

20기 정규세션

ToBig's 19기 강의를자

이혁준

18. Generative Basic

Contents

Unit 01 | Why study generative modeling?

Unit 02 | How do generative models work compare to GAN?

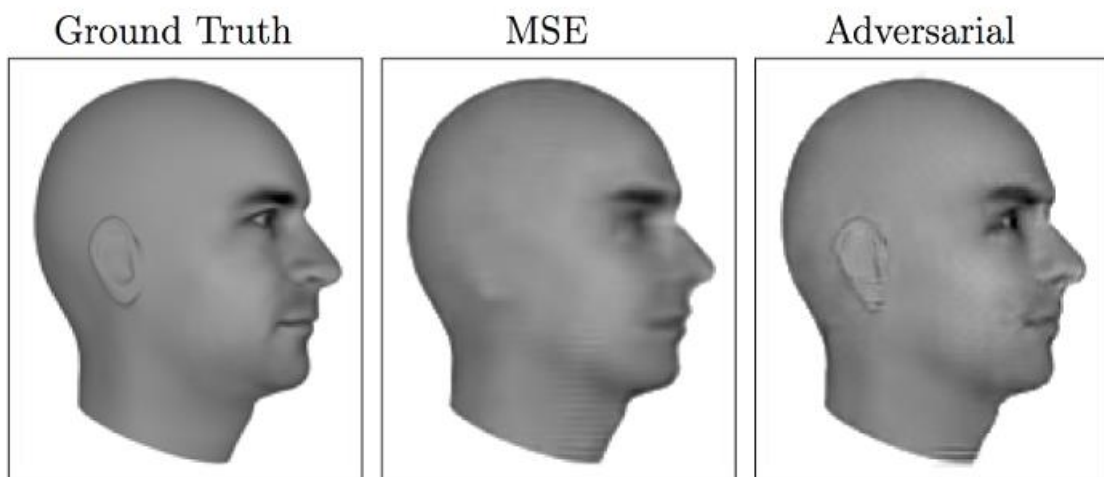
Unit 03 | How do GANs work?

Unit 04 | Assignments

Unit 01 | Why study generative modeling?

- 생성모델을 활용한 training과 sampling은 고차원 probability distribution을 조작하고, 표현하는데 좋은 도구임.
- 발생 가능성이 있는 미래를 simulate하는 데이터를 제공할 수 있으며 이는 강화학습(Reinforcement learning)에 여러가지 방식으로 적용될 수 있음.
- Missing data에 대해 학습할 수 있으며, 이는 semi-supervised learning이 가능함.
 - (Semi-supervised learning) Label의 개수를 줄이는 방식의 학습 방식으로, 다수의 unlabeled example을 학습함으로써 모델 generalization을 도울 수 있음.

Unit 01 | Why study generative modeling?



Lotter et al. (2015)

- MSE loss로 학습한 이미지의 경우 조금씩 다른 이미지들의 평균에 해당
 - 다른 가능성 무시
 - 하지만 이 경우, ear, eye가 blur
- GAN loss를 활용한 경우 다양한 가능성들을 이미 파악하고 있기 때문에, 현실적이고 상세한 결과를 생성한걸 확인할 수 있음.

- Multi-modal 학습에 용이하게 사용됨.
 - 하나의 input은 다수의 정답을 가질 수 있음.
 - ML 모델을 학습하는 전통적인 방식인 MSE의 경우 하나의 정답을 파악하는 것만 가능
 - GAN loss를 활용하면 이 간극을 줄일 수 있음.

Unit 01 | Why study generative modeling?

- 확률 분포에서 생성된 이미지 데이터를 요구하는 경우가 다수 존재함.
 - Single Image Super-resolution
 - Goal) 저해상도 이미지를 고해상도 품질로 이해하는 것
 - 생성 모델은 주어진 이미지 갖는 정보보다 많은 내용을 이해하고, 생성하는데 사용됨.
 - Create art
 - 사용자의 input, needs에 맞는 realistic한 이미지를 생성할 수 있음.
 - Image-to-image translation

