


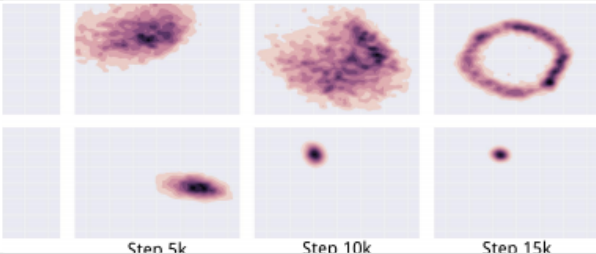
Generative Adversarial Nets

| | |
|---------------------------|-----------------------|
| ≡ 태그 | CVGANGenerative Model |
| ≡ Conference/Journal Name | NeurIPS |
| ≡ Year | 2014 |

초짜 대학원생의 입장에서 이해하는 Unrolled Generative Adversarial Networks (1)


Unrolled Generative Adversarial Networks - L. Metz et al. 2016을 바탕으로 작성한 리뷰 오늘 소개할 논문에서 풀고자 하는 문제는 GAN의 고질적 문제 중 하나인 불안정한 학습의 안정화(stabilize)입니다. GAN의 불안정성은 크게 세 가지 문제로 정리되는데요: Mode collapsing or dropping (generator) Generator and discriminator oscillating during


 <https://jaejunyoo.blogspot.com/2017/02/unrolled-generative-adversarial-network-1.html>



GAN의 기초(GAN,DCGAN,WGAN,CGAN)

작성자 : 광운대학교 전자통신공학과 신민정 생성모델링은 가지고있는 데이터 분포에서 sampling한 것 같은 새로운 데이터를 만드는 것입니다. 생성모델링은 판별모델링과 비교하면 이해하기 쉽습니다. 판별모델링은 label y가 필요한 supervised learning이고, 생성모델링은 label y가 필요없는 unsupervised learning입니다. [label y가 있는 경우도 존

 <https://velog.io/@tobigs-gm1/basicofgan>



1. Technical Novelty (Contribution)

- 통계적 방식에서 기인한 모델링이 아닌 Generator와 Discriminator라는 두 네트워크가 min-max game을 통해 optimal point로 향할 수 있음을 수학적으로 증명.
- vae등 기존의 모델에서 intractable한 부분을 새로운 네트워크 구조로 해결.
- RBM, DBM,GNS,NCE 등등..의 고전적 생성모델과의 대조.

| | Deep directed graphical models | Deep undirected graphical models | Generative autoencoders | Adversarial models |
|-------------------|--|---|--|--|
| Training | Inference needed during training. | Inference needed during training. MCMC needed to approximate partition function gradient. | Enforced tradeoff between mixing and power of reconstruction generation | Synchronizing the discriminator with the generator. Helvetica. |
| Inference | Learned approximate inference | Variational inference | MCMC-based inference | Learned approximate inference |
| Sampling | No difficulties | Requires Markov chain | Requires Markov chain | No difficulties |
| Evaluating $p(x)$ | Intractable, may be approximated with AIS | Intractable, may be approximated with AIS | Not explicitly represented, may be approximated with Parzen density estimation | Not explicitly represented, may be approximated with Parzen density estimation |
| Model design | Nearly all models incur extreme difficulty | Careful design needed to ensure multiple properties | Any differentiable function is theoretically permitted | Any differentiable function is theoretically permitted |

