InfoGAN and beyond

Sergio Calvo Ordoñez   Kristina Nikolić   Szilvia Ujváry

InfoGAN

InfoGAN [1] learns disentangled and interpretable representations by maximizing the mutual information between a subset of the latent variables and the GAN generated sample. This is done through the addition of an extra term to the objective function.

Background: GANs and Mutual Information

GANs are trained by a two-player minimax game between Discriminator D and Generator G with value function $V_{\text{GAN}}(D, G)$:

$$\min_G \max_D V_{\text{GAN}}(D, G) = \mathbb{E}_{x \sim p_{\text{real}}(x)}[\log D(x)] + \mathbb{E}_{z \sim p_z(z)}[\log(1 - D(G(z)))]$$

The mutual information (MI) $I(X; Y)$ between random variables X (all images) and Y (real/fake label) is:

$$I(X; Y) = H(X) - H(X|Y) = D_{KL}(P_{(X,Y)} || P_X \otimes P_Y)$$

As the MI is intractable, we use a variational lower bound on $I(c; G(z, c))$:

$$I(c; G(z, c)) \geq \mathbb{E}_{x \sim G(z,c)}[\mathbb{E}_{c' \sim P(c|x)}[\log Q(c'|x)]] + H(c) = L_I(G, Q)$$

In GANs G minimizes the same variational lower bound on $I(X; Y)$ [5]:

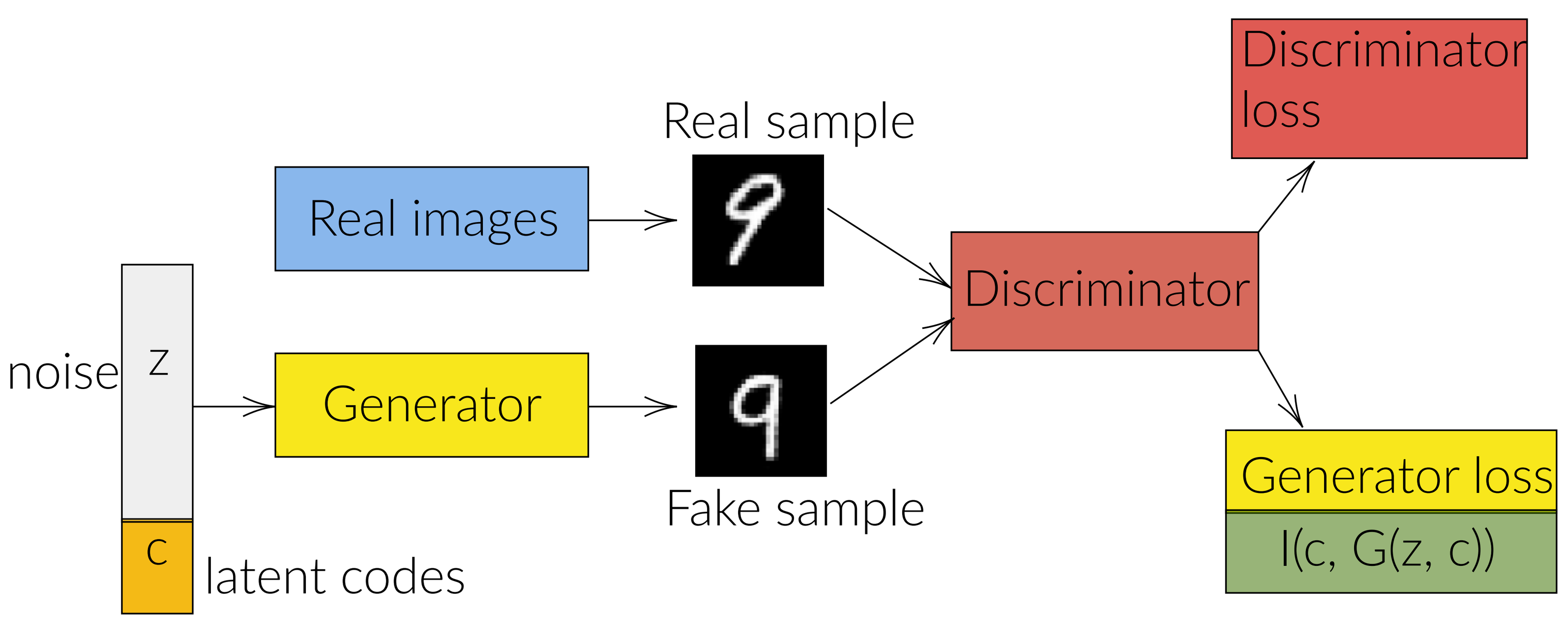
$$I(X; Y) \geq \mathbb{E}_{x \sim p_{\text{all images}}(x)} \mathbb{E}_{y \sim p_{\text{is x real?}}(y|x)}[\log q(y|x)] + H(Y)$$

$$= \mathbb{E}_{x \sim p_{\text{real}}(x)}[\log q(y = 1|x)] + \mathbb{E}_{x \sim p_{\text{fake}}(x,z)}[\log(1 - q(y = 1|x))] + H(Y)$$

