Paper summary

AI VISION Lab

1. 공부한 논문의 제목, 게재된 학회 혹은 저널 등 논문 기본 정보를 적으세요.
   - **A. 이름: Deep Unsupervised Learning using Nonequilibrium Thermodynamics**
   - **B. 저널: ICML**
   - **C. 도메인: Diffusion**
   - **D. 출판연도: 2015**
   - **E. 저자: Jascha Sohl-Dickstein et al.**

2. 논문에서 제안한 알고리즘 및 프레임워크에 대해 본인이 이해한대로 다이어그램을 그려보세요. **논문 Figure를 그대로 따라 그리면 안됩니다.**
   - A. **이 논문은 VAE의 장점을 흡수하되, 단점을 보완하는 논문임.**
   - B. **[선행연구] VAE전, Generative model들의 경우, Inference(X -> Z) 과정과, Model의 Probability 설계는 독립된 구조였음**
   - C. **[선행연구] 그러나, VAE의 경우, Inference – z를 통한 Generate 과정을 결합하여 해당 구조를 깨뜨리되, 우수한 성능을 보였음.**
   - D. **다만, VAE의 단점은 Learning 과정에서 Inference + Generate과정이 결합되었다는 점임(학습이 한꺼번에 진행)**
   - E. **Diffusion에서는 이런 한계를 해결하고자 output 까지의 Process를 Forward + Reverse process로 구분함.**
   - F. **[Forward Process] 해당 과정에 대해서는 학습을 하지 않으며, $Input: x_0$ *를 $Gaussian\ Kernel$ 에 통과시키는 과정으로 구성됨.***
   - G. **이는, $q(x_t) * \pi\,(t \mid t-1) = q(x_{t+1})$, t는 0, T(T = 1000)의 값. 이때, Kernel은 $\pi \sim N(0, I)$를 따르며, $q(x_{0,1,2,\dots,T})$가 Gaussian Distribution을 따르도록 유도함. Markov Chain을 따르므로 해당 값은** $q(x_{0,1,2,\dots,T}) = q(x_0)\Pi_1^T q(x_t \mid x_{t-1})$, 이때, $q(x_t \mid x_{t-1})$은 $N(x_t; x_{t-1}\sqrt{1-\beta_t} I\beta_t)$, β_t 는 Noise

H. **[Reverse Process] Notation: $p(x_0)$:Sample 결과라고 할 수 있음. 해당 값을 구하기 위한 과정은, Gaussian Distribution을 따른다는 점에 의거하여(β가 충분히 작다면 = T가 충분히 크다면 Gaussian Distribution을 따름- Kernel이 Gaussian을 따르므로) μ, Σ 만으로 표현할 수 있다는 장점이 있음.**

I. 그러나 해당 과정을 위해서는 우선, Model Probability를 구해야 하며, 이는 다음과 같은 과정에 의해 구할 수 있음.

J. $p(x_0) = \int dx_{0,1,2,\ldots,T} * p_{0,1,2,\ldots,T}$ 그러나, p에 대한 Joint Distribution은 계산이 불가능하므로 Tractable한 q에 대한 식으로 변환시키되, p와 q에 대해서, 동일한 Task (t시점일 때, $q(x_t \mid x_{t-1})/p(x_{t-1} \mid x_t$, 심지어 Diffusion의 경우 Dimension이 다른 Generative Model과 달리 감소하지도 않아서 훨씬 어려움)

K. Training(Loss Function)의 경우 Log Likelihood를 기반으로 학습하며, 이때 Intractable한 부분을 처리하기 위해서 KL-Divergence로 변환해주는 과정을 거침. (이때, 직접 Loss를 구하는 것이 불가능하다는 점 때문에, Jensen's Inequality를 통한, ELBO를 구하는 형태로 진행)

3. 본인이 생각하는 이 논문의 장점이 무엇이라고 생각하나요? **논문 Contribution bullet을 그대로 따라 적으면 안됩니다.**
   A. **Diffusion 모델의 초기 과정 (Forward + Reverse)를 제안함.**
   B. Sampling(AIS) + Nonequilibrium Theory = 이미 존재했던 이론. 그러나 여러 이론 들을 결합해 Generative model에 적용했다는 점이 이 논문의 가장 큰 장점이라고 생각함.