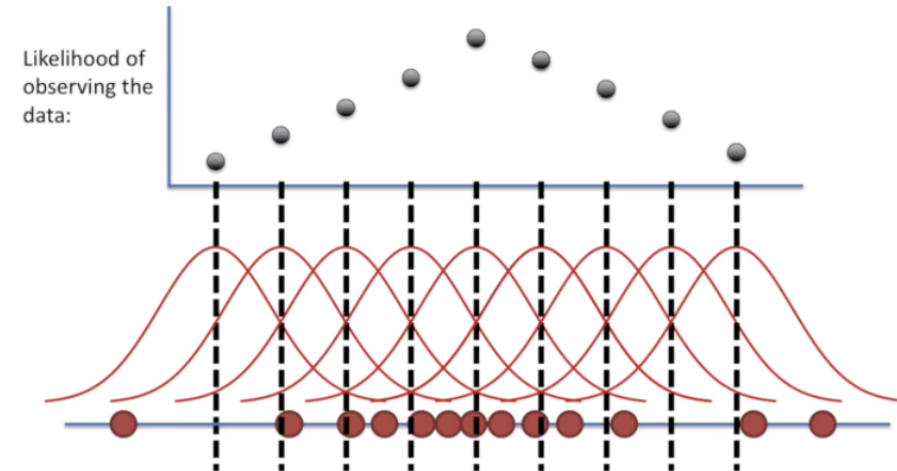
Maximum Likelihood Estimation

(Basic Idea) parameter로 정의되는 확률분포를 추정할 수 있는 모델을 정의하는 것.

Likelihood function, 가능도 $\prod_{i=1}^m p_{\text{model}}(x^{(i)}; \theta)$

관측값이 있을 때 그 중에서 가장 가능성이 높은 값을 측정하는 함수

주로 $L(\theta|D)$ 로 표현하여, 특정 파라미터에 따라 측정되는 값임을 의미



출처: StatQuest with Josh Starmer

➡ 훈련 데이터에 likelihood를 최대화하는 모델을 선택하는 것이 목표!

Maximum Likelihood Estimation

Likelihood function

$$\prod_{i=1}^m p_{\text{model}}(\mathbf{x}^{(i)}; \boldsymbol{\theta})$$

$$\boldsymbol{\theta}^* = \arg \max_{\boldsymbol{\theta}} \prod_{i=1}^m p_{\text{model}}(\mathbf{x}^{(i)}; \boldsymbol{\theta}) \quad (1)$$

$$= \arg \max_{\boldsymbol{\theta}} \log \prod_{i=1}^m p_{\text{model}}(\mathbf{x}^{(i)}; \boldsymbol{\theta}) \quad (2)$$

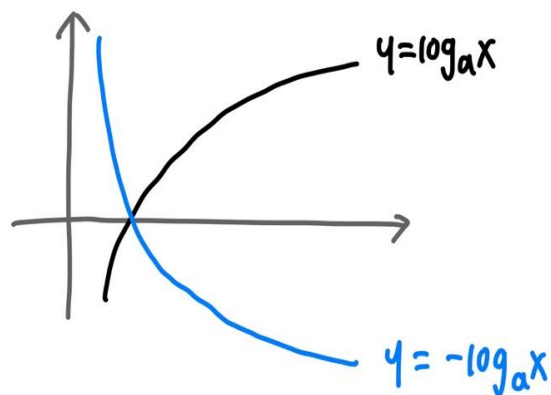
$$= \arg \max_{\boldsymbol{\theta}} \sum_{i=1}^m \log p_{\text{model}}(\mathbf{x}^{(i)}; \boldsymbol{\theta}). \quad (3)$$

- Log space에서는 곱셈을 덧셈으로 계산할 수 있기 때문에, 자연로그를 이용

KL divergence

- 정보 엔트로피
 - 정보량
 - 놀람의 정도
 - 정보이론의 기본은 있음직한 일들이 발생했을 때 그 메시지는 적은 정보를 담고 있지만, 희박한 이벤트가 발생했을 때는 높은 정도의 정보를 담고 있음.
 - $I(x_i) = -\log_a p(x_j)$
 - $p(x_j)$: x_j 가 발생할 확률
 - a : 측정하는 정보에 따라 다르게 정해지는 값