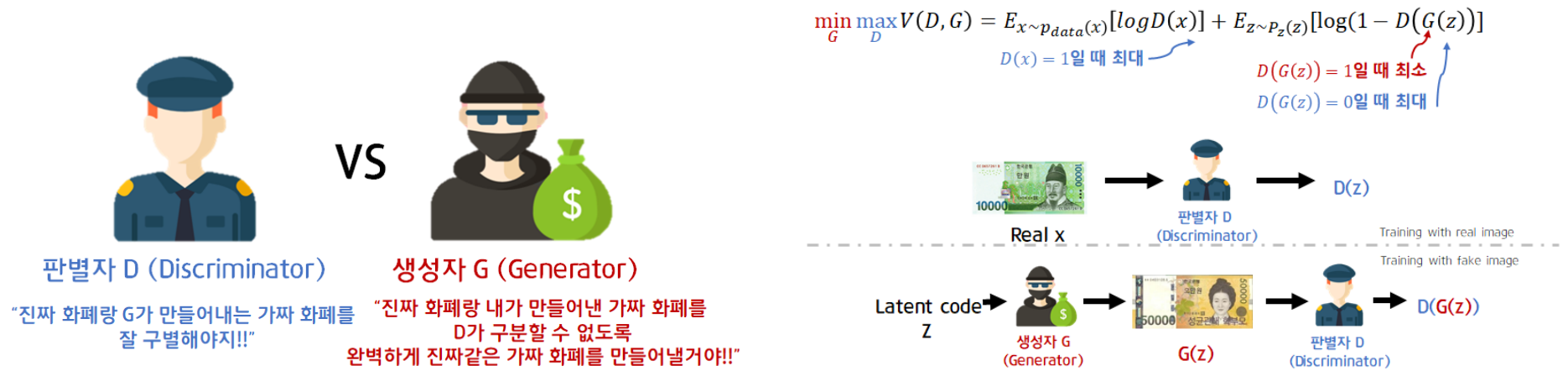
2. Problem Setup

- vae와 그 이후에 탄생하였던 Markov chain 이나 approximate inference등 MLE관점의 통계적 모델링을 사용.
- 그러나 이러한 방식은 활성화함수, dropout 등 기존의 딥러닝 네트워크의 성능을 향상시킬 수 있는 방식등을 적용하기에 어려운 프레임워크를 가짐.

3. Methods (or Solutions)

Notation

- p_g : generator를 통해 만들어지는 data distribution
- $p_z(z)$: Input으로 사용되는 noise z 의 prior
- $G(z; \theta_g)$: 모델 파라미터 θ_g 를 가지는 Generator 에 Input z 를 넣음. [맵핑]
- $D(x; \theta_d)$: 모델 파라미터 θ_d 를 가지는 Discriminator에 Input x 를 넣음. output은 fake or real에 대한 확률값[0,1]으로 [맵핑]
- D 는 fake or real을 가장 잘 판별할수 있도록 학습, G 는 D 를 가장 잘 속일 수 있도록 학습.



How to train

Algorithm 1 Minibatch stochastic gradient descent training of generative adversarial nets. The number of steps to apply to the discriminator, k , is a hyperparameter. We used $k = 1$, the least expensive option, in our experiments.

for number of training iterations **do**

for k steps **do** *매 epoch 상 K번 D를 학습하고 G를 학습한다.*

- Sample minibatch of m noise samples $\{z^{(1)}, \dots, z^{(m)}\}$ from noise prior $p_g(z)$.
- Sample minibatch of m examples $\{x^{(1)}, \dots, x^{(m)}\}$ from data generating distribution $p_{data}(x)$.
- Update the discriminator by ascending its stochastic gradient:

$$\nabla_{\theta_d} \frac{1}{m} \sum_{i=1}^m \left[\log D(x^{(i)}) + \log(1 - D(G(z^{(i)}))) \right].$$

