of the paper are summarised in the following points:

- derivation of a variational lower bound for the maximal mutual information of a generaive model belonging in a certain family, see equation 6;
- definition and bounds estimation for the network capacity for a variational autoencoder, see equation 9;
- association of the two main properties of VAE, generation quality and disentanglement representation, to two different information concepts, respectively Mutual Information and network capacity;
- proposal of a new learning principle for unsupervised models: the Capacity-Constrained InfoMax, see equation 10, that allows both to learn a disentangled representation while maintaining optimal generative performance.

The work is divided as follows: in the second section we describe briefly the VAE and its variants; in the third and fourth sections we describe the variational infomax method and related work. We conclude the paper with the experimental results and the conclusions.

2 Background

The aim of this section is to describe VAE, understand principal issues of the ELBO objective and describe the two most relevant approaches to overcome such issues.

2.1 Notation and preliminary definitions

We use calligraphic letters (i.e. $\mathcal{X}$) for sets, capital letters (i.e. X) for random variables, and lower case letters (i.e. x) for their samples. With abuse of notation we denote both the probability and the corresponding density with the lower case letters (i.e. $p(x)$).

f-Divergence Given two random distributions $p(x)$ and $q(x)$, the f-divergence

$$D_f(p(x)||q(x)) = \int f\Big(\frac{p(y)}{q(y)}\Big)q(y)dy \tag{1}$$

is an (intuitive) measure of the distance between the distributions p and q. In the case $f(x) = x \log x$, D is called Kullback-Leibler (KL) divergence.

Mutual Information and Capacity Given a channel $Z \to X$ with X and Z random variables, jointly distributed according to $p(x, z)$ and with marginals $p(x)$ and $p(z)$. The mutual information

$$I(X, Z) = D_{KL}(p(x, z)||p(x)p(z)),$$

is a measure of the reduction of uncertainty in X due to the knowledge of Z, and the capacity

$$C(X, Z) = \sup_{p(z)\in\mathcal{P}} I(X, Z)$$

is the maximal information that can be shared for a fixed generator $p(x|z)$.

2.2 Variational autoencoder

From now on let us assume that the unknown distribution of the data $p(x)$ coincides with the empirical one $p_D(x)$, and that the distribution of the latent representation $p(z)$ is known. In this context the VAE is a model solving the following optimisation problem: find the generative model $p_\theta(x,z) \in \mathcal{P}_\theta$, specified by the parameters θ of the associated neural network, maximising the ELBO objective

$$ELBO_{\theta,\phi} = \mathbb{E}_{p(x)}[-D_{KL}(q_\phi(z|x)||p(z)) + \mathbb{E}_{q(z|x)}[\log p(x|z)]], \tag{2}$$

a lower bound of the unfeasible-to-compute marginal likelihood $\mathbb{E}_{p(x)}[\log p_\theta(x)]$. The ELBO objective is optimized by a regularized autoencoder, with encoder and decoder parameterizing, respectively, the inference and generative distributions, $q_\phi(z|x)$ and $p_\theta(x|z)$, with $\phi \in \Phi$, $\theta \in \Theta$ and regularizer defined by the *rate* term $D_{KL}(q_\phi(z|x)||p(z))$, measuring the excess number of bits required to encode samples from the encoder using the optimal code designed for $p(z)$.

2.3 Uninformative representation issue

As underlined in the introduction, the main issue of VAE is that the representations are not really informative of the input data and in the worst case, it is learned a Z-independent generative model $p_\theta(x|z) = p_\theta(x)$. Such issues are intrinsic in the ELBO objective, equation 2, that can reach the optimum when $D_{KL}(q_\phi(z|x)||p(z)) = 0$ (Zhao et al., 2017). The latter case means that the representation is completely uninformative, indeed the rate term, which can be rewritten as

$$D_{KL}(q_\phi(z|x)||p(z)) = I_q(X,Z) + D_{KL}(q_\phi(z)||p(z)),$$

is a penalty on the encoding information, and is zero when $I_q(X,Z) = 0$, with $q_\phi(z|x) = q_\phi(z) = p(z)$, i.e. when q_ϕ does not encode any information about the input x.

We now describe the two most relevant models that try to overcome the uninformative representation issue.

InfoVAE In (Zhao et al., 2017) the InfoVAE family of models was proposed, a generalisation of the VAE model optimising the objective

$$-\alpha I_q(X,Z) - \lambda D_{KL}(q_\phi(z)||p(z)) + \mathbb{E}_{p(x)}[\mathbb{E}_{q(z|x)}[\log p(x|z)]],$$

with α and λ two real positive hyper-parameters.

The main advantage of this definition is that it is possible to consider separately the two components of the rate term. In particular, in (Zhao et al., 2017) it was observed that by eliminating the information penalty ($\alpha = 0$), the generative performance of the model improves and the representation results are more informative.

β-VAE In (Higgins et al., 2017), starting from the observation that the optimal case is rare, but most of the learned features by VAE are not disentangled, it is proposed an opposite approach: put a high penalty to the rate term, in order to constrain the model to learn the most informative property of the data, and then have a disentangled represention of the data. The β-VAE family is a particular case of InfoVAE where $\alpha = \lambda \gg 1$. This idea, that at first sight looks counter-intuitive, is based on the observation that by the additive property of the KL-divergence

$$D_{KL}(q_\phi(z|x)||p(z)) = \sum_{i=1}^{dim(\mathcal{Z})} D_{KL}(q_\phi(z_i|x)||p(z_i)) \tag{3}$$

pushing the penalty associated with the rate is equivalent to penalising the informativeness of most features, leaving few features containing the relevant information. Starting from a bits-back coding argument, a similar conclusion was derived in (Chen et al., 2016).

