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 InfoVAE-GAN, 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 Autencoder [5], which implements a mapping between the data and an estimated latent space. The VAE's loss function maximizes the lower bound on the marginal log-likelihood of the data, while accurately reconstructing the data from the mapping of the latent space using the Kullback-Leibler (KL) divergence. Learning interpretable and disentangled representations have been considered in $\beta-$VAE [6] by setting a large penalty on the KL divergence term in order to encourage the independence between latent variables. On the other hand β-VAE sacrifices the quality of data reconstruction when inducing disentangled representations, [7]. β-TCVAE model introduced the usage of the total correlation (TC) penalty, which is a measure of multivariate mutual independence. The TC penalty was used in various VAE frameworks [8] for inducing disentangled representations. However, TC is biased and is zero only if estimated on the whole dataset, [9]. Meanwhile, reducing the bias to zero is impossible for a large-scale dataset. The drawback of VAE based approaches is that they generally produce blurred and unclear images when compared to Generative Adversarial Networks (GANs) [10]. Few research efforts have been devoted to use GANs for disentangled representations [11], and with mixed results.

This research study has the following contributions :

1) A novel two-stage training algorithm where the inference model is estimated separately from the generator.
2) A data-free log-likelihood optimization approach able to learn an accurate inference model from a GAN.

2. BACKGROUND AND RELATED WORKS

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

$$\begin{aligned}\mathcal{L}(\phi,\theta) =& \mathbb{E}_{q_\theta(\mathbf{z}|\mathbf{x})}[\log p_\phi(\mathbf{x}|\mathbf{z})] - D_{KL}(q_\theta(\mathbf{z}|\mathbf{x})||p(\mathbf{z})) \\ &\leq \log p(\mathbf{x})\end{aligned} \tag{1}$$

where $\mathbf{x}$ and $\mathbf{z}$ are the input data and the corresponding latent variables, and the conditional distributions $q_\theta(\mathbf{z}|\mathbf{x})$ and $p_\phi(\mathbf{x}|\mathbf{z})$, are implemented by the Encoder and Decoder networks, of parameters θ and ϕ, 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:

$$\begin{aligned}\min_G \max_D V(D,G) = &\mathbb{E}_{\mathbf{x}\sim\, pd(\mathbf{x})}[\log D(\mathbf{x})] \\ &+ \mathbb{E}_{\mathbf{z}\sim\, p(\mathbf{z})}[\log[1 - D(G(\mathbf{z}))]].\end{aligned} \tag{2}$$

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 was shown to stabilize adversarial distribution matching [3]. However, these methods only focus on improving the generation capability and do not design suitable objective functions for inducing 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 $\mathbf{x} \in \mathbb{R}^d$ represent the observed random variables sampled from the empirical data distribution $\mathbb{P}_{\mathbf{x}}$. One of the goals of our model is to train the Generator to approximate the true data distribution $\mathbb{P}_{\mathbf{x}}$. Let us assume three underlying generative factors $\mathbf{z}, \mathbf{c}, \mathbf{d}$, corresponding to random, continuous and discrete variables, which are sampled from three independent prior distributions $\mathbf{z} \sim \mathcal{N}(\mathbf{I}^z, \mathbf{\Sigma}^z)$, $\mathbf{c} \sim \mathcal{N}(\mathbf{I}^c, \mathbf{\Sigma}^c)$, $\mathbf{d} \sim Cat(k = K, p = 1/K)$, where Cat denotes the Categorical distribution and $\mathcal{N}$ is he Gaussian distribution. Let us consider that the data generated $\mathbf{x}'$ is produced by a generator $G_\psi(\mathbf{z}, \mathbf{d}, \mathbf{c})$, implemented by a neural network with trainable parameters ψ, and $\mathbb{P}_G$ to represent the distribution of data generated by G. The generation process is defined as: $\mathbf{d} \sim p(\mathbf{d}),\ \mathbf{z} \sim p(\mathbf{z}), c \sim p(\mathbf{c}), \mathbf{x} \sim p_\psi(\mathbf{x}|\mathbf{z}, \mathbf{d}, \mathbf{c})$.

