INFOVAEGAN: LEARNING JOINT INTERPRETABLE REPRESENTATIONS BY INFORMATION MAXIMIZATION AND MAXIMUM LIKELIHOOD

Fei Ye and Adrian G. Bors

Department of Computer Science, University of York, York YO10 5GH, UK

ABSTRACT
Learning disentangled and interpretable representations is an important step towards accomplishing comprehensive data representations on the manifold. In this paper, we propose a novel representation learning algorithm which combines the inference abilities of Variational Autoencoders (VAE) with the generalization capability of Generative Adversarial Networks (GAN). The proposed model, called InfoVAEGAN, consists of three networks: Encoder, Generator and Discriminator. InfoVAEGAN aims to jointly learn discrete and continuous interpretable representations in an unsupervised manner by using two different data-free log-likelihood functions onto the variables sampled from the generator’s distribution. We propose a two-stage algorithm for optimizing the inference network separately from the generator training. Moreover, we enforce the learning of interpretable representations through the maximization of the mutual information between the existing latent variables and those created through generative and inference processes.

Index Terms— Hybrid VAE-GAN generative models, Disentangled representations, Mutual information.

1. INTRODUCTION
Unsupervised disentangled representation learning is a challenging task in any machine learning application. Most studies consider disentangled representation to be a data decomposition into sets of statistically and syntactically independent variables. Such data sets are assumed to be semantically distinct and to represent different categories of data characteristics. Learning disentangled representations that may capture semantic meaningful information can allow to explicitly edit images and is useful for a variety of tasks [1, 2, 3]. Enabling disentangled representations can overcome overfitting during the training, leading to better generalization in models, [4].

One of the most popular generative models is the Variational Autoencoder (VAE). VAEs [5] aim to maximize a lower bound on the marginal log-likelihood of the data:

\[
\mathcal{L}(\phi, \theta) = \mathbb{E}_{q_\phi(z|x)}[\log p_\phi(x|z)] - D_{KL}(q_\phi(z|x) || p(z)) \leq \log p(x)
\]

where \(x\) and \(z\) are the input data and the corresponding latent variables, and the conditional distributions \(q_\phi(z|x)\) and \(p_\phi(x|z)\), are implemented by the Encoder and Decoder networks, of parameters \(\theta\) and \(\phi\), respectively. These networks are trained using the Stochastic Gradient Descent (SGD) algorithm.

Generative adversarial networks (GAN). GANs also consist of two network components: Generator and Discriminator which are trained for playing a Minimax game, defined by the following loss:

\[
\min_G \max_D V(D,G) = \mathbb{E}_{x \sim p_d(x)}[\log D(x)] + \mathbb{E}_{z \sim p(z)}[\log(1 - D(G(z)))]
\]

2. BACKGROUND AND RELATED WORKS

Variational autoencoder (VAE). VAEs [5] aim to maximize a lower bound on the marginal log-likelihood of the data:

\[
\mathcal{L}(\phi, \theta) = \mathbb{E}_{q_\phi(z|x)}[\log p_\phi(x|z)] - D_{KL}(q_\phi(z|x) || p(z)) \leq \log p(x)
\]

where \(x\) and \(z\) are the input data and the corresponding latent variables, and the conditional distributions \(q_\phi(z|x)\) and \(p_\phi(x|z)\), are implemented by the Encoder and Decoder networks, of parameters \(\theta\) and \(\phi\), respectively. These networks are trained using the Stochastic Gradient Descent (SGD) algorithm.

Generative adversarial networks (GAN). GANs also consist of two network components: Generator and Discriminator which are trained for playing a Minimax game, defined by the following loss:

\[
\min_G \max_D V(D,G) = \mathbb{E}_{x \sim p_d(x)}[\log D(x)] + \mathbb{E}_{z \sim p(z)}[\log(1 - D(G(z)))]
\]
While the discriminator network is trained to distinguish between real and fake data, the generator aims to produce more realistic data that can fool the discriminator. GANs are challenging to control and may generate unexpected results.

**Hybrid models.** Hybrid models attempt to address the drawbacks of GANs and VAEs, by combining their architectures. These models usually have three components: an Encoder for mapping data into the latent space, a Generator to recover data from the latent space, and a Discriminator to distinguish real from fake data. Adversarial learning can be performed in the data space, latent space [2], or on their joint spaces.

