type 연구

AI summary ✦ Try on this page

(이과대학 수학) 양민석 Nov 25

Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, Yoshua Bengio

GAN 구조 학습

RESEARCH

OVERVIEW

generative model G: 이미지 생성

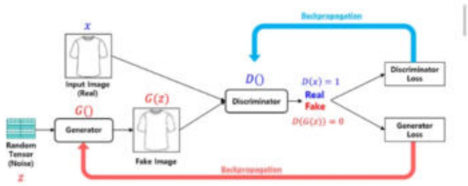
discriminative model D: 이미지 구별

G는 D의 실수를 최대화 시키는게 목표이다. minimax tow-player game과 동일하다.

CONTENT

1. Introduction

딥러닝은 class label을 mapping하는 모델에서 Backpropagation과 Dropout 등을 사용하여 뛰어난 성과를 거두었다. 하지만 deep generative model에서는 MLE에서 intractable 확률 계산을 approximating하는 문제에 어려움을 겪어왔다. 그래서 저자는 새로운 generative model을 제시해서 이러한 어려움을 피해나갈 수 있었다.



adversarial nets에서 discriminative model은 샘플이 model distribution으로부터 왔는지 아닌지 확인하고 generative model은 fake이미지를 생성해서 속이는 역할을 한다.

논문에서 저자는 generative model을 counterfeiters(화폐 위조자), discriminative model을 경찰에 비유했다.

generative model은 랜덤 노이즈 Z를 multilayer perceptron에 넣어 fake image를 만들고 discriminative model은 multilayer perceptron에 데이터를 넣어 모방 데이터인지 진짜 데이터인지 확인한다. generative model과 discriminative model 2개를 합쳐 adversarial nets라고 한다. discriminative loss와 generator loss 2개를 둬서 각각에 대해 Backpropagation을 사용하고 dropout algorithms를 사용하였다. 여기서 중요한 것은 appropximate inference나 Markov chain을 사용하지 않았다는 점이다.

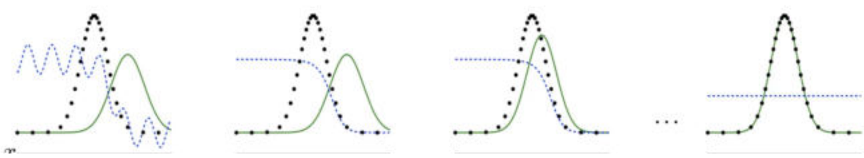
2. Adversarial nets

$$p_g : generator's\ distribution\ over\ data\ x$$
$$p_z(z) : input\ noise\ variables$$
$$G(z;\theta_g) : mapping\ to\ data\ space\ with\ parameters\ \theta_g$$
$$D(x;\theta_d) : represents\ probability\ that\ x\ came\ from\ the\ data\ rather\ than\ p_g\ =\ single\ scaler$$
$$< Train >$$
$$D : maximize\ the\ probability\ of\ assigning\ the\ correct\ label\ to\ both\ training\ examples\ and\ samples\ from\ G$$
$$G : minimize\ log(1 - D(G(z)))$$

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

- Discriminator는 진짜 데이터는 1로 판단하여 첫 번째 항을 0으로, 두 번째 가짜 데이터를 0으로 판단하여 두 번째 항도 0으로 만드는 것이 목표이다.
- 즉, Discriminator가 얻을 수 있는 이상적인 최대값은 0이다.
- 반대로 Generator는 두 번째 항의 D(G(z)) = 1로 판단하게 하여 두 번째 항 전체를 -∞ 만드는 것이 목표이다.
- 즉, Generator가 얻을 수 있는 이상적인 최솟값은 -∞이다.
- 학습을 계속하다보면 Discriminator도 어느 것이 진짜 인지 가짜인지 구분하기 힘들어져 D(x) = 0.5에 수렴하게 된다.

2.1 Theoretical analysis of adversarial nets



(a) (b) (c) (d)

검은색: p_data, 파란색: discriminative distribution, 초록색: generative distribution

- 처음에는 원본 데이터에 대해서는 1에 가까운 값을 fake 이미지에 대해서는 0에 가까운 값으로 구별을 한다.

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

- D의 결과값이 위의 수식으로 수렴한다. 추후 증명예정.
- 그러다가 generative distribution이 점점 원본 데이터의 분포와 같아지게 되고 가면 갈수록 진짜와 가짜를 구별해내지 못하게 되어 0.5로 수렴한다.
- 결론적으로 P_data = P_g일 때 global optimum을 가진다.

