[5주차] 논문리뷰

DDPM

1. Introduction

💡 논문에서 다루고 있는 주제가 무엇인지와 해당 주제의 필요성이 무엇인가

논문에서 제안하는 방법이 기존 방법의 문제점에 대응되도록 제안 되었는가

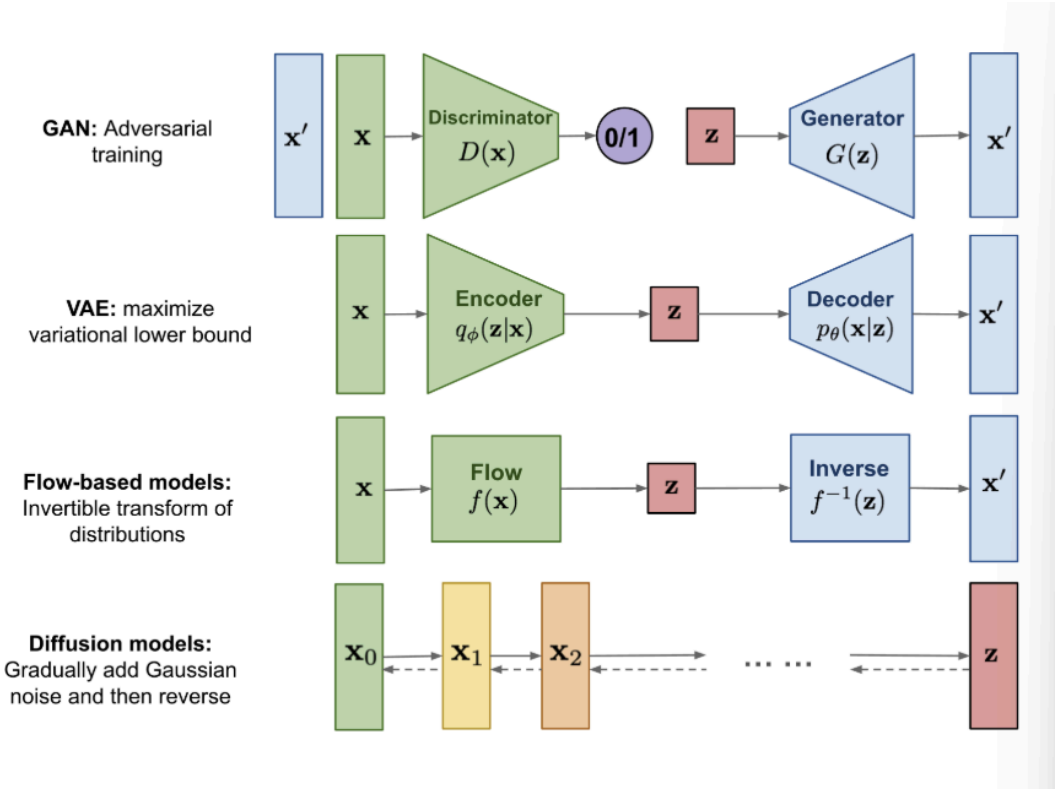
논문이 다루는 분야

- Diffusion 계열 생성형 모델 (발전 역사가 궁금해서 GPT 도움 받아 표로 정리)

연도	모델/기술 이름	주요 기여 또는 특징	관련 논문/출처
2015	Diffusion Probabilistic Models (DDPM 초기 개념)	확률적 그래픽 모델 기반의 초기 Diffusion 개념 정립	Sohl-Dickstein et al. (2015)
2020	DDPM (Denoising Diffusion Probabilistic Models)	Gaussian noise를 점진적으로 제거하는 방식, 이미지 생성 성능에서 GAN과 경쟁	Ho et al. (2020), NeurIPS
2021	DDIM (Denoising Diffusion Implicit Models)	DDPM보다 더 빠른 샘플링 가능, non-Markovian 구조 도입	Song et al. (2021), ICLR
2021	Score-based Generative Modeling (SDE 기반)	확률적 미분방정식(SDE) 기반으로 연속적인 시간 표현	Song et al. (2021), ICLR
2022	GLIDE	텍스트 조건 기반 생성 + Diffusion, CLIP을 활용한 조건 강화	Nichol et al. (2022), OpenAI
2022	Latent Diffusion Models (LDM)	고해상도 이미지 생성을 위한 latent 공간에서의 Diffusion 수행 → 계산 효율성 향상	Rombach et al. (2022), CVPR
2022	Imagen	T5 기반의 텍스트 인코더 + Diffusion 조합, 매우 높은 FID 성능 기록	Saharia et al. (2022), Google Brain
2022	Stable Diffusion	오픈소스 LDM 기반 모델, 커스터마이징과 Fine-tuning에 최적화	Rombach et al. (2022), CompVis
2023	ControlNet	조건 기반 제어 기능 (포즈, 윤곽 등)을 Diffusion에 통합	Lvmin Zhang et al. (2023), Stanford