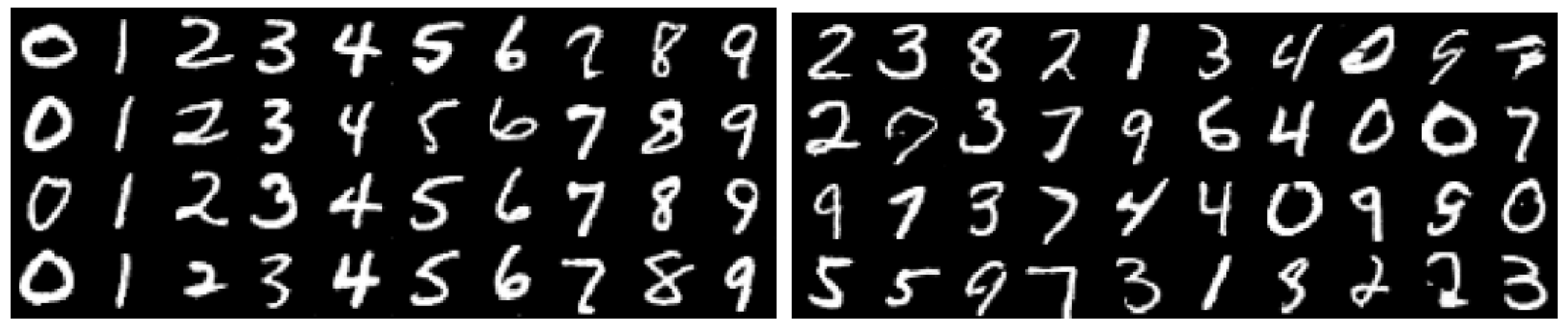
Method: Information Regularizing GANs

Maximize the MI between Generator output $G(z, c)$ and latent codes c as a regularizer:

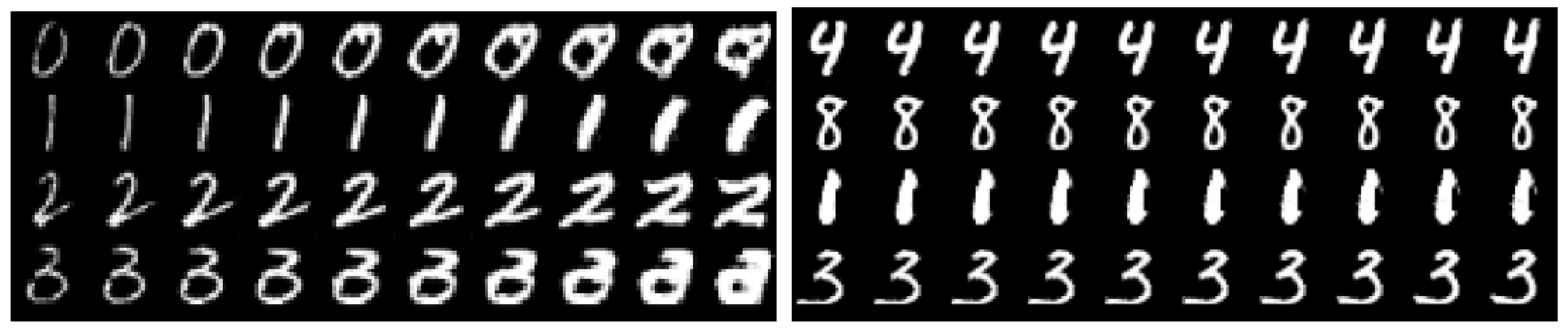
$$\min_G \max_D V_{I\text{-GAN}}(D, G) = V_{\text{GAN}}(D, G) - \lambda L_I(G, Q)$$