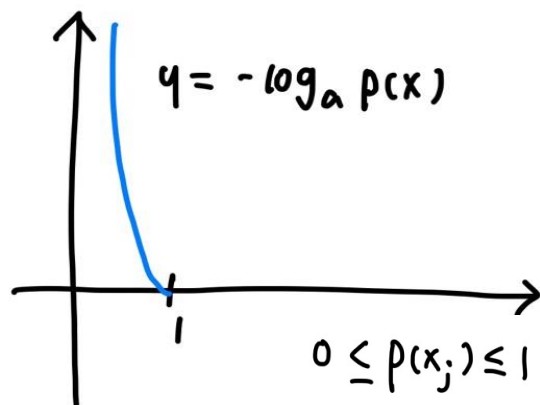
KL divergence



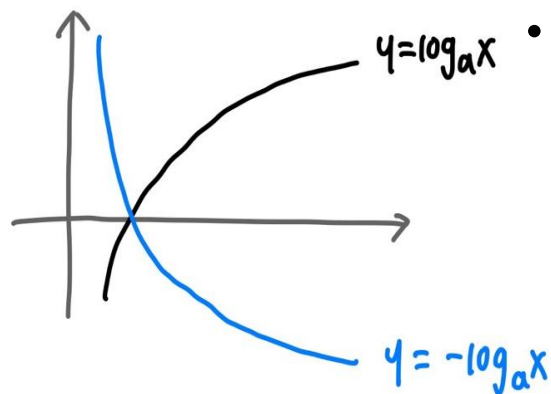
- 정보 엔트로피
- 정보량
- 예시)

	독일 승	한국 승	비김
확률	81%	5%	14%

- 독일이 이기는 경우의 정보량 $-\log_2 0.81 = 0.304$
- 한국이 이기는 경우의 정보량 $-\log_2 0.05 = 4.3219$ (매우 놀라운 사건)
- 비기는 경우의 정보량 $-\log_2 0.14 = 2.8365$

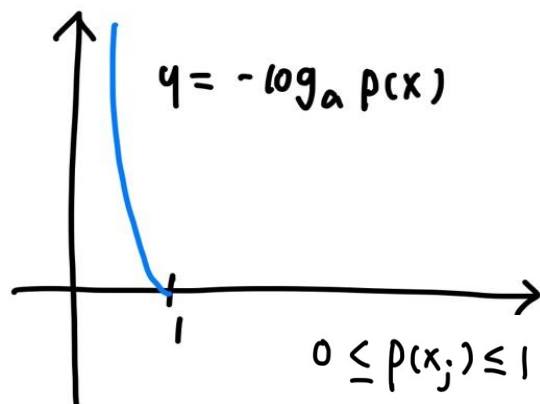


KL divergence



- 정보 엔트로피
- 엔트로피
- 가능한 모든 결과의 평균 정보량

$$H(x) = E\{I(x_j)\} = \sum_{j=1}^n p(x_j) \log_2 p(x_j)$$



- 예시)
 - 한국 vs 독일 축구경기의 엔트로피
$$0.81 * (-\log_2 0.81) + 0.05(-\log_2 0.05) + 0.14(-\log_2 0.14) = 0.8595$$
 - 스웨덴 vs 멕시코 축구경기의 엔트로피
$$0.36 * (-\log_2 0.36) + 0.34(-\log_2 0.34) + 0.3(-\log_2 0.3) = 1.5809$$

KL divergence

- **Cross-Entropy**

$$H(P, Q) = \sum_{i=1}^k P_i * \log_2 \frac{1}{Q_i}$$

- Q는 예측값, P는 실제값
- 모델링을 통하여 구한 분포값인 확률 분포 Q를 이용하여 실제 확률분포 P를 예측함.
- 즉, cross-entropy는 두 분포 사이에 존재하는 정보량을 구하는 것

KL divergence

- 정의
 - $D_{KL}(P||Q)$
 - 확률 분포 P와 Q 간의 차이
 - P는 사후, 즉 실제 확률 분포
 - Q는 사전 분포를 의미
- 해석
 - 실제 분포 P가 주어 졌을 때, Q를 모델로 사용하는 경우 예상되는 정보 엔트로피
 - 실제 데이터 분포 P를 정확히 알 수 없지만, Q와의 차이를 minimize하는 방식으로 추정할 수 있음.

