Paper summary

AI VISION Lab

1. 공부한 논문의 제목, 게재된 학회 혹은 저널 등 논문 기본 정보를 적으세요.
   - **A. 이름: Palette: Image-to-Image Diffusion Models**
   - **B. 저널: SIGGRAPH**
   - **C. 도메인: i2i, Diffusion**
   - **D. 출판연도: 2022**
   - **E. 저자: Google Brain Team**

2. 논문에서 제안한 알고리즘 및 프레임워크에 대해 본인이 이해한대로 다이어그램을 그려보세요. **논문 Figure를 그대로 따라 그리면 안됩니다.**
   A. 구조가 특별하기 보다는 하나의 모델의 구조에서 다양한 Task(Image Inpainting, Uncropping, Colorization, JPEG Restoration에 모두 적용할 수 있는 구조를 사용 (ADM의 U-Net Architecture를 가져옴, 256*256)

   B. 기존 모델들은 Unconditional한 Diffusion p(x)를 학습하는 것을 목표로 했으나, 일반적인 경우 Unconditional Diffusion Model이 존재하는 모든 데이터셋을 학습하여 다양한 Task의 결과를 얻어내는 과정은 상당히 어렵기 때문에, p(x)가 아닌, p(y|x)를 목표로 함.

   C. 이때, Condition을 주입하는 방법은 Concatenation.

   D. 결과적으로 다양한 Task의 condition을 줬을 때, condition을 주입하지 않고 학습한 기존 모델들과 달리 다양한 Task(Task마다 다른 Condition주입)에서 높은 성능을 보임.

3. 본인이 생각하는 이 논문의 장점이 무엇이라고 생각하나요? **논문 Contribution bullet을 그대로 따라 적으면 안됩니다.**
   A. L1 vs L2에 대한 실험

   B. 다양한 option에 따른 모델 결과 (Unconditional vs Conditional)

   C. 양적, 질적 관점에서 모두 고려한 결과물

D. 다양한 Low-level Task측면에서 여전히 고려되고 있는 아키텍처 구조

4. 이 논문을 읽으면서 느낀 점, 혹은 배운 점이 있으면 적어보세요.
   A. 다양한 실험 – metrics 기반으로 진행되는 과정에서 어떤 방식으로 Cloud Removal 실험을 진행하면 좋을지 생각할 수 있는 논문

   B. 또한, 다양한 i2i (i2sb, DBIM)모델들이 해당 논문의 구조(ADM -> Palette로 발전되는 모델 구조)를 기반으로 모델을 작성하고 있기 때문에 모델 설계를 하는 측면에서도 도움이 되었다.

5. 이 논문의 한계점이 있다면 무엇이라고 생각하나요?
   A. Sampling 속도

6. 본인의 연구에 접목시켜볼 점이 있을지 생각하고 적어보세요.
   A. L1 vs L2에 대한 직접적인 성능비교 (L1의 경우 – 다양성 측면에서 낮은 점수를 보였으나 강건한 sample을 보여줌): 실제 Target을 가진 상태에서 Restoration을 목표로 하는 Cloud Removal을 하는 경우 L1을 사용하는 것이 Uncertainty를 상승시키는데 더 효과적이라 판단됨

   B. U-Net의 Attention Layer 성능 비교 실험을 통해 확인. (Global Attention의 결과로 blur한 형태의 output이 나온다고 판단. 그러나 여전히 Global Attention이 Replaced Layer(3x3, 5x5, 7x7 – Receptive Field 확장, Local Attention – head 개수:4개) 보다 성능이 좋게 나오고 있다는 점. 지금 사용하는 모델에서 – Attention을 대체하기 보다는 계산적으로 light한 layer를 사용하는 방향성 고려.

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

날짜: 2025-08-21

이름: 신준원

Palette : Image-to-Image Diffusion models

Introduction

Image-to-Image translation ⇒ ex) inpainting, JPEG restoration 등...
single input & multiple output ⇒ 어려운 task

↳ 일반적으로, output 분포 (conditioned) 을 학습하도록 함.

↳ output의 true 분포 학습 = 목표.

↳ tractable하게 condition 주입.

해당논문 ⇒ Diffusion이 ImageNet에서도 GAN을 능가한 성능을 보인것을 앎.