We conclude this section observing that although InfoVAE and $\beta-$VAE approaches are antithetic, in both the cases the hyper-parameter λ associated to the KL divergence term $D_{KL}(q_\phi(z)||p(z))$, is bigger than 1; as we will see in the next section, this factor is controlling the capacity of the network.

3 The Model

3.1 The Variational InfoMax

Assuming known the distribution associated to the two random variables $p(x)$ and $p(z)$, the InfoMax objective is defined as: find the joint distribution $p_\theta(x, z) \in \mathcal{P}_\theta := \{p_\theta(x, z) : \mathbb{E}_{p(z)}[p_\theta(x|z)] = p(x), \quad \mathbb{E}_{p(x)}[p_\theta(z|x)] = p(z)\}$ maximising the mutual information $I_\theta(X, Z) = D_{KL}(p_\theta(x, z)||p(x)p(z))$, i.e. find $\theta^* \in \Theta$ s.t. $I_{\theta^*} \geq I_\theta$ for any $\theta \in \Theta$.

Since the definition via KL divergence is computationally intractable, it is necessary to re-write the mutual information as

$$I_\theta(X, Z) = h_\theta(X) - h_\theta(X|Z), \tag{4}$$

where $h_\theta(X) = -\mathbb{E}_{p_\theta(x)}[\log p_\theta(x)]$ is the entropy of X, and $h_\theta(X|Z) = -\mathbb{E}_{p_\theta(x,z)}[\log p_\theta(x|z)]$ is the conditional entropy $h_\theta(X|Z)$. Since $p_\theta(x, z) \in \mathcal{P}_\theta$ the entropy $h_\theta(X) = h(X)$ is constant, and in order to maximise the mutual information it is sufficient to minimise the conditional entropy.

Excluding some special cases (Bell & Sejnowski, 1997), minimising the conditional entropy is unfeasible, so it is necessary to consider an associated variational problem: for any $q_\phi(z|x)$ such that $q_\phi(z) = p(z)$ and $\phi \in \Phi = \Theta$, learn the generative model $p_\theta(x|z)$ minimising the reconstruction accuracy term $\mathbb{E}_{p(x)}[\mathbb{E}_{q_\phi(z|x)}[\log(p_\theta(x|z))]]$. Indeed, the following variational objective:

$$I_{\theta,\phi}(X, Z) = h(X) + \mathbb{E}_{p(x)}\mathbb{E}_{q(z|x)}[\log p_\theta(x|z)] \qquad \text{s.t. } q_\phi(z) = p(z) \tag{5}$$

is a lower bound of $I_{\theta^*}(X, Z)$ and is maximal when $q(z|x) = p_\theta(z|x) = p_{\theta^*}(z|x)$, (see the Appendix).

Unfortunately, the formulation in equation 5 is still unfeasible to compute, because it requires that $q_\phi(z) = p(z)$, but by the butterfly architecture of the autoencoder, $q_\phi(z)$ tends to be uniformly distributed on the space $\mathcal{Z}$. For this reason, the model is trained maximising the following relaxed form:

$$VIM_{\theta,\phi} = \mathbb{E}_{p(x)}\mathbb{E}_{q(z|x)}[\log p_\theta(x|z)] - \lambda D(q_\phi(z)||p(z)), \tag{6}$$

where it is introduced a term $D(q_\phi(z)||p(z))$ encouraging the empirical distribution $q_\phi(z)$ to be close, according to the metric D, to $p(z)$. In the following, in order to avoid any confusion the variational autoencoder trained maximising equation 6 will be dubbed VIMAE.

From now on let us assume, $D = D_{KL}$. Under this condition the VIM objective coincides with the objective optimised in InfoVAE, and the regularizer is approximated via the Maximum Mean Discrepancy (MMD) (Zhao et al., 2017) defined as:

$$\text{MMD}(q(z), p(z)) = \sup_{f:\|f\|_{\mathcal{H}_k}\leq 1} \mathbb{E}_{p(z)}[f(Z)] - \mathbb{E}_{q(z)}[f(Z)] \tag{7}$$

where $\mathcal{H}_k$ is the Reproducing Kernel Hilbert Space associated to a positive definite kernel $k(\cdot, \cdot) : \mathcal{Z} \times \mathcal{Z} \to \mathbb{R}_+$.

Encoding channel In VAE we observed that an uninformative representation was caused by the non-informativeness of the encoding map $q_\phi(z|x)$. Since from equation 6 it is not clear how $q_\phi(z|x)$ behaves, we consider an equivalent representation, (see the Appendix):

$$VIM_{\theta,\phi} = -D_{KL}(p(x)||p_\theta(x)) - (\lambda - 1)D_{KL}(q_\phi(z)||p(z)) + I_{\theta,\phi}(X, Z). \tag{8}$$

From equation 8 we see that the infomax objective, equation 6, can be read as a composition of three sub-objectives: find a generative model $p_\theta(x|z)$, with marginal resembling the visible distribution $p(x)$ (first term); maximise the (unbounded) variational mutual information (third term); and learn an inferred distribution $q_\phi(z, x)$ close to the generative model $p_\theta(x, z)$. Then the optimum is obtained by $q_\phi(x, z) = p_\theta(x, z)$ such that $I_\theta(X, Z)$ is maximal, confirming the validity of the approximation made above.

