IB-GAN: Disentangled Representation Learning with Information Bottleneck Generative Adversarial Networks

Insu Jeon, Wonkwang Lee, Myeongjang Pyeon, Gunhee Kim
Vision & Learning Lab, Dept. of Computer Science and Engineering, Seoul National University, Seoul, South Korea

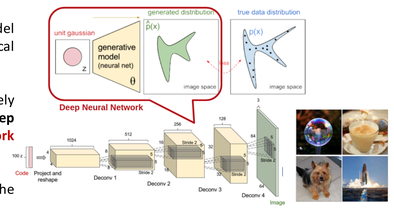
1. Introduction

Deep Generative Models

GOAL: Modeling a probability distribution model similar to the true data distribution using the empirical data.

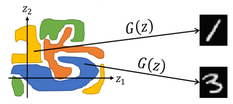
Unlike traditional generative models, which often rely on the classical probability distributions, **the deep generative models employ a deep neural network (DNN)** to model the generative distribution.

The generative model G(z) learns how to map the latent vector z to the data.

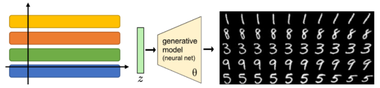


DNN enables generative modeling for complex data domain.

Disentangled Representation Learning



(a) **entangled** latent space and mapped images



(b) **disentangled** latent space

(c) rotation and thickness factors captured in the MNIST dataset

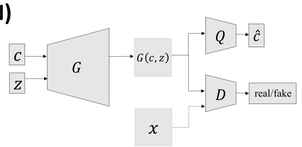
Conventional Deep Generative Model such as GAN (generative adversarial network) has no restrictions on utilizing the latent space. Thus, *the latent representation z in the deep generator G(z) in GAN can be used in a highly entangled way*.

Disentangled Representation Learning is a research field that simultaneously learns the generative model and the latent representation to control the factor of generated output images. **Disentanglement** is often described as statistical independence of the data generative factors, which aligns with the human intuition of interpretability.

2. Motivation and Background

Information Maximizing GAN (InfoGAN)

$$\min_G \max_D \mathcal{L}_{\text{InfoGAN}}(D,G) = \mathcal{L}_{\text{GAN}}(D,G) - I(c, G(z,c))$$
$$\mathcal{L}_{\text{GAN}}(D,G) = \mathbb{E}_{p(x)}[\log(D(x))] + \mathbb{E}_{p(z)}[\log(1-D(G(z)))]$$



InfoGAN (Chen et al. 2018): unsupervised GAN-based model for disentangled representation learning based on the GAN's objective and maximizing Mutual Information (MI) between generated images and the extra latent variable c sampled from some predefined distribution p(c).

GANs: a popular deep generative model approach trains two modules, the generator G and the discriminator D, alternatively using min-max (or adversarial) optimization.

- Generator G(z): generate data samples similar to x
  Discriminator D(x): distinguish real or fake data
- z ~ p(Z) : normal latent distribution
- c ~ (C) : some (categorical or continuous) distribution

Pros: By maximizing the MI term, the major concepts' interpretable representation is automatically aligned to the latent variable c without any supervision.
Cons: The generated data is only inverted to a fixed prior distribution p(c), which is not flexible.

Information Bottleneck Principle

$$\max_{q_\phi(z|x)} \mathcal{L}_{\text{IB}} = I(Z,Y) - \beta I(Z,X)$$

- X and Y : the input and target variable
- I(·, ·) : Mutual Information (MI)
- β : a Lagrange multiplier controls the trade-off

IB principle (Tishby et al., 1999; Alemi et al. 2017, 2018) provides an intuitive meaning for the good latent representation from the perspective of information theory.

GOAL: obtaining a compressive representation Z from the input X while maintaining the predictive information about the target Y as much as possible. The learned representation Z can act as minimal sufficient statics of X for predicting Y.

3. Information Bottleneck GAN (IB-GAN)

Information Bottleneck GAN's objective

GOAL: designing a new GAN-based unsupervised model that combines the pros of InfoGAN and IB principle for the disentangled representation learning.

Information Bottleneck GAN (IB-GAN) inspired by the IB principle introduces the minimization of mutual information (MI) into InfoGAN's objective, which constrains the latent representation z.

$$\min_G \max_D \mathcal{L}_{\text{IB-GAN}}(D,G) = \mathcal{L}_{\text{GAN}}(D,G) - \left[I^L(z, G(z)) - \beta I^U(z, G(z))\right]$$

IB-GAN maximizes the shared MI between the generator G and the input code z while also constraining the MI, which acts as a constraining mechanism that InfoGAN misses in the perspective of the IB-principle. IB-GAN combines the IB principle into InfoGAN for the disentangled representation learning. $I^L(\cdot,\cdot)$, $I^U(\cdot,\cdot)$: the lower and upper-bound of MI respectively. β > 0 is a trade-off coefficient, which controls the maximum amount of information shared by the Generated image G(z) and the latent code z.