KL divergence

-

Taxonomy of deep generative models

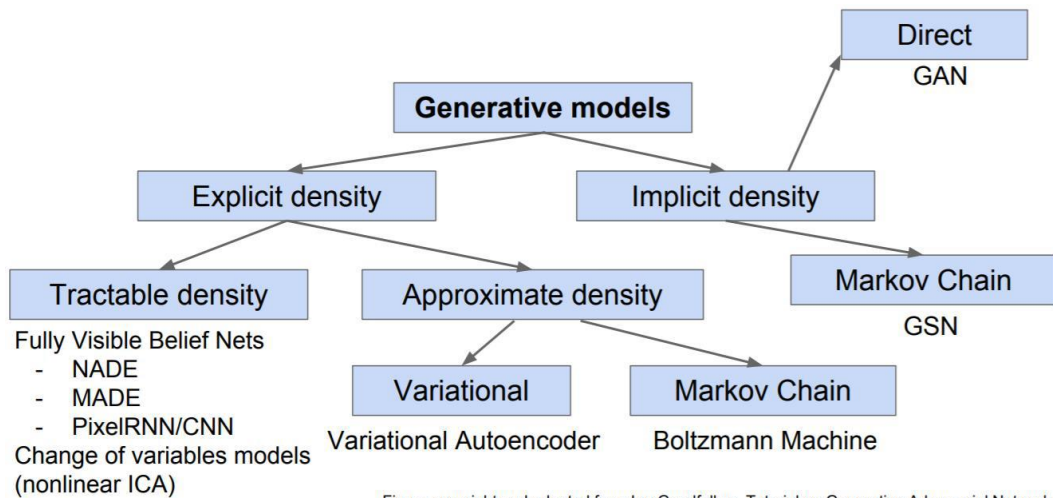


Figure copyright and adapted from Ian Goodfellow, Tutorial on Generative Adversarial Networks, 2017.

Model		특징
Explicit density	Tractable density	모델의 사전분포를 가정하여 기존 값으로부터 데이터 분포를 추정
	Approximate density	사전분포를 근사 시켜 데이터 분포를 추정
Implicit density		데이터의 확률 분포를 모르는 상태에서 샘플링을 반복하여 특정 확률 분포에 수렴 시켜 추정

Comparing GANs

- GANS은 기존 생성모델의 다음 단점을 보완하고자 함.
 - (FVBNs) 주어진 x 의 차원에 비례하는 과정으로 샘플을 생성하는 것이 아니라 병렬로 생성할 수 있음.
 - (Boltzmann machine) 생성기는 이전 모델에 비해 적은 제약조건을 가짐
 - (GSNs) Markov chain 필요 없음
 - (VAE) Variational bound가 필요 없으며, universal 추정기로 사용할 수 있음.
- GANs의 단점
 - 내시 균형 문제를 해결하도록 학습 해야 하는데, 이는 오히려 기존의 objective function을 최적화하는 것보다 어려운 문제
 - 죄수의 딜레마