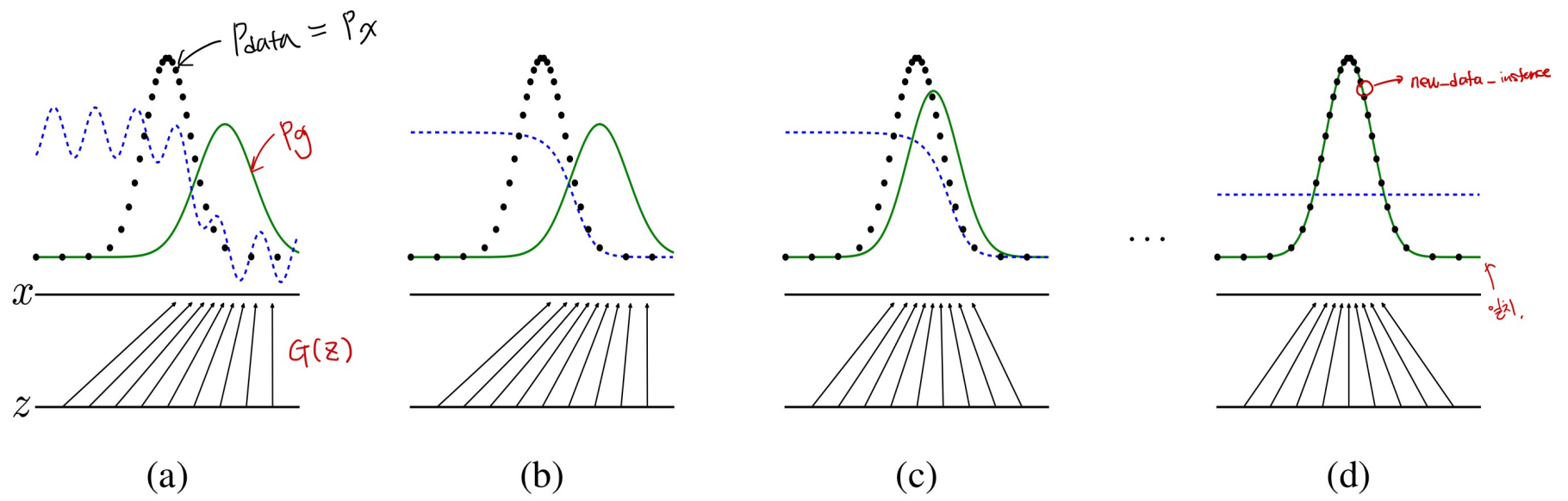
end for

- Sample minibatch of m noise samples $\{z^{(1)}, \dots, z^{(m)}\}$ from noise prior $p_g(z)$.
- Update the generator by descending its stochastic gradient:

$$\nabla_{\theta_g} \frac{1}{m} \sum_{i=1}^m \log(1 - D(G(z^{(i)}))).$$

end for

The gradient-based updates can use any standard gradient-based learning rule. We used momentum in our experiments.



- (a) : Adversarial pair가 수렴에 가까워졌다고 가정시, D는 부분적으로 정확한 분류기(but G가 아직 실제 데이터의 분포와 다르다)
- (b) : D가 최적의 분류기가 된 상태
- (c) : G가 실제 데이터의 분포로 수렴
- (d) : p_g 와 p_{data} 가 일치함. p_g 의 검은점 사이에 있는 데이터는 instance로 추출할 수 있다.(why? 실제 분포를 완벽히 모방했기 때문에)

Find global optimality $p_g = p_{data}$

$$E_{x \sim p_{data}(x)} [\log D(x)] + E_{z \sim p_z(z)} [\log (1 - D(g(z)))]$$

$$\begin{aligned} V(G, D) &= E_{x \sim p_{data}(x)} [\log D(x)] + E_{z \sim p_z(z)} [\log (1 - D(g(z)))] \\ &= \int_x p_{data}(x) \log D(x) dx + \int_z p_z(z) \log (1 - D(g(z))) dz \\ &= \int_x p_{data}(x) \log D(x) dx + \int_x p_g(x) \log (1 - D(x)) dx \end{aligned}$$

$$\begin{aligned} \frac{\partial V(G, D)}{\partial D} &= \frac{p_{data}(x)}{D(x)} - \frac{p_g(x)}{1 - D(x)} = 0 \\ &= \frac{(1 - D(x))p_{data}(x) - D(x)p_g(x)}{D(x)(1 - D(x))} = 0 \\ &= p_{data}(x) - D(x)p_{data}(x) - D(x)p_g(x) = 0 \\ &= p_{data}(x) - D(x)[p_{data}(x) + p_g(x)] = 0 \\ p_D(x) &= \frac{p_{data}(x)}{p_{data}(x) + p_g(x)} \end{aligned}$$

$$C(G) = \max_D V(G, D) = E_{x \sim p_{data}(x)} [\log D^*(x)] + E_{z \sim p_z(z)} [\log (1 - D^*(g(z)))]$$

$$D^*_G(x) = \frac{p_{data}(x)}{p_{data}(x) + p_g(x)} \quad \left(\begin{aligned} &= E_{x \sim p_{data}(x)} \left[\log \frac{p_{data}(x)}{p_{data}(x) + p_g(x)} \right] + E_{x \sim p_g(x)} \left[\log \frac{p_g(x)}{p_{data}(x) + p_g(x)} \right] \\ &= E_{x \sim p_{data}(x)} \left[\log \frac{2 \cdot p_{data}(x)}{p_{data}(x) + p_g(x)} \right] + E_{x \sim p_g(x)} \left[\log \frac{2 \cdot p_g(x)}{p_{data}(x) + p_g(x)} \right] - \log(4) \end{aligned} \right)$$