연도	모델/기술 이름	주요 기여 또는 특징	관련 논문/출처
2023	SDXL (Stable Diffusion XL)	안정성과 해상도 향상, 텍스트 이해력 강화	Stability AI (2023)
2024	Sora (Video Diffusion, 발표됨)	OpenAI에서 개발한 비디오 생성용 Diffusion 모델 (텍스트→비디오)	OpenAI Sora (2024, 발표 기준)

- 생성형 모델 계열 구분



해당 task에서 기존 연구 한계점

- 디퓨전 모델은 그동안 좋은 품질의 이미지를 생성할 수 있다는 근거가 부족했음
  → 이 모델에서 좋은 품질 생성 가능함을 보이게 됨
- 언급된 기존 연구 (좋은 성능을 보여왔음)
  - Generative adversarial networks (GANs)
  - autoregressive models
  - flows
  - variational autoencoders (VAEs)
  - energy-based modeling

    - score matching
- 전반적으로 기존의 연구들도 좋은 성과를 보여왔다고 말함
- Introduction에서 언급했듯, 모델을 어떻게 구상하는지에 따라 foward 과정과 sampling 과정이 각기 다른 모델을 사용한 효과를 준다고 함 → 요 때 GAN보다 좋은 성능 내기도 한다고 비교
- 로그 가능도 측면에서 energy-based modeling, score matching와 비교(더 낫다)
- 복원 과정을 살펴봤을 때, autoregressive model과 비슷하게 점진적으로 학습하지만, 픽셀 순서 학습이 아닌 노이즈 크기 순서로 학습함을 밝힘.

논문의 contributions

- 논문의 작업 그래프로 정리
  → 노이즈를 추가해서 데이터를 조금씩 손상시키고
  → 반대로 노이즈를 제거하며 원래 데이터를 복원하는 모델을 학습
  ⇒ 노이즈의 크기가 작고, 가우시안 분포를 따르면 복원과정 역시 조건부 가우시안 형태로 설계할 수 있어서 신경망으로 쉽게 구현할 수 있음

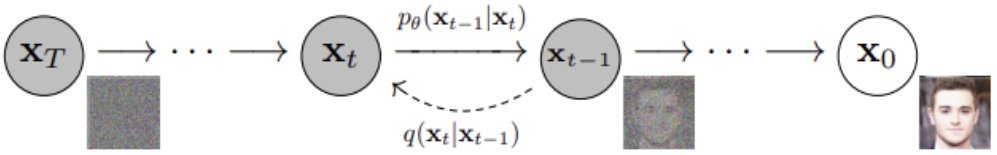


Figure 2: The directed graphical model considered in this work.

- **(핵심기여)** 특정한 방식으로 디퓨전 모델을 설계(parameterize)하면,
    - 학습 과정은 denoising score matching 기법처럼 되고
    - 샘플링(생성) 과정은 annealed Langevin dynamics 방법처럼 작동함을
    - 수학적으로 발견함. 요 때 가장 좋은 품질의 샘플 생성.
- 품질 좋은 이미지 만들어도 **로그 가능도(log-likelihood)** 지표에서는 기존의 디퓨전 모델들과 비교해 경쟁력이 낮은 경향이 있음.
  ⇒ 그래도 energy-based modeling나 score matching 보다는 좋음
- 무손실 압축길이의 대부분이 눈치채기 힘든 미세한 디테일 설명에 쓰임을 발견했음
  ⇒ 그래서 손실 압축의 관점에서 분석했음
  ⇒ 모델이 픽셀 순서대로 복원하는 게 아니라, 노이즈의 크기(정보의 정밀도) 정도에 따른 순서대로 이미지를 단계적으로 복원함을 발견했음
    - 무손실 압축 : 데이터를 원본 그대로 복원할 수 있도록 압축하는 방법

- 무손실 압축 실이 : 무손실 압축 시 필요한 최소 비트 수
- 무손실 압축에 쓰인 비트 중 상당수가 지나치게 미세한 정보 표현에 사용되어서 그럴 필요가 없다는 생각으로 손실 압축 관점에서 다시 분석한 것.
- progressive decoding : 데이터를 점점 단계적으로 정교하게 복원해나가는 방식
- bit ordering : 정보를 복원할 때 어떤 순서로 데이터를 처리할 지 정의하는 방식
  - 기존 autoregressive model : 이전까지 복원된 정보들을 기반으로 다음 정보를 한 단계씩 예측. 픽셀 순서대로.
  - 디퓨전 모델 : 노이즈 크기 순서대로(정밀도가 낮는 순서대로. 큰 윤곽 → 디테일)
- 디퓨전 모델이 기존 오토리그레시브 모델처럼 데이터를 순차적으로 복원하긴 하지만, 픽셀 순이 아닌 노이즈 크기(정밀도 정도)에 따라 복원한다는 점에서 더 일반적이고 강력한 구조를 가진다는 것.

