D3PM

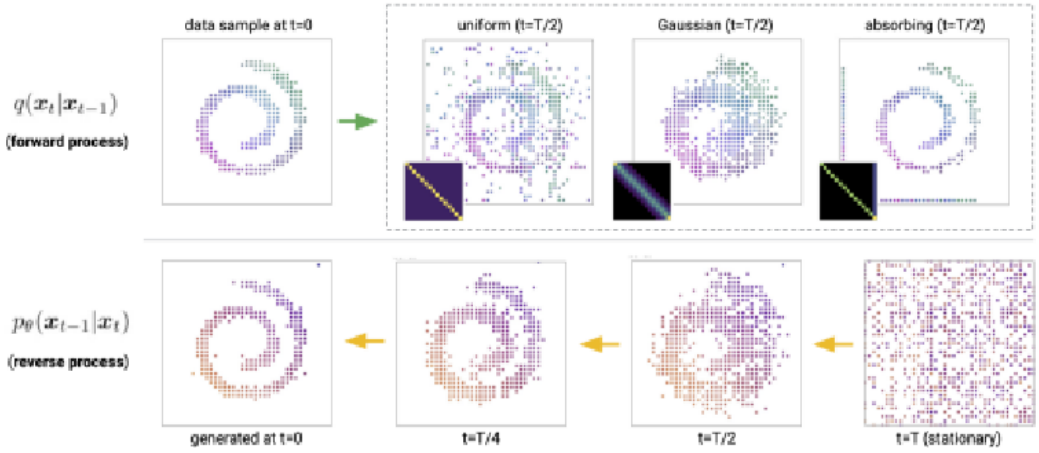
Abstract

- Denoising diffusion probabilistic models (DDPMs)
  - 연속 상태 공간에서 이미지 및 파형 생성에서 인상적인 결과를 보여줌
- 이 논문에서는 **Discrete Denoising Diffusion Probabilistic Models(D3PMs)**을 소개
  - 연속 공간에서 가우시안 커널을 모방하는 전이 행렬, 임베딩 공간에서 nearest neighbors을 기반으로 하는 행렬, 그리고 absorbing states를 도입하는 행렬을 포함함
  - 전이 행렬의 선택이 이미지와 텍스트 도메인에서 결과가 개선되게 하는 중요한 디자인 결정임을 보여줌
  - variational lower bound와 auxiliary cross entropy loss를 결합한 새로운 loss function를 소개함
  - 텍스트의 경우, 이 모델 클래스는 LM1B에서 큰 어휘로 확장하면서 character-level 텍스트 생성에서 강력한 결과를 얻음
  - CIFAR-10 이미지 데이터 세트에서, 논문의 모델은 샘플 품질을 접근하고, 연속 공간 DDPM 모델의 log-likelihood를 초과함

1. Introduction

- 최근에 diffusion model은 이미지 및 오디오 생성을 위한 매력적인 대안으로 등장하여 더 적은 추론 단계로 GAN과 비교할 수 있는 샘플 품질과 autoregressive model들과 비교할 수 있는 log-likelihoods을 달성함
- Diffusion model은 미리 정의된 forward process를 reverse 하기 위해 훈련된 parameterized Markov chain입니다.
  - 순방향 프로세스는 훈련 데이터를 순차적으로 손상시켜 순수한 잡음으로 변환하는 확률적 과정

- Diffusion model은 최대 likelihood와 score matching과 관련된 안정적인 목표를 사용하여 훈련되며, parallel iterative refinement을 사용하여autoregressive model보다 빠른 샘플링을 허용함
- Diffusion model들은 이산과 연속 상태 공간에서 제안되었지만, 가장 최근의 연구는 연속 상태 공간에서 작동하는 Gaussian diffusion process에 집중됨
- 이산 상태 공간을 갖는 확산 모델은 텍스트 및 이미지 segmentation 도메인에서 탐구되었지만, 대규모 텍스트 또는 이미지 생성을 위한 경쟁력 있는 모델 클래스로 아직 입증되지 않았음
- 이 연구의 목표는 더 구조화된 categorical corruption process를 사용하여 이산 diffusion 모델을 개선하고 확장하는 것