2.2 Theoretical Results

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** ---------- 1

  for k steps **do** ---------- 2

    • Sample minibatch of m noise samples $\{z^{(1)}, \ldots, z^{(m)}\}$ from noise prior $p_g(z)$. ---------- 3

    • Sample minibatch of m examples $\{x^{(1)}, \ldots, x^{(m)}\}$ from data generating distribution $p_{data}(x)$. ---------- 4

    • Update the discriminator by ascending its stochastic gradient:

$$\nabla_{\theta_d} \frac{1}{m} \sum_{i=1}^{m} \left[\log D\left(x^{(i)}\right) + \log\left(1 - D\left(G\left(z^{(i)}\right)\right)\right) \right].$$ ---------- 5

  end for

  • Sample minibatch of m noise samples $\{z^{(1)}, \ldots, z^{(m)}\}$ from noise prior $p_g(z)$.

  • Update the generator by descending its stochastic gradient:

$$\nabla_{\theta_g} \frac{1}{m} \sum_{i=1}^{m} \log\left(1 - D\left(G\left(z^{(i)}\right)\right)\right).$$ ---------- 6

end for

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

1. epoch만큼 반복문 수행
2. k번만큼 discriminator gradient ascending 시행(max) → 저자는 k = 1로 둠
3. 1번만큼 generator gradient descend 시행(min) → 실제로는 gradient ascend 사용

2.3 Global Optimality

Proposition 1) G가 고정된 경우, 최적의 discriminator D,

$$D_G^*(x) = \frac{P_{data}(x)}{P_{data}(x) + P_g(x)}$$

Pf) $V(G,D) = \int_x P_{data}(x)\log D(x)\,dx + \int_z P_z(z)(1 - D(g(z)))\,dz$

$= \int_x [P_{data}(x)\log D(x)) + P_g(x)(1 - \log D(x))]\,dx$

For $\forall (a,b) \in \mathbb{R}^2 \setminus \{0,0\}$ $\quad y \to a\log y + b\log(1-y)$

극미분: $\frac{a}{y} - \frac{b}{1-y} = 0, \quad a(1-y) - by = 0, \quad y(a+b) = a, \quad \boxed{y = \frac{a}{a+b}}$

$\therefore D_G^*(x) = \frac{P_{data}(x)}{P_{data}(x) + P_g(x)}$

$C(G) = \max_D V(G,D)$

$= E_{x\sim P_{data}}[\log D_G^*(x)] + E_{z\sim P_z}[\log(1 - D_G^*(G(z)))]$

$= E_{x\sim P_{data}}[\log D_G^*(x)] + E_{x\sim P_g}[\log(1 - D_G^*(x))]$

$= E_{x\sim P_{data}}\left[\log \frac{P_{data}(x)}{P_{data}(x)+P_g(x)}\right] + E_{x\sim P_g}\left[\log \frac{P_g(x)}{P_{data}(x)+P_g(x)}\right]$

Theorem1) The global minimum of the virtual training criterion $C(G)$ is achieved if and only if $P_g = P_{data}$. At that point $C(G)$ achieves the value $-\log 4$. G입장에서는 C(G)를 최소화해야 함

Pf) ($\Leftarrow$) $P_g = P_{data} \to D_G^*(x) = \frac{1}{2}$

$C(G) = \log\frac{1}{2} + \log\frac{1}{2} = -\log 4$

$D_{KL}(P\|Q) = \sum P(i)\log\frac{P(i)}{Q(i)}$

$JSD(P\|Q) = \frac{1}{2}D_{KL}(P\|M) + \frac{1}{2}D_{KL}(Q\|M) \geq 0, \quad M = \frac{P+Q}{2}$

($\Rightarrow$)

$C(G) = C(G) + \log 4 - \log 4$

$= -\log 4 + E_{x\sim P_{data}}\left[\log\frac{2\times P_{data}(x)}{P_{data}(x)+P_g(x)}\right] + E_{x\sim P_g(x)}\left[\log\frac{2\times P_g(x)}{P_{data}(x)+P_g(x)}\right]$

$= -\log 4 + KL\left(P_{data} \middle\| \frac{P_{data}+P_g}{2}\right) + KL\left(P_g \middle\| \frac{P_{data}+P_g}{2}\right)$

$= -\log 4 + 2 \times JSD(P_{data}\|P_g) \geq -\log 4$

Where $P_{data} = P_g$

- 2개의 증명을 통해 minmax problem이 global minimum에서 unique solution을 가지고 어떠한 조건이 만족하면 그 solution의 값으로 수렴한다.

2.4 Convergence of Algorithm1

Proposition 2. *If G and D have enough capacity, and at each step of Algorithm 1, the discriminator is allowed to reach its optimum given G, and p_g is updated so as to improve the criterion*

$$\mathbb{E}_{x\sim p_{data}}[\log D_G^*(x)] + \mathbb{E}_{x\sim p_g}[\log(1 - D_G^*(x))]$$

then p_g converges to p_{data}

Proof. Consider $V(G,D) = U(p_g, D)$ as a function of p_g as done in the above criterion. Note that $U(p_g, D)$ is convex in p_g. The subderivatives of a supremum of convex functions include the derivative of the function at the point where the maximum is attained. In other words, if $f(x) = \sup_{\alpha\in\mathcal{A}} f_\alpha(x)$ and $f_\alpha(x)$ is convex in x for every α, then $\partial f_\beta(x) \in \partial f$ if $\beta = \arg\sup_{\alpha\in\mathcal{A}} f_\alpha(x)$. This is equivalent to computing a gradient descent update for p_g at the optimal D given the corresponding G. $\sup_D U(p_g, D)$ is convex in p_g with a unique global optima as proven in Thm 1, therefore with sufficiently small updates of p_g, p_g converges to p_x, concluding the proof. ■