4. 이 논문을 읽으면서 느낀 점, 혹은 배운 점이 있으면 적어보세요.
   A. 매번 읽을 때마다, 그렇구나 하고 읽었던 부분에 대해서 깊게 공부하고, 실제 모델의 의미에 대해서 추가적으로 파악해 나가는 과정속에서 부족했던 개념을 채워 나간다는 느낌이 듦.

5. 이 논문의 한계점이 있다면 무엇이라고 생각하나요?
   A. 복잡한 Loss Function (Lower Bound)의 존재는 학습을 어렵게 만듦.
   B. Sampling까지 필요한 시간 (T=1000)이 너무 긺

6. 본인의 연구에 접목시켜볼 점이 있을지 생각하고 적어보세요.
   A. Diffusion 선행연구

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

날짜: 2025-06-28

이름: 신준원

~~6.0~~ 6.28 Deep Unsupervised Learning using Nonequilibrium Thermodynamics

Ars. Fokker-planck. Kolmogorov

문제 의식.

기존의 Generative model → flexibility & tractability.

- tractability : 계산가능?
- flexibility : 어떤 분포나 하려고 표현가능? → trade-off.

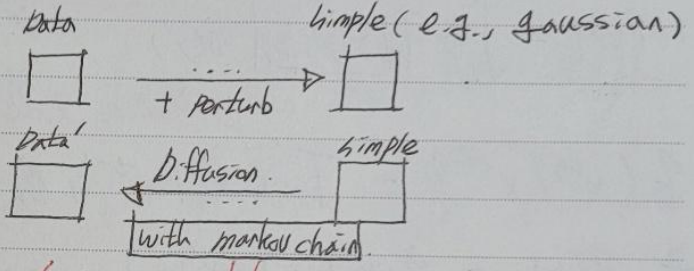
계산이 쉽다면 → 분포가 단순
분포 복잡하면 → 계산 불가능

→ 복잡한 분포 학습하는 Generative model?

→ 존재하는 다양한 방안 고려 ⇒ trade-off 해결 실패.

제안하는 새로운 모델 (간략설명)

- 모든 데이터 분포 ↑
- 샘플링 속도 ↑
- 다른 분포와 결합이 쉬움
- log likelihood 계산 쉬움



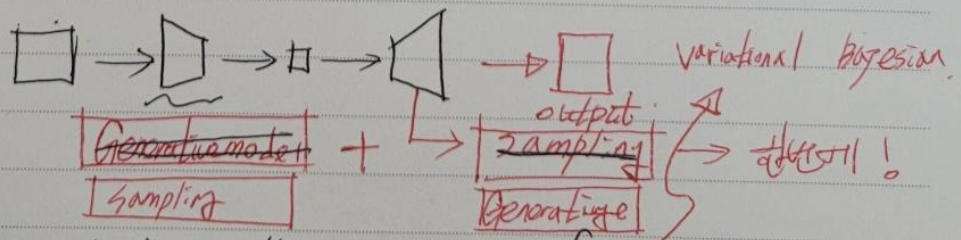
(non-equilibrium statistical physics)
+ Sequential MC

→ step 여러개로 구성, 정확을 한번에 학습X.
→ noise → 여러 step으로 나뉘어 학습 ⇒ 다루기 쉬움.

다른 연구의 관계성

The wake-sleep algorithm → Probability model의 학습과 Sampling 과정을 추론과정과 → 이미지를 생성.
(Inference)

VAE 의 등장



본 논문에서는 효율하게 "한번에" 생성할 수 있음 (+VAE와 차이점 존재)

① 물리학 (quasi-static process) 기반 학습 + annealed importance sampling
② 다른분포와 계산 쉬움
③ lower (수학적)
④ upper & lower bound 제공

