(H. Kim) InfoGAN: Interpretable Representation Learning by Information Maximizing Generative Adversarial Nets

Prof.	□
작성일시	@March 11, 2022 1:48 PM
최종 편집	@March 11, 2022 1:48 PM
Keywords	GANs Generarive model study Paper review / Study
Reviewer	
Reviewed	☑
Paper (해당시)	
발표일자	
발표영상 (해당시)	
추가 자료	
PPT (해당시)	
Note (추가info)	
담당자	
Column	

Summary

1. wavelet pooling and unpooling : preserve information of the content to the transfer network
2. progressive stylization : stylize the image faithfully

Abstract

> This paper describes InfoGAN, an information-theoretic extension to the Generative Adversarial Network that is able to learn disentangled representations in a completely unsupervised manner. InfoGAN is a generative adversarial network that also maximizes the mutual information between a small subset of the latent variables and the observation. We derive a lower bound of the mutual information objective that can be optimized efficiently. Specifically, InfoGAN successfully disentangles writing styles from digit shapes on the MNIST dataset, pose from lighting of 3D rendered images, and background digits from the central digit on the SVHN dataset. It also discovers visual concepts that include hair styles, presence/absence of eyeglasses, and emotions on the CelebA face dataset. Experiments show that **InfoGAN learns interpretable representations** that are competitive with representations learned by existing supervised methods.

InfoGAN : disentangled representation을 비지도로 학습

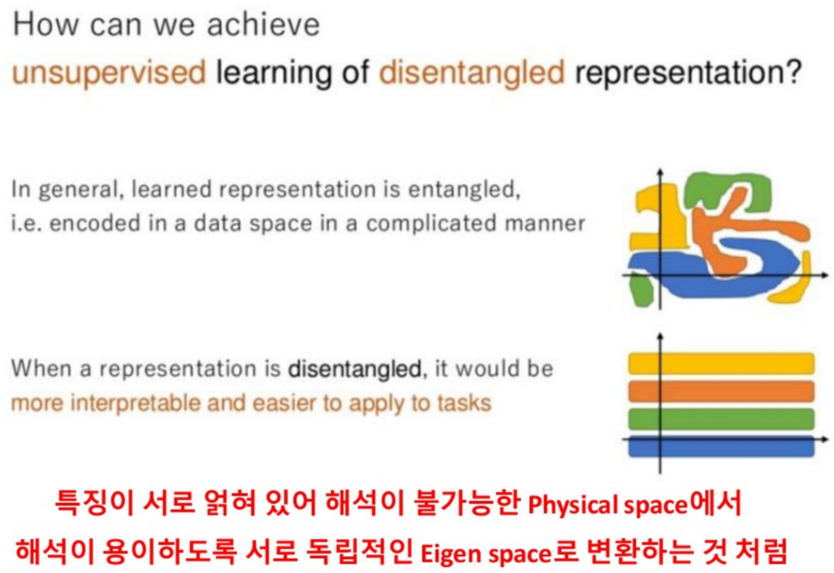
원리 : maximize the mutual information(between latent variable과 observation)

Introduction

- Unsupervised learning : unlabelled data로 부터 value를 추출
    - representation learning : goal = use unlabelled data to learn a representation(중요한 semantic feature를 가지는 representation)
    - disentangled representation : represent the salient attribute of a data instance(relevant한 ask에 대해 helpful)
- InfoGAN

- maximize the mutual information(MI) between a fixed small subset of the GAN’s noise variable and the observation : interpretable & meaningpul representation 학습 가능
- GANs가 base

▼ 조금 더 이해해보기[교수님 블로그 참고]



How can we achieve
unsupervised learning of disentangled representation?

In general, learned representation is entangled,
i.e. encoded in a data space in a complicated manner

When a representation is disentangled, it would be
more interpretable and easier to apply to tasks

특징이 서로 얽혀 있어 해석이 불가능한 Physical space에서
해석이 용이하도록 서로 독립적인 Eigen space로 변환하는 것 처럼

위 사진처럼 representation이 entagled되어 있지 않고, 아래 사진처럼 평평하게 disentangled되어 있다면, more interpretable하다는 것을 알 수 있습니다. 이처럼, representation을 학습할 때, 좀 더 좋은 성질을 갖도록(disentagled하도록) 제약을 주고싶다!라는 motivation에서 나온게 infoGAN입니다!