The GAN framework

Generative Adversarial Networks

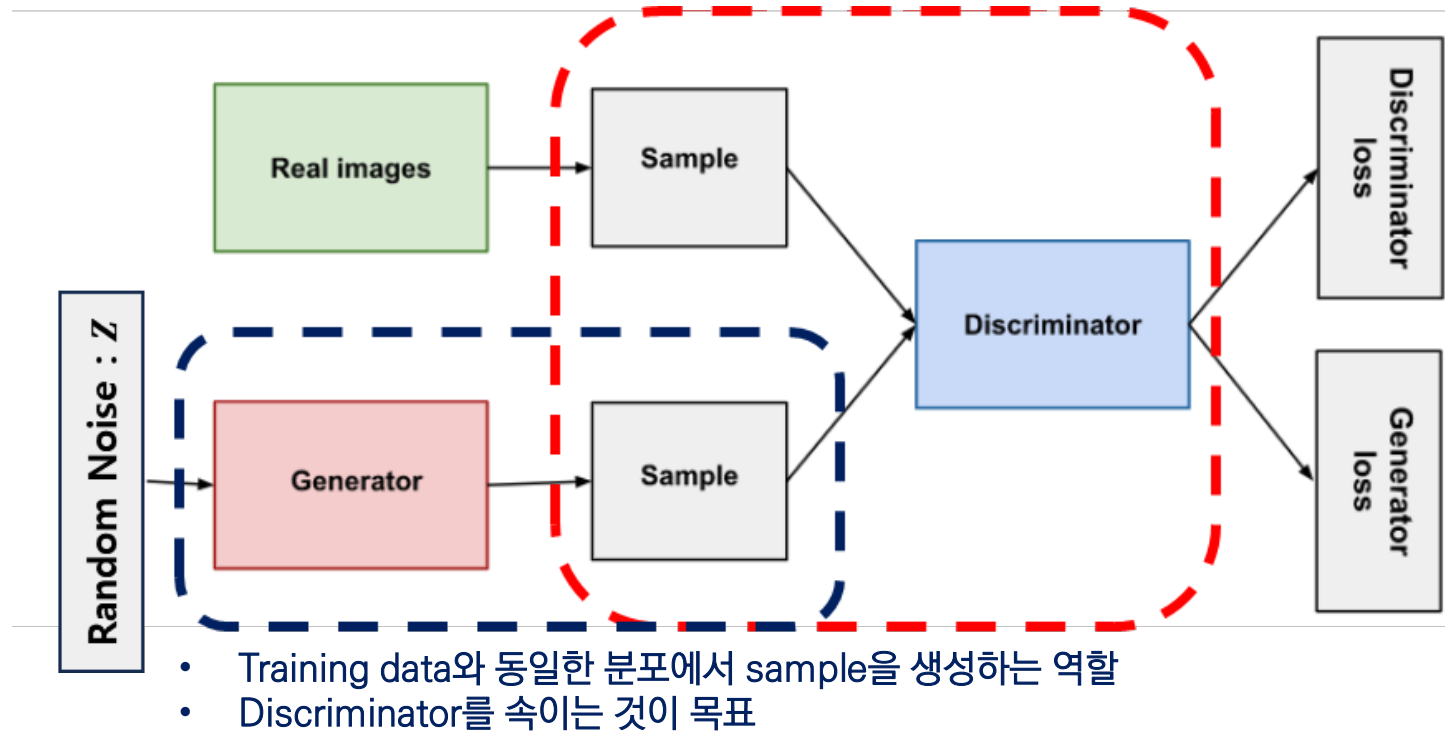
생산적

적대

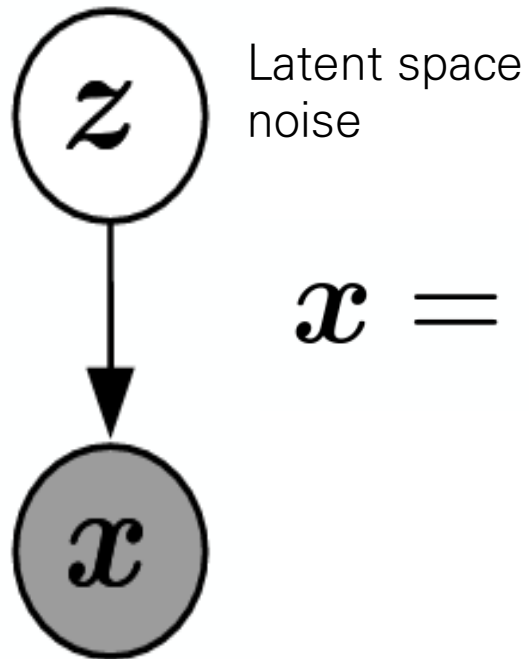
신경망

The GAN framework

- 생성된 이미지가 진짜인지 가짜인지 구분하는 역할
- 기존의 지도학습과 유사하게 real or fake (0 or 1)을 학습하는 것이 목표



Generator Network



$$x = G(z; \theta^{(G)})$$

- 미분가능한 함수 (only requirements)
- z 사이즈에 상관없이 훈련 가능
 - 최소 데이터 x 의 차원의 크기를 가져야 함.
- 주어진 x 를 가우시안 분포를 따르게 만들 수 있음.