3.2 Channel capacity

In a channel with variational mutual information $I_{\theta,\phi}$ as defined in equation 5 the (variational) capacity $C_{\theta,\phi}(X, Z)$, is defined as

$$C_{\theta,\phi}(X, Z) = \sup_{\theta,\phi,p(z)\in\mathcal{P}} I_{\theta,\phi}(X, Z). \tag{9}$$

If $\mathcal{P}$ is the space of all distributions on $\mathcal{Z}$, the capacity of the network coincides with the variational mutual information of the model trained minimising only the reconstruction loss. In the latter case, as observed above, it is not guaranteed to learn the generator with $p_\theta(z) = p(z)$. This is because, given two equally informative generative models $p_\theta(x|z)$ and $p_{\theta'}(x|z)$, with encoder respectively $q_\phi(z|x)$ and $q_{\phi'}(z|x)$, with $h(q_\phi(z)) < h(q_{\phi'}(z))$, then $I_{\theta,\phi} < I_{\theta',\phi'}$. From such observation we deduce that bounding the entropy of the representation Z is the way to bound the capacity without penalising the encoding information, and that the penalty introduced in equation 6 is actually a bound to the capacity itself. So, the VIM objective can be defined as a variational approximation of the Capacity-Constrained InfoMax:

$$\max_{Z \in \mathcal{Z}} I(X, Z) - \lambda C(X, Z), \tag{10}$$

given a set of equally informative generators, learn the one having the minimal capacity. The idea of the Capacity-Constrained InfoMax is similar to the idea of the Information Bottleneck, from which was derived the β-VAE: constrain the capacity of the network in order to learn only the relevant features of the input data. The difference with the Information Bottleneck,

$$\max_{Z \in \mathcal{Z}} I(X, Z) - \lambda I_q(X, Z). \tag{11}$$

lies in the second term: the network capacity instead of the encoding information. This choice allows the network to learn a good representation (small capacity) while maintaining good generative performance (high mutual information). Indeed, as shown in equation 8, the generative performance is associated to the informativeness of $q_\phi(z|x)$ and $p_\theta(x|z)$.

In order to test the assumption that it is sufficient to bound the entropy of Z, instead of the encoding mutual information, to learn a good representation, in the experiments (see below) we consider the cases Z is Normal (VIMAE-n) or Logistic (VIMAE-l) distributed. We choose to compare the popular Normal distribution with the Logistic one for two reasons: the Logistic has less entropy than a Gaussian distribution and because it is a common assumption in natural science to suppose that the hidden factors of the visible data are logistically distributed (Hyvärinen et al., 2009).

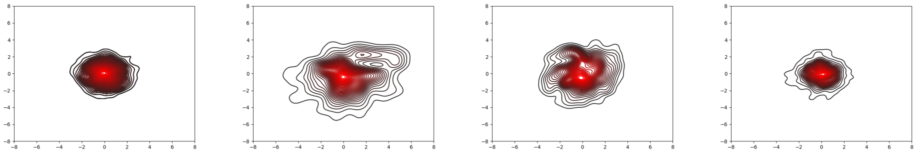
4 Related work

Autoencoder literature Autoencoder models are one of the most used family of neural networks to extract features in an unsupervised way (Bengio et al., 2013), and their relationship with Information Theory is well-established from the first unregularised autoencoders (Baldi & Hornik, 1989). The classical unregularised autoencoders, minimising the reconstruction loss $\mathbb{E}_{p(x)}[\mathbb{E}_{q_\phi(z|x)}[-\log p_\theta(x|z)]]$, are maximizing an unbounded information, i.e. they are looking for a solution in the space $\tilde{\mathcal{P}}_\theta = \{p_\theta : p_\theta(x) = p(x)\}$. A solution in this wide space is good only for reconstruction performance because Z contains all the possible information that can be stored in the space $\mathcal{Z}$, but from this representation it is impossible to sample, because the prior is unknown; and moreover such representation, in general, is not robust to input noise (Vincent et al., 2008).

Many regularised models have been proposed, but the most well known is VAE, that minimises the expected code length of communicating x. As we observed in the previous sections, it is not guaranteed that the method finds a useful representation, and in the second section we illustrated two principal ways to improve VAE.

The objective in equation 6 was derived independently in (Tolstikhin et al., 2017) and (Zhao et al., 2017). Particularly relevant is the derivation in (Tolstikhin et al., 2017) because it allows us to describe an informative model $p_\theta(x, z)$ as the one minimising the transport cost between the original and generated data.

Finally, we underline that in case we wish to consider a Jensen-Shannon divergence in equation 6 it is necessary to consider an adversarial network model, discriminating the true samples $z \sim p(z)$ from the fake sampled by $q_\phi(z)$ (Goodfellow et al., 2014). In the latter case the obtained model is equivalent to the Adversarial AutoEncoder (Makhzani et al., 2015). We conclude by remarking that in all the cases cited above the Infomax objective was never maximised using a prior $p(z)$ different from a Gaussian.