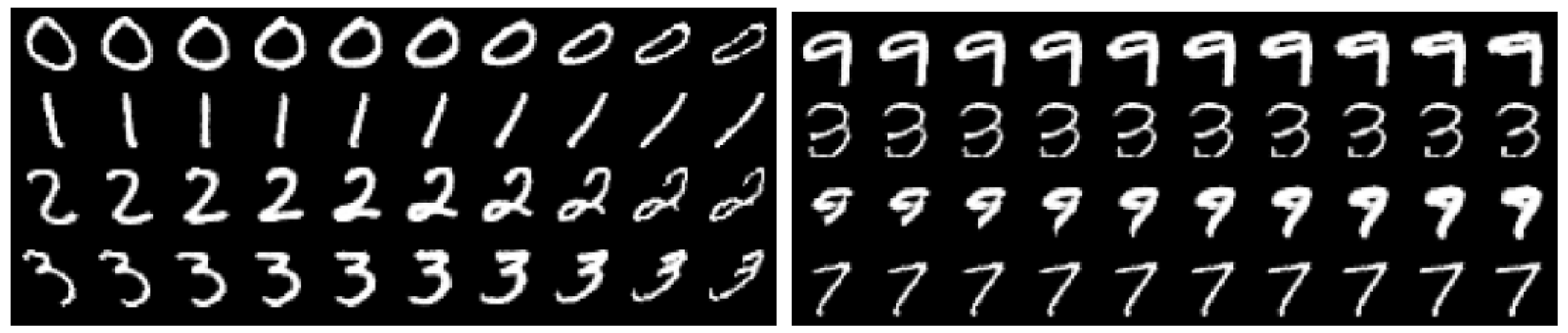
InfoGAN vs Vanilla GAN



(a) Varying c_1 on InfoGAN - **digit**   (b) Varying c_1 on a Vanilla GAN



(a) Varying c_2 on InfoGAN - **thickness**   (b) Varying c_2 on a Vanilla GAN



(a) Varying c_3 on InfoGAN - **tilt**   (b) Varying c_3 on a Vanilla GAN

Stability Analysis and Mutual Information

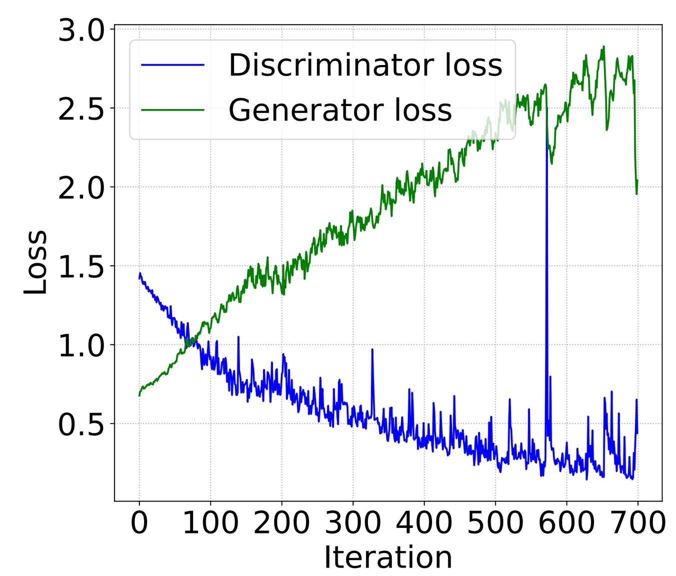


Figure 1. Training loss curves for both networks in our InfoGAN for MNIST.

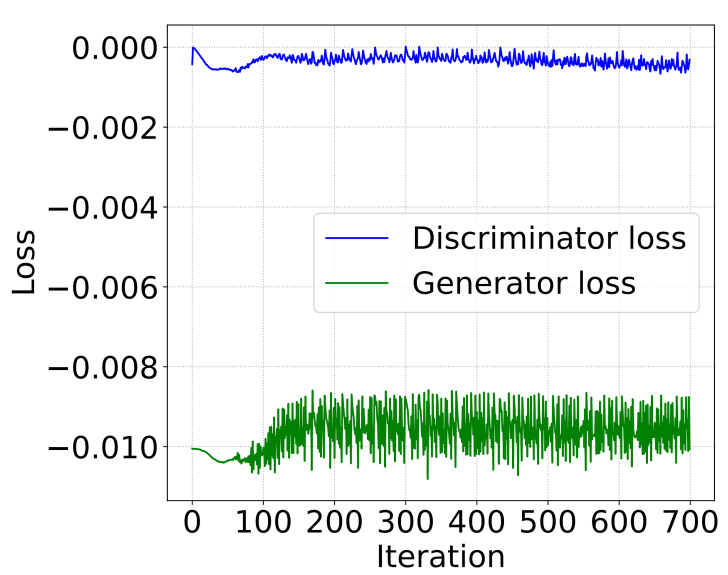


Figure 2. Training loss curves for both networks in our Info-WGAN for MNIST.