4. Approximation of IB-GAN objective

Variational Lower and Upper Bounds of Mutual Information

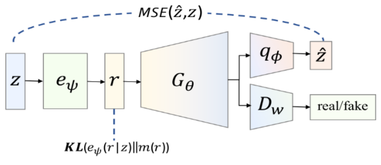
$$\mathbb{E}_{p_\theta(x|z)}[\log q_\phi(z|x)] - H(z) = I^L(z, G(z)) \le I(c, G(z))$$

The variational lower-bound of MI is formulated by introducing the reconstructor model q(z|x) similar to InfoGAN. This can be achieved by minimizing the reconstruction error of the input z and the approximation sampled from the reconstructor model q(z|G(z)).

$$I(z, G(R(z))) \le I^U(z, R(z)) = \mathbb{E}_{e_\psi(r|z)p(z)}[\text{KL}(e_\psi(r|z)||m(r))]$$

We developed a new formulation of variational upper-bound of MI based on the Markov property: if any generative process follows Z– > R– > X, then I(Z, X) <= I(Z, R). Essentially, This introduces an intermediate stochastic encoder model e(r|z) and the KL divergence constraint term.

Approximation of IB-GAN's objective and architecture



Algorithm 1 IB-GAN training algorithm

Input: batch size B, hyperparameters β, and the learning rates η_ϕ, η_θ, η_ψ, and η_w of the parameter of reconstructor, generator, encoder, and discriminator model respectively.
while not converged **do**
Sample $\{z^1,\dots,z^B\} \sim p(z)$
Sample $\{x^1,\dots,x^B\} \sim p(z)$
Sample $\{r^1,\dots,r^B\} \sim e_\psi(r|z^i)$ for $i \in \{1\dots B\}$
Sample $\{x^1_\theta,\dots,x^B_\theta\} \sim p_\theta(x|r^i)$ for $i \in \{1\dots B\}$
Sample $\{\hat{z}^1,\dots,\hat{z}^B\} \sim q_\phi(\hat{z}|x^i_\theta)$ for $i \in \{1\dots B\}$
$g_\phi \leftarrow \nabla_\phi \frac{1}{B}\sum_i (\hat{z}^i - z^i)^2$
$g_w \leftarrow \nabla_w \frac{1}{B}\sum_i \log\sigma(D_w(x^i_\theta)) + \log(1-\sigma(D_w(x^i)))$
$g_\theta \leftarrow \nabla_\theta \frac{1}{B}\sum_i \log\sigma(D_w(x^i_\theta)) - (\hat{z}^i - z^i)^2$
$g_\psi \leftarrow \nabla_\psi \frac{1}{B}\sum_i \log\sigma(D_w(x^i_\theta)) - (\hat{z}^i - z^i)^2 + \beta KL(e_\psi(r|z^i)||m(r))$
$\phi \leftarrow \phi - \eta_\phi g_\phi$; $w \leftarrow w - \eta_w g_w$;
$\theta \leftarrow \theta - \eta_\theta g_\theta$; $\psi \leftarrow \psi - \eta_\psi g_\psi$
end while

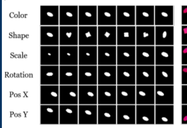
$$\min_{G,q_\phi,e_\psi} \max_D \mathcal{L}_{\text{IB-GAN}}(D,G,q_\phi,e_\psi)$$
$$= \mathcal{L}_{\text{GAN}}(D,G) - \{\mathbb{E}_{p(z)}[\mathbb{E}_{p_\theta(x|r)e_\psi(r|z)}[\log q_\phi(z|x)] - \beta\text{KL}(e_\psi(r|z)||m(r))]\}$$

An approximation of IB-GAN's objective adopts an intermediate representation encoder e(r|z) and the KL constraint term derived from the variational upper-bound of MI in the IB principle. The representation encoder e(r|z) and the prior m(r) are assumed to be Gaussian. The MSE loss is computed by the decoder q(z|x), similar to InfoGAN.

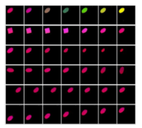
The resulting architecture of IB-GAN is partly analogous to that of β-VAE, a state-of-the-art disentangled representation learning model based on VAE (Kingma and Welling 2013; Rezende et al. 2014), as well, but it does not suffer from the shortcoming of the β-VAE such as generating blur output image.