$p(z) = \mathcal{N}(0, 1)$ VAE, 0.70 β-VAE, 0.53 VIMAE, 0.08

Figure 1: 2-d learned representations. Under each plot the model name is followed by the respective MMD value.

Information theoretic literature Information theory is strongly related with neural networks, and not only with autoencoders. Originally the InfoMax objective was applied to a self-organised system with a single hidden layer, (Bell & Sejnowski, 1997; Linsker, 1989) where the bound in the capacity was given by the numbers of hidden neurons. More recently, the (naive) InfoMax has given way to a new information-theoretic principle: the Information-Bottleneck (IB) (Tishby et al., 2000). The idea of this principle is that a feed-forward neural network trained for task T tends to learn a minimal sufficient representation of the data, maximising the following objective:

$$\max_{Z} I(Z, T) - \beta I(X, Z). \tag{12}$$

Although it was shown that in the general case this principle does not hold true (Saxe et al., 2018), the principle was used as a regularisation technique with success both in unsupervised (Higgins et al., 2017) and supervised (Alemi et al., 2016) settings. We observe that the VIM, equation 6, and IB, equation 12, coincide in the case of deterministic encoder, where the encoding information is the entropy of Z.

5 Experiments

Here we empirically evaluate the VIMAE. The section is divided into three parts: in the first part we compare the ability of VAE, β-VAE and VIMAE to infer the representation, $z \sim \mathcal{N}(0, I)$. Such an experiment is to evaluate the entropy of $Z \sim q(Z)$ and then, as observed in section 3.2, an indirect way to estimate capacity of the network ($C(X, Z) \propto h(Z)$). In the second part we evaluate the reconstruction and generative performance of the models. Indeed, the combination of the two tasks is estimation of the mutual information of the generative model $I_\theta(X, Z)$, see equation 8. In the third part we evaluate the robustness to noise and generalisation property of the learnt representation, observing that an informative model with small capacity is the best one for these tasks.

5.1 The entropy of Z

Experiments in this part are performed with an autoencoder trained with the MNIST data-set, a collection of 70k monocromatic handwritten digits, where both the inference and generative distribution are modelled by 3-layer deep neural nets with 256 hidden units in each layer and $\mathcal{Z} = \mathbb{R}^2$.

In figure 1 are plotted the 2d representations learnt by the different methods and we observe that VAE and β-VAE are not able to learn a hidden representation fitting the prior $p(z)$, with $h(q(z)) > h(p(z))$. These results show that the representation learnt by ELBO is not a small capacity one and that the divergence penalty introduced in equation 6 is a bound for the entropy.

5.2 The model information

The experiments in these final sections were performed with the same settings and autoencoder models used in (Tolstikhin et al., 2017), an architecture similar to the DCGAN (Radford et al., 2015) with batch normalization (Ioffe & Szegedy, 2015) (more details given in the Appendix). We consider four data-sets: MNIST and CIFAR10, two standard data-sets with ground-truth labels; Omniglot, a data-set of 1623 characters from 50 alphabets, 30 training and 20 evaluation, where each character appears 80 times, to evaluate the informativeness of the model and the quality of the learned representation; in the Appendix, we also consider CelebA (Liu et al., 2015), consisting

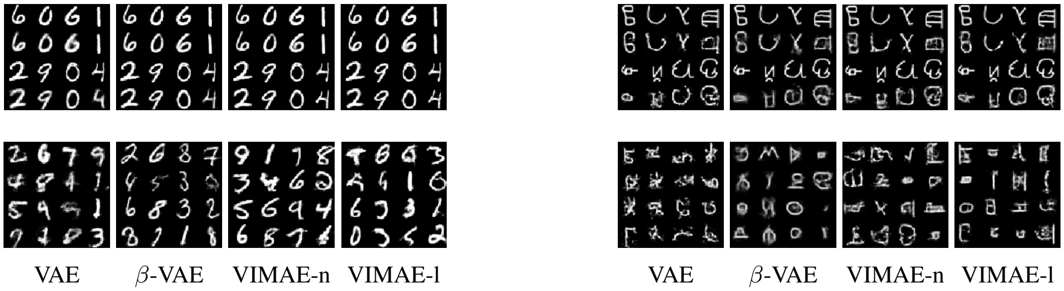


Figure 2: Test reconstruction (top) and random generative samples (bottom) of the different methods with MNIST (left) and Omniglot (right). In test reconstructions, the odd rows are the original data.

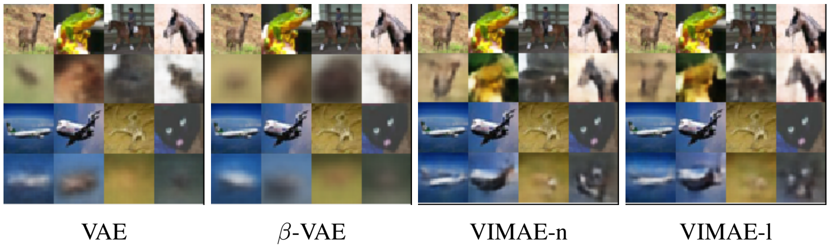


Figure 3: Test reconstruction, CIFAR 10. Odd rows are the original data.

of roughly of 203k faces of 64×64 resolution, in order to compare the generative quality of the pictures. After considering many parameters for β and λ, we choose, in accordance with what was suggested in (Tolstikhin et al., 2017), $\beta = \lambda = 10$ for MNIST and Omniglot and $\beta = \lambda = 100$ for CelebA and CIFAR10 experiments.