Training Procedure

- SGD 알고리즘을 사용하여(Adam) 동시에 두 샘플에 대해 훈련
 - 훈련 샘플
 - 생성 샘플
- 최적) 한 네트워크에 대해서 k step을 동작하면, 다른 한 네트워크에서는 모든 step에 대해 동작

Objective function

$$\min_G \max_D V(D, G) = \min_{\theta_g} \max_{\theta_d} \left[\mathbb{E}_{x \sim p_{data}} \log \underbrace{D_{\theta_d}(x)}_{\substack{\text{Discriminator output} \\ \text{for real data } x}} + \mathbb{E}_{z \sim p(z)} \log (1 - \underbrace{D_{\theta_d}(G_{\theta_g}(z))}_{\substack{\text{Discriminator output} \\ \text{for generated fake data } G(z)}}) \right]$$

Discriminator outputs likelihood in (0,1) of real image

- Discriminator은 기본적으로 분류기에 해당
- p_{data} : 실제 데이터 분포 sample
- D_{θ_d} : 실제 sample에 대한 discriminator 분류 값
- $p(z)$: 실제 데이터 분포 sample
- $D_{\theta_d}(G_{\theta_g}(x))$: fake sample에 대한 디스크리미네이터의 샘플 값

Objective function : Discriminator

$$\max_D V(D, G) = \max_{\theta_d} \left[\mathbb{E}_{x \sim p_{data}} \log D_{\theta_d}(x) + \mathbb{E}_{z \sim p(z)} \log (1 - D_{\theta_d}(G_{\theta_g}(z))) \right]$$

\swarrow
 V 최대화하는
 D 찾기

\downarrow
 G 고정
(given)

진짜 이미지

가짜 이미지

Objective function : Generator

$$\max_{\theta_d} \left[\cancel{\mathbb{E}_{x \sim p_{data}} \log D_{\theta_d}(x)} + \mathbb{E}_{z \sim p(z)} \log (1 - D_{\theta_d}(G_{\theta_g}(z))) \right]$$

학습하지 않음

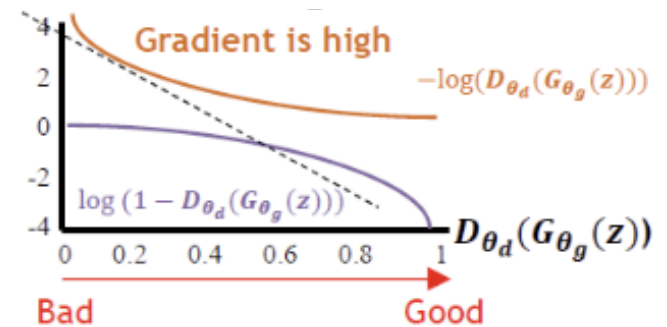


$$\min_G V(D, G) = \min_{\theta_g} \left[\mathbb{E}_{z \sim p(z)} \log(1 - D_{\theta_d}(G_{\theta_g}(z))) \right]$$

D 고정
(given)



$$\max_{\theta_g} \left[\mathbb{E}_{z \sim p(z)} \log(D_{\theta_d}(G_{\theta_g}(z))) \right]$$



Optimizing : Discriminator 최적해

$$D^*(x) = \arg \max_D V(D) = E_{x \sim p_{data}} [\log D(x)] + E_{x \sim p_g} [\log(1 - D(x))]$$

$$p_{data} = p_{generate=g}$$

$$V(G, D) = \int_x p_{data}(x) [\log D(x)] + p_g(x) [\log(1 - D(x))] dx$$

$$D^* = \arg \max_D [p_{data}(x) \log D(x) + p_g(x) \log(1 - D(x))]$$

$$= a \log y + b \log(1 - y)$$

$$\frac{a}{y} + \frac{-b}{1-y} = \frac{a - (a+b)y}{y(1-y)} \quad \Rightarrow \quad \frac{a - (a+b)y}{y(1-y)} = 0$$

$$D^* = \frac{p_{data}(x)}{p_{data}(x) + p_g(x)}$$

Unit 03 | How do GANs work?