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 $\mathbb{P}_G$ to the data distribution $\mathbb{P}_{\mathbf{x}}$. By considering the Kantorovich-Rubinstein duality [12], the optimal transport adversarial learning is defined as:

$$\min_G \max_{D \in \Theta} \mathbb{E}_{\mathbf{x} \sim \mathbb{P}_{\mathbf{x}}}[D(\mathbf{x})] - \mathbb{E}_{\mathbf{x}' \sim \mathbb{P}_G} D(\mathbf{x}')] \tag{3}$$

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

$$\begin{aligned} &\min_G \max_D \mathbb{E}_{\mathbf{x} \sim \mathbb{P}_{\mathbf{x}}}[D(\mathbf{x})] - \mathbb{E}_{\mathbf{x}' \sim \mathbb{P}_G}[D(\mathbf{x}')] \\ &+ \lambda \mathbb{E}_{\tilde{\mathbf{x}} \sim \mathbb{P}_{\tilde{\mathbf{x}}}}[(\|\nabla_{\tilde{x}} D(\tilde{\mathbf{x}})\|_2 - 1)^2], \end{aligned} \tag{4}$$

where $\mathbb{P}_{\tilde{\mathbf{x}}}$ is defined as sampling uniformly along straight lines between pairs of data sampled from $\mathbb{P}_{\mathbf{x}}$ and $\mathbb{P}_G$.

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 $\mathbf{z}$ and $\mathbf{u} = (\mathbf{d}, \mathbf{c})$, 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 $\mathbf{x}' \sim G(\tilde{\mathbf{z}}, \tilde{\mathbf{d}}, \tilde{\mathbf{c}})$ be the generated data where $\tilde{\mathbf{z}}, \tilde{\mathbf{d}}, \tilde{\mathbf{c}}$ are latent variables sampled from the prior distributions $p(\tilde{\mathbf{z}})$, $p(\tilde{\mathbf{d}})$, $p(\tilde{\mathbf{c}})$.*

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

$$\log p_\psi(\mathbf{x}') = \iiint \log p_\psi(\mathbf{x}'|\mathbf{d}, \mathbf{c}, \tilde{\mathbf{z}})\, p(\mathbf{d}, \mathbf{c})\, p(\tilde{\mathbf{z}})\, d\mathbf{d}\, d\mathbf{c}\, d\tilde{\mathbf{z}}. \tag{5}$$

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

$$\log p_\psi(\mathbf{x}') \geq \mathbb{E}_{q_{\omega,\xi}(\mathbf{d},\mathbf{c},\mathbf{z}|\mathbf{x}')} \left[\log \frac{p_\psi(\mathbf{x}', \tilde{\mathbf{d}}, \tilde{\mathbf{c}}, \mathbf{z})}{q_\xi(\mathbf{z}|\mathbf{x}') q_\omega(\mathbf{d}, \mathbf{c}|\mathbf{x}')} \right]. \tag{6}$$

The scheme for optimizing both $q_\omega(\mathbf{d}, \mathbf{c}|\mathbf{x}')$ and $q_\xi(\mathbf{z}|\mathbf{x}')$, 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, we can derive more accurate inference models. Aligning two joint distributions by using adversarial learning would also be harder to achieve than matching two single distributions individually. Unlike in InfoGAN [11], the proposed model has a full inference mechanism, which enables the inference of both meaningful and nuisance latent representations, benefiting many down-stream tasks such as data reconstructions and interpolations.

Proposition 1 *For a given inference model, we can estimate the testing data log-likelihood :*

$$\begin{aligned}&\log p_\psi(\mathbf{x}_t) \geq \mathbb{E}_{q_{\omega,\xi}(\mathbf{d},\mathbf{c},\mathbf{z}|\mathbf{x}_t)}[\log p_\psi(\mathbf{d},\mathbf{c},\mathbf{z}|\mathbf{x}_t)]\\&- D_{KL}(q_\omega(\mathbf{d}|\mathbf{x}_t)||p(\mathbf{d})) - D_{KL}(q_\omega(\mathbf{c}|\mathbf{x}_t)||p(\mathbf{c}))\\&- D_{KL}(q_\xi(\mathbf{z}|\mathbf{x}_t)||p(\mathbf{z})) = \mathcal{L}(\psi,\xi,\omega;\mathbf{x}_t)\end{aligned} \tag{7}$$

where $\mathbf{x}_t$ represent testing data. The model implementing $p_\psi(\mathbf{x}_t)$ 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:*