2. 제안 방법론

💡 Introduction에서 언급된 내용과 동일하게 작성되어 있는가

Introduction에서 언급한 제안 방법이 가지는 장점에 대한 근거가 있는가

제안 방법에 대한 설명이 구현 가능하도록 작성되어 있는가

Background (수식 단위로 정리)

▼ **(1) Reverse Diffusion Process**

$$p_\theta(\mathbf{x}_{0:T}) := p(\mathbf{x}_T) \prod_{t=1}^{T} p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t), \qquad p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t) := \mathcal{N}(\mathbf{x}_{t-1}; \boldsymbol{\mu}_\theta(\mathbf{x}_t, t), \boldsymbol{\Sigma}_\theta(\mathbf{x}_t, t)) \quad (1)$$

- reverse process는 학습 가능한 가우시안 조건부 확률로 구성된 마르코프 체인
- 초기 상태는 표준 정규분포에서 시작
- 각 단계는 이전 단계를 예측하는 정규분포로 표현됨
- 평균과 분산은 신경망이 예측한다
- 모델이 노이즈 제거 과정을 익힘

▼ **(2) Forward Diffusion Process)**

$$q(\mathbf{x}_{1:T}|\mathbf{x}_0) := \prod_{t=1}^{T} q(\mathbf{x}_t|\mathbf{x}_{t-1}), \qquad q(\mathbf{x}_t|\mathbf{x}_{t-1}) := \mathcal{N}(\mathbf{x}_t; \sqrt{1-\beta_t}\mathbf{x}_{t-1}, \beta_t\mathbf{I}) \tag{2}$$

- 원본 데이터에 점차 노이즈를 더해가는 고정된 마르코프 체인
- 원본 데이터 x0에 노이즈를 반복적으로 추가하여 최종 노이즈 상태 xT를 생성
- 역방향 과정을 추정하기 위한 기준 분포
- 학습 목적상 완전히 정의되어 있어야 함

▼ (3) Loss Function (Basic Variational Bound)

$$\mathbb{E}\left[-\log p_\theta(\mathbf{x}_0)\right] \leq \mathbb{E}_q\left[-\log \frac{p_\theta(\mathbf{x}_{0:T})}{q(\mathbf{x}_{1:T}|\mathbf{x}_0)}\right] = \mathbb{E}_q\left[-\log p(\mathbf{x}_T) - \sum_{t>1} \log \frac{p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t)}{q(\mathbf{x}_t|\mathbf{x}_{t-1})}\right] =: L \tag{3}$$

- 기본 손실 함수인 변분 바운드를 정의한 식
- 샘플링 확률 pθ(x0)의 음의 로그를 직접 계산하지 못해서
  ⇒ 변분 추론을 통해 하한을 최적화함.

▼ (4) Closed-form Forward Sampling

$$q(\mathbf{x}_t|\mathbf{x}_0) = \mathcal{N}(\mathbf{x}_t; \sqrt{\bar{\alpha}_t}\mathbf{x}_0, (1-\bar{\alpha}_t)\mathbf{I}) \tag{4}$$

- t단계의 샘플 xt를 원본 데이터 x0과 가우시안 노이즈로부터 직접 샘플링
- x0과 노이즈 ϵ~N(0,I)로부터 직접 xt를 만듦
- 데이터 샘플링 및 노이즈 레벨 설정이 가능해짐

▼ (5) Decomposed Variational Loss

$$\mathbb{E}_q\Big[\underbrace{D_{\mathrm{KL}}(q(\mathbf{x}_T|\mathbf{x}_0) \parallel p(\mathbf{x}_T))}_{L_T} + \sum_{t>1} \underbrace{D_{\mathrm{KL}}(q(\mathbf{x}_{t-1}|\mathbf{x}_t, \mathbf{x}_0) \parallel p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t))}_{L_{t-1}} \underbrace{-\log p_\theta(\mathbf{x}_0|\mathbf{x}_1)}_{L_0}\Big] \tag{5}$$

- 수식(3)-(변분바운드)을 KL divergence의 합으로 재작성한 형태
- 각 항은 정방향/역방향 분포의 차이를 비교함
- 전체 손실을 3개의 항으로 분리
  - LT: xT와 prior 차이
  - Lt−1: 정방향 사후와 역방향 예측 차이
  - L0: 최종 복원 에러

- 모든 항이 가우시안 간 KL이므로 닫힌 형태로 손실을 계산할 수 있게 되어 학습 안정성과 효율성 有

▼ **(6) Forward Posterior Distribution**

$$q(\mathbf{x}_{t-1}|\mathbf{x}_t, \mathbf{x}_0) = \mathcal{N}(\mathbf{x}_{t-1}; \tilde{\boldsymbol{\mu}}_t(\mathbf{x}_t, \mathbf{x}_0), \tilde{\beta}_t \mathbf{I}), \tag{6}$$

