



[논문리뷰] InfoGAN [2016]

Introduction

✨ 해당 논문은 **Unsupervised manner**에서 **disentangled representations**을 학습하는 것이 가능한 **Information-theoretic**으로 확장한 GAN인 InfoGAN을 소개합니다.

InfoGAN은 generative adversarial Network이며,
변수와 관찰로부터의 부분집합에서 **mutual information**을 최대화합니다.

InfoGAN은 효율적으로 최적화하는 것을 목표로 하는 **mutual information**의 하한을 도출합니다.

실험을 통하여, InfoGAN은 기존 **Supervised Method**에 의해 학습된 표현을 통합적인 표현으로 학습한다는 것을 증명하였습니다.

Unsupervised Representation을 통하여, 핵심적인 특징을 나타내는 **disentangled representation**을 학습할 수 있다면, **classification**등의 **downstream task**에서 유용하게 사용하는 것이 가능합니다.

정리하자면,

1. GAN의 생성자에 Input이 되는 noise 변수들의 어떤 고정된 부분집합과 관측 데이터간에 상호정보를 최대화하게 함으로써, 해석 가능하며 풍부한 의미를 갖는 representation 학습할 수 있게 하는 목적함수를 수정하였습니다.
2. 비지도 방식으로 학습된 표현은 기존의 레이블 정보를 사용하는 지도학습 방식으로 학습된 표현과 유사한 퍼포먼스를 보여주며, 상호정보를 cost로 하는 생성 모델이 learning disentangled representations에 대하여, 좋은 접근법이 될 수 있음을 시사합니다.

Mutual Information for Inducing Latent Codes

GAN의 생성자는 noise에 대해 상관없이 vector z 를 사용합니다.

결과적으로 **Generator**의 분포에 맞지 않는 Noise는 의미론적 특징을 가지지 않는다고 응답합니다.

본 논문에서는 **noise vector**를 **two parts**로 **decompose**하는 것입니다.

1. Z : Source of incompressible Noise [분해 불가능한 Vector]
2. 데이터 분포의 핵심적인 의미론적 특징에 대응되는 latent code [부분집합]

✨ 수학적 설명은 하단의 **LINK**를 참고하여 작성하였습니다.

상호정보량의 정의

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

⊗는 단순히 곱을 의미한다고 생각하면 된다. 스칼라면 $P_X(x)P_Y(y)$ 처럼 그냥 막 곱하면 되는데, P_X, P_Y 는 distribution이기 때문에 ⊗로 표기해 놓은 것이다. "확률 변수 X, Y 가 서로 독립이라 가정했을 때의 joint distribution과 true joint distribution 간의 거리가 얼마나 되나"라는 뜻이다.

정의를 통해 일단 수식적인 의미부터 살펴보자. 일단 D_{KL} 은 Kullback-Leibler divergence이고 두 분포 사이의 거리를 의미한다. 두 분포가 같다면 최소값인 0을 가진다.

그럼 상호정보량이 최소일 때에는 $P_{(X,Y)}$ 와 $P_X \otimes P_Y$ (즉, $P_{X,Y}(x, y)$ 와 $P_X(x)P_Y(y)$)가 같은 분포라고 생각할 수 있겠다. 따라서 P_X, P_Y 가 서로 독립일 때 최소값을 가진다.

말로 풀어 써보면 "Y가 어떤 값으로 정해졌을 때의 X의 불확정성" 정도로 생각할 수 있는데, 이는 극단적인 상황을 생각해 보면 금방 이해할 수 있다.

만약 X 와 Y 가 독립이라면 $Y = 3$ 이라는 정보는 X 를 결정하는 데에 아무 영향을 주지 않는다. X 는 여전히 불확정적이다.

(위의 1번 예시를 생각해 보자.)

하지만 만약 $P(X \neq Y) = 0$ 이라면?(X 랑 Y 가 항상 같다면?) Y 가 어떤 값으로 결정되면 X 도 무조건 결정할 수 있다. 그러면 불확정성이 완전히 깨져버리고 Y 가 X 의 완전한 정보를 제공한다. 이 때 상호정보량은 양으로 발산한다. (상호정보량은 음이 될 수 없다!)

정의를 이용해 다음과 같이 계산할 수 있다.

$$I(X; Y) = \int_{\mathcal{Y}} \int_{\mathcal{X}} p(x, y) \log \left(\frac{p(x, y)}{p(x)p(y)} \right) dx dy$$

"불확정성"이라는 것은 엔트로피 H 를 이용해 표현할 수 있다.

$$H(X) = - \int_{\mathcal{X}} p(x) \log p(x) dx$$

이 값은 $p(x)$ 가 모든 x 에 대해 같아지면 가장 불확정적이므로 최댓값을 갖는다.