- Background : Generative Adversarial Networks(GAN)
    - Generative Adversarial Networks
    - GAN min-max problem :

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

Method

1. Mutual Information for inducing latent codes

    - 기존 GAN은 noise vector z만 사용(noise will be used by the generator in a highly entangled way → z는 semantic feature에 해당하지 않는다, 즉, noise z와 semantic feature간의 연관관계를 찾기 어렵다.)
    - InfoGAN은 input을 두 개로 분해 : z (incompressible noise) + c (latent code, 두드러진 semantic feature를 target, concatenation of all c_i)
    - propose information-theoretic regularization : latent code c와 generator distribution $G(z,c)$ 사이의 MI가 높아져야함(즉, $I(c;G(z,c))$가 높아져야한다.)
    - Information theory
        - X와 Y사이의 MI(H는 entorpy) :

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

        - 독립적인 경우 : $I(X;Y) = 0$,

          deterministic, invertible function으로 related된 경우 : maximal MI
        - information-regularized minmax game :

$$\min_{G} \max_{D} V_I(D,G) = V(D,G) - \lambda I(c;G(z,c))$$

2. Variational mutual information maximization

    - $I(c;G(z,c))$를 바로 maximize하는 것은 어려우니, auxiliary distribution $Q(c|x)$가 $P(c|x)$에 근사하도록 함으로써, lower bound를 구할 수 있다.

$$\begin{aligned} I(c;G(z,c)) &= H(c) - H(c|G(z,c)) \\ &= \mathbb{E}_{x\sim G(z,c)}[\mathbb{E}_{c'\sim P(c|x)}[\log P(c'|x)]] + H(c) \\ &= \mathbb{E}_{x\sim G(z,c)}[\underbrace{D_{\mathrm{KL}}(P(\cdot|x) \parallel Q(\cdot|x))}_{\geq 0} + \mathbb{E}_{c'\sim P(c|x)}[\log Q(c'|x)]] + H(c) \\ &\geq \mathbb{E}_{x\sim G(z,c)}[\mathbb{E}_{c'\sim P(c|x)}[\log Q(c'|x)]] + H(c) \end{aligned}$$

    - Lemma 5.1

Lemma 5.1 *For random variables X, Y and function $f(x, y)$ under suitable regularity conditions:* $\mathbb{E}_{x\sim X, y\sim Y|x}[f(x, y)] = \mathbb{E}_{x\sim X, y\sim Y|x, x'\sim X|y}[f(x', y)]$.

By using Lemma A.1, we can define a variational lower bound, $L_I(G, Q)$, of the mutual information, $I(c; G(z, c))$:

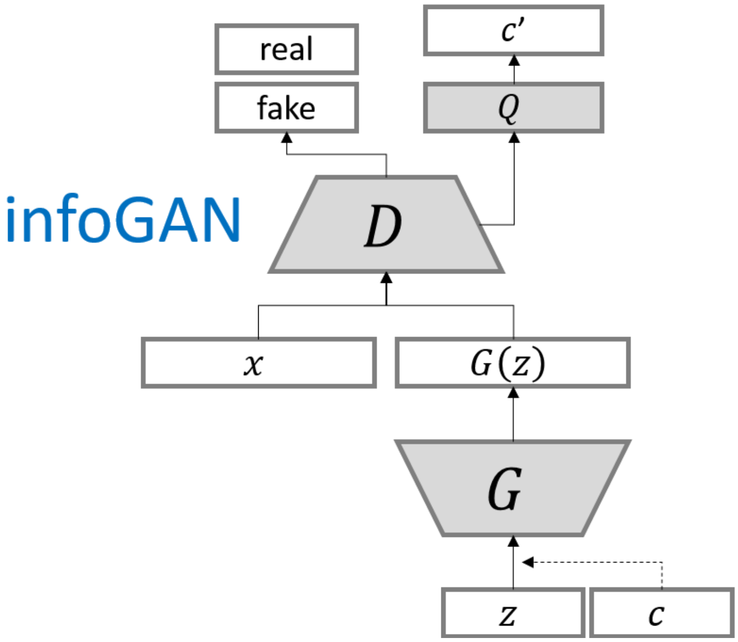
$$\begin{aligned} L_I(G, Q) &= E_{c\sim P(c), x\sim G(z,c)}[\log Q(c|x)] + H(c) \\ &= E_{x\sim G(z,c)}[\mathbb{E}_{c'\sim P(c|x)}[\log Q(c'|x)]] + H(c) \\ &\le I(c; G(z, c)) \end{aligned} \tag{5}$$

- objective function :

$$\underset{G,Q}{min}\,\underset{D}{max}\,V_{InfoGAN}(D, G, Q) = V(D, G) - \lambda L_I(G, Q)$$