The goal of this section is to evaluate the informativeness of the learnt generative model $p_\theta(x|z)$. From what was described above, the reconstruction loss or the generation quality alone are not reliable metrics, because the reconstruction loss is an estimation of the variational mutual information $I_{\theta,\phi}(X, Z)$, and the generation quality is an estimation of $D_{KL}(p_\theta(x)||p(x))$ that, as observed in section 2.3, it is possible to minimise with an uninformative generator. But, according to equation 8 the combination of the two task performances is a good empirical estimation of $I_\theta(X, Z)$; indeed, by generative experiments, we require that $q(z) = p(z)$, so the KL divergence term in equation 8 comes for free.

Reconstruction and generative performances From figure 2 we observe that in the MNIST and Omniglot experiments, all the models are able to reconstruct, without big differences, the input data. Slightly different is the behaviour in the CIFAR10 experiments shown in figure 3, where the ELBO-based models suffer, in particular β-VAE; that result is not surprising and is in agreement with the theoretical observation that β-VAE is penalising the encoding mutual information $I_q(X, Z)$.

The models that we are considering are defined as generative models, so giving a sample $z \sim p(z)$ they should be able to generate a new data x similar to the original one. In figure 2 (bottom) are plotted the generated samples of MNIST and Omniglot, obtained from the different models. We observe that VAE and $\beta-$VAE do not generate good samples. Such qualitative results are confirmed by the experiments with the CelebA data-set (Liu et al., 2015), see the Appendix, where it is observed that if the generative difference between the two VIMAEs is small, the difference with the VAE counterparts is high. Such behaviour is in agreement with what was observed until now: the ELBO based model does not learn a good generative network, and the good reconstruction is simply associated to a large entropy of Z, instead of an informative generative model $p_\theta(x|z)$.

5.3 Generalisation property

We defined a good representation as the one containing the relevant properties of the visible data and able to generalise from the task for which was trained. In order to evaluate such quality, following the

Table 1: Semi-supervised classification CIFAR10.

	accuracy (%)		
Method	$\nu = 0$	$\mathcal{N}(0, 0.3^2)$	$\mathcal{B}(0.2)$
VAE	30	25	16
β-VAE	29	26	19
VIMAE-n	29	28	**23**
VIMAE-l	**32**	**34**	**23**

,

Table 2: Semi-supervised classification, MNIST.

	accuracy (%)				
Method	$\nu = 0$	$\nu = \mathcal{N}(0, \sigma^2)$		$\nu = \mathcal{B}(p)$	
		0.2	0.4	0.2	0.5
VAE	80	77	70	72	52
β-VAE	92	86	82	91	84
VIMAE-n	**93**	**92**	86	**92**	86
VIMAE-l	**93**	**92**	**88**	**92**	**87**

Table 3: Semi-supervised classification, Omniglot (random sampling: 20%).

accuracy (%)				
$\nu = 0$	$\nu = \mathcal{N}(0, \sigma^2)$		$\nu = \mathcal{B}(p)$	
	0.2	0.4	0.2	0.5
22	22	17	22	16
21	21	22	19	17
22	**23**	**24**	22	**22**
24	**23**	20	**23**	**22**

approach proposed in (Rifai et al., 2011), we evaluate the accuracy of an SVM directly trained on the learned features of the data. Proceeding as in (Zhao et al., 2017), we train the M1+TSVM (Kingma et al., 2014) and use the semi-supervised performance over 1000 (100 for Omniglot) samples as an approximate metric to verify the relevance and the quality of the learned representation. In order to evaluate the robustness of the learned features, we performed the same algorithm on the representation associated to corrupted data, i.e. $z \sim q(z|x + \nu)$, considering two types of noise: Gaussian and mask. In the Gaussian case, we add to each pixel a ν value sampled from $\mathcal{N}(0, \sigma^2)$ with $\sigma \in \{0.2, 0.3, 0.4\}$, and in the masking case a fraction ν of the elements is forced to be 0: each pixel is masked according to a Bernoulli distribution $\mathcal{B}(p), p \in \{0.2, 0.5\}$. Higher classification performance suggests that the learned representation contains the relevant information and, in case of corrupted input data, that it is robust. In the Omniglot case by the challenge of the task (the test alphabet was never seen in the training) we consider a 5-character data-set, split into 300 (60 × 5) for training and 100 for evaluation.

From the classification scores listed in tables 1- 3, we see that the ELBO-based model learnt good representations for clean data, but not when corrupted data is given as input. This is particularly clear in the Bernoulli case, that is a noise different from the one seen in the training. Particularly relevant is the behaviour of the two VIMAEs: they are comparable in the cases of clean data and small noise, but the one with big capacity, VIMAE-n, suffers in large noise, while the one with small capacity, VIMAE-l, is the most robust and in some challenging cases, see table 1, the noise helps to improve the model accuracy. Such a result is consistent with the idea that a small capacity network is learning the relevant factors of the input data, that are the only ones robust to the input noise.