어떤 다른 확률변수가 condition으로 들어왔을 때도 불확정성을 표현할 수 있다.

$$\begin{aligned} H(X|Y) &= - \int_{\mathcal{X}} \int_{\mathcal{Y}} p(x, y) \log \frac{p(x, y)}{p(y)} dy dx = \int_{\mathcal{X}} \int_{\mathcal{Y}} p(x, y) \log p(x|y) dy dx \\ &= \mathbb{E}_{x \sim P_X} [\mathbb{E}_{y \sim P_Y} [\log P(X|Y)]] \end{aligned}$$

그럼 상호정보량을 엔트로피를 이용해서 표현이 가능해진다.

$$I(X; Y) = H(X) - H(X|Y) = \mathbb{E}_{X \sim P_X} [\mathbb{E}_{Y \sim P_{Y|X}} [\log P(X|Y)]] + H(X)$$

Variational Information Maximization

이제 위 식에 $X \leftarrow c, Y \leftarrow G(z, c)$ 를 대입하면 상호정보량을 계산할 수 있다.

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

하지만 여기서 장벽이 하나 더 생긴다. 우리가 $P(c|x)$ 를 알고 있나?? 어떤 이미지의 code c 를 알아내는 것이 목표인데, 알아내고 싶은 분포를 이용해 수식을 계산한다?? 이걸 말이 안 된다. 그래서 이 논문의 저자들은 Auxiliary Distribution Q 를 두어 $P(c|x)$ 를 직접적으로 이용하는 것을 피했다.

$P(c|x)$ 는 관측된 x 에 대한 c 의 분포이고 $P(c)$ 는 그냥 c 의 분포다. 만약 $P(c|x)$ 를 이미 알고 있다면 관측된 이미지 x 를 가장 잘 표현할 수 있는 c 를 아는 것이나 다름 없다. 그럼 학습이 의미가 없다.

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

값을 구하기 힘든 식의 계산을 lower bound를 이용해 해결하는 모습을 볼 수 있다. 이를 **Variational Information Maximization**이라 한다.

빨간색으로 줄 친 부분을 보면 $P(c|x)$ 를 필요로 하는 부분을 **Law of total Expectation**을 이용해 계산했다.

$$\mathbb{E}_Y [\mathbb{E}_X [X|Y]] = \mathbb{E}[X]$$

아담의 정리라고도 부른다. 논문에서는 이를 더 일반화해 사용했다.

lemma 5.1 증명

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

아랫첨자가 주인공이니 크게 썼다.

앞문 간에, 이를 통해 c 를 데이터 x 에 상관없이 마구잡이로 샘플링 해 계산해도 원하는 대로 학습이 된다는 것을 예상할 수 있다. 샘플링 할 c 는 진짜 해당 step에서 생성된 이미지에 상관 없이 "아무거나" 뽑으면 된다. 진짜 아무거나!!

따라서 최종 목표는

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

가 된다.

Implementation

일반적인 DCGAN에서 stride 2가 없어서 추가되어, 크기가 28x28이 되었습니다.

C.1 MNIST

The network architectures are shown in Table 1. The discriminator D and the recognition network Q shares most of the network. For this task, we use 1 ten-dimensional categorical code, 2 continuous latent codes and 62 noise variables, resulting in a concatenated dimension of 74.

Table 1: The discriminator and generator CNNs used for MNIST dataset.

discriminator D / recognition network Q	generator G
Input 28×28 Gray image	Input $\in \mathbb{R}^{74}$
4×4 conv. 64 IRELU, stride 2	FC, 1024 RELU, batchnorm
4×4 conv. 128 IRELU, stride 2, batchnorm	FC, $7 \times 7 \times 128$ RELU, batchnorm
FC, 1024 IRELU, batchnorm	4×4 upconv, 64 RELU, stride 2, batchnorm
FC, output layer for D , FC, 128-batchnorm-IRELU-FC, output for Q	4×4 upconv, 1 channel stride 2

Conclusion

해당 논문은 Information Maximizing 방식(상호정보량)을 도입한 InfoGAN이라는 representation Learning Algorithm을 제안합니다. 이전의 방식과 비교했을 때,

InfoGAN은 완벽한 Unsupervised Algorithm으로 학습하고, 복잡한 데이터셋등에서도 Interpretable하고 disentangled representations가 가능합니다.


상호정보량의 핵심 아이디어는 VAE와 같은 다른 생성모델에도 확장가능하다는 것을 의미합니다.

learning hierarchical latent representations, semi-supervised learning과 같은 분야에 후속연구로 진행될 수 있음을 시사합니다.

Reference

InfoGAN: Interpretable Representation Learning by Information Maximizing Generative Adversarial Nets


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

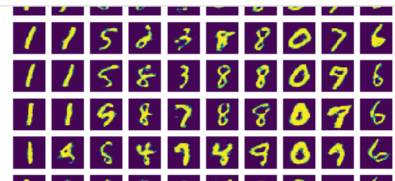
 <https://arxiv.org/abs/1606.03657>



[학부생의 딥러닝] GANs | InfoGAN : Information maximizing GAN

InfoGAN - Tensorflow 구현, PyTorch 구현 레퍼런스 - InfoGAN - Interpretable Representation Learning by Information Maximizing Generative Adversarial Nets : <https://arxiv.org/abs/1606.03657> - 상호정..

 <https://haawron.tistory.com/10>



<https://velog.io/@changdaeoh/InfoGAN-review>