InfoGAN: Interpretable Representation Learning by Information Maximizing Generative Adversarial Nets

by Xi Chen, Yan Duan, Rein Houthooft, John Schulman, Ilya Sutskever, Pieter Abbeel
UC Berkeley, Department of Electrical Engineering and Computer Sciences
OpenAI

Presented by
Marharyta Dekret & Diana Grygorian
Unsupervised learning of disentangled representation

Usually, learned representation is entangled (encoded in complicated manner)

When representation is disentangled, it would be easier to apply to tasks
Disentangling information

man with glasses - man without glasses + woman =
Supervised Learning

“to learn is to recognize”

Unsupervised Learning

“to learn is to replicate”
Plan

- GAN
- Mutual Information for Inducing Latent Codes
- Variational Mutual Information Maximization
- Implementation
- Experiments
GAN – Generative Adversarial Network

**Generator network - G**

*transform*

\[ z \sim P_{\text{noise}}(z) \rightarrow \text{sample } G(z) \]

**Goal:**

\[ \max \rightarrow D(G(z)) \]

\[ \max \log D(G(z)) \]

**Discriminator network - D**

*distinguish between true P_{\text{data}} and P_G*

**Goal:**

\[ \max \rightarrow D(x)(1 - D(G(z))) \]

\[ \max \log D(x) + \log(1 - D(G(z))) \]
GAN

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

where

$$D(x) = \frac{P_{data}(x)}{P_{data}(x) + P_G(x)}$$
Generative adversarial networks: Revisit

\[ Z \sim N(0, I) \]
Big momentum - both participants of the game respond too sharply and correct their behavior so much that they make it even worse.
Mutual Information for Inducing Latent Codes

♦ decompose the input noise vector into two parts:
  ♦ $z$ – incompressible noise
  ♦ $c$ – latent code

Denote the set of $c_1, c_2, ..., c_L$ - assume a factored distribution given by:

$$P(c_1, c_2, ..., c_L) = \prod_{i=1}^{L} P(c_i)$$
Mutual Information for Inducing Latent Codes

\[ G(z, c) \text{ – generator network with both noise } z \text{ and the latent code } c \]

\[ \diamond \text{ In standard GAN: } P_G(x|c) = P_G(x) \]

Information-theoretic regularization:

\[ I(c; G(z, c)) \]

Mutual information between latent code \( c \) and generator distribution \( G(z, c) \) should be high
Theory

Mutual information between $X$ and $Y$ $\rightarrow I(X; Y)$

measures the “amount of information” learned from knowledge of random variable $Y$ about the other random variable $X$

The mutual information can be expressed as the difference of two entropy terms:

$$I(X; Y) = H(X) - H(X|Y) = H(Y) - H(Y|X)$$

is the reduction of uncertainty in $X$ when $Y$ is observed
Information-regularized minimax game:

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

Problem!

$I(c; G(z, c))$

is hard to minimize directly because of access to the posterior $P(c|x)$.

Given $x \sim P_G(x)$, $P_G(c|x)$ should have small entropy.
Solution: Variational Mutual Information Maximization [5]

Define an auxiliary distribution \( Q(c|x) \) to approximate \( P(c|x) \):

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

\[
= E_{x \sim G(z, c)} \left[ E_{c' \sim P(c|x)} [\log P(c'|x)] \right] + H(c)
\]

\[
= E_{x \sim G(z, c)} \left[ D_{KL} (P(\cdot|x)||Q(\cdot|x)) \right] + E_{c' \sim P(c|x)} [\log Q(c'|x)] + H(c) 
\]

\[
\geq E_{x \sim G(z, c)} \left[ E_{c' \sim P(c|x)} [\log Q(c'|x)] \right] + H(c) 
\]

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

$$E_{x \sim X, y \sim Y | x}[f(x, y)] = E_{x \sim X, y \sim Y | x, x' \sim X | y}[f(x', y)].$$

It removes the need to sample from the posterior.
With lemma define a variational lower bound, $L_I(G, Q)$, of the mutual information, $I(c; G(z, c))$:

Remind the lemma: $E_{x \sim x, y \sim y | x} [f(x, y)] = E_{x \sim x, y \sim y | x, x' \sim x | y} [f(x', y)]$

$$L_I(G, Q) = E_{c \sim P(c), x \sim G(z; c)} [\log Q(c | x)] + H(c)$$

$$= E_{x \sim G(z; c)} [E_{c' \sim P(c | x)} [\log Q(c' | x)]] + H(c)$$

$$\leq I(c; G(z, c))$$

Previously we proved $I(c; G(z, c)) \geq E_{x \sim G(z, c)} [E_{c' \sim P(c | x)} [\log Q(c' | x)]] + H(c)$
In particular, $L_I$ can be maximized w.r.t. $Q$ directly and w.r.t. $G$ via the reparametrization trick.

$L_I(G, Q)$ can be added to GAN’s objectives with no change to GAN’s training procedure.

When the variational lower bound attains its maximum

$$L_I(G, Q) = H(c)$$

for discrete latent codes, the bound becomes tight and the maximal mutual information is achieved.
The resulting algorithm - Information Maximizing Generative Adversarial Networks (InfoGAN).

$$\min_{G,Q} \max_D V_{InfoGAN} = V(D, G) - \lambda L_1(G, Q)$$

minimax game with a variational regularization of mutual information and a hyperparameter $\lambda$
Implementation

- Parametrize the **auxiliary distribution** $Q$ as a neural network
- $Q$ and $D$ share all convolutional layers
- There is one final fully connected layer to output parameters for the conditional distribution $Q(c|x)$
◊ Categorical latent code $c_i$ - the natural choice of softmax nonlinearity to represent $Q(c_i|x)$.

◊ Continuous latent code $c_j$ - treating $Q(c_j|x)$ as a factored Gaussian is sufficient.

◊ An extra hyperparameter $\lambda$, it’s easy to tune and simply setting to 1 is sufficient for discrete latent codes.
Experiment goals:

1 - investigate if mutual information can be maximized efficiently

2 - evaluate if can learn disentangled and interpretable representations by varying only one latent factor at a time in order to assess if varying such factor results in only one type of semantic variation in generated images
Mutual Information Maximization

- $c \sim \text{Cat}(K = 10, \ p = 0.1)$ quickly maximized to $H(c) \approx 2.30$

- GAN - the generator is not explicitly encouraged to maximize the mutual information with the latent codes

- $Q$ reasonably approximates the true posterior $P(c \mid x)$
Manipulating latent codes on MNIST

(a) Digit type

(b) No clear meaning

(c) Rotation

(d) Width
Manipulating latent codes on 3D Faces

(a) Pose (angle)

(b) Elevation

(c) Lighting

(d) Width
Manipulating latent codes on 3D Chairs

(a) Rotation

(b) Width
Manipulating latent codes on SVHN

(a) Continuous code

(b) Discrete code
Conclusions

◊ In contrast to other approaches, InfoGAN is unsupervised and learns interpretable and disentangled representations.

◊ InfoGAN adds only negligible computation cost on top of GAN and is easy to train.

◊ The core idea of using mutual information to induce representation can be applied to other methods, which is a promising area of future work.
Possible extensions:

- learning hierarchical latent representations
- improving semi-supervised learning with better codes
- using InfoGAN as a high-dimensional data discovery tool.
References


Questions?