- 정방향 과정의 조건부 사후분포를 정의한 식
- 수식 (5)의 Lt−1 항 계산에 사용됨

▼ **(7) Posterior Mean and Variance**

$$\text{where} \quad \tilde{\boldsymbol{\mu}}_t(\mathbf{x}_t, \mathbf{x}_0) := \frac{\sqrt{\bar{\alpha}_{t-1}}\beta_t}{1-\bar{\alpha}_t}\mathbf{x}_0 + \frac{\sqrt{\alpha_t}(1-\bar{\alpha}_{t-1})}{1-\bar{\alpha}_t}\mathbf{x}_t \quad \text{and} \quad \tilde{\beta}_t := \frac{1-\bar{\alpha}_{t-1}}{1-\bar{\alpha}_t}\beta_t \tag{7}$$

- 수식 (6)의 평균과 분산을 수식화한 것
- 평균: x0와 xt의 선형 결합
- 분산: 노이즈에 비례하여 조정됨

Diffusion models and denoising autoencoders

- 확산모델 구현에서 많은 자유도 가짐
  - 선택지
    - 정방향 과정에서의 분산 βt\beta_tβt 값들
    - 역방향 과정의 모델 구조
    - 역방향 과정에서의 가우시안 분포의 파라미터화 방식
- 디퓨전 모델과 denoising score matching 사이의 연결 제시
  ⇒ simplified, weighted variational bound를 유도하며, 학습 목적함수로 사용할 수 있음.
- 단순성과 실험적 성능이 본 논문에서 제시한 모델 설계의 정당성

▼ **Forward process and LT**

- 정방향 과정의 분산 βt를 reparameterization를 통해 학습 가능한 변수로 처리할 수 있다는 사실을 무시하고, 고정된 상수값으로 설정
- 근사 사후분포 q에 학습 가능한 파라미터가 없고 따라서 손실함수 LT는 학습 중 상수값이 되어 무시할 수 있게 됨

▼ **Reverse process and L1:T −1**

- 역방향 과정에 대한 설정
  ⇒ 정방향 사후 평균과 역방향 평균의 차이를 최소화하는 손실을, 원본 데이터 x0와 노이즈 ϵ을 기반으로 재구성하고, 이를 통해 모델이 직접 평균을 예측하는 대신 노이즈 ϵ을 예측하도록 파라미터화함

1. 분산

$$\boldsymbol{\Sigma}_\theta(\mathbf{x}_t, t) = \sigma_t^2 \mathbf{I}$$

- 학습하지 않고, 고정된 값으로 설정
  - x0~N(0,I)일 때 최적 / 엔트로피 상한

$$\sigma_t^2 = \beta_t$$

  - x0가 단일한 점으로 고정될 때 최적 / 엔트로피 하한

$$\sigma_t^2 = \tilde{\beta}_t = \frac{1-\bar{\alpha}_{t-1}}{1-\bar{\alpha}_t}\beta_t$$

2. 평균
   - 손실함수 기반으로 설계

$$L_{t-1} = \mathbb{E}_q \left[\frac{1}{2\sigma_t^2} \| \tilde{\boldsymbol{\mu}}_t(\mathbf{x}_t, \mathbf{x}_0) - \boldsymbol{\mu}_\theta(\mathbf{x}_t, t) \|^2 \right] + C \tag{8}$$

   - 가장 직관적인 형태는 정방향 사후분포의 평균을 예측하는 것
   - 이 수식을 확장해서 계산 가능한 실용적 형태로 전개

$$L_{t-1} - C = \mathbb{E}_{\mathbf{x}_0, \boldsymbol{\epsilon}} \left[\frac{1}{2\sigma_t^2} \left\| \tilde{\boldsymbol{\mu}}_t \left(\mathbf{x}_t(\mathbf{x}_0, \boldsymbol{\epsilon}), \frac{1}{\sqrt{\bar{\alpha}_t}} (\mathbf{x}_t(\mathbf{x}_0, \boldsymbol{\epsilon}) - \sqrt{1-\bar{\alpha}_t}\boldsymbol{\epsilon}) \right) - \boldsymbol{\mu}_\theta(\mathbf{x}_t(\mathbf{x}_0, \boldsymbol{\epsilon}), t) \right\|^2 \right] \tag{9}$$

     - xt를 x0과 노이즈 ϵ의 함수로 바꿔서, 손실을 더 쉽게 계산할 수 있게 만든 식

$$\mathbb{E}_{\mathbf{x}_0, \boldsymbol{\epsilon}} \left[\frac{1}{2\sigma_t^2} \left\| \frac{1}{\sqrt{\alpha_t}} \left(\mathbf{x}_t(\mathbf{x}_0, \boldsymbol{\epsilon}) - \frac{\beta_t}{\sqrt{1-\bar{\alpha}_t}} \boldsymbol{\epsilon} \right) - \boldsymbol{\mu}_\theta(\mathbf{x}_t(\mathbf{x}_0, \boldsymbol{\epsilon}), t) \right\|^2 \right] \tag{10}$$

- 타깃 평균값을 xt와 ϵ만으로 표현. 어떤 평균을 가져야 하는지에 대한 정답을 학습하게 됨.