Optimizing : Generator 최적해

$$\min_G \max_D V(D, G) = \min_G (D^*, G)$$

$$V(D^*, G) = \mathbb{E}_{x \sim p_{data}} [\log(D^*(x))] + \mathbb{E}_{x \sim p_G} [\log(1 - D^*(x))]$$

$$= \mathbb{E}_{x \sim p_{data}} \left[\log \left(\frac{p_{data}(x)}{p_{data}(x) + p_g(x)} \right) \right] + \mathbb{E}_{x \sim p_G} \left[\log \left(1 - \frac{p_{data}(x)}{p_{data}(x) + p_g(x)} \right) \right]$$

$$= \int_x p_{data}(x) \left[\log \left(\frac{p_{data}(x)}{p_{data}(x) + p_g(x)} \right) \right] dx + \int_x p_g(x) \left[\log \left(\frac{p_g(x)}{p_{data}(x) + p_g(x)} \right) \right] dx$$

$$= -\log 4 + \log 4 + \int_x p_{data}(x) \log \left(\frac{p_{data}(x)}{p_{data}(x) + p_g(x)} \right) dx + \int_x p_g(x) \log \left(\frac{p_g(x)}{p_{data}(x) + p_g(x)} \right) dx$$

$$= -\log 4 + \int_x p_{data}(x) \log \left(\frac{2 \cdot p_{data}(x)}{p_{data}(x) + p_g(x)} \right) dx + \int_x p_g(x) \log \left(\frac{2 \cdot p_g(x)}{p_{data}(x) + p_g(x)} \right) dx$$

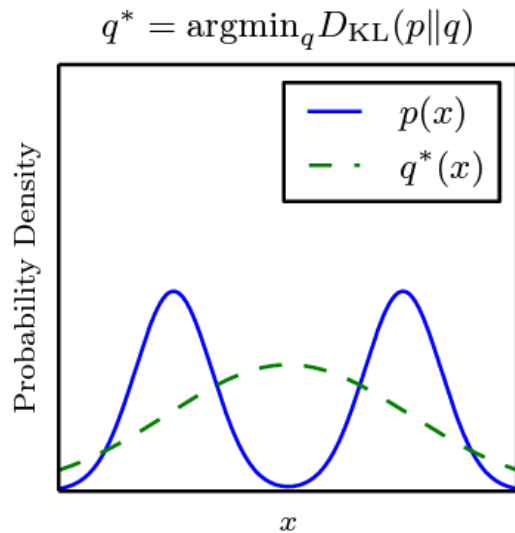
$$= -\log 4 + \textcolor{blue}{KL}(p_{data} || \frac{p_{data} + p_g}{2}) + \textcolor{blue}{KL}(p_g || \frac{p_{data} + p_g}{2})$$

$$= -\log 4 + 2 \cdot \textcolor{red}{JSD}(p_{data} || p_g)$$

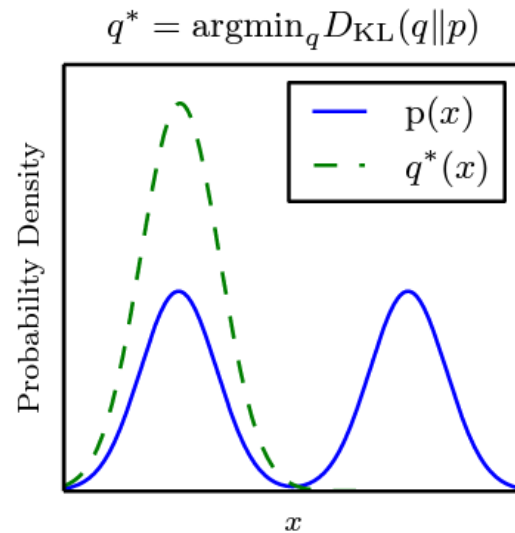
$$\textcolor{blue}{KL}(P || Q) = \int P(x) \log \frac{P(x)}{Q(x)}$$

$$\textcolor{red}{JSD}(P || Q) = \frac{1}{2} KL(P || M) + \frac{1}{2} KL(Q || M)$$

Is the choice of divergence a distinguishing feature of GANs?