6 Conclusion

We propose a VAE-based generative model, VIMAE, optimising Variational InfoMax, the variational form of the Capacity-Constrained InfoMax, a principle suggesting to learn the maximally informative generative model and maintain a bounded capacity. We show both theoretically and with computational experiments, the difference with the Information Bottlenck, and we observe that it is possible to learn a good generative model while maintaining an informative hidden representation. In accord to the theoretical analysis, we observed in the numerical experiments that the reconstruction and generative qualities are not orthogonal to the general and robust representation issue, because the former are associated to the mutual information of the model and the latter to the capacity of the model itself.

REFERENCES

Alexander A Alemi, Ian Fischer, Joshua V Dillon, and Kevin Murphy. Deep variational information bottleneck. *arXiv preprint arXiv:1612.00410*, 2016.

Pierre Baldi and Kurt Hornik. Neural networks and principal component analysis: Learning from examples without local minima. *Neural networks*, 2(1):53–58, 1989.

Anthony J Bell and Terrence J Sejnowski. The independent components of natural scenes are edge filters. *Vision research*, 37(23):3327–3338, 1997.

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

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

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.

Laurent Dinh, Jascha Sohl-Dickstein, and Samy Bengio. Density estimation using real nvp. *arXiv preprint arXiv:1605.08803*, 2016.

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

Irina Higgins, Loic Matthey, Arka Pal, Christopher Burgess, Xavier Glorot, Matthew Botvinick, Shakir Mohamed, and Alexander Lerchner. beta-vae: Learning basic visual concepts with a constrained variational framework. In *International Conference on Learning Representations*, volume 3, 2017.

Geoffrey E Hinton, Simon Osindero, and Yee-Whye Teh. A fast learning algorithm for deep belief nets. *Neural computation*, 18(7):1527–1554, 2006.

Aapo Hyvärinen, Jarmo Hurri, and Patrick O Hoyer. *Natural image statistics: A probabilistic approach to early computational vision.*, volume 39. Springer Science & Business Media, 2009.

Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *arXiv preprint arXiv:1502.03167*, 2015.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

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

Durk P Kingma, Shakir Mohamed, Danilo Jimenez Rezende, and Max Welling. Semi-supervised learning with deep generative models. In *Advances in neural information processing systems*, pp. 3581–3589, 2014.

Ralph Linsker. An application of the principle of maximum information preservation to linear systems. In *Advances in neural information processing systems*, pp. 186–194, 1989.

Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of the IEEE international conference on computer vision*, pp. 3730–3738, 2015.

Chris J Maddison, John Lawson, George Tucker, Nicolas Heess, Mohammad Norouzi, Andriy Mnih, Arnaud Doucet, and Yee Teh. Filtering variational objectives. In *Advances in Neural Information Processing Systems*, pp. 6573–6583, 2017.

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

Alec Radford, Luke Metz, and Soumith Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv preprint arXiv:1511.06434*, 2015.

Danilo Jimenez Rezende, Shakir Mohamed, and Daan Wierstra. Stochastic backpropagation and approximate inference in deep generative models. *arXiv preprint arXiv:1401.4082*, 2014.

Salah Rifai, Grégoire Mesnil, Pascal Vincent, Xavier Muller, Yoshua Bengio, Yann Dauphin, and Xavier Glorot. Higher order contractive auto-encoder. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pp. 645–660. Springer, 2011.

Andrew Michael Saxe, Yamini Bansal, Joel Dapello, Madhu Advani, Artemy Kolchinsky, Brendan Daniel Tracey, and David Daniel Cox. On the information bottleneck theory of deep learning. 2018.

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

Ilya Tolstikhin, Olivier Bousquet, Sylvain Gelly, and Bernhard Schoelkopf. Wasserstein auto-encoders. *arXiv preprint arXiv:1711.01558*, 2017.

Pascal Vincent, Hugo Larochelle, Yoshua Bengio, and Pierre-Antoine Manzagol. Extracting and composing robust features with denoising autoencoders. In *Proceedings of the 25th international conference on Machine learning*, pp. 1096–1103. ACM, 2008.

Oriol Vinyals, Charles Blundell, Timothy Lillicrap, koray kavukcuoglu, and Daan Wierstra. Matching networks for one shot learning. In D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett (eds.), *Advances in Neural Information Processing Systems 29*, pp. 3630–3638. Curran Associates, Inc., 2016. URL `http://papers.nips.cc/paper/6385-matching-networks-for-one-shot-learning.pdf`.

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

Relationship between encoding, decoding and variational Information

Defined $q(z,x) := q_\phi(z|x)p(x)$ the encoding distribution and $p_\theta(x,z) := p_\theta(x|z)p(z)$ the decoding one. The encoding $I_q(X,Z)$, decoding $I_\theta(X,Z)$ and the variational $I_{\theta,\phi}(X,Z)$ information are defined respectively:

$$\begin{aligned} I_q(X,Z) &= D_{KL}(q(z,x)||q(z)p(x)) = h(X) - h_q(X|Z) \\ I_\theta(X,Z) &= D_{KL}(p_\theta(z,x)||p(z)p(x)) = h(X) - h_\theta(X|Z) \\ I_{\theta,\phi}(X,Z) &= h(X) - \mathbb{E}_{q(z,x)}[-\log p_\theta(x|z)] \end{aligned}$$