$$\boldsymbol{\mu}_\theta(\mathbf{x}_t, t) = \tilde{\boldsymbol{\mu}}_t\left(\mathbf{x}_t, \frac{1}{\sqrt{\bar{\alpha}_t}}(\mathbf{x}_t - \sqrt{1-\bar{\alpha}_t}\boldsymbol{\epsilon}_\theta(\mathbf{x}_t))\right) = \frac{1}{\sqrt{\alpha_t}}\left(\mathbf{x}_t - \frac{\beta_t}{\sqrt{1-\bar{\alpha}_t}}\boldsymbol{\epsilon}_\theta(\mathbf{x}_t, t)\right) \quad (11)$$

- 모델이 평균을 직접 예측하지 않고, 노이즈 ϵ을 예측하게 만든 파라미터화

$$\mathbb{E}_{\mathbf{x}_0,\boldsymbol{\epsilon}}\left[\frac{\beta_t^2}{2\sigma_t^2\alpha_t(1-\bar{\alpha}_t)}\left\|\boldsymbol{\epsilon} - \boldsymbol{\epsilon}_\theta(\sqrt{\bar{\alpha}_t}\mathbf{x}_0 + \sqrt{1-\bar{\alpha}_t}\boldsymbol{\epsilon}, t)\right\|^2\right] \quad (12)$$

- 모델이 예측한 노이즈 ϵθ와 실제 노이즈 ϵ 간의 오차를 측정하는 손실 함수

▼ Data scaling, reverse process decoder, and L0

- 이미지를 정수 값 {0,1,...,255}로 표현한다고 가정하고, 이를 [−1,1] 범위로 선형 스케일링함
  ⇒ 역방향 과정 입력이 정규화되어 학습이 더 잘 이루어짐
  ⇒ 이산 데이터를 연속적 분포로 잘 처리할 수 있음(VAE 에서 검증됨
  ⇒ 실제 로그우도 계산이 가능해짐, 마지막 단계에서 더 현실적인 이미지 복원이 가능해짐

$$p_\theta(\mathbf{x}_0|\mathbf{x}_1) = \prod_{i=1}^{D}\int_{\delta_-(x_0^i)}^{\delta_+(x_0^i)} \mathcal{N}(x; \mu_\theta^i(\mathbf{x}_1, 1), \sigma_1^2)\,dx$$
$$\delta_+(x) = \begin{cases}\infty & \text{if } x = 1\\ x + \frac{1}{255} & \text{if } x < 1\end{cases} \qquad \delta_-(x) = \begin{cases}-\infty & \text{if } x = -1\\ x - \frac{1}{255} & \text{if } x > -1\end{cases} \quad (13)$$

- 독립적인 이산 디코더 pθ(x0 ∣ x1)
- 변분 오토인코더(VAE)나 자기회귀 모델에서 흔히 사용되는 디코딩 방식과 유사
  - 노이즈를 따로 추가하지 않아도 됨
  - 데이터 스케일링의 Jacobian을 계산하지 않아도 됨
  - 손실 없는 이산 표현을 사용할 수 있음

▼ Simplified training objective

- 복잡한 KL 기반 손실 대신, 모델이 노이즈를 얼마나 정확히 예측했는지를 측정하는 단순한 MSE 손실을 사용함.
- 성능이 비슷하거나 더 좋고, 구현도 간단하며, 특히 강한 노이즈에 대한 복원 성능이 더 좋아짐
- 학습 속도가 빨라지고, 노이즈 제거 능력이 향상되어 샘플 품질 개선에 기여

$$L_{\text{simple}}(\theta) := \mathbb{E}_{t,\mathbf{x}_0,\boldsymbol{\epsilon}}\left[\left\|\boldsymbol{\epsilon} - \boldsymbol{\epsilon}_\theta(\sqrt{\bar{\alpha}_t}\mathbf{x}_0 + \sqrt{1-\bar{\alpha}_t}\boldsymbol{\epsilon}, t)\right\|^2\right] \tag{14}$$

- 더 단순한 손실 함수로도 좋은 결과를 얻을 수 있음을 발견한 것

3. 실험 및 결과

💡 Introduction에서 언급한 제안 방법의 장점을 검증하기 위한 실험이 있는가

Dataset

- CIFAR-10, CelebA-HQ, LSUN

Baseline

1. 기본 설정
   - 확산 단계 수T=1000
     → 이는 기존 연구와 비교를 쉽게 하기 위해
   - 정방향 노이즈 분산 βt = 선형증가하도록 고정
     → 입력 데이터를 [-1, 1] 범위로 정규화했을 때 적당한 수준의 노이즈를 추가하는 효과
2. 역방향 과정 모델링
   - U-Net 구조 사용 (PixelCNN++ 기반, 마스킹은 없음)
   - Group Normalization 사용
   - 시간 t정보는 Transformer의 사인/코사인 위치 임베딩으로 주입
   - 16×16 크기 feature map에서 self-attention 사용

결과

1. 샘플 품질 평가
   - CIFAR-10에서 다음 지표 측정:
     - Inception Score (IS)
     - Fréchet Inception Distance (FID)
     - Negative Log Likelihood (NLL): 무손실 부호화 길이

- 결과:

Table 1: CIFAR10 results. NLL measured in bits/dim.