- 논문의 모델은 이산 데이터 (이미지 포함)를 연속 공간으로 변환하거나 임베딩할 필요가 없으며, forward process에서 사용되는 전이 행렬에 구조나 도메인 지식을 포함 시킬 수 있음
- 이런 유연성을 활용하여 크게 개선된 결과를 얻을 수 있음
    - 텍스트 데이터에 적합한 구조화된 corruption processes를 개발하며, 토큰 간의 유사성을 이용하여 점진적인 corruption과 denoising을 가능하게 함
    - [MASK] 토큰을 삽입하는 corruption processes를 탐구하여 autoregressive 및 mask-based generative 모델들과 유사한 모델과의 유사성을 찾을 수 있음
    - 우리는 continuous diffusion models에 의해 활용된 locality에서 영감을 얻어 양자화된 이미지에 대한 discrete diffusion models을 연구합니다. 이는 더 유사한 상태로 우선적으로 확산되고 이미지 영역에서 훨씬 더 나은 결과를 가져오는 discrete corruption process의 특정 선택으로 이어집니다
- 기술적 및 개념적 기여

- 새로운 구조화된 diffusion 모델을 디자인하는 것을 넘어, D3PMs의 훈련을 안정화하는 새로운 auxiliary loss을 소개하고 상호 정보를 기반으로 한 여러 noise schedule들을 개발하여 성능을 향상 시킴
- character-level의 텍스트 생성에서 텍스트 생성을 위한 다양한 non-autoregressive baseline을 매우 능가하고, discrete diffusion model을 큰 어휘와 긴 시퀀스 길이로 성공적으로 확장함
- CIFAR-10 이미지 데이터 세트에서 log-likelihood와 샘플 품질에서 Ho et al.의 Gaussian diffusion model에 접근하거나 능가하는 강력한 결과를 얻었음

2. Background: diffusion models

- 확산 모델은 forward와 reverse Markov process에 의해 characterize된 latent variable generative model임
- forward process 데이터 x0~q(x0)를 증가하는 noisy latent variable들의 시퀀스로 corrput함
- learned reverse Markov process는 점진적으로 latent variable들을 데이터 분포를 향해 denoise함
- time step T가 무한대로 가면, forward process와 the reverse process 모두 동일한 함수를 형태를 공유하며, forward process와 동일한 class의 분포들로부터 학습된 reverse process를 사용할 수 있음

3. Diffusion models for discrete state space

- 이후부터는 이산 상태 공간을 갖는 확산 모델의 일반적인 클래스를 Discrete Denoising Diffusion Probabilistic Models (D3PMs)로 참조.

3.1 Choice of Markov transition matrices for the forward process

- 이미지 및 텍스트를 포함한 대부분의 현실 세계 이산 데이터의 경우에는 전이 행렬 Qt에 domain-dependent 구조를 추가하여 forward corruption process과 learnable reverse denoising process을 제어하는 것이 합리적이라고 주장

1. **Uniform**

다른 state으로의 전이 확률이 균일하기 때문에, 이 논문에서는 이 diffusion instance를 D3PM-uniform으로 동등하게 참조합니다.

2. **Absorbing state**
    - 각 토큰이 동일하게 유지되거나 어떤 확률 βt로 [MASK]로 transition되는 흡수 상태를 가진 전이 행렬을 고려 → 이것은 범주 간에 특별한 관계를 부과하지는 않지만, uniform diffusion과 비슷하게 corrputed 토큰을 원래 토큰과 구별할 수 있게 함
    - stationary 분포는 균일하지 않지만 모든 mass가 [MASK] 토큰에 있음
    - 이미지의 경우, 회색 픽셀을 [MASK] absorbing 토큰으로 재사용함
3. **Discretized Gaussian**
    - 다른 상태로 균일하게 transition하는 대신, 순서형 데이터의 경우 이산화된, truncated 가우시안 분포를 사용하여 연속 공간 확산 모델을 모방하는 것을 제안 → 이것은 전이 행렬이 doubly stochastic인 정규화를 선택하게 하며, 균일한 stationary 분포를 가지게 됨 → 이 전이 행렬은 더 높은 확률을 가진 유사한 state로 transition하며, 이미지와 같은 양자화된 순서형 데이터에 적합함
4. **Token embedding distance**
    - D3PM 프레임워크의 generality을 입증하기 위해 임베딩 공간에서 유사성을 사용하여 forward 프로세스를 안내하고, 균일한 stationary 분포를 유지하면서 유사한 임베딩을 갖는 토큰 간에 더 자주 전환하는 doubly-stochastic 전이 행렬을 구성하는 방법을 탐구함