$$\begin{aligned} KL(P \parallel Q) &= \int_{-\infty}^{\infty} P \cdot \log \frac{P}{Q} \\ &= \int_x p_{data}(x) \log \frac{2 \cdot p_{data}(x)}{p_{data}(x) + p_g(x)} dx + \int_x p_g(x) \log \frac{2 \cdot p_g(x)}{p_{data}(x) + p_g(x)} - \log(4) \\ &= KL \left(p_{data} \parallel \frac{p_{data}(x) + p_g(x)}{2} \right) + KL \left(p_g \parallel \frac{p_{data}(x) + p_g(x)}{2} \right) - \log(4) \\ &= 2 * JSP(p_{data} \parallel p_g) - \log(4) \end{aligned}$$

$$\text{if } p_{data} = p_g, \text{ optimal } -\log(4)$$

4. Experiments

Dataset

- MNIST DATA SET

Experimental Set-Up

- G : ReLu, Sigmoid
- D : maxout, dropout

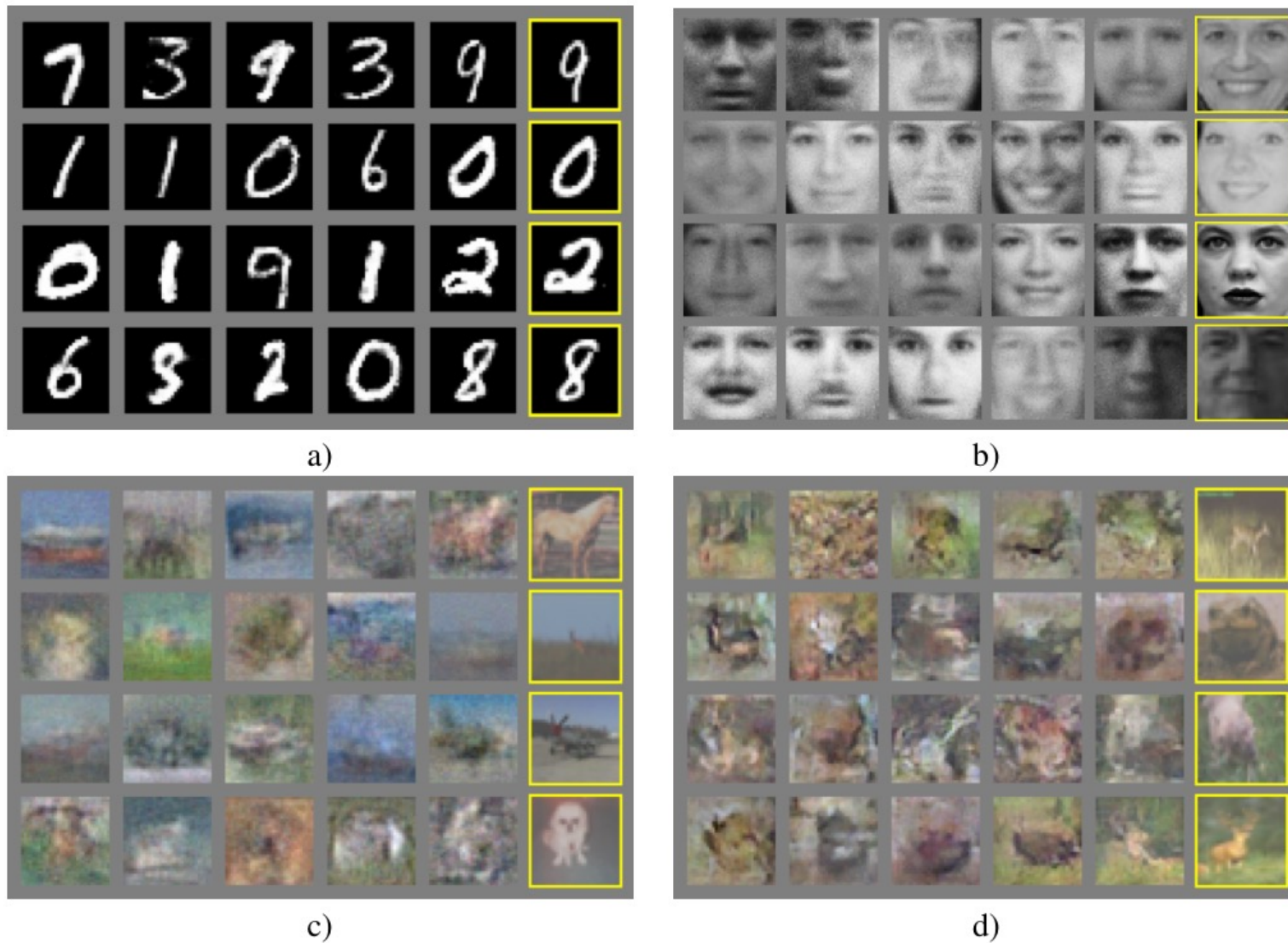