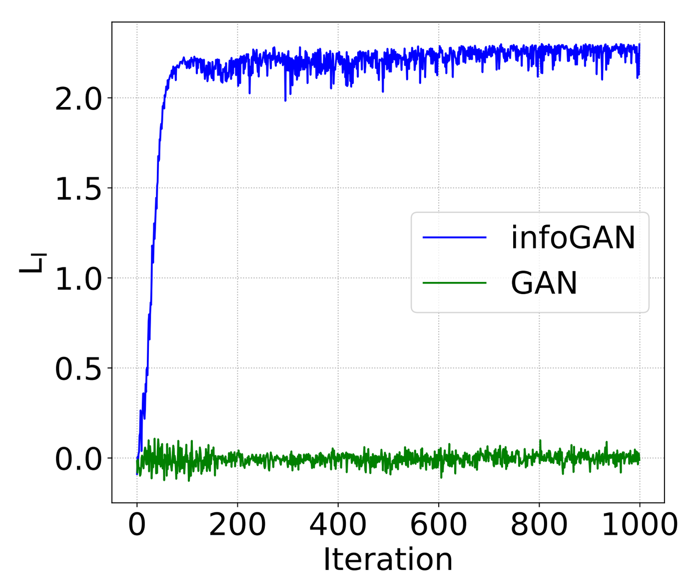


Figure 3. L_I for discrete code c_1.

Findings

- InfoGAN disentangles data features, Vanilla GAN does not.
- Info-WGAN training is more stable than InfoGAN.
- L_I for the discrete code increases up to the entropy $H(c_1)$=2.3.
- Reproducing experiment on CelebA dataset did not yield expected results.

Info-WGANs

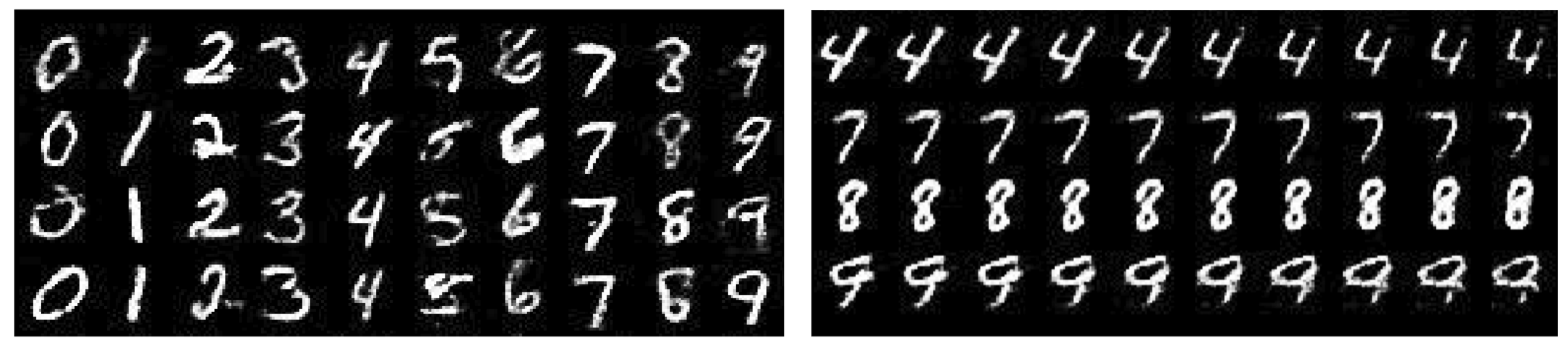
Wasserstein GANs (WGANs) [2] optimise Wasserstein distance. Benefits:

1. This metric is less prone to model collapse and vanishing gradients.
2. Introduction of weight clipping to enforce Lipschitz constraint.

$$\min_G \max_D V_{\text{WGAN}}(D, G) = \mathbb{E}_{x \sim p_{\text{real}}}[D(x)] - \mathbb{E}_{z \sim p(z)}[D(G(z))]$$

$$\Downarrow$$

$$\min_G \max_D V_{I\text{-WGAN}}(D, G) = V_{\text{WGAN}}(D, G) - \lambda L_I(G, Q)$$