3.2 Noise schedules

- discretized Gaussian diffusion
    - 이산화하기 전 variance of the Gaussian을 선형적으로 증가하는 것을 탐색
- uniform diffusion
    - 우리는 코사인 함수로의 전환의 cumulative probabilit을 설정하는 코사인 스케줄을 사용
- a general set of transition matrices Qt
    - 이전에 제안된 schedule들이 직접 적용되지 않을 수 있습니다

3.3 Parameterization of the reverse process

- neural network nnθ(xt)를 사용하는 것에 집중해

$$\widetilde{p}_\theta(\widetilde{\boldsymbol{x}}_0|\boldsymbol{x}_t)$$

의 로짓을 예측함 → 이를 q(xt−1|xt, x0)와 x0의 one-hot representations에 대한 총체을 결합하여 다음과 같은 parameterization을 얻음:

$$p_\theta(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t) \propto \sum_{\tilde{\boldsymbol{x}}_0} q(\boldsymbol{x}_{t-1}, \boldsymbol{x}_t|\tilde{\boldsymbol{x}}_0)\tilde{p}_\theta(\tilde{\boldsymbol{x}}_0|\boldsymbol{x}_t).$$

- x0-parameterization에서

$$\widetilde{p}_\theta(\widetilde{\boldsymbol{x}}_0|\boldsymbol{x}_t)$$

가 모든 probability mass를 원래 값인 x0에 놓을 경우 KL divergence

$$D_{\mathrm{KL}}[q(x_{t-1}|x_t, x_0) \parallel p_\theta(x_{t-1}|x_t)]$$

가 0이 됨

- 주어진 state xt에서 최적의 reverse process는 q(xt|xt−1)이 0이 아닌 state들로의 transition만을 고려함 ⇒ Qt의 sparsity 패턴은 pθ(xt−1|xt)에서 이상적인 reverse transition 확률의 sparsity 패턴을 결정함
- 이런 parameterization은 학습된 reverse 확률 분포 pθ(xt−1|xt)가 Markov transition 행렬 Qt의 선택에 의해 지시된 올바른 sparsity 패턴을 갖도록 자동으로 보장한다. 이 parameterization을 통해

$$p_\theta(x_{t-k}|x_t) = \sum q(x_{t-k}, x_t|\tilde{x}_0)\tilde{p}_\theta(\tilde{x}_0|x_t)$$

를 예측해 inference를 수행할 수 있음(k steps at a time)

- 마지막으로 순서형 이산 데이터를 모델링 할 때 neural net의 출력으로 직접적으로

$$\widetilde{p}_\theta(\widetilde{\boldsymbol{x}}_0|\boldsymbol{x}_t)$$

의 로짓을 예측하는 대신에 다른 옵션은 a truncated discretized logistic distribution로 확률을 모델링하는 것 ⇒ 이는 reverse model에 추가적인 순서형 inductive bias를 제공하고 이미지를 위한 FID와 log-likelihood 점수를 높임

3.4 Loss function

- reverse process의 x0-parameterization를 위한 보조 denoising 목표를 소개 → 각 타임 스텝마다 데이터 x0의 좋은 예측을 장려
- 이것을 negative variational lower bound와 결합하여 대체 손실 함수를 산출함:

$$L_\lambda = L_{\text{vb}} + \lambda\, \mathbb{E}_{q(\boldsymbol{x}_0)} \mathbb{E}_{q(\boldsymbol{x}_t|\boldsymbol{x}_0)} [-\log \widetilde{p}_\theta(\boldsymbol{x}_0|\boldsymbol{x}_t)].$$

4. Connection to existing probabilistic models for text

- D3PM 프레임워크와 기존의 probabilistic and language 모델링 접근 사이의 흥미로운 연결에 대해 설명
- BERT는 단일 단계 diffusion model임
- Autoregressive model들은 (이산) diffusion model들임
- (Generative) Masked Language-Models (MLMs)들은 diffusion model들임