3. Implementation

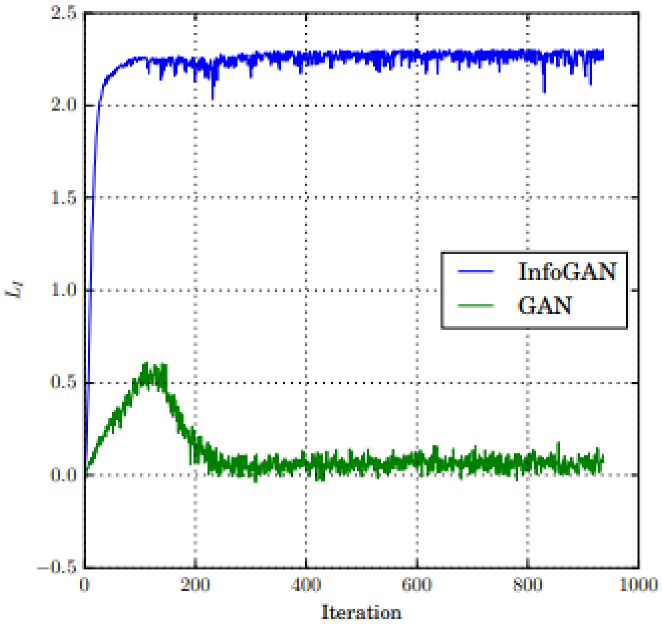
- auxiliary distribution Q = neural network
- categorical latent code c_i = softmax nonlinearity to represent $P(c_j|x)$
- continuous latent code c_j = true posterior $P(c_j|x)$에 따라 다양, $Q(c_j|x)$를 factored Gaussian 으로 사용하는 것으로 충분
- λ = 1로 설정, 작은 λ는 $\lambda L_I(G, Q)$가 GAN objectives의 동일한 scale인 것을 보장
- GAN은 훈련시키기 어려우므로, DC-GAN으로 introduce된 existing techniques 사용



Experiment & Result

goal : MI가 잘 maximize 되는가, InfoGAN이 disentangled 하고 interpretable representation을 학습할 수 있는가(한가지 latent vector를 바꿨을 때, 한 가지의 semantic variation만 발생하는지 확인)

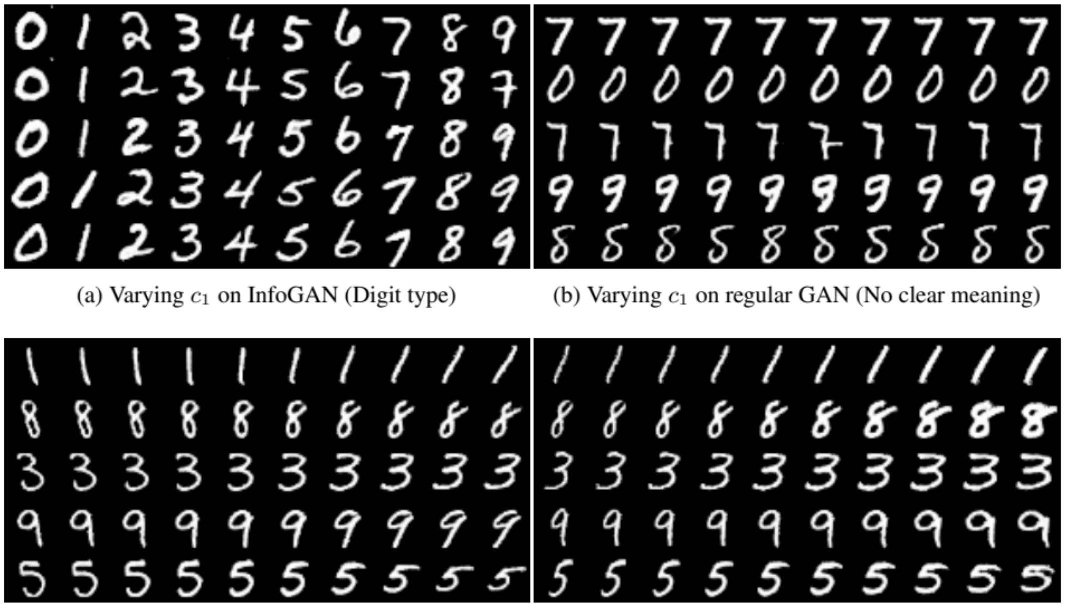
1. Mutual Information Maximization(c and $G(z,c)$)
    - MNIST dataset과 latent code $c \sim Cat(K=10, p=0.1)$와 함께 학습
    - $L_I(G,Q)$가 $H(c) \approx 2.30$으로 빠르게 maximized, 즉, bound가 tight하고 MI가 maximal
    - GAN : Q reasonable approximates the true posterior $P(c|x)$ = latent code와 generated image 사이의 MI가 적음 ⇒ generator가 latent code를 활용한다는 보장 x



2. Disentangle Representation
    - c_1 : capture drastic change in shape(caterical code), c_2, c_3 : capture continuous variation(각각 rotation, width)
    - MNIST :
        - generator는 단순히 digit을 stretch하거나 rotate하지 않지만, 대신 자연스럽게 보이기 위해서 thickness or stroke style 같은 다른 디테일을 조절한다.