Maximum likelihood



Reverse KL

P: 데이터 분포, 가우시안 분포 2개를 합쳐서 하나의 분포로 정의 (bimodal)
Q: single 가우시안 분포

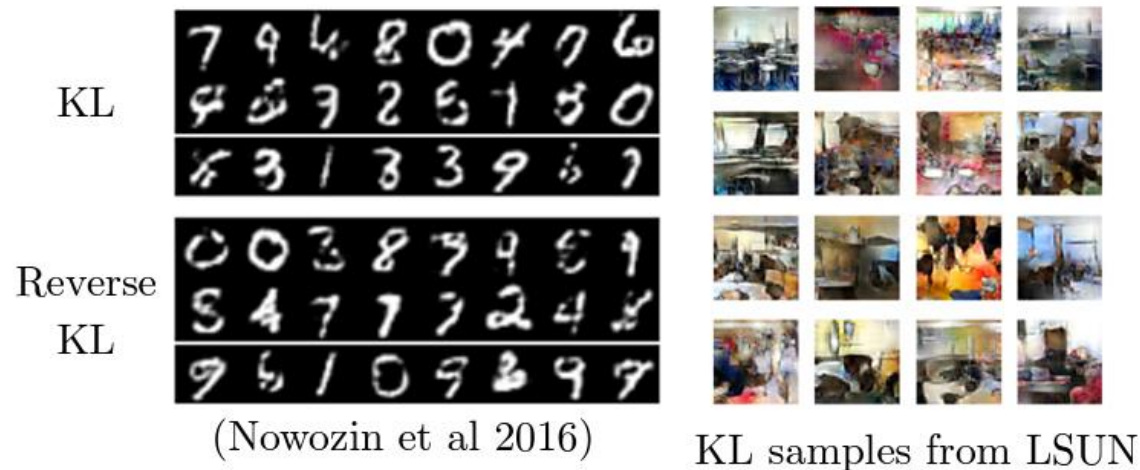
- Maximum likelihood
 - Q가 p의 분포를 따르는 것
 - 모델 분포는 두 mode의 데이터에 대해 평균값을 갖는 가우시안 분포를 가지고 있음.
- Reverse KL
 - P가 Q의 분포를 따르는 것
 - Reverse KL의 경우 기존 데이터가 발생한 확률을 따라가는 형태이기 때문에, unusual한 샘플은 생성하지 않음.
 - Maximum likelihood에 비해 결과가 더 좋아 보임.

만약 Discriminator가 optimal하다면, generator gradient는 maximum likelihood와 동일하다.

- 증명) Goodfellow, I. J. (2014). On distinguishability criteria for estimating generative models. In International Conference on Learning Representations, Workshops Track.

Is the choice of divergence a distinguishing feature of GANs?

증거 1)



F-GAN, model train to minimize KL

증거 2)

Reverse KL로 학습한 GANs 모델의 경우에도 mode collapse 문제를 해결하지 못함.

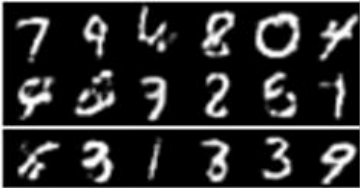
잘 생성할 수 있는 특정 숫자만 잘 할 수 있다.

Is the choice of divergence a distinguishing feature of GANs?


증거 1)

KL

Reverse KL



어떤 Divergence를 선택해도
GAN 성능에 많은 영향을 미치지 않음.



(Nowozin et al 2016) KL samples from LSUN

GANs 모델의 경우에도 mode
| 못함.

F-GAN, model train to minimize KL

1.

Read the paper and organize it in any format such as notation pdf, ppt, etc.

[GANs\(NIPs2014, Ian Goodfellow\) https://arxiv.org/abs/1406.2661](https://arxiv.org/abs/1406.2661)

2.

How to evaluate GANs? Search for some metrics and review them as well.

Reference

Lotter, W., Kreiman, G., and Cox, D. (2015). Unsupervised learning of visual structure using predictive generative networks. arXiv preprint arXiv:1511.06380.

최대우도법(MLE)-공돌이의 수학정리노트 <https://angeloyeo.github.io/2020/07/17/MLE.html>

정보 이론 <https://bskyvision.com/entry/%EC%A0%95%EB%B3%B4%EC%9D%B4%EB%A1%A0-%EC%A0%95%EB%B3%B4%EB%9F%89%EA%B3%BC-%EC%97%94%ED%8A%B8%EB%A1%9C%ED%94%BC%EC%9D%98-%EC%9D%98%EB%AF%B8>

KL divergence https://angeloyeo.github.io/2020/10/27/KL_divergence.html

<https://hwiyoung.tistory.com/408>

KAIST EE, AI602: Recent Advances in Deep Learning; Generative Models II: Explicit Density Models

https://ko.wikipedia.org/wiki/%EA%B0%95%ED%99%94_%ED%95%99%EC%8A%B5



들어주셔서 감사합니다.