VAE : generator + sampler 서로의 장점 바탕으로 학습
diffusion : 식도 쉽고 결합하여 효과적인 학습이 가능

Algorithm

forward process

goal: Complex ~~data~~ distribution ⟶ simple distribution

data distribution : $q(x^0)$ ⇝ $\pi(y)$

input — Sample → Gaussian

with **markov chain kernel**

$$\pi(y) = \int dy' \, T_\pi(y|y'; \beta) \cdot \pi(y') \quad / \; y'+1 = y$$

β : Diffusion Rate

$$q(x^{(t)}|x^{(t-1)}) = T_\pi(x^t|x^{t-1}; \beta_t)$$

$\therefore \pi(x) = q(x^0) \cdot q(x^1|x^0) \cdots\cdots q(x^T|x^{T-1})$

$$q(x^{(0,\cdots T)}) = q(x^0) \cdot \prod_{t=1}^{T} q(x^t|x^{t-1}) \quad : \text{whole forward process}$$

$$q(x^t|x^{t-1}) = \mathcal{N}(x^t; x^{t-1}\sqrt{1-\beta_t} \cdot I \cdot \beta_t) \quad : \text{gaussian kernel}$$

- 생략.

q^0 → forward 수식 / p^T 수식 → Diff Reverse

Reverse Process → training target

$$p(x^T) = \pi(x^T)$$

$$p(x^{0,\dots,T}) = p(x^T) \prod_{t=1}^{T} p(x^{t-1} | x^t)$$

최종결과 $p(x^0)$

$$p(x^{t-1} | x^t) = \mathcal{N}(x^{t-1} ; f_\mu(x^t, t), f_\Sigma(x^t, t))$$

~~with $f_b(x^t, t)$~~ 계산불가능

계산량 ∝ T

model probability → $p(x^0) = \int dx^{(1 \dots T)} \cdot p(x^{(0 \dots T)})$

$$p(x^0) = \int dx^{(1 \dots T)} p(x^{0 \dots T}) \frac{q(x^{(1 \dots T)} | x^0)}{q(x^{(1 \dots T)} | x^0)}$$

← Jarzynski equality

$$p(x^0) = \int dx^{(1 \dots T)} \cdot q(x^{(1 \dots T)} | x_0) \frac{p(x^{0 \dots T})}{q(x^{(1 \dots T)} | x^0)}$$

$$p(x^0) = \int dx^{(1 \dots T)} \cdot q(x^{(1 \dots T)} | x_0) \, p(x^T) \prod_{t=1}^{T} \frac{p(x^{t-1} | x^t)}{q(x^t | x^{t-1})}$$

$(\beta \to \infty)$ → 만약 서로 분포 동일하다면 $p(x^0)$는 → q만으로 계산가능

즉, ~~noise~~ 등 step당 noise가 → 0 수준으로, 서로 동일.
(이론적으로는 ...)

$\mathbb{E}_{x^0 \sim p_{data}}[\log p_\theta(x^0)]$

Training (log likelihood) $\rightarrow \int dx^0 p_{data}(x^0) \cdot \log p_\theta(x^0)$ 계산.

$L = \int dx^0 q(x^0) \cdot \log p(x^0)$ $p(x^0)$는 앞서 정의 → 대입시.

$L = \int dx^0 q(x^0) \cdot \log \left[\int dx^{(1\cdots T)} q(x^{(1\cdots T)} | x^0) \cdot p(x^T) \prod_{t=1}^{T} \frac{p(x^{t-1}|x^t)}{q(x^t|x^{t-1})} \right]$

Jensen's inequality $f(E[x]) \le E[f(x)]$ - 오목

$f(E[x]) \ge E[f(x)]$ - 볼록 . (log함수)

즉, $\log E[x] \ge E[\log x]$

→ log는 integral 안과 밖을 수 있다는 것!

Fixed 이유, $q(x^{(1\cdots T)}|x^0) \Rightarrow \frac{q(x^{(0\cdots T)})}{q(x^0)}$

x에 대한 $dx^{(1\cdots T)}$

$L \ge \int dx^{(1\cdots T)} q(x^{(1\cdots T)}|x^0) \cdot dx^0 q(x^0) \log p(x^T) \prod_{t=1}^{T} \frac{p(x^{t-1}|x^t)}{q(x^t|x^{t-1})}$

$L \ge \int dx^{(0\cdots T)} q(x^{(0\cdots T)}) \cdot \log \left[p(x^T) \cdot \prod_{t=1}^{T} \frac{p(x^{t-1}|x^t)}{q(x^t|x^{t-1})} \right]$

→ log → 곱 ~ 합의 관계.