    - generalizable test : latent code를 -1~1 → -2~2로 변경(기존엔 훈련하지 않았던 wide region)했지만, 여전히 meaningful stylization을 get

    ▼ Fig. 2



(a) Varying c_1 on InfoGAN (Digit type)

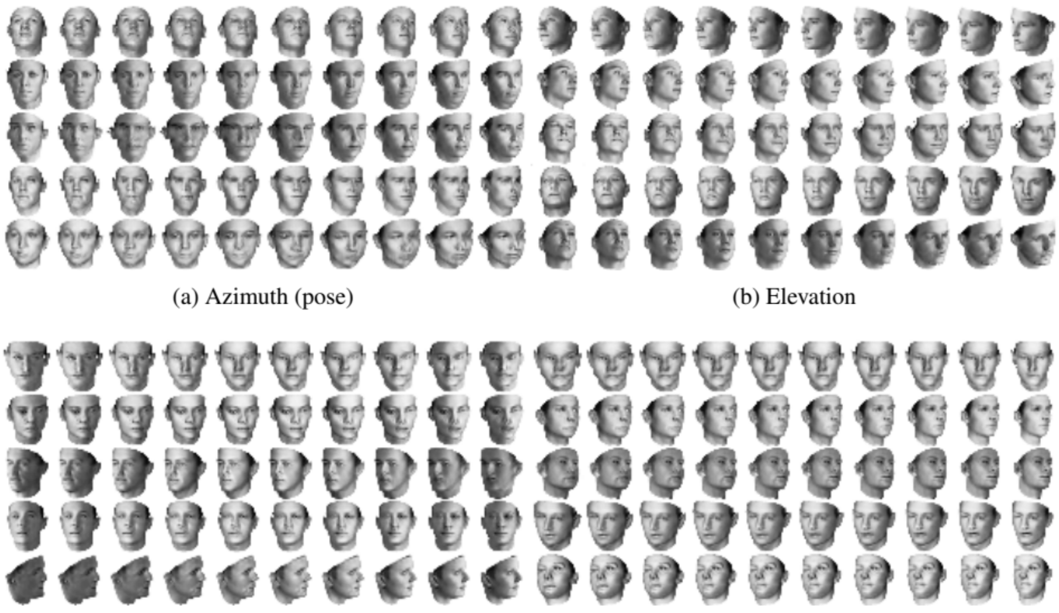
(b) Varying c_1 on regular GAN (No clear meaning)

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

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

- Faces dataset :
    - DC-IGN이 supervision을 사용하여 latent factor를 pose, elevation, ligthing로 continuous latent variable로 표현하는 방법을 배운다.
    - InfoGAN은 pose, elevation, ligthing을 recover하는 disentangled representation을 배움

    ▼ Fig. 3



(a) Azimuth (pose)

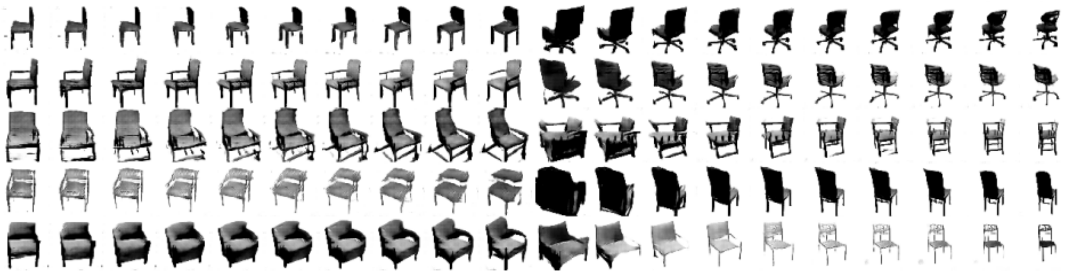
(b) Elevation

(c) Lighting

(d) Wide or Narrow

- Chairs dataset
    - DC-IGN이 rotation을 나타내는 continuous code를 배울 수 있다
    - InfoGAN은 다시 continuous code로 같은 concept를 학습할 수 있고, signle continuous code를 이용하여 다른 너비의 같은 타입의 체어 사이를 연속적으로 interpolate 할 수 있다.

▼ Fig. 4



(a) Rotation

(b) Width

3. Conclusion : InfoGAN은 기존 연구와는 다르게 unsupervised representation learning이며 interpretable and disentangled representation을 학습한다. 기존 GAN에서 negligible computation cost만 추가하여 훈련도 쉽다.

References

1. InfoGAN 교수님 블로그 : https://jaejunyoo.blogspot.com/2017/03/infogan-1.html