Model	IS	FID	NLL Test (Train)
Conditional			
EBM [11]	8.30	37.9	
JEM [17]	8.76	38.4	
BigGAN [3]	9.22	14.73	
StyleGAN2 + ADA (v1) [29]	**10.06**	**2.67**	
Unconditional			
Diffusion (original) [53]			≤ 5.40
Gated PixelCNN [59]	4.60	65.93	3.03 (2.90)
Sparse Transformer [7]			**2.80**
PixelIQN [43]	5.29	49.46	
EBM [11]	6.78	38.2	
NCSNv2 [56]		31.75	
NCSN [55]	8.87 ± 0.12	25.32	
SNGAN [39]	8.22 ± 0.05	21.7	
SNGAN-DDLS [4]	9.09 ± 0.10	15.42	
StyleGAN2 + ADA (v1) [29]	$\mathbf{9.74} \pm 0.05$	3.26	
Ours (L, fixed isotropic $\mathbf{\Sigma}$)	7.67 ± 0.13	13.51	≤ 3.70 (3.69)
Ours (L_{simple})	9.46 ± 0.11	**3.17**	≤ 3.75 (3.72)

  - FID = **3.17**
  - 이는 기존 논문들보다도 우수한 샘플 품질임

Figure 3: LSUN Church samples. FID=7.89

Figure 4: LSUN Bedroom samples. FID=4.90

2. 역방향 파라미터화 및 학습 목표 손실 실험

Table 2: Unconditional CIFAR10 reverse process parameterization and training objective ablation. Blank entries were unstable to train and generated poor samples with out-of-range scores.

Objective	IS	FID
$\tilde{\mu}$ prediction (baseline)		
L, learned diagonal Σ	7.28±0.10	23.69
L, fixed isotropic Σ	8.06±0.09	13.22
$\|\tilde{\mu} - \tilde{\mu}_\theta\|^2$	–	–
ϵ prediction (ours)		
L, learned diagonal Σ	–	–
L, fixed isotropic Σ	7.67±0.13	13.51
$\|\tilde{\epsilon} - \epsilon_\theta\|^2$ (L_{simple})	**9.46±0.11**	**3.17**

- 정규 손실로는 뮤 예측보다 ϵ예측이 성능이 더 좋음
- 분산까지 학습시키면 불안정하고 성능 나쁨

3. 점진적 압축 (Progressive coding)
    a. Progressive lossy compression
        - log-likelihood는 다른 모델보다 낮음에도 불구하고, 샘플 품질이 매우 우수하므로, DDPM은 손실 압축에 탁월한 inductive bias를 갖는다

$$\mathbf{x}_0 \approx \hat{\mathbf{x}}_0 = \left(\mathbf{x}_t - \sqrt{1-\bar{\alpha}_t}\boldsymbol{\epsilon}_\theta(\mathbf{x}_t)\right)/\sqrt{\bar{\alpha}_t} \tag{15}$$

        - xt를 알고 있을 때, 역방향 과정을 일부만 수행해도 x0를 근사 가능

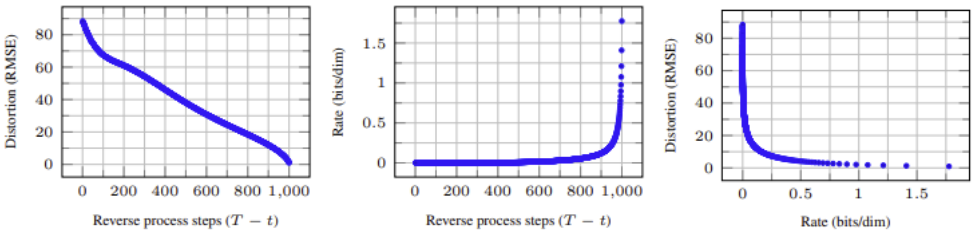


Figure 5: Unconditional CIFAR10 test set rate-distortion vs. time. Distortion is measured in root mean squared error on a $[0, 255]$ scale. See Table 4 for details.

        - (왼쪽) 역방향 단계가 줄어들수록 복원 오류(RMSE)는 급격히 증가함
        - (가운데) 역방향 초반 단계는 거의 정보를 추가하지 않음
        - (오른쪽) 낮은 rate에서 대부분의 distortion이 발생하고, 이후는 미세한 개선만

⇒ 대부분의 비트가 시각적으로 중요하지 않은 정보 복원에 사용되고 있음

b. Progressive generation

- progressive decompression from random bits 개념을 통해 점진적 생성 과정을 수행

Figure 6: Unconditional CIFAR10 progressive generation ($\hat{\mathbf{x}}_0$ over time, from left to right). Extended samples and sample quality metrics over time in the appendix (Figs. 10 and 14).

Figure 7: When conditioned on the same latent, CelebA-HQ 256×256 samples share high-level attributes. Bottom-right quadrants are $\mathbf{x}_t$, and other quadrants are samples from $p_\theta(\mathbf{x}_0|\mathbf{x}_t)$.