5. Text generation

1. Character-level generation on text8

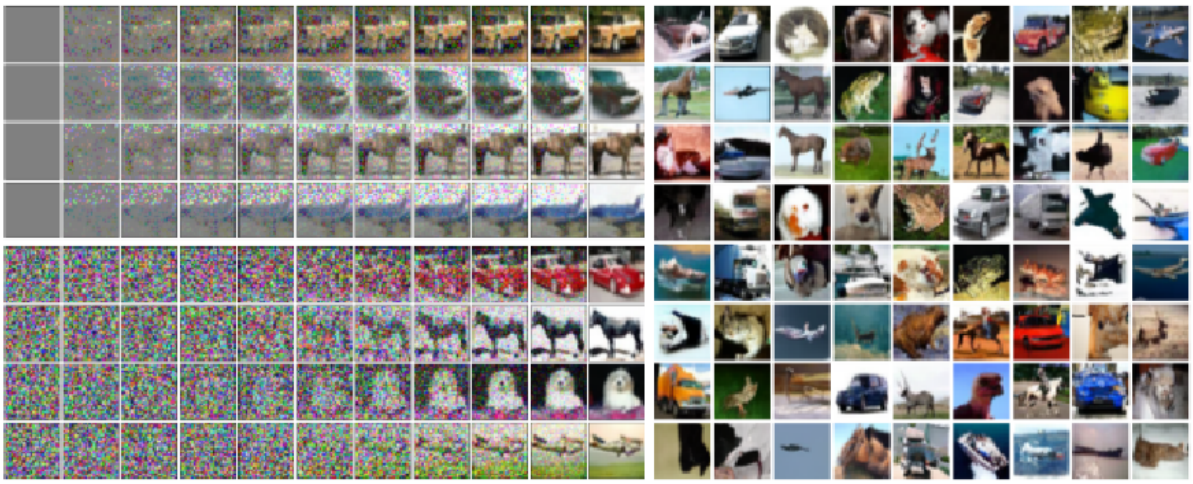
Model	Model steps	NLL (bits/char) (↓)	Sample time (s) (↓)
Discrete Flow [49] (8×3 layers)	-	1.23	0.16
Argmax Coupling Flow [20]	-	1.80	0.40 ± 0.03
IAF / SCF [57]‡	-	1.88	0.04 ± 0.0004
Multinomial Diffusion (D3PM uniform) [20]	1000	≤ 1.72	26.6 ± 2.2
D3PM uniform [20] (ours)	1000	$\leq 1.61 \pm 0.02$	3.6 ± 0.4
D3PM NN (L_{vb}) (ours)	1000	$\leq 1.59 \pm 0.03$	3.1474 ± 0.0002
D3PM mask ($L_{\lambda=0.01}$) (ours)	1000	$\leq 1.45 \pm 0.02$	3.4 ± 0.3
D3PM uniform [20] (ours)	256	$\leq 1.68 \pm 0.01$	0.5801 ± 0.0001
D3PM NN (L_{vb}) (ours)	256	$\leq 1.64 \pm 0.02$	0.813 ± 0.002
D3PM absorbing ($L_{\lambda=0.01}$) (ours)	256	$\leq 1.47 \pm 0.03$	0.598 ± 0.002
Transformer decoder (ours)	256	1.23	0.3570 ± 0.0002
Transformer decoder [1]	256	1.18	-
Transformer XL [10]†	256	1.08	-
D3PM uniform [20] (ours)	20	$\leq 1.79 \pm 0.03$	0.0771 ± 0.0005
D3PM NN (L_{vb}) (ours)	20	$\leq 1.75 \pm 0.02$	0.1110 ± 0.0001
D3PM absorbing ($L_{\lambda=0.01}$) (ours)	20	$\leq 1.56 \pm 0.04$	0.0785 ± 0.0003

- D3PM에 대해, D3PM absorbing model이 uniform과 NN diffusion model을 능가하면서 가장 성능이 좋게 나옴

2. Text generation on LM1B

- LM1B에 대한 실험
- 전체적으로 mask diffusion (D3PM absorbing)이 비교적 잘 수행되며 동일한 크기의 유사한 autoregressive model의 성능에 접근하고 훨씬 적은 단계로 scale하는 반면, uniform diffusion은 성능이 현저히 떨어짐
- D3PM NN model은 log likelihoods의 측면에서 uniform model에 비해 성능이 좋지 않음

6. Image generation



- 왼쪽 위: progressive sampling at t = 1000, 900, 800, ..., 0 for D3PM absorbing
- 왼쪽 아래: D3PM Gauss + logistic
- 오른쪽: D3PM Gauss + logistic model → best model

7. Related Work

- diffusion generative models
- denoising autoencoders

8. Discussion

- 한계
  - 다른 non-autoregressive generative model들과 같이, 논문의 모델은 텍스트 생성을 위한 Transformer XL과 같은 강한 strong autoregressive model에 비해 여전히 부족하고, continuous
    diffusion model들은 여전히 이미지 품질에 있어서 더 강한 결과들을 산출함
  - 사용하는 evaluation metrics에서 오는 한계