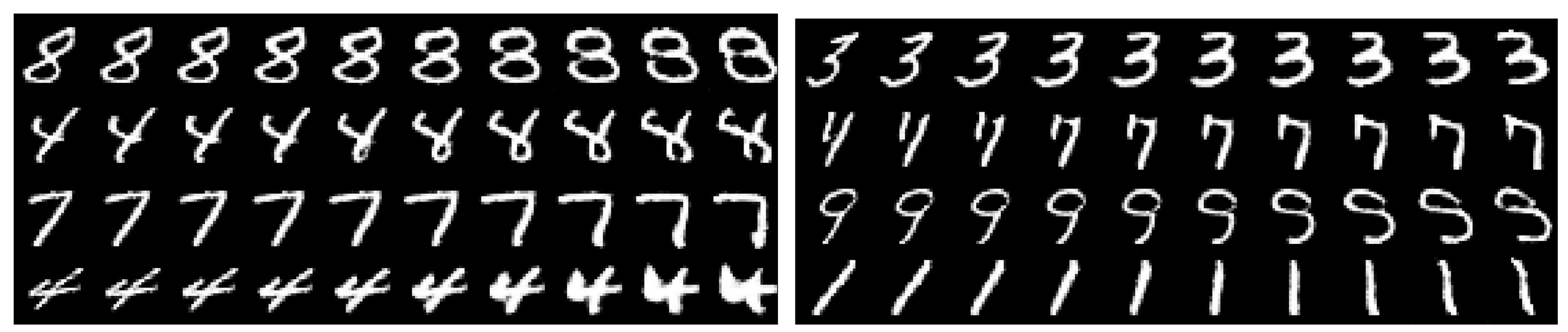
(a) Varying c_1 on Info-WGAN - **digit**   (b) Varying c_{cont} on Info-WGAN - ?

Findings: Info-WGAN performs satisfactorily with discrete latent codes although it finds difficulties interpreting the continuous ones.

MINE + GANs

MINE (Mutual Information Neural Estimator) [3] is a lower bound on the MI, obtained from the Donsker-Varadhan representation of the KL divergence by restricting function T to be parametrized by a neural net.

$$D_{KL}(P_{(X,Y)} || P_X \otimes P_Y) \geq \sup_{\theta \in \Theta} \mathbb{E}_{P_{X,Y}}[T_\theta] - \log \mathbb{E}_{P_X \otimes P_Y}[e^{T_\theta}]$$



(a) Varying c_2 on MineGAN - **thickness**   (b) Varying c_3 on MineGAN - **tilt**

Findings: MineGAN is only able to learn the continuous codes and is harder to train: tricks were needed to stabilize the MI.

Future work: Disentanglement of VAEs

Variational Autoencoders (VAEs) [4]:

- Have a more continuous and smooth latent space.
- Provide a more structured and interpretable latent space.
- Can perform interpolation in the latent space.

$$L_{\text{VAE}+c} = \mathbb{E}_{c \sim Q(c|x), z \sim Q(z|x)} \left[\ln \frac{P(z)P(c)P(x|z, c)}{Q(c|x)Q(z|x)} \right] \to L_{I\text{-VAE}} = L_{\text{VAE}} + \lambda L_I$$

- **Can we predict the appropriate number of latent codes in an unsupervised manner for different datasets?**

[1] Chen, X., Duan, Y., Houthooft, R., Schulman, J., Sutskever, I., and Abbeel, P. Infogan: Interpretable representation learning by information maximizing generative adversarial nets. In *Advances in Neural Information Processing Systems*, pp. 2172– 2180, 2016
[2] Arjovsky, M., Chintala, S., and Bottou, L. Wasserstein GAN. *arXiv preprint arXiv:1701.07875*, January 2017
[3] Mohamed Ishmael Belghazi, Aristide Baratin, Sai Rajeswar, Sherjil Ozair, Yoshua Bengio, Aaron Courville, and R Devon Hjelm. Mine: Mutual information neural estimation.
[4] Diederik P Kingma and Max Welling. Auto-encoding variational bayes, 2013.
[5] Huszár F. (2016, August 4). InfoGAN: using the variational bound on mutual information (twice). Blog post. *https://www.inference.vc/infogan-variational-bound-on-mutual-information-twice/*. Accessed 16 March 2023