Figure 2: Visualization of samples from the model. Rightmost column shows the nearest training example of the neighboring sample, in order to demonstrate that the model has not memorized the training set. Samples are fair random draws, not cherry-picked. Unlike most other visualizations of deep generative models, these images show actual samples from the model distributions, not conditional means given samples of hidden units. Moreover, these samples are uncorrelated because the sampling process does not depend on Markov chain mixing. a) MNIST b) TFD c) CIFAR-10 (fully connected model) d) CIFAR-10 (convolutional discriminator and “deconvolutional” generator)

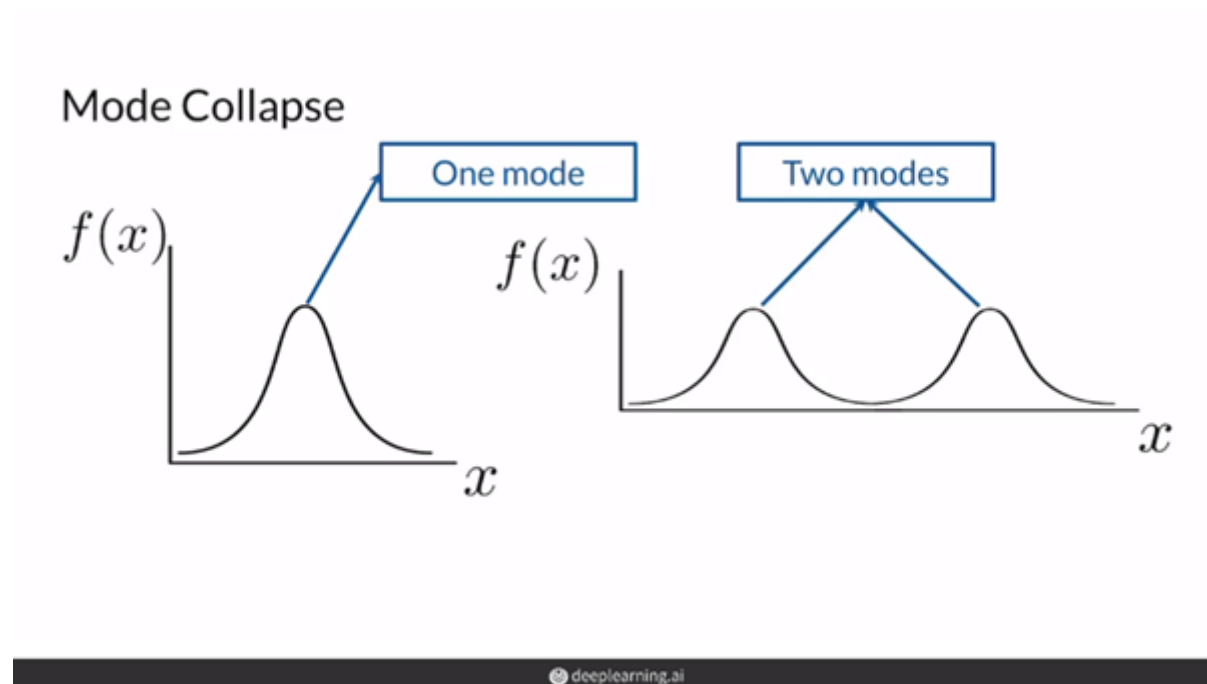
Problem

- mode collapse
 - 기존 GAN에서 풀고자하는 문제는 min-max다. 그러나 실제 학습을 할때 알고리즘상에서 G와 D를 한번씩 번갈아면서 가중치 업데이트가 이루어지기 때문에 neural network는 아래와 같은 max-min problem을 구별하지 못할 것이다.