Lately, the likelihood estimation as a regularization term from the empirical data distribution \( P \) disentangled representations. Our paper is the first to propose an appropriate objective function for training a hybrid VAE-GAN method for learning both continuous and discrete disentangled representations.

### 3. THE INFOVAEGAN MODEL

The proposed InfoVAEGAN model is made up of three networks: Encoder, Generator and Discriminator.

#### 3.1. Generation from prior distributions

Let \( x \in \mathbb{R}^d \) represent the observed random variables sampled from the empirical data distribution \( P_x \). One of the goals of our model is to train the Generator to approximate the true data distribution \( P_x \). Let us assume three underlying generative factors \( z, c, d \), corresponding to random, continuous and discrete variables, which are sampled from three independent prior distributions \( z \sim \mathcal{N}(\mathbf{1}, \mathbf{S}_z), c \sim \mathcal{N}(\mathbf{1}, \mathbf{S}_c), d \sim \text{Cat}(k = K, p = 1/K) \), where \( \text{Cat} \) denotes the Categorical distribution and \( \mathcal{N} \) is the Gaussian distribution. Let us consider that the data generated \( x' \) is produced by a generator \( G_{\psi}(z, d, c) \), implemented by a neural network with trainable parameters \( \psi \), and \( P_G \) to represent the distribution of data generated by \( G \). The generation process is defined as: 

\[
d \sim p(d), \quad z \sim p(z), \quad c \sim p(c), \quad x \sim p_{\psi}(x|z, d, c).
\]

For the Discriminator network we use the Earth-mover distance, as in the Wasserstein GAN (WGAN) model [10], which is defined as the optimal path of transporting information mass from the generator distribution \( P_G \) to the data distribution \( P_x \). By considering the Kantorovich-Rubinstein duality [12], the optimal transport adversarial learning is defined as:

\[
\min_G \max_D \mathbb{E}_{x \sim P_x}[D(x)] - \mathbb{E}_{x' \sim P_G}[D(x')]
\]

where \( \Theta \) represents a set of 1-Lipschitz functions. We introduce a gradient penalty term [13], to enforce the Lipschitz constraint, resulting in:

\[
\min_G \max_D \mathbb{E}_{x \sim P_x}[D(x)] - \mathbb{E}_{x' \sim P_G}[D(x')] + \lambda \mathbb{E}_{x \sim P_x}[\|\nabla_x D(x)\|^2 - 1]^2,
\]

where \( P_x \) is defined as sampling uniformly along straight lines between pairs of data sampled from \( P_x \) and \( P_G \).

#### 3.2. Data-free log-likelihood optimization

In this section, we introduce two data-free log-likelihood optimization functions, which are used to learn the disentangled latent representations \( z \) and \( c = (d, c) \), respectively. Instead of maximizing the sample log-likelihood, as commonly used in the VAE framework [5], we optimize the log-likelihood function by deriving a lower bound over the data samples drawn from the generator distribution.

**Definition 1** Let \( x' \sim G(z, d, c) \) be the generated data where \( z, d, c \) are latent variables sampled from the prior distributions \( p(z), p(d), p(c) \).

**Definition 2** Let \( q_{\omega}(d, c|x), q_{\xi}(z|x) \) represent two independent conditional distributions implemented by two inference models. Let us define \((d, c)\) as interpretable representations which model discrete and continuous meaningful variations of the data and \( z, x' \) as the observed variables. Let us define a latent variable model \( p_{\psi}(x'|z, d, c) = p_{\psi}(x'|z, d, c)p(d, c)p(z) \). Then, the log-likelihood of \( p_{\psi}(x') \) is defined as:

\[
\log p_{\psi}(x') = \int \int \int \log p_{\psi}(x'|d, c, z) p(d, c) p(z) \, dd \, dc \, dz.
\]

This expression is intractable and can be rewritten by considering its Evidence Lower Bound (ELBO), as:

\[
\log p_{\psi}(x') \geq \mathbb{E}_{q_{\omega}(d, c|x')} \left[ \log \frac{p_{\psi}(x', d, c, z)}{q_{\xi}(z|x')q_{\omega}(d, c|x')} \right].
\]

The scheme for optimizing both \( q_{\omega}(d, c|x') \) and \( q_{\xi}(z|x') \), without updating the Generator, is very efficient.

