InfoGAN: Interpretable Representation Learning by Information Maximizing GAN

Authors: **Xi Chen**, **Yan Duan**, **Rein Houthooft**, **John Schulman**, **Ilya Sutskever**, **Pieter Abbeel**
Affiliations: Berkeley EECS, OpenAI
Link: arXiv

Abstract

This paper describes InfoGAN, an information-theoretic extension to GAN that is able to learn disentangled representations in an unsupervised manner. The modification is that the InfoGAN also maximizes the mutual information between a small subset of the latent variables and the observation.

Method

To encourage the GAN to learn interpretable and meaningful representations, the InfoGAN maximizes the mutual information between a fixed small subset of the GAN's noise variables and the observations.

Recap. For a GAN, the goal is to learn a general distribution $P_G(x)$ that matches the real data distribution $P_{\text{data}}(x)$. The minimax game is given by

$$\min_G \max_D V(D,G) = \mathbb{E}_{x\sim P_{\text{data}}}[\log D(x)] + \mathbb{E}_{z\sim \text{noise}}[\log(1-D(G(z)))]$$

Mutual information for inducing latent codes

To encourage more meaningful representation of the input noise vector z, we decompose the input noise vector into two parts: (i) z, which is treated as source of incompressible noise; (ii) c, which is called the latent code and will target the salient structured semantic features of the data distribution.

In case the generator $G(z,c)$ ignores the additional latent codes, we request that there should be high mutual information between latent codes c and the generator distribution $G(z,c)$, i.e., $I(c;G(z,c))$ should be high. From which we have a information-regularized minimax game

$$\min_G \max_D V_I(D,G) = V(D,G) - \lambda I(c;G(z,c))$$

Variational Mutual Information Maximization

In practice, the mutual information term $I(c;G(z,c))$ is hard to maximize directly as it requires access to the posterior $P(c \mid x)$. However, we could obtain a lower bound of it by

defining an auxiliary distribution $Q(c \mid x)$ to approximate $P(c \mid x)$:

$$\begin{aligned} I(c; G(z,c)) &= H(c) - H(c \mid G(z,c)) \\ &= \mathbb{E}_{x \sim G(z,c)} \left[\mathbb{E}_{c' \sim P(c|x)} \left[\log P(c' \mid x) \right] \right] + H(c) \\ &= \mathbb{E}_{x \sim G(z,c)} \left[D_{KL}(P(\cdot \mid x) \,||\, Q(\cdot \mid x)) + \mathbb{E}_{c' \sim P(c|x)} \left[\log Q(c' \mid x) \right] \right] + H(c) \\ &\geq \mathbb{E}_{x \sim G(z,c)} \left[\mathbb{E}_{c' \sim P(c|x)} \left[\log Q(c' \mid x) \right] \right] + H(c) \end{aligned}$$

This technique of lower bounding mutual information is known as Variational Information Maximization.

Lemma. For random variables X, Y and function $f(x, y)$ under suitable regularity conditions:

$$\mathbb{E}_{x \sim X, y \sim Y|x}[f(x,y)] = \mathbb{E}_{x \sim X, y \sim Y|x, x' \sim X}[f(x', y)]$$

Using this lemma, we can define a variational lower bound $L_I(G, Q)$, of the mutual information, $I(c; G(z, c))$:

$$\begin{aligned} L_I(G,Q) &= \mathbb{E}_{c \sim P(c), x \sim G(z,c)} \left[\log Q(c \mid x) \right] + H(c) \\ &= \mathbb{E}_{x \sim G(z,c)} \left[\mathbb{E}_{c' \sim P(c|x)} \left[\log Q(c' \mid x) \right] \right] + H(c) \\ &\leq I(c; G(z,c)) \end{aligned}$$

Th author pointed out that $L_I(G, Q)$ is easy to approximate with Monte Carlo simulation. In particular, L_I can be maximized w.r.t. Q directly and w.r.t. G via the reparameterization trick. Hence, InfoGAN is defined as this minimax game:

$$\min_{G,Q} \max_{D} V_{\text{InfoGAN}}(D, G, Q) = V(D, G) - \lambda L_I(G, Q)$$

Results

Mutual information maximization

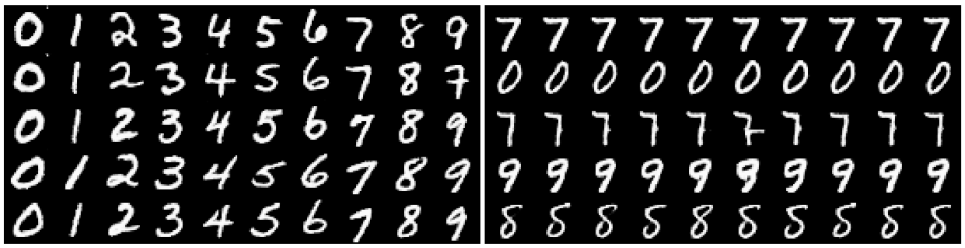
Experiment results show that while in regular GANs, there is no guarantee that the generator will make use of the latent codes, InfoGAN can quickly maximize the mutual information.

Disentangled representation

In Figure 1, an InfoGAN is trained on MNIST with a discrete code $c_1 \sim \text{Cat}(K = 10, p = 0.1)$ and two continuous codes $c_2, c_3 \sim \text{Unif}(-1, 1)$.

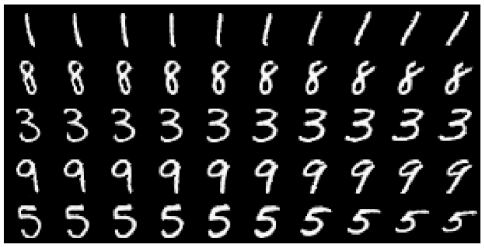
Conclusion

In contrast to previous approaches, InfoGAN does not require supervision and learns interpretable and disentangled representations on challenging datasets. Other extensions to this work include: learning hierarchical latent representations, improving semi-supervised learning with better codes, and using InfoGAN as a high-dimensional data discovery tool.

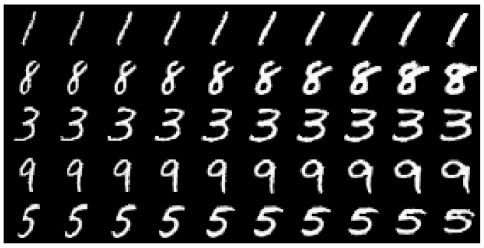


(a) Varying c_1 on InfoGAN (Digit type)

(b) Varying c_1 on regular GAN (No clear meaning)



(c) Varying c_2 from -2 to 2 on InfoGAN (Rotation)



(d) Varying c_3 from -2 to 2 on InfoGAN (Width)

Figure 1: Manipulating latent codes on MNIST

Mutual Information

In information theory, mutual information between X and Y, $I(X;Y)$, measures the "amount of information" learned from knowledge of random variable Y about the other random variable X. It can be expressed as the difference of two entropy terms:

$$I(X;Y) = H(X) - H(X \mid Y) = H(Y) - H(Y \mid X)$$

An intuitive interpretation of $I(X;Y)$ is the reduction of uncertainty in X when Y is observed.