Paper summary

AI VISION Lab

1. 공부한 논문의 제목, 게재된 학회 혹은 저널 등 논문 기본 정보를 적으세요.
   A. 이름: Generative Modeling by Estimating Gradients of the Data Distribution
   B. 저널: NeurIPS
   C. 도메인: Generative model, NCSN
   D. 출판연도: 2019.7.12
   E. 저자: Yang Song, Stefano Ermon

2. 논문에서 제안한 알고리즘 및 프레임워크에 대해 본인이 이해한대로 다이어그램을 그려보세요. **논문 Figure를 그대로 따라 그리면 안됩니다.**

   A. [선행연구] Generative model을 설계하는 가장 큰 요인은 P_{data}분포를 따르는 데이터를 Sampling하고자 함에 있음. 같은 목표를 가지고 설계된 다양한 Generative model은 크게 두가지(논문 출판 시기 기준) 나눌 수 있다.

      i. Log-likelihood based: $p(x;\theta) = \frac{1}{z(x)}\, q(x;\theta)$ 의 형태를 구하고자 함. (Transformer-Auto Regression에서는, Chain형식으로 계산되는 구조로 (output -> 다음의 input) $z(x)$를 정의할 필요가 없다는 장점이 있음. 다음 예시로는 Energy based model같은 Surrogate loss 기반의 접근방식으로, 가장 큰 특징으로는 $z(x)$를 구하는 것이 어렵다는 점 때문에 이를 우회하는 구조를 가지고 있음.)

      ii. GAN: Likelihood based가 아니며 따라서, PDF를 구하지 않는 구조로, divergences를 최소화하는 형식으로 계산함.

   B. [선행연구] 앞서 두가지 model들은 특정한 구조를 사용해야 한다는 한계(예: Transformer, EBM 등)와 학습이 불안정하며, Implicit한 구조로 인해 다른 모델과의 성능 비교가 어렵다는 한계를 가짐(예: GAN)

   C. 본 논문에서는 이러한 한계를 해결할 수 있는, 새로운 Generative model의 구조를 제안함. (NCSNs, Noise Conditional Score Networks)

D. Model의 설계에서 목표로 하는 것은. 모든 데이터 공간(Low – High dimensional space)에서 일관적인 성능을 뽑아내는 것이라고 할 수 있음. 그러나 해당 과정을 위한 가장 큰 문제는 Low-dimensional에 있는 Data의 양이 매우 적다는 점임 (심지어 Score Function은 데이터의 미분 + log 로 구성됨 -> 즉, Low dimension이 실제로 들어가는 경우 값이 사라지는 경우도 있음) 이러한 문제는 단순히 Model의 low-dimension에 대한 표현력이 낮다는 것을 의미할 뿐만 아니라 Sampling(이 논문에서는 Langevin Dynamics) 성능을 일관적이지 못하게 함.

E. 따라서, Low-Dimensional한 경우에도 강건한 모델의 성능을 위해 두가지 방안을 제안함.
   i. Perturbed Data
   ii. Annealed Langevin Dynamics

F. 앞서 이야기한 바와 동일하게, Low-dimension에 대한 표현력이 떨어지는 문제를 data + noise를 통해 해결하고 있음. (따라서, Denoising Score Matching을 목적함수로 이용)

G. Langevin Dynamics는 Score Function(p_{data})만 있다면, Sampling을 할 수 있다는 장점이 있으나, 논문에서는 Annealed Langevin Dynamics가 필요하다고 설명하고 있음. $p_{data} = \pi * p_1(x) + (1-\pi) * p_2(x)$(e.g., $\pi = 0.7$)로 p_{data} 분포를 구성했을 때, 이상적인 결과는 Score Function (p_{data})의 Vector field에서 p_1의 분포가 p_2보다 선명한 경우라고 할 수 있음. 그러나, 한번의 Langevin Dynamics에서는 Score Function를 사용해서 구하다 보니, 결과적으로 $\pi : Scale\ Factor$가 제거되는 문제가 발생함. 즉, p_{data}를 잘 설명하지 못하고 있다고 할 수 있음.

H. 따라서, Annealed Langevin Dynamics가 필요한 것인데, 간단하게 이야기하자면 Langevin Dynamics에 대해서 나온 분포를 토대로 다시 Langevin Dynamics를 진행하고 다시 진행한다면 (논문에서, 10번) 모든 Dimension을 잘 반영하고 있는 model sampling이 나올 것이라는 내용임.