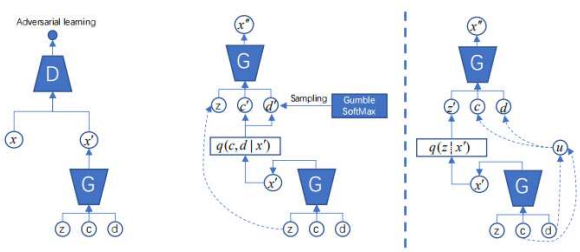
$$\log p_\psi(\mathbf{x}_t) = p_\psi(\mathbf{x}_t|\mathbf{d},\mathbf{c},\mathbf{z})q_\omega(\mathbf{d},\mathbf{c}|\mathbf{x}_t)q_\xi(\mathbf{z}|\mathbf{x}_t) \tag{8}$$

Then we define the model log-likelihood as :

$$\log p_\psi(\mathbf{x}_t) = \log \mathbb{E}_{q_{\omega,\xi}(\mathbf{d},\mathbf{c},\mathbf{z}|\mathbf{x}_t)}\left[\frac{p_\psi(\mathbf{x}_t,\mathbf{d},\mathbf{c},\mathbf{z})}{q_{\omega,\xi}(\mathbf{d},\mathbf{c},\mathbf{z}|\mathbf{x}_t}\right] \tag{9}$$

According to the Jensen inequality, we have :

$$\begin{aligned}&\log p_\psi(\mathbf{x}_t) \geq \mathbb{E}_{q_{\omega,\xi}(\mathbf{d},\mathbf{c},\mathbf{z}|\mathbf{x}_t)}\left[\log\frac{p_\psi(\mathbf{x}_t,\mathbf{d},\mathbf{c},\mathbf{z})}{q_{\omega,\xi}(\mathbf{d},\mathbf{c},\mathbf{z}|\mathbf{x}_t)}\right]\\&= \mathbb{E}_{q_{\omega,\xi}(\mathbf{d},\mathbf{c},\mathbf{z}|\mathbf{x}_t)}\left[\log\frac{p_\psi(\mathbf{d},\mathbf{c},\mathbf{z}|\mathbf{x}_t)p(\mathbf{d})p(\mathbf{c})p(\mathbf{z})}{q_\omega(\mathbf{d}|\mathbf{x}_t)q_\omega(\mathbf{c}|\mathbf{x}_t)q_\xi(\mathbf{z}|\mathbf{x}_t)}\right]\\&= \mathbb{E}_{q_{\omega,\xi}(\mathbf{d},\mathbf{c},\mathbf{z}|\mathbf{x}_t)}[\log p_\psi(\mathbf{d},\mathbf{c},\mathbf{z}|\mathbf{x}_t)] - D_{KL}(q_\omega(\mathbf{d}|\mathbf{x}_t)||p(\mathbf{d}))\\&- D_{KL}(q_\omega(\mathbf{c}|\mathbf{x}_t)||p(\mathbf{c})) - D_{KL}(q_\xi(\mathbf{z}|\mathbf{x}_t)||p(\mathbf{z})).\end{aligned} \tag{10}$$



(a) Generator. (b) Inference models.

Fig. 1. Unsupervised learning structures in generative models, where $\mathbf{c}$ and $\mathbf{d}$ are continuous and discrete variables, while $\mathbf{z}$ represents Gaussian noise.