Assuming $\theta^* \in \Theta$ is the parameter associated to the maximal decoding information, $I_\theta(X,Z) \leq I_{\theta^*}(X,Z)$ for any $\theta \in \Theta$, it follows that for any $q_\phi(x,z) \in \mathcal{P}_\theta$, i.e. for any $q_\phi(z) = \mathbb{E}_{p(x)}[q_\phi(x|z)] = p(z)$ and $\phi \in \Phi \subset \Theta$,

$$I_{\theta^*}(X,Z) \geq I_q(X,Z).$$

Then a lower bound of I_q is a lower bound of I_{θ^*}. By property of KL-divergence we have that for any $p_\theta(x|z)$ the following relationship holds:

$$\mathbb{E}_{q(z,x)}[-\log p_\theta(x|z)] = h_q(X|Z) + \mathbb{E}_{q(z)}[D_{KL}(q_\phi(x|z)||p_\theta(x|z))] \tag{13}$$

From equation 13 and the definition of the variational information $I_{\theta,\phi}$ we deduce that:

$$I_{\theta^*}(X,Z) \geq I_q(X,Z) \geq I_{\theta,\phi}(X,Z)$$

We conclude observing that if $\Theta = \Phi$, at optimum the three information terms above are equal, and then $q_\phi(x,z) = p_\theta(x,z)$.

Observation: If $\Phi \subset \Theta$, the variational mutual information can be at most equal to I_{q^*}, the maximal I_q, but it is not guaranteed that $I_{q^*} = I_{\theta^*}$.

Derivation of equation 8

In (Zhao et al., 2017), it is observed that equation 6 can be written as follows:

$$\begin{aligned} VIM_{\theta,\phi} = &- D_{KL}(p(x)||p_\theta(x)) - \mathbb{E}_{p(x)}[D_{KL}(q_\phi(z|x)||p_\theta(z|x))] - \\ &- (\lambda - 1)D_{KL}(q_\phi(z)||p(z)) + I_q(X,Z). \end{aligned}$$

From equation above, to verify that equation 8 is correct, it is sufficient to show that

$$I_{\theta,\phi}(X,Z) = I_q(X,Z) - \mathbb{E}_{p(x)}[D_{KL}(q_\phi(z|x)||p_\theta(z|x))]. \tag{14}$$

The equation 14 follows by the property of the autoencoder and equation 14.

More precisely, by equation 13 we have that

$$I_{\theta,\phi}(X,Z) = I_q(X,Z) - \mathbb{E}_{q(z)}D_{KL}(q_\phi(x|z)||p_\theta(x|z))$$

and, by AutoEncoder architecture, as observed in section 3, $p_\theta(z) = \mathbb{E}_{p(x)}[p_\theta(z|x)] = q(z)$. Then the following equation holds

$$\mathbb{E}_{q(z)}D_{KL}(q_\phi(x|z)||p_\theta(x|z)) = \mathbb{E}_{p(x)}D_{KL}(q_\phi(z|x)||p_\theta(z|x)).$$

Indeed,

$$\begin{aligned} &\mathbb{E}_{q(z)}D_{KL}(q_\phi(x|z)||p_\theta(x|z)) = \int q(z) \int q(x|z) \log \frac{q(x|z)}{p_\theta(x|z)} dxdz \\ &= \int\int q(x)q(z|x) \log \frac{q(x|z)q(z)}{p_\theta(x|z)(z)} dxdz = \int p(x) \int q(z|x) \log \frac{q(z|x)}{p_\theta(z|x)} dxdz = \\ &= \mathbb{E}_{p(x)}D_{KL}(q_\phi(z|x)||p_\theta(z|x)). \end{aligned}$$

Further Details on Experiments

In all the experiments in section 5.2 we considered the latent space $\mathcal{Z} = \mathbb{R}^d$, for all the models we choose the prior $p(z)$ to be a Gaussian with zero mean and identity covariance, only in VIMAE-l we choose the prior $p(z)$ to be a logistic with mean zero and identity variance. We choose $p_\theta(x|z)$ to be similar to DCGAN with batch normalization and $q_\phi(z|x)$ to be a convolutional deep neural network. The entire models are trained end to end by Adam (Kingma & Ba, 2014) with $\alpha = 10^{-3}, \beta_1 = 0.5, \beta_2 = 0.999$. We considered a deterministic decoder and we approximate the reconstruction loss with the L_2 loss, i.e. $\mathbb{E}_{p(x)}[\mathbb{E}_{q(z|x)}[-\log p_\theta(x|z)]] = \|x - x_g\|_2^2$, with x_g indicating the generated datum. In VIMAE case, while training we were adding a pixel-wise Gaussian noise truncated at 0.01 to all the images before feeding them to encoder, in order to make the encoder random. In VAE and $\beta-$VAE case, instead we used the standard reparameterization trick (Kingma & Welling, 2013).

In the following we describe the data-sets considered and the associated neural networks, we follow the same description given in (Tolstikhin et al., 2017) since we used the same neural nets.

MNIST and Omniglot

MNIST is a data-set containing 70k grey-scale handwritten digits and associated labels of resolution 28×28, subdivided in three subsets: train (50k), validation (10k) and test (10k).
Omniglot is a data-set containing 1623 different handwritten characters of resolution 28×28 from 50 different alphabets. Each of the 1623 characters was drawn online via Amazon's Mechanical Turk by 20 different people. In the same fashion as done in (Vinyals et al., 2016) we considered an augmented version where each character is rotated respectively by 90, 180, 270 degrees, in this way each character appears 80 times. The Omniglot data-set although has the same resolution of the MNIST, for this reason we use the same network, it is more challenging because, it is more entropic then MNIST, in fact the classes move from 10 to 1623 and the test classes are never seen in the training.