5. Experiment

Experiment on (color) dSprites and dataset



(a) dSprites

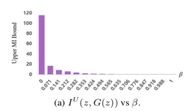


(b) Color-dSprites

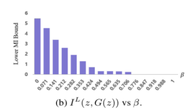
Models	dSprites	Color-dSprites
GAN	0.40 ± 0.05	0.35 ± 0.04
InfoGAN	0.61 ± 0.03	0.55 ± 0.08
IB-GAN	$\mathbf{0.80 \pm 0.07}$	$\mathbf{0.79 \pm 0.05}$
VAE	0.61 ± 0.04	0.59 ± 0.06
β-VAE	0.69 ± 0.09	0.74 ± 0.06
FactorVAE	$\mathbf{0.81 \pm 0.07}$	$\mathbf{0.82 \pm 0.06}$
β-TCVAE	0.79 ± 0.06	0.80 ± 0.07

(c)

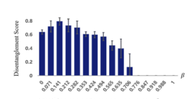
Latent generative factors captured by IB-GAN on (a) dSprites (Matthey et al. 2017) and (b) Color-dSprites (Burgess et al. 2018; Locatelloet al. 2018) dataset. (c) Comparison of disentanglement metric (Kim and Mnih 2018) in the dSprites and the color dSprites dataset. In our experiment, the disentanglement scores (Kim and Mnih 2018) of IB-GAN exceed those of GAN (Goodfellow et al. 2014), VAE (Kingma and Welling 2013; Rezende et al. 2014), and InfoGAN (Chen et al. 2016), and are comparable to those of β-VAE.



(a) $I^U(z, G(z))$ vs β.



(b) $I^L(z, G(z))$ vs β.



(c) Disentanglement score vs β.

The effects of parameter β on (a)-(b) the converged upper and lower-bound of MI and (c) on the disentanglement metric scores (Kim and Mnih 2018). (a)-(b) shows that introducing the upper MI bound $I^U(z, G(z))$ can constrain the lower MI bound; the gap between two MI bounds decreases as beta increases. The average disentanglement score varies according to β in (c), supporting that the upper MI bound affects the disentanglement-promoting behavior in IB-GAN.

Experiment on CelebA and 3D Chairs dataset

Latent generative factors captured by IB-GAN on CelebA (Liu et al. 2015) and 3D Chairs (Aubry et al. 2014)

(a) Azimuth

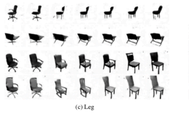
(b) Skin Tone

(a) Azimuth

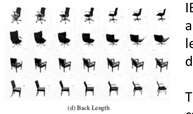
(b) Scale

(c) Gender

(d) Smile



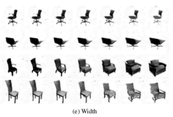
(c) Leg



(d) Back Length

(e) Hair Length

(f) Hair Color



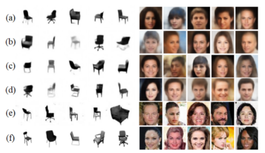
(e) Width

IB-GAN discovers various human recognizable attributes such as azimuth, skin tone, gender, smile, hair length, and hair color on CelebA dataset.

IB-GAN also disentangles azimuth, scales, leg types, back length, and width on 3D Chairs dataset.

These factors were hardly captured by the original InfoGAN (Chen et al. 2016; Higgins et al. 2017a; Kim and Mnih 2018; Chen et al. 2018), demonstrating the effectiveness of the proposed IB-GAN model for disentangled representation learning.

Randomly generated samples on CelebA and 3D Chairs dataset



Models	CelebA	3D Chairs
VAE	129.7	56.2
β-VAE	131.0	91.3
FactorVAE	109.7	44.7
β-TCVAE	125.0	57.3
GAN	8.4	27.9
InfoGAN	9.3	25.6
IB-GAN	**7.4**	**25.5**

(g)

Comparison of the quality of sample on CelebA and 3D Chairs dataset:
(a) VAE (Kingma and Welling 2013; Rezende et al. 2014) , (b) β-VAE (Higgins et al. 2017a), (c) FactorVAE (Kim and Mnih 2018), (d) β-TCVAE (Chen et al. 2018), and (e) IB-GAN and (f) real images. We could observe that the images obtained from IB-GAN are often sharper and realistic than those obtained from β-VAE and its variants.
(g) FID scores (Heusel et al. 2017) on CelebA and 3D Chairs dataset. The lower FID score, the better quality and diversity of samples. IB-GAN can produce diverse and qualitative image samples, while capturing various factors of variations.

6. Discussion and Conclusion

- IB-GAN is a novel GAN-based model for unsupervised learning of disentangled representation. IB-GAN can be seen as an extension to the InfoGAN, supplementing an information-constrained mechanism that InfoGAN lacks in the perspective of Information Bottleneck (IB) theory.
- The IB-GAN achieves comparable disentanglement results to existing state-of-the-art VAE-based models and produces a better quality of samples than standard GAN and InfoGAN.
- The approach of constructing the variational upper bound of generative MI by introducing an intermediate stochastic representation is a universal methodology. It may advance the design of other generative models based on the generative MI in the future.