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 $\mathbf{u} = (\mathbf{d},\mathbf{c})$, while we want to maximize the MI between the joint latent variable $\mathbf{u}$ and the decoder output, $\mathrm{I}(\mathbf{u},\mathrm{G}(\mathbf{z},\mathbf{u}))$. 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 $p(\mathbf{u}|\mathbf{x})$. In order to address this problem, we define an auxiliary distribution $W(\mathbf{u}|\mathbf{x})$ to approximate the true posterior and then derive a lower bound on the mutual information, expressed by using the marginal entropy $H(\mathbf{u})$, and the conditional entropy $H(\mathbf{u}|G(\mathbf{z},\mathbf{u}))$:

$$\begin{aligned}&\mathrm{I}(\mathbf{u},G(\mathbf{z},\mathbf{u})) = H(\mathbf{u}) - H(\mathbf{u}|G(\mathbf{z},\mathbf{u}))\\&= \iint G(\mathbf{z},\mathbf{u})p(\mathbf{u}|\mathbf{x})\log\frac{p(\mathbf{u}|\mathbf{x})}{W(\mathbf{u}|\mathbf{x})}\,d\mathbf{x}d\mathbf{u}\\&\quad+\iint G(\mathbf{z},\mathbf{u})p(\mathbf{u}|\mathbf{x})\log W(\mathbf{u}|\mathbf{x})d\mathbf{x}d\mathbf{u} + H(\mathbf{u})\\&= \mathbb{E}_{\mathbf{x}\sim G(\mathbf{z},\mathbf{u})}D_{KL}[p(\mathbf{u}|\mathbf{x})||W(\mathbf{u}|\mathbf{x})]\\&\quad+\mathbb{E}_{\mathbf{x}\sim G(\mathbf{z},\mathbf{u})}[\mathbb{E}_{\mathbf{u}\sim p(\mathbf{u},\mathbf{x})}[\log W(\mathbf{u}|\mathbf{x})]] + H(\mathbf{u})\\&\geqslant \mathbb{E}_{\mathbf{x}\sim G(\mathbf{z},\mathbf{u})}[\mathbb{E}_{\mathbf{u}\sim p(\mathbf{u},\mathbf{x})}[\log W(\mathbf{u}|\mathbf{x})]] + H(\mathbf{u}) = \mathcal{L}_{MI}\end{aligned} \tag{11}$$

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

The graph structure of the InfoVAEGAN is shown in Fig. 1, where $q_\omega(\mathbf{d}|\mathbf{x})$ and $q_\omega(\mathbf{c}|\mathbf{x})$ 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 $q_\xi(\mathbf{z}|\mathbf{x})$, is implemented by a neural network with trainable parameters ξ, as it can be seen in the lower part of the left side of Fig. 1b. The Generator is shown in Fig. 1a.

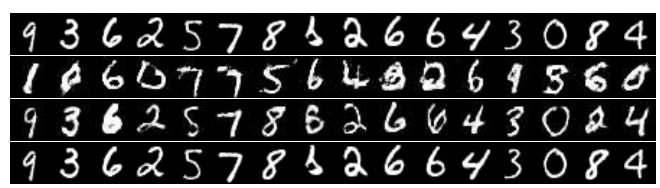
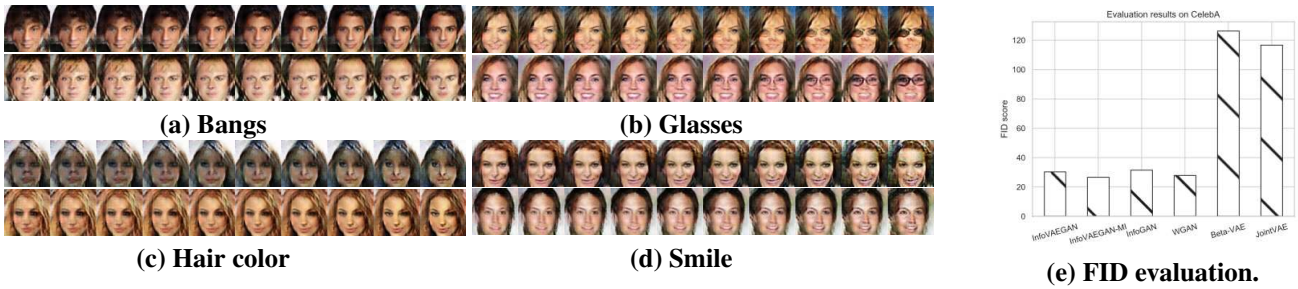


Fig. 2. Reconstruction results on each row: real images, reconstructions by ALI [15], InfoGAN [11] and InfoVAEGAN.