### 4. THE THEORETICAL FRAMEWORK

In existing hybrid methods, the inference model and the generator network are trained jointly by using a single objective function. However, in the proposed InfoVAEGAN model we have independent optimization procedures for the inference and generation. This choice has many advantages. For instance, the training of the inference model implemented by the Encoder, does not interfere with the optimization of the Generator, which results in a stable training procedure. When the Generator approximates the true data distribution exactly, the training of the inference model implemented by the Encoder, does not interfere with the optimization of the Generator, which results in a stable training procedure. When the Generator approximates the true data distribution exactly, the training of the inference model implemented by the Encoder, does not interfere with the optimization of the Generator, which results in a stable training procedure. When the Generator approximates the true data distribution exactly, the training of the inference model implemented by the Encoder, does not interfere with the optimization of the Generator, which results in a stable training procedure.
Proposition 1: For a given inference model, we can estimate the testing data log-likelihood:

\[
\log p_\psi(x_i) \geq \mathbb{E}_{q_{\omega,c}(d,c,z|x_i)} \left[ \log p_\psi(d, c, z|x_i) \right] - D_{KL}(q_\omega(d|x_i)||p(d)) - D_{KL}(q_c(c|x_i)||p(c)) 
\]

where \( x_i \) represent testing data. The model implementing \( p_\psi(x_i) \) combines the two inference models and a Generator.

Proof 1: We combine the two inference models for continuous and discrete variables, and a Generator into a single model:

\[
\log p_\psi(x_i) = \log p_\psi(x_i|d, c, z) q_\omega(d, c|x) q_\xi(z|x) 
\]

Then we define the model log-likelihood as:

\[
\log p_\psi(x_i) = \mathbb{E}_{q_{\omega,c}(d,c,z|x_i)} \left[ \frac{p_\psi(x_i, d, c, z)}{q_{\omega,c}(d, c, z|x_i)} \right] 
\]

According to the Jensen inequality, we have:

\[
\log p_\psi(x_i) \geq \mathbb{E}_{q_{\omega,c}(d,c,z|x_i)} \left[ \log \frac{p_\psi(x_i, d, c, z)}{q_{\omega,c}(d, c, z|x_i)} \right] 
\]

\[
= \mathbb{E}_{q_{\omega,c}(d,c,z|x_i)} \left[ \log \frac{p_\psi(x_i, d, c, z)}{q_{\omega,c}(d, c, z|x_i)} \right] - D_{KL}(q_\omega(d|x_i)||p(d)) - D_{KL}(q_c(c|x_i)||p(c)) 
\]

\[
\geq \mathbb{E}_{q_{\omega,c}(d,c,z|x_i)} \left[ \log \frac{p_\psi(x_i, d, c, z)}{q_{\omega,c}(d, c, z|x_i)} \right] - D_{KL}(q_\omega(d|x_i)||p(d)) - D_{KL}(q_c(c|x_i)||p(c)) 
\]

\[
(10) 
\]

5. MUTUAL INFORMATION MAXIMIZATION FOR INTERPRETABLE REPRESENTATIONS

In the proposed InfoVAEGAN model, we transfer the underlying characteristic information of continuous and discrete latent variables during the decoder-generation process by using the Mutual Information (MI) maximization. Let us denote the joint latent variables by \( u = (d, c) \), while we want to maximize the MI between the joint latent variable \( u \) and the decoder output, \( I(u, G(z, u)) \). According to the research study from [14] it is difficult to optimize the mutual information directly, given that it needs to access the information represented by the true posterior \( p(u|x) \). In order to address this problem, we define an auxiliary distribution \( W(u|x) \) to approximate the true posterior and then derive a lower bound on the mutual information, expressed by using the marginal entropy \( H(u) \), and the conditional entropy \( H(u|G(z, u)) \):

\[
I(u, G(z, u)) = H(u) - H(u|G(z, u)) 
\]

\[
= \int \int G(z, u) p(u|x) \log \frac{p(u|x)}{W(u|x)} \, dx \, du 
\]

\[
+ \int \int G(z, u) p(u|x) \log W(u|x) \, dx \, du + H(u) 
\]

\[
= \mathbb{E}_{x \sim G(z, u)} D_{KL}[p(u|x)||W(u|x)] 
\]

\[
+ \mathbb{E}_{x \sim G(z, u)} [\mathbb{E}_{u \sim p(u, x)} \log W(u|x)] + H(u) 
\]

\[
\geq \mathbb{E}_{x \sim G(z, u)} \mathbb{E}_{u \sim p(u, x)} [\log W(u|x)] + H(u) = \mathcal{L}_{MI} 
\]