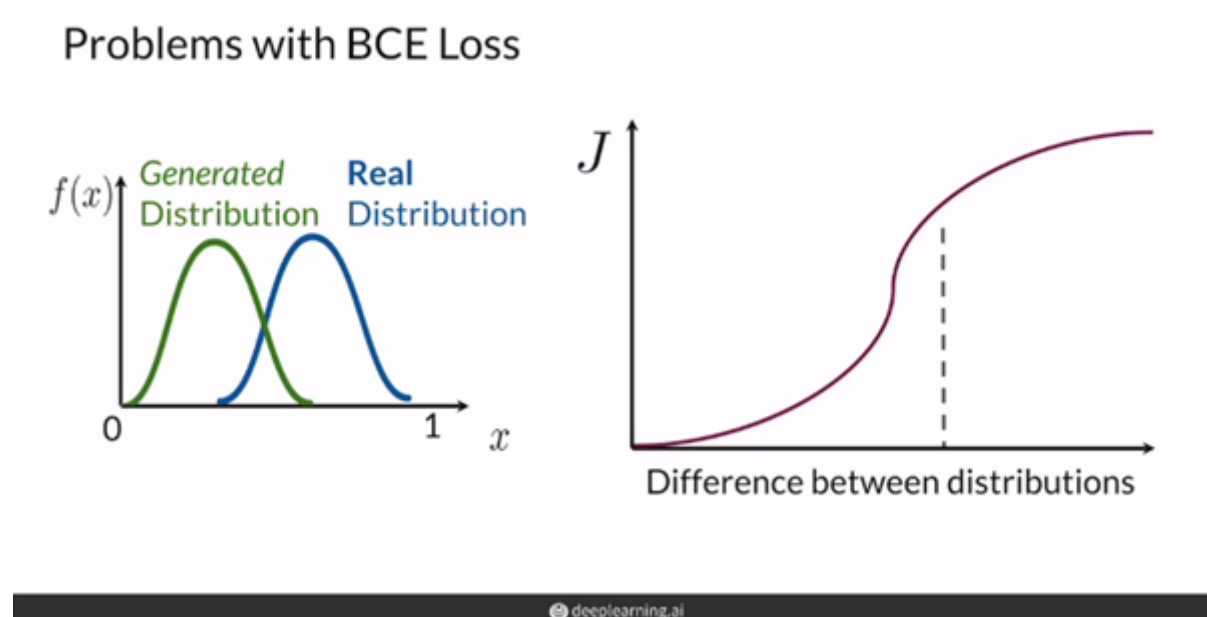
$$G^* = \min_G \max_D V(G, D)$$

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

- 결국 maxmin의 식에서 문제가 발생하게 되는데 G 에 대한 minimization 문제가 선생되기때문에 Generator의 입장에서는 현재 고정되어있는 D가 가장 헛갈려 할 수 있는 샘플 하나만을 보내주면 그만이다. 즉 V를 최소화하는 mode 값 하나만을 내보내게된다.
- 그렇게 되면 학습 이후 Generator를 어떤 noise값을 태워서 뽑아내보면 다 1만 나오는 기현상을 만나게된다.



- gradient vanishing
 - Discriminator와 Generator의 학습이 더 어려움
 - 학습초기에는 Discriminator의 성능이 좋지 않기 때문에 문제가 되지 않음.
 - 그러나 판별자가 잘 학습될수록 0에 가까운 gradient를 발생시켜 더이상 학습 진행 안됨.



5. Thoughts

Strength of the Paper

- 그간의 통계적 모델을 대체하는 혁신적인 방식

Weakness of the Paper (aka things could have been improved)

- 실험이 아쉽다. 여러가지 역동적인 실험을 좀더 보여줄 수 있을거라 생각하는데 아마 gan이 가진 약점을 숨기기위해 그러지 않았을까 싶다.
- mode collapse, gradient vanishing problem등 이후 이것들을 해결하기 위해 여러 gan들이 제안됨.