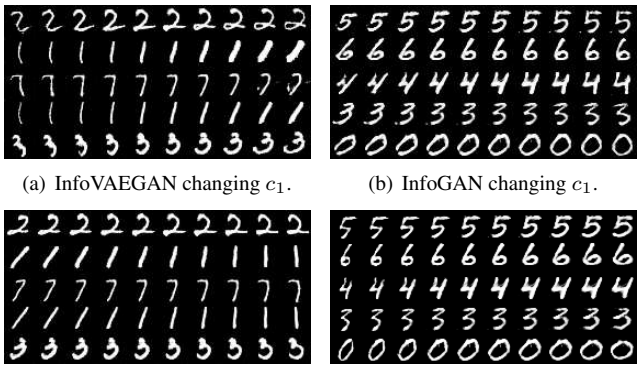
6. EXPERIMENTAL RESULTS

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 Gumble-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.



(a) Bangs **(b) Glasses** **(c) Hair color** **(d) Smile** **(e) FID evaluation.**

Fig. 3. We change a single latent variable in the latent space from -1 to 1 while fixing all other latent variables for CelebA dataset in (a)-(d). FID evaluation when using CelebA database for training is provided in (e).



(a) InfoVAEGAN changing c_1. (b) InfoGAN changing c_1. (c) InfoVAEGAN changing c_2. (d) InfoGAN changing c_2.

Fig. 4. Generation results when changing the continuous variables c_1 and c_2 from -1 to 1.

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.

Table 1. Unsupervised classification results for M runs.

MNIST				
Method	K	M	Mean	Best
InfoVAEGAN	1	4	95.42	96.15
JointVAE [18]	1	4	71.53	87.32
SubGAN [19]	20	1	/	90.81
InfoGAN [11]	1	1	/	93.35
GMVAE [20]	30	1	/	89.27
GMVAE [20]	16	1	/	87.82
AAE [21]	16	1	/	90.45
CatGAN [22]	30	1	/	95.73
DEC [23]	10	1	/	84.30
PixelGAN [24]	30	1	/	94.73

Table 2. Disentanglement evaluation on the dSprites.

Methods	M	Score
InfoVAEGAN	10	0.79
Beta-VAE [6]	10	0.73
FactorVAE [7]	10	0.82
JointVAE [18]	10	0.69

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

[1] Fei Ye and Adrian G. Bors, "Deep mixture generative autoencoders," *IEEE Transactions on Neural Networks and Learning Systems*, pp. 1–15, 2021.

[2] Fei Ye and Adrian G Bors, "Learning joint latent representations based on information maximization," *Information Sciences*, vol. 567, no. 8, pp. 216–236, 2021.

[3] Fei Ye and Adrian G Bors, "Learning latent representations across multiple data domains using lifelong vaegan," in *Proc. European Conference on Computer Vision (ECCV), vol. LNCS 12365*, 2020, pp. 777–795.

[4] A. Alemi, I. Fischer, J. Dillon, and K. Murphy, "Deep variational information bottleneck," *Proc. Int. Conf. of Learning Representation (ICLR), arXiv preprint arXiv:1612.00410*, 2017.

[5] D. P Kingma and M. Welling, "Auto-encoding variational Bayes," *arXiv preprint arXiv:1312.6114*, 2013.

[6] I. Higgins, L. Matthey, A. Pal, C. Burgess, X. Glorot, M. Botvinick, S. Mohamed, and A. Lerchner, "β-VAE: Learning basic visual concepts with a constrained variational framework," in *Proc. Int. Conf. on Learning Representations (ICLR)*, 2017.

[7] H. Kim and A. Mnih, "Disentangling by factorising," *Proc. Int. Conf. on Machine Learning (ICML)*, vol. PMLR 80, pp. 2649–2658, 2018.

[8] S. Gao, R. Brekelmans, G. Ver Steeg, and A. Galstyan, "Auto-encoding total correlation explanation," in *Proc. of Int. Conf. on Artificial Intelligence and Statistics (AISTATS), vol. PMLR 89*, 2019, pp. 1157–1166.