I. 해당 논문에서는 결과적으로, 기존의 Generative model의 한계를 어느정도 해결하면서 우수한 성능(CIFAR-10, MNIST에서 GAN대비)의 model을 제안함. (안정적인 학습- GAN보다, Scalable – 다른 Auto Regressive, EBM 대비)

3. 본인이 생각하는 이 논문의 장점이 무엇이라고 생각하나요? **논문 Contribution bullet을 그대로 따라 적으면 안됩니다.**
   A. 문제점을 정확히 파악하고 (Generative model의 한계, Langevin dynamics) 실험적으로 보여줌과 동시에 해결방안을 제시함. 그렇게 제시한 해결방안이 실제 실험에서 우수한 성능임을 보임으로써 우수한 논문임을 보임.

4. 이 논문을 읽으면서 느낀 점, 혹은 배운 점이 있으면 적어보세요.
   A. 기존 모델의 문제점을 파악하고 -> 보완점을 제시한 것인지, 새로운 것을 적용하기 위해 문제점을 찾은 것인가가 궁금함.

5. 이 논문의 한계점이 있다면 무엇이라고 생각하나요?
   A. Sampling – 10번(NFE: 10) 즉, 속도가 느릴 수밖에 없음
   B. 뿐만 아니라, Score Matching! = Model. (서로 미분한 값에 대한 비교이므로)

6. 본인의 연구에 접목시켜볼 점이 있을지 생각하고 적어보세요.
   A. Diffusion 선행 연구

7. 본 Summary를 작성하는 과정에서 생성형AI를 사용했나요?
   A. 아니요

날짜: 2025-06-27

이름: 신준원

2025.06.27 Generative modeling by Estimating Gradients of Data distribution

Langevin Dynamics

- 분자 동역학 + SDE (확률 미분 방정식)
- 일종의 MCMC 시뮬레이션

ex) score matching

① explicit 한 구조 필요.

Introduction

당시 (2019) Generative model
- likelihood-based → ex) Autoregressive, surrogate loss
  - → chain rule 기반 Factorization 하여도 P(x)를 표현 가능. (Z 필요X)
- GAN → Implicit 한 구조.

단점 :
- likelihood-based → special 구조 ex) RNN, transformer or surrogate loss
- GAN : Implicit → 학습 불안정
  - 심지어 GAN A model, GAN B model → 학습, 비교 어려움.

추가 ⇒ 목적함수 → noise contrastive estimation, minimum probability flow 등.
→ low Dimensional data에만 효과적.

기존의 Generative model method : 한계 or 제한 존재

① high dimension 만 성능↑, low dimension 만 성능↑ → 모든 Dimension에서 좋은 성능 필요.

high Dimensional vs low Dimensional

개념 : 차원 크기.

→ Generative modeling 관점

→ 이미지의 차원 (ex) mnist → 784 pixel) 로 고차원이지만 실제 변동, 중요한 pixel (차원) 는 low Dimensional.

즉, low Dimensional에 대해서 제대로 표현해줄 필요가 있음

① high Dimensional

→ 또한 필요
즉, 이론적으로는 모든 Dimension에 대해 설명가능한 Generative model 제시

그동안 low Dimension 에 대한 Score Matching은 하지 않음.

→ low Dimension에 대한 부분 → 비어있음 (또는 0)

따라서 본 논문에서는 noise + Data 를 통해 학습

→ large noise ⇒ low Dimension화

(Sampling) Langevin dynamic 에 적용시

→ noise 를 크게 한번에 학습 진행 (low Dimension)

→ 점차 noise 를 줄여가며 학습 진행 (High Dimension)

모든 dimension 에 대하여 값을 가진 sample 생성

+ 제약 X, 모든 Score network에 적용가능 / 최적화된 비교 진행 가능

likelihood-based 한계 GAN 한계

Score based generative modeling → matrix

$$\frac{1}{2}\mathbb{E}_{p_{data}}\left[\| S_\theta(x) - \nabla_x \log p_{data}(x)\|_2^2\right]$$

$$\approx \mathbb{E}_{p_{data}}\left[\mathrm{tr}(\nabla_x S_\theta(x)) + \frac{1}{2}\|S_\theta(x)\|_2^2\right]$$

식을 전개 후 상수처리

notation 이유

Condition: Jacobian of $S_\theta(x)$ 가 대칭 행렬

→ Vector valued function

$S_\theta^*(x) = \nabla_x \log p_{data}(x)$

계산에 대한 비용 감소를 위

tr: trace → $\sum_{i=1}^{d} \frac{\partial}{\partial x_i} \log p_\theta(x) \rightarrow \mathrm{tr}(\nabla_x^2 \log p_\theta(x))$

Background

Generative modeling → P_{data} 의 분포에서 Sampling 하는 것

sampling을 하기 위해선? → PDF (P_{data})를 알아야함.