- 큰 스케일 특성(예: 윤곽선, 구조)은 초기에 먼저 생성됨
- 세부 특성(예: 눈썹, 머리카락)은 마지막에 생성됨
- 개념 압축(conceptual compression)과 유사한 성질

c. Connection to autoregressive decoding

$$L = D_{\mathrm{KL}}(q(\mathbf{x}_T) \,\|\, p(\mathbf{x}_T)) + \mathbb{E}_q\left[\sum_{t\geq 1} D_{\mathrm{KL}}(q(\mathbf{x}_{t-1}|\mathbf{x}_t) \,\|\, p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t))\right] + H(\mathbf{x}_0) \qquad (16)$$

- autoregressive decoding 방식처럼, 이전 단계를 현재 상태로부터 조건부로 복원함
  ⇒ 하지만 DDPM의 순서는 고정된 좌표 순서가 아닌, 노이즈를 통해 정의된 순서

4. 이미지 보간 (Interpolation)

- 두 이미지를 정방향으로 보낸 후 잠재공간에서 선형보간 후 역방향으로 되돌려 이미지 복원
- 결과:

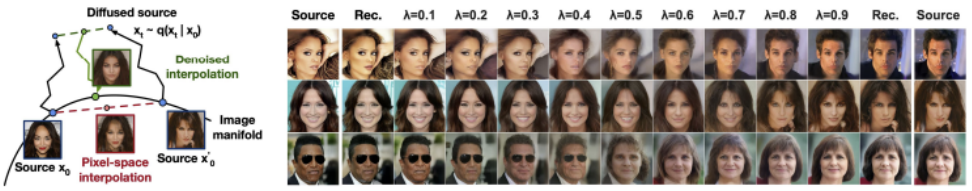


Figure 8: Interpolations of CelebA-HQ 256x256 images with 500 timesteps of diffusion.

- 표정, 자세, 피부톤, 배경 등은 자연스럽게 보간됨
- 안경 여부는 보간되지 않음 (모델이 이를 강하게 인코딩하지 않았음)

4. Related Work

💡 Introduction에서 언급한 기존 연구들에 대해 어떻게 서술하는가
제안 방법의 차별성을 어떻게 표현하고 있는가

- 기존의 생성 모델들과 유사점
  - Flows, VAE ↔ 잠재 변수 기반 구조, 마르코프 체인 기반 구조를 가짐
- 확산 모델만의 강점
  - forward process 학습이 없어 간단
  - 확률 분포의 log-likelihood 계산이 가능
- 이론적으로 다른 기법들과 연결됨
  (forward과정과 sampling과정이 각각 매칭됨)
  - Denoising Score Matching ≈ Variational Inference
  - Sampling ≈ Langevin Dynamics
  - 이론적으로 Energy-Based Model과도 연결
- rate–distortion 곡선을 한 번에 평가하는 방식
  → annealed importance sampling과 유사
- 확산 모델의 디코딩 과정
  → 오토리그레시브 모델과 유사
  → But, 노이즈 기반 순서를 따라서 더 유연
  → subscale 또는 비고정 순서의 디코딩 모델로 확장 가능

5. 결론 (배운점)

💡 연구의 의의 및 한계점, 본인이 생각하는 좋았던/아쉬웠던 점 (배운점)

- 디퓨전 모델임에도 고품질 이미지 샘플을 생성할 수 있게 됨
- 확산 모델과 다른 여러 기법 / 개념들 간의 연관성을 밝히고, 점진적 손실 압축 등의 현상을 찾아냄
  - variational inference
  - denoising score matching
  - annealed Langevin dynamics
  - autoregressive model
  - progressive lossy compression
- 멀티 모달과 엮었을 때의 유용성 등을 더 탐구할 것

- 디퓨전 모델의 발전 흐름을 먼저 찾아보고 본 논문을 보니, 본 논문이 어느 정도 위치에 있는 논문인지, 어떤 기여를 했고 디퓨전 모델계에 어떤 흐름을 가져왔는지 등을 생각하면서 읽을 수 있었다.
- 생성 분야가 GAN VAE Flow-based Diffusion으로 크게 나뉨을 알게 되었다. 여기에 요즘은 Transfomer도 이미지 생성에 사용되고 있음을 알게 되었다.
- 그러다보니 본 논문의 contribution이 더 잘 보이는 것 같다.
- background로 이해해야 할 수식이 많았다. 구체적인 내용까지 모두 이해한 것은 아니지만, 수식간의 관계, 왜 이렇게 전개되는지 등을 중심으로 이해했다. 추후 더 세부적으로 이해할 수 있으면 좋겠다.
- 디퓨전 모델이 트랜스포머 모델과도 엮여서 새로운 모델이 나왔다고 알고 있다.