We choose $\mathcal{Z} = \mathbb{R}^8$, and $\beta = \lambda = 10$, we used mini-batches of size 100 and trained the model for 80 epochs. Both encoder and decoder used fully convolutional architectures with 4×4 convolutional filters.
Encoder:

$$\begin{aligned} x \in \mathbb{R}^{28\times 28} &\to \text{Conv}_{128} \to \text{BN} \to \text{ReLu} \\ &\to \text{Conv}_{256} \to \text{BN} \to \text{ReLu} \\ &\to \text{Conv}_{512} \to \text{BN} \to \text{ReLu} \\ &\to \text{Conv}_{1024} \to \text{BN} \to \text{ReLu} \end{aligned}$$

Decoder:

$$\begin{aligned} z \in \mathbb{R}^{8} &\to \text{FC}_{7\times 7\times 1024} \\ &\to \text{FSConv}_{512} \to \text{BN} \to \text{ReLu} \\ &\to \text{FSConv}_{256} \to \text{BN} \to \text{ReLu} \to \text{FSConv}_{1} \end{aligned}$$

Where Conv_k stands for a convolution with k filters, FSConv_k for the fractional strided convolution with k filters, BN for batch normalization, ReLU for the rectified linear units, and FC_k for the fully connected layer mapping to $\mathbb{R}^k$. All the convolutions in the encoder used vertical and horizontal strides 2 and `SAME` padding.

CelebA and CIFAR10

CelebA is a data-set with 202 599 faces images. We preprocessed the images by first taking a 140×140 center crops and then resizing to the 64×64 resolution and we consider the last 20k images as test subset.
CIFAR10 is a dataset consisting of of 60k 32×32 colour images in 10 classes, with 6k images per class. There are 50k training images and 10k test images.

For these data-sets we choose the same network with the same hyper-parameters, $\lambda = 100$ and $\mathcal{Z} = \mathbb{R}^{64}$. We used mini-batches of size 100 and trained the model for 60 epochs. Both encoder and decoder used a fully convolutional architectures with 5×5 convolutioanal filters.
Encoder:

$$\begin{aligned} x \in \mathbb{R}^{64\times 64\times 3} &\to \text{Conv}_{128} \to \text{BN} \to \text{ReLu} \\ &\to \text{Conv}_{256} \to \text{BN} \to \text{ReLu} \\ &\to \text{Conv}_{512} \to \text{BN} \to \text{ReLu} \\ &\to \text{Conv}_{1024} \to \text{BN} \to \text{ReLu} \end{aligned}$$

Decoder:

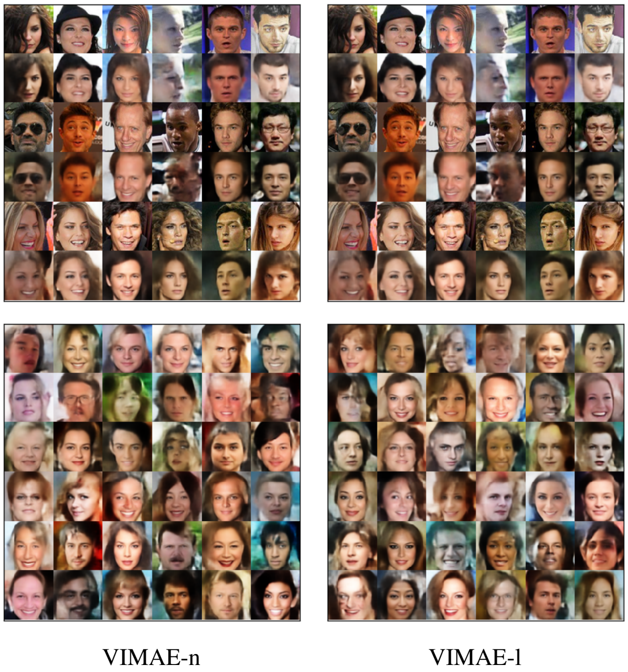
$$\begin{aligned} z \in \mathbb{R}^{64} &\to \text{FC}_{8\times 8\times 1024} \\ &\to \text{FSConv}_{512} \to \text{BN} \to \text{ReLu} \\ &\to \text{FSConv}_{256} \to \text{BN} \to \text{ReLu} \\ &\to \text{FSConv}_{128} \to \text{BN} \to \text{ReLu} \to \text{FSConv}_{1} \end{aligned}$$

CelebA Experiment

In section 5.2 we evaluate the generative performance qualitatively, on relatively simple grey-scale data-sets. In order to consider a more challenging data-set and quantitatively compare the generative performances of the trained models we evaluate the Frechet Inception Distance (FID) on CelebA based on 10^4 samples. From table 4, we observe, in agreement on what observed in section 5.2 and figure 4 that the difference between the two VIMAE models is minimal, instead it is big the difference with the ELBO based models (β-VAE is not listed in table 4, because it does not converge).

Table 4: FID scores for generated samples on CelebA (smaller is better)

Method	FID
VAE	82
VIMAE-l	56
VIMAE-n	55



VIMAE-n VIMAE-l

Figure 4: Test reconstruction (top) and random samples (bottom) of the two VIMAE models, $\lambda = 10$.