이때 P_{data} 를 구할 수 있으면 $p_\theta(x) = \frac{1}{Z(\theta)} \cdot q_\theta(x)$

일반적인 확률분포 = 1 → 총합 확률 합을 1로 만들어주는 역할 $Z(\theta)$

그런데 $Z(\theta)$ 를 구하기 어려움 → 따라서 원하는 다양한 Solution 존재.

Denoising Score matching

가장 큰 장점 → $tr(\nabla_x S_\theta(x))$ 를 매우 편한 대체 가능

? 어떻게 대체하는가 $tr(\nabla_x S_\theta(x))$

noise $\sim N(0, I\sigma)$ Gaussian, $\tilde{x} = x + noise$

이때 → noise가 Gaussian을 따르고 $tr(\nabla_x S_\theta(x)) \to \frac{\|x - \tilde{x}\|}{\sigma^2}$

Objective → $\frac{1}{2} E_{q_\sigma(\tilde{x}|x) p_{data}(x)} \left[\| S_\theta(\tilde{x}) - \nabla_{\tilde{x}} \log q_\sigma(\tilde{x}|x) \|_2^2 \right]$

Sliced Score Matching

장점 ⇒ Add noise 필요없음.

단점 → 계산량 ×4

↳ 위를 2번을 해결하는 다른 방식의 Score Matching

$E_{(v)} E_{p_{data}} \left[v^T \nabla_x s_\theta(x) v + \frac{1}{2} \| s_\theta(x) \|_2^2 \right]$

P_v : Simple distribution of Random Vector

e.g. multivariate normal

Sampling with Langevin dynamics

Langevin dynamics → $\nabla_x \log p(x)$ 만으로 Sampling 가능

$$\tilde{x}_t = \tilde{x}_{t-1} + \frac{\epsilon}{2}\nabla_x \log p(\tilde{x}_{t-1}) + \sqrt{\epsilon}\, z_t$$

→ Diffusion forwarding과 유사함

① $\tilde{x}_0 \sim \pi(x)$, π → prior distribution

② $z \sim N(0, I)$

③ $\tilde{x}_T = p(x)$ ($\epsilon \to 0$ & $T \to \infty$)

④ Metropolis-Hastings 필요 ($\epsilon > 0$ & $T < \infty$)

→ error 수정위해 필요 (그러나 skip하는 경우도 존재)

⇒ ϵ이 충분히 작고, T가 충분히 크다면 해당 error 무시가능

→ 본 논문에서도 무시

Metropolis-Hastings 란?

-one of MCMC

(notation은 대부분 해당 알고리즘을 따르고 있음)

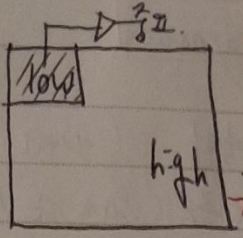
Sampling 하려는 분포 $\pi(x)$ ($p_{data}(x)$

x_t 시점에서 생성가능한 분포 $= q(y \mid x_t)$

x_t → y 로의 진행 (판단기준 $\pi(x)$가 크거나 작은.)

The manifold hypothesis

real world data. →



즉, 우리의 high-dim에서는 low에서만 학습 가능

score-based generative model

문제점 1. $\nabla_x \log p_{data}(x)$ = high Dimensional
따라서 X가 low Dimensional 한 경우 표현이려움

문제점 2. General 한 Score Matching objective function의 경우
→ $d \in \mathbb{R}^n$ 의 꼴에 의해 학습함.
즉, $d \in \mathbb{R}^n$ 이 아닌 e.g) low-dimension 경우 문제가능

e.g.) Resnet with CIFAR.

Sliced Score Matching (장점 → noise X. 단점 ⇒ $d \in \mathbb{R}^n$ 에서만 가능)

→ X 그대로 입력했을때 = 학습 불안정

→ X + noise = 학습 안정.

문제점 정리

→ low-dimensional 에서도 학습 가능하도록.

Data 부족.

⊕ Score → vector.