6. 기타

Reference

- https://www.youtube.com/watch?v=H45lF4sUgiE&t=206s

- https://velog.io/@hanlyang0522/DDPM-Denoising-Diffusion-Probabilistic-Models-%EB%85%BC%EB%AC%B8-%EB%A6%AC%EB%B7%B0

CLIP

0. 멀티모달에 대한 기본적인 이해

멀티 모달

- 여러 종류의 데이터를 함께 이해하고 처리하는 인공지능 기술
- 다양한 모달리티 간의 연관성과 의미의 연결을 배워서 활용
- 구조
  - Early Fusion
    - 서로 다른 모달리티의 원본데이터를 결합해서 처리
    - 예: 이미지 픽셀 + 텍스트 임베딩을 concat
  - Late Fusion
    - 각 모달리티를 따로 처리한 후, 마지막에 결합
    - 예: CNN + RNN 결과를 나중에 합쳐서 판단
  - Joint Embedding
    - 서로 다른 모달리티를 같은 임베딩 공간에 매핑
    - 예: CLIP – 이미지와 텍스트를 동일한 공간에 표현

멀티 모달 학습 기술의 흐름

시기	주요 기술/특징	한계
1990s~2010s	이미지와 텍스트의 단순한 연관 학습	일반화 어려움
2016~2018	CNN + 텍스트 메타데이터 예측	약한 지도 학습
2019~2020	트랜스포머 + 자연어 기반 학습 (VirTex, ConVIRT)	작은 데이터셋, 낮은 성능
2021	CLIP: contrastive learning + 대규모 데이터	대규모 제로샷 가능
2022~	GPT계열 확장 → 멀티모달 LLM 시대	multimodal reasoning 가능

멀티 모달 활용 예시

- 모달리티 종류
  - 시각(Vision) - 이미지, 영상
  - 언어(Language) - 텍스트, 음성
  - 청각(Audio) - 음악, 소리, 음성
  - 센서 데이터 - 위치 정보, 생체신호

문제	사용되는 모달리티	예시 모델
이미지 캡셔닝	이미지 + 텍스트	Show and Tell (2015)
텍스트로 이미지 검색	이미지 + 텍스트	CLIP (2021)
텍스트로 이미지 생성	텍스트 → 이미지	DALL·E, Stable Diffusion
이미지 보고 질문 답변	이미지 + 텍스트	Flamingo, BLIP, GPT-4V
음성 인식	음성 → 텍스트	Whisper, DeepSpeech

1. Introduction

논문에서 다루고 있는 주제가 무엇인지와 해당 주제의 필요성이 무엇인가

논문에서 제안하는 방법이 기존 방법의 문제점에 대응되도록 제안 되었는가

논문이 다루는 분야

- 멀티 모달
- 이미지 & 텍스트 멀티 모달

개념

- zero-shot transfer
  - 모델이 한 번도 학습하지 않은 태스크나 클래스에 대해, 라벨링된 데이터 없이도 추론하거나 문제를 해결하는 능력
  - Few-shot learning → 적은 양의 라벨로 학습 후 적용
  - Transfer learning → 기존 학습된 지식을 새로운 태스크에 활용

기존 연구 한계점 & 논문의 contributions

- 웹 스케일의 텍스트 데이터를 통한 학습이 기존 고품질 NLP 데이터 라벨링보다 학습에 있어서 훨씬 효과적이다.
  - raw text로부터 직접 학습 → NLP 분야 혁신
  - autoregressive / masked language modeling 등도 성능 꾸준히 향상
  - text-to-text 방식의 입출력 인터페이스 개발
    → 특정 태스크 없이도 다양한 데이터셋에 zero-shot transfer 가능
  - GPT-3 → 특정 데이터셋 훈련 없이도 다양 태스크에 맞춤 모델 수준 성능
- But, 컴퓨터 비전에서 여전히 사람이 라벨링한 데이터셋 사용해서 학습
  - 웹 텍스트로부터 직접 학습하는 돌파구!!
  - 기존 연구
    - Mori
      → 이미지와 짝지어진 텍스트에서 명사와 형용사를 예측하는 모델을 통해 이미지 검색 개선 시도
    - 수많은 연구
      → 이미지-텍스트 pair을 활용한 학습,
      → n-gram 예측, caption 기반 학습
    - 최신 연구
      → Transformer, 마스킹 언어 모델링, contrastive learning을 활용한 시도
    - natural language supervision 사용한 이미지 표현 학습
      → 드물고, 성능도 기존보다 좋지 않음
      → Li : ImageNet에서 top-1 정확도가 11.5%

        - 다른 연구들 weak supervision 사용하거나, 클래스가 고정되어 있음
          → 동적인 출력 구조가 없으며, 제로샷 적용에 한계있음
        - 기존 모델들은 대부분 학습에 사용된 이미지가 20만개 이하였음
- 본 논문에서는 웹에서 수집한 4억개의 이미지-텍스트 쌍 사용
- 다양한 컴퓨터 비전 태스크에 대해 유연하고 효율적인 zero-shot transfer 가능하게 함
- 라벨이 전혀 없는 상황에서도, 완전 지도학습 모델과 비슷하거나 더 뛰어난 성능을 보이기도 함