(11)

where the auxiliary distribution \( W(u|x) \) is implemented by the Encoder. In practice, we sample a pair of latent variables \( d, c \) from \( q_{\omega}(d, c|x) \). We estimate the mutual information by means of the lower bound \( \mathcal{L}_{MI} \), from (11), while the last term \( H(u) \) represents the marginal entropy of the latent variables.

The graph structure of the InfoVAEGAN is shown in Fig. 1, where \( q_{\omega}(d|x) \) and \( q_{\omega}(c|x) \) are implemented by the same network except for the last layer which is different for the inference of each latent variable. The inference network, representing \( q_\xi(z|x) \), is implemented by a neural network with trainable parameters \( \xi \), as it can be seen in the lower part of the left side of Fig. 1b. The Generator is shown in Fig. 1a.

Fig. 2. Reconstruction results on each row: real images, reconstructions by ALI [15], InfoGAN [11] and InfoVAEGAN.

6. EXPERIMENTAL RESULTS

In the following we evaluate the performance of InfoVAEGAN on the MNIST dataset [16], representing images of handwritten digits. In order to learn the discrete latent variable which captures different styles of handwritten digits we use a categorical vector sampled from \( Cat(K = 10, p = 0.1) \) and two continuous variables, sampled from the uniform distribution \( U(-1, 1) \), as latent variables. The reconstruction results for the images from MNIST, shown in the first row from Fig. 2, by ALI [15], InfoGAN [11], and InfoVAEGAN, are provided in the following rows of images, respectively. For the proposed InfoVAEGAN approach, the discrete latent variables are sampled from the Gumbel-softmax distribution, while the continuous latent variables are sampled from the Gaussian distribution, whose mean and diagonal covariance are parameterized by the Encoder. From these results it can be observed that InfoVAEGAN provides better digit image reconstructions than InfoGAN or ALI.
We modify the continuous codes $c_1$, $c_2$ within the range $[-1, 1]$ and fix the other latent variables. The generative results for MNIST dataset are shown in Figures 4a and 4c for InfoVAEGAN, while for InfoGAN are provided in Figures 4b and 4d, when modifying $c_1$ and $c_2$. It can be observed that by varying the latent codes in InfoVAEGAN, we generate images showing meaningful characteristics such as rotations or a variety of handwriting styles. We also consider a 10-dimensional vector for the discrete and continuous latent variables in order to model underlying changing factors in the CelebA dataset [17]. We change a single latent variable in the images generated by InfoVAEGAN while fixing the others. The results shown in Figures 3a-d indicate variations in face image representations such as bangs, glasses, hair colour and in smiling.

The results when using InfoVAEGAN in unsupervised classification on the MNIST dataset, when compared with other methods, are provided in Table 1. Most unsupervised learning methods adopt mixture deep learning models ($K$ represents the number of components) requiring significantly more parameters. We can observe that InfoVAEGAN achieves higher accuracy than InfoGAN [11], and other models.

We investigate the disentanglement ability of the proposed approach by using the metric from [7] and the dataset dSprites [25]. The results are reported in Table 2, where all other results are cited from [18]. The proposed approach achieves a competitive disentanglement score when compared with the current state of the art. We also use the Fréchet Inception Distance (FID) [26] to evaluate the quality of the generated images when considering the CelebA dataset in Fig. 3e, where InfoVAEGAN-MI denotes that the proposed approach does not use the mutual information (MI) loss. These results show that the proposed approach can balance well the disentanglement ability and image generation quality.

7. CONCLUSION

In this paper, we introduce InfoVAEGAN, a new deep learning approach for learning jointly discrete and continuous interpretable representations. InfoVAEGAN optimizes separately the inference model and the generator providing advantages over other hybrid methods. The proposed approach is a good tool to provide inference mechanisms when considering any generative GAN model without the need of any real data. In addition, InfoVAEGAN can generate high-quality interpretable data variations which can successfully be used for disentangled and interpretable representation learning.
8. REFERENCES