↳ data와 SM 사이의 비교를 통해서

→ high Dimension 부분에서만 비교한 부분을 학습가능

Slow mixing of Langevin dynamics

문제 → low dimension에 취약함. (그래서 구역이 나뉘짐)

결과 → Langevin dynamics 성능 좋은 ($p_{data} \approx p_\theta$) 어렵다.

→ $p_{data} = \pi p_1(x) + (1-\pi) p_2(x)$ 이때 $\pi \sim N(0, I)$

score function → $\log p_{data} = \log p_1(x) + \log p_2(x)$

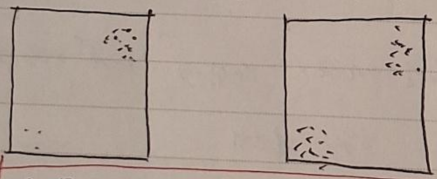
→ $\nabla \log p_{data} = \nabla \log p_1(x) + \nabla \log p_2(x)$

score function은 두개의 distribution으로 표현되므로.

→ π에 해당하는 부분이 사라지는 결과를 얻을 수 있음

~~그렇다면 π~~

결과적으로 → mixing 했을때



3:7 → π가 생략되므로 Langevin에서는 → 5:5처럼 결과가 나옴

가속 방법 → Langevin 한계 존재

→ Annealed 하게 해보자

noise Conditional Score Network

발상 → noise 추가해서 score modeling 하면 좋겠다

① 모든 case + noise → low manifold 현상 아니라 full dimensional space 에 존재가능.

② large distribution of Gaussian.
→ low dimension 에 효과적임.

③ multiple noise level.
→ True Data distribution 으로 수렴하게 만듦.

직관화한 Network (training + Sampling)

① 다양한 noise + data 사용

② 한개의 network, 모든 noise level에 대해서 ~~추정~~

① Anneal down the noise level

Detail

$\{\sigma_i\}_{i=1}^{L}$, $\left[\sigma_1 > \sigma_2 > \cdots > \sigma_L\right.$

$\left.\frac{\sigma_1}{\sigma_2} = \cdots = \frac{\sigma_{L-1}}{\sigma_L} > 1\right]$ ~ 같은 비율 → $\sigma_1 \cdot \square = \sigma_2$

$$q_\sigma(x) \triangleq \int p_{data}(t)\mathcal{N}(x|t, \sigma^2 I)\,dt$$

$$\forall \sigma \in \{\sigma_i\}_{i=1}^{L} : \boxed{s_\theta(x, \sigma)} \approx \nabla_x \log q_\sigma(x)$$

(noise Conditional SN) NCSN → 가능 → $x \in \mathbb{R}^D$

→ ~~focus~~ focus on Image generation

- U-Net
- Instance normalization (Conditional)

Learning NCSNs via Score Matching

Sliced $\times \frac{1}{2}$ = denoising

어떤 Score Matching : denoising Score Matching → 속도!

denoising Score Matching → $\frac{1}{2} E_{p_{data}(x)} E_{\tilde{x} \sim N(x, \sigma^2 I)} \left[\left\| s_\theta(\tilde{x}, \sigma) + \frac{\tilde{x} - x}{\sigma^2} \right\|_2^2 \right]$

↓

sigma (모든 level) 처리필요

$$L(\theta; \{\sigma_i\}_{i=1}^{L}) \triangleq \frac{1}{L} \sum_{i=1}^{L} \lambda(\sigma_i) \, l(\theta; \sigma_i).$$

$\lambda(\sigma_i)$ → Coefficient function

$s_\theta(x, \sigma) \approx s_\theta^*(x, \sigma)$

경험적인 기준 ← 수치 → $\lambda(\sigma) = \sigma^2$.

Langevin dynamics inference

$$\tilde{x}_t \leftarrow \tilde{x}_{t-1} + \frac{\alpha_i}{2} s_\theta(\tilde{x}_{t-1}, \sigma_i) + \sqrt{\alpha_i} z_t$$

$\alpha_i = \epsilon \cdot \sigma_i^2 / \sigma_L^2$ (점차 감소) → $\sigma_L \approx 0$

σ_1 일때 잘 동작했다면 → low dim일때 주의 ↑
(noise ↑)

☆ $\{\sigma_1\}$을 기반으로 $\{\sigma_2\}$ 분포 확인

즉, $\{\sigma_1\}$도 $\{\sigma_2\}$에서는 high Dimensional 상태.

(L=10 step)

VAE, GAN 의 성능을 뛰어넘은 어려움 → MCMC 기반의 다양한 접근 시도