그러나, 실제 Task에서 보았을때 ⇒ GAN만큼 다재다능할까?

⇒ ① Colorization / ② un-cropping / ③ inpainting / ④ JPEG restoration 기준 ⇒ 관련 검증

↳ palette model ⇒ 특정한 Loss, Condition 없이도 4가 task 성능 우수!

↳ 기존 diffusion model

↳ train unconditional model. (난이도↑)

⇒ use in Conditional Task

결과적, Conditional한 train ⇒ for multiple Task

Palette

Condition Diffusion model 기반 (denoising process. Conditional on an input signal)

↳ $P(y|x)$

↳ e.g.) Color, Grayscale

↳ 1 or 5 second.

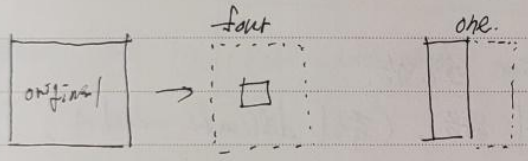
Appendix) Detail

① human evaluation (Image-To-Image !)

Colorization : grayscale (x), RGB (y)

Inpainting : Rectangular mask (10 ~ 40 %) + 0. (random) with Algimple.

Uncropping : four direction or one direction



JPEG restoration : (5, 30)

Learning process (Denoising)

output : y noise : γ

$$\left\| f_\theta(x, \sqrt{\gamma}y + \sqrt{1-\gamma}\epsilon, \gamma) - \epsilon \right\|_p^p$$

for destoration.

γ : noise indicator

$p=1 \Rightarrow L_1$ ↳ L_1 hallucination 적음 ; diversity 적음.

$p=2 \Rightarrow L_2$ L_2 diversity ↑.

Architecture

based on 256×256 ADM + ① Class condition 사용 X

② Concat 을 통한 Conditioning 사용.

Evaluation

ia: ⇒ evaluation 이 매우 어려움.

① FID

② human evaluation (3s, 5s)

③ Qualitative evaluation

④ PSNR, SSIM (reference based score)

⊕ Benchmark dataset 의 부재.

본 논문에서는 ⇒ Imagenet 기반, task (4개) 에 대한 evaluation이 가능한 protocol 제안

↳ pixel-level evaluation (예: PSNR, SSIM) 회피.

⇒ PSNR, SSIM = blur Image에 대해서 높은 점수.

따라서, human standard 와 맞지 않음. (특히 difficult task시)

(with Res50)

⇒ 사용한 metrics IS, FID, CA, PD

└ (Perceptual Distance)

Feature space 상 유클리드 거리.

Sample diversity는 : SSIM & Lpips 사용

human evaluation : 2AFC → 3s와 5s (⇒ fool rate)

Experiments

task별 비교.

Palette ⇒ 특정 task x. vs 특정 Task 특화 모델

실험 : Palette 수.

Self-attention layer in V-Net

이전 연구(12)에서 self-attention을 사용한 이유

⇒ Unseen 해상도 Generalization에 효과적임

⇒ 본 논문에서는 해상도 layer (self-attention)의 영향력을 파악하려 실험.

목표1) trade-off (Large context dependency ↔ resolution robustness)

- Baseline ⇒ 32, 8, 16 (Global) → 각각의 Q.
- Local ⇒ 32, 8, 16 (local)
  ↳ 각각의 Query Block의 크기.
- w/o attention ResBlock ⇒ Attention 대신 2개의 ResBlock 추가 (Receptive field ↑)
- w/o attention + Dilated Conv Block (3x3, 7x7)

결과: Global self attention > Dilated Conv > Conv > Local attention

L_1 Loss vs L_2 Loss

Quantitave Compare L_1 vs L_2 (sample Quality)

Lower SSIM, High Lpips] = Diversity ↑. $\frac{L_2}{\text{Diversity}}$ < $\frac{L_1}{\text{Diversity}\downarrow}$

↳ mode collapse의 원인?

Uncondition vs Condition

Multi Tasking 관점에서.

⇒ Unconditional model (trained) ⇒ task 구분 = ~~학습 noise로~~ output 구분

왜? 다양한 Data를 학습하기에 어려움.

결과적, Conditional model이 더 좋은 결과 보임.

↳ Palette.