In practice, adversarial nets represent a limited family of p_g distributions via the function $G(z;\theta_g)$, and we optimize θ_g rather than p_g itself. Using a multilayer perceptron to define G introduces multiple critical points in parameter space. However, the excellent performance of multilayer perceptrons in practice suggests that they are a reasonable model to use despite their lack of theoretical guarantees.

- 명제2: G, D가 충분한 용량을 갖고 Algorithm1 각 step에서 discriminator은 주어진 G에 대하여 위에서 제시한 최적의 값으로 도달하고 p_g를 업데이트 해서 criterion을 개선하면 p_g → p_data로 수렴한다.
- 정리하자면 주어진 Generator에 대해 D가 optim으로 수렴하고, p_g를 업데이트하면 p_g → p_data로 수렴한다는 이야기이다.

$$V(G,D) = U(p_g, D) : p_g\text{의 함수로 생각}$$
$$U(p_g, D) : convex\ in\ p_g$$
$$\text{볼록 함수에서 최상부의 도함수} = \text{최대값에 도달한 지점의 도함수}$$
$$if\ f(x) = sup_{\alpha\in A} f_\alpha(x)\ and\ f_\alpha(x)\ is\ convex\ in\ x\ for\ every\ \alpha, then\ \partial f_\beta(x) \in \partial f\ if\ \beta = arg\ sup_{\alpha\in A} f_\alpha(x)$$

- 정확하게 이해가 가지는 않는다.
- p_g에 대한 gradient descent update와 동일하다.
- convex 하므로 global optimum을 가지고 p_g는 결국 p_data로 수렴하게 된다.
- Subderivative 정의

수학에서 **하방미분**(subdifferential, subderivative)은 미분을 일반화하여 미분가능하지 않은 볼록 함수에 적용할 수 있도록 하는 방법이다. 볼록 최적화 등 볼록 함수를 연구하는 해석에서 중요하게 사용된다.

정의 [편집]

볼록함수 $f : I \to \mathbb{R}$가 있을 때, I의 점 x_0에서의 **하방미분계수**는,

$$f(x) - f(x_0) \geq c(x - x_0)$$

가 I의 모든 점 x에 대해 성립하게 하는 실수 c를 가리킨다.

3. Experiments

- G: ReLU(Rectifier linear activations) + sigmoid 혼합 사용
- D: maxout 사용, 학습시킬 때 dropout 사용

- 장점: backpropagation만 사용, inference 필요없음, 이미지가 더 선명함
- 단점: p_g(x)가 명시적으로 존재하지 않음, D와 G가 균형을 잘 맞춰서 성능이 향상되어야함. 만약 G가 D가 너무 발전하기 이전에 발전해버리면 G가 z 데이터를 너무 많이 붕괴시키기 때문이다.

단어장

adversarial: 적대적인

pitted: 싸우게 하다

counterfeiters: 화폐 위조자

SEARCH

REFER

#	Title	Year	link
[1]			

Proposition 1) G가 고정된 경우, 최적의 discriminator D,

$$D_G^*(x) = \frac{P_{data}(x)}{P_{data}(x) + P_g(x)}$$

Pf) $V(G, D) = \int_x P_{data}(x) \log D(x))dx + \int_z P_z(z)(1 - D(G(z)))dz$

$= \int_x [P_{data}(x) \log D(x)) + P_g(x)(1 - \log D(x))]dx$

For $\forall (a,b) \in \mathbb{R}^2 \setminus \{0,0\}$ $\quad y \to a \log y + b \log(1-y)$

→ 미분: $\frac{a}{y} - \frac{b}{1-y} = 0$, $\quad a(1-y) - by = 0$, $\quad y(a+b) = a$, $\boxed{y = \frac{a}{a+b}}$

$$\therefore D_G^*(x) = \frac{P_{data}(x)}{P_{data}(x) + P_g(x)}$$

$$C(G) = \max_D V(G, D)$$

$$= E_{x \sim P_{data}}[\log D_G^*(x)] + E_{z \sim P_z}[\log(1 - D_G^*(G(z)))]$$

$$= E_{x \sim P_{data}}[\log D_G^*(x)] + E_{x \sim P_g}[\log(1 - D_G^*(x))]$$

$$= E_{x \sim P_{data}}\left[\log \frac{P_{data}(x)}{P_{data}(x) + P_g(x)}\right] + E_{x \sim P_g}\left[\log \frac{P_g(x)}{P_{data}(x) + P_g(x)}\right]$$