[9] R. Lopez, J. Regier, M. Jordan, and N. Yosef, "Information constraints on auto-encoding variational Bayes," in *Adv in Neur Inf Proc Sys (NIPS)*, 2018, pp. 6117–6128.

[10] M. Arjovsky, S. Chintala, and L. Bottou, "Wasserstein generative adversarial networks," in *Proc. Int. Conf. on Machine Learning (ICML), vol. PMLR 70*, 2017, pp. 214–223.

[11] X. Chen, Y. Duan, R. Houthooft, J. Schulman, I. Sutskever, and P. Abbeel, "InfoGAN: Interpretable representation learning by information maximizing generative adversarial nets," in *Advances in Neural Inf. Proc. Systems (NIPS)*, 2016, pp. 2172–2180.

[12] C. Villani, *Optimal transport: Old and New*, Springer, 2008.

[13] I. Gulrajani, F. Ahmed, M. Arjovsky, V. Dumoulin, and A. C. Courville, "Improved training of Wasserstein GANs," in *Advances in Neural Inf. Proc. Systems (NIPS)*, 2017, pp. 5767–5777.

[14] Fei Ye and Adrian G. Bors, "Lifelong learning of interpretable image representations," in *Proc. Int. Conf. on Image Processing Theory, Tools and Applications (IPTA)*, 2020, pp. 1–6.

[15] V. Dumoulin, I. Belghazi, B. Poole, O. Mastropietro, A. Lamb, M. Arjovsky, and A. Courville, "Adversarially learned inference," *Proc. Int. Conf. on Learning Rep. (ICLR), arXiv preprint arXiv:1606.00704*, 2017.

[16] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recog.," *Proc. of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.

[17] Z. Liu, P. Luo, X. Wang, and X. Tang, "Deep learning face attributes in the wild," in *Proc. of IEEE Int. Conf. on Computer Vision (ICCV)*, 2015, pp. 3730–3738.

[18] E. Dupont, "Learning disentangled joint continuous and discrete representations," in *Advances in Neural Inf. Proc. Systems (NIPS)*, 2018, pp. 710–720.

[19] J. Liang, J. Yang, H.-Y. Lee, K. Wang, and M.-H. Yang, "Sub-GAN: An unsupervised generative model via subspaces," in *Proc. of the European Conf. on Computer Vision (ECCV), vol. LNCS 11215*, 2018, pp. 698–714.

[20] N. Dilokthanakul, P. Mediano, M. Garnelo, M. Lee, H. Salimbeni, K. Arulkumaran, and M. Shanahan, "Deep unsupervised clustering with Gaussian mixture variational autoencoders," *arXiv preprint arXiv:1611.02648*, 2016.

[21] A. Makhzani, J. Shlens, N. Jaitly, I. Goodfellow, and B. Frey, "Adversarial autoencoders," in *Proc. ICLR-workshop, arXiv preprint arXiv:1511.05644*, 2016.

[22] J. T. Springenberg, "Unsupervised and semi-supervised learning with categorical generative adversarial networks," *Proc. Int. Conf. on Learning Representations (ICLR), arXiv preprint arXiv:1511.06390*, 2016.

[23] J. Xie, R. Girshick, and A. Farhadi, "Unsupervised deep embedding for clustering analysis," *Proc. Int. Conf. on Machine Learning (ICML), arXiv preprint arXiv:1511.06335*, vol. PMLR 48, pp. 478–487, 2016.

[24] A. Makhzani and B. Frey, "PixelGAN autoencoders," in *Adv in Neur Inf Proc Sys (NIPS)*, 2017, pp. 1972–1982.

[25] L. Matthey, I. Higgins, D. Hassabis, and A. Lerchner, "dSprites: Disentanglement testing Sprites dataset, https://github.com/deepmind/dsprites-dataset/," 2017.

[26] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter, "GANs trained by a two time-scale update rule converge to a local Nash equilibrium," in *Adv. in Neural Inf. Proc. Syst. (NIPS)*, 2017, pp. 6626–6637.