$p(x^T) = \pi(x^T) \approx q(x^T)$

$L = \int \sum_{t=1}^{T} \frac{p(x^{t-1}|x^t)}{q(x^t|x^{t-1})} + \int dx^T \cdot q(x^T) \log p(x^T)$

$-H_p(x^T)$ → $H(x) = -\int p(x) \log p(x) dx$

$-H_p(x^T) + \int \sum_{t=1}^{T} \frac{p(x^{t-1}|x^t)}{q(x^t|x^{t-1})}$

↳ 가장자리의 비정상적인 효과 (최종목표) $q(x^1|x^0)$

Edge effect ⇒ $L=0$ → 목표 forward 과 동일하도록 설정

$$p(x^0|x^1) = q(x^1|x^0) \cdot \frac{\pi(x^0)}{\pi(x^1)} = T_\pi(x^0|x^1;\beta)$$

$$\left\{ L \geq \sum_{t=2}^{T} \int dx^{(0\cdots T)} \cdot q^{(0\cdots T)} \log\left[\frac{p(x^{t-1}|x^t)}{q(x^t|x^{t-1})}\right] \right\} + \int dx^0 \cdot dx^1 q(x^0, x^1) \log\left[\frac{q(x^1|x^0)\pi(x^0)}{q(x^1|x^0)\pi(x^1)}\right]$$

ELBO

$t=1$ → $p(x^0|x^1) = \dfrac{q(x^1|x^0)}{q(x^1|x^0)} \cdot \dfrac{\pi(x^0)}{\pi(x^1)}$ → Lec.

그래프

Posterior with markov chain

↳ $p(x_0)$ → x_0 우리가 알고싶은 (input)

$p(x_{t-1}|x_t)$ 를 알아야 함 ⇒ 계산 필요 ⇒ x_0를 condition으로

markov chain (forward) → $q(x_t|x_{t-1}) = q(x_t|x_{t-1}, x_0)$

$$\Rightarrow L \geq \sum_{t=2}^{T} \int dx^{(0\cdots T)} q^{(0\cdots T)} \log\left[\frac{p(x^{t-1}|x^t)}{q(x^t|x^{t-1}, x_0)}\right] - H_p(x^T)$$

T 항 무관함.

$$\frac{q(x^t|x^{t-1}, x_0)}{} = \frac{q(x^t, x^{t-1}|x^0)}{q(x^{t-1}|x_0)} \cdot \frac{q(x^{t-1}|x^t, x_0)}{q(x^t, x^{t-1}|x_0)} \quad \frac{1}{q(x^t|x_0)}$$

$$= \frac{q(x^{t-1}|x^t, x_0)}{q(x^{t-1}|x_0)} \cdot \frac{q(x_0)}{q(x_t|x_0)}$$

$$q(x^t, x^{t-1}|x_0) = q(x^{t-1}|x^t, x_0) \cdot q(x^t|x_0)$$

$$q(x^t|x^{t-1}, x_0) = \frac{q(x_{t-1}|x_t, x_0) \cdot q(x_t|x_0)}{q(x_{t-1}|x_0)}$$

Posterior

$$q(x_{t-1}|x_t, x_0) = \frac{q(x^t|x^{t-1}, 0) \cdot q(x_{t-1}|x_0)}{q(x_t|x_0)}$$

L_0를 Condition의 제시 → 해당값 P 학습을 위해 사용. (곱 $\frac{q}{p}$ → (q-p)확대)

$$L \geq K = \sum_{t=2}^{T} \int dx^{(0\ldots T)} q(x^{0\ldots T}) \log \left[\frac{p(x^{t-1}|x^t)}{q(x^{t-1}|x^t, x^0)} \right] \left[\frac{q(x^{t-1}|x^0)}{q(x^t|x^0)} \right] - H_p(x^T)$$

항끼리 재정렬 과정

$$L \geq K = \sum_{t=2}^{T} \int dx^{(0\ldots T)} q(x^{0\ldots T}) \log \left[\frac{p(x^{t-1}|x^t)}{q(x^{t-1}|x^t, x^0)} \right] + \sum_{t=2}^{T} \left[H_q(x^t|x^0) - H_q(x^{t-1}|x^0) \right]$$

$H_q(x^2|x^0) - H_q(x^1|x_0) + H_q(x^3|x^0) - H_q(x^2|x^0) \rightarrow H_q(x^T|x^0) - H_q(x^1|x_0)$ 만 남음

$$H_q(x^T|x^0) - H_q(x^1|x^0) - H_p(x^T)$$

$D_{KL}(q\|p) \rightarrow \mathbb{E}(\log q - \log p)$ ← $\mathbb{E}(p - \log q)$

$$L \geq K = -\sum_{t=2}^{T} \int dx^0 \cdot dx^t q(x^0, x^t) \cdot D_{KL}\left(q(x^{t-1}|x^t, x^0) \| p(x^{t-1}|x^t) \right) +$$

→ 간략히 표현한 식 "K" → Variational Bayesian Method (VAE) 유도

⊕ T 충분히 큰 경우 무한 → 증명하기 가능 ($p \Rightarrow q$)

⊕ 결과적으로 충분한 T → Gaussian distribution

⇒ m, σ에 대한 regression task와 동일.

B Scheduling. (noise scheduling)

$B_1 \Rightarrow$ over fitting 예방 (작은 값) – 고정

$B_{2\ldots T} \Rightarrow$ 증가함.

→ denoising or inpainting 시 필요한 작업.

multiplying Distributions, and Computing Posteriors

$$\tilde{p}(x_0) \propto p(x_0) \cdot r(x_0)$$ → Diffusion에서는 관련!

$r(x_0)$: intermediate distribution function → noise

$$\tilde{p}(x^{0 \dots T}) = \frac{1}{\tilde{Z}_T} \cdot p(x^T) \cdot r(x^T)$$

↓ 부분 (x^t)

$$\tilde{p}(x^t) = \frac{1}{\tilde{Z}_t} \cdot p(x^t) \cdot r(x^t)$$

이때 t step 에 관해서.

$$\tilde{p}(x_t) = \int dx^{t+1} \tilde{p}(x^t | x^{t+1}) \cdot \tilde{p}(x^{t+1})$$

$$\frac{p(x^t) \cdot r(x^t)}{\tilde{Z}_t} = \int dx^{t+1} \tilde{p}(x^t | x^{t+1}) \cdot \frac{p(x^{t+1}) \cdot r(x^{t+1})}{\tilde{Z}_{t+1}}$$

$\tilde{Z}_t$: 상수

$$p(x^t) = \int dx^{t+1} \tilde{p}(x^t | x^{t+1}) \frac{\tilde{Z}_t \, r(x^{t+1})}{\tilde{Z}_{t+1} \, r(x^t)} \cdot p(x^{t+1})$$

$$\rightarrow p(x^t) = \int dx^{t+1} p(x^t | x^{t+1}) \cdot p(x^{t+1})$$ 이므로 p, $\tilde{p}$의 관계 확인 가능

$$\tilde{p}(x^t | x^{t+1}) = p(x^t | x^{t+1}) \cdot \frac{\tilde{Z}_{t+1} \, r(x^t)}{\tilde{Z}_t \, r(x^{t+1})}$$

weight ratio

↳ 이 값은 비례하며, 곱에 영향 X (mean에도 영향 가능)

$$\tilde{p}(x^t | x^{t+1}) = \frac{1}{\tilde{Z}_t(x^{t+1})} \cdot p(x^t | x^{t+1}) \cdot r(x^t)$$

$\frac{1}{\tilde{Z}_t(x^{t+1})}$: normalization

forward process / Reverse process 사이에서

→ reweight 필요

based AIS algorithm.