# Paper summary

AI VISION Lab

1. 공부한 논문의 제목, 게재된 학회 혹은 저널 등 논문 기본 정보를 적으세요.
   - **A. 이름: 신준원**
   - **B. 저널: CVPR**
   - **C. 도메인: Diffusion**
   - **D. 출판연도: 2023**
   - **E. 저자: Jing Nathan Yan, Jiatao Gu, Alexander M. Rush**

2. 논문에서 제안한 알고리즘 및 프레임워크에 대해 본인이 이해한대로 다이어그램을 그려보세요. **논문 Figure를 그대로 따라 그리면 안됩니다.**
   - A. [선행연구] Attention Block은 U-Net, Transformer를 사용할 때 빠짐없이 사용됨. 그러나, 해당 Block은 고해상도에서 높은 연산량, 파라미터 수를 요구한다는 단점이 있음.

   - B. 따라서, 해당 논문에서는 Attention Block 없이 학습가능한 Architecture를 제안하고 있음.

   - C. SSM (State Space Model) Block을 제안하고 있음. Transformer의 경우 Patchify를 요구하고 있는데, 이 과정에서 High-frequency 표현, Patch 결합 과정의 불안정성 등의 문제가 있다고 판단하고 있음. 따라서, Patch화 하지 않은 상태에서, Attention Block을 제거할 수 있다는 점에서 SSM Block은 우수한 선택이라 할 수 있음. RNN 형태의 구조와 닮았으나, SSM의 Equation에 따라, 연산량을 감소시킬 수 있음(sub-quadratic)과 동시에, Convolutional한 계산이 가능함.

   - D. 또한, Input Image(Noised)를 직접 사용하지 않고, Latent Diffusion의 Encoder를 사용, 이를 Flatten화 해서 사용함(DiT와 동일- Patch, MLP).

   - E. Condition(C, T)는 Scale-Shift 형태로 적용하며, Attention은 Elemental-wise sum + Mul 형태로 처리하고 있음.

   - F. 결과적으로, Linear Decoder를 사용해, Input 과 동일한 Size로 통일해줌(DiT

와 동일 -> Noise, Covariance Prediction).

3. 본인이 생각하는 이 논문의 장점이 무엇이라고 생각하나요? **논문 Contribution bullet을 그대로 따라 적으면 안됩니다.**
   A. Attention Block을 제거했음에도, 비슷한 성능을 뽑았다는 점은 유의미하다고 판단됨.

4. 이 논문을 읽으면서 느낀 점, 혹은 배운 점이 있으면 적어보세요.
   A. 고해상도 문제 Attention의 연산량이 높다는 점을 문제로, 마치, U-Net을 Transformer로 대체했던 것처럼, 이를 효과적으로 대처했다는 점에서, 현재 정답이라고 여겨지는 구조에 한 번씩 의문을 가져보는 시도가 필요해 보임.

5. 이 논문의 한계점이 있다면 무엇이라고 생각하나요?
   A. Diffusion Transformer, DiT-XL 대비 연산량은 증가했으나, 성능적으로는 부족한 모습을 보인다는 점이 아쉬움.

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

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

날짜: 2025-07-16

이름: 신준원

Diffusion models without Attention 07/16

$\boxed{\text{introduction}}$

Diffusion → 연산량이 너무 높다. ⇒ Diffusion State space model (DiffuSSM)

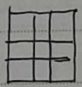
연산량 많은 Attention ⇒ State space model backbone으로 대체

↓

U-Net, transformer 모두 필요.

① representation Compression

→ patchifying.



단점 → high-frequency ↓ (spatial information)

(trade-off) structural integrity

↓

나눠서. 진행 ⇒ 전체구조 파악에 불리.

→ multi-scale resolution

attention 연산량 ↓.

~~Channel~~ → 계산영향

resolution

단점 → spatial detail ↓ (down-sample)

∘ Generate. 가짜 (up-sample)

→ sub-quadratic space

(one of attention approximation method)

② DiffuSSM

→ gated state space model (GSM) backbone 활용

→ (LRA, Audio → 사용됨)

↓

$\boxed{\text{Sequence model}}$ 성능 향상 이미 존재.

⊕ hourglass architecture

↓

VAE (Latent model)의 encoder 사용.

(sequence 구조 유지)

→ z (representation)를 flatten 해서 처리.

## $\boxed{\text{State space models (SSM)}}$

input sequence of scalars $= u_1, u_2, u_3, \dots u_L$ / $\underline{\text{output}} = y_1, y_2, \dots, y_L$

(N×1)      1×N.

equation => $\underset{N\times N}{x_k} = \underset{N\times N}{\bar{A}} x_{k-1} + \underset{N\times 1}{\bar{B}} u_k$ , $\underset{1\times N}{y_k} = \underset{1\times N}{\bar{C}} x_k$

장점 : long convolutional 연산 가능 (with linear equation)

특성 : ① $\underline{\text{FFT}}$ (Fast Fourier Transform)

↳ $\underline{\text{sub-quadratic}}$ Algorithm.

↳ $O(N^2L)$ ⇒ 해당 Algorithm $O(L \log L)$

) → long-sequence handling에 용이함

② $\underline{\bar{A}, \bar{B}, \bar{C}}$ (discrete-time values)

↳ Continuous-time state-space 에서 얻는다면.

⇒ stable & effective한 접근 가능

따라서, 본문에서는 ⇒ $\underline{A.B.C}$ 를 Continuous한 상태에서 discrete 하게 변경하는 구조 사용 (S4D)

⊕ Bidirectional SSM layer 사용

⇒ $\underline{\text{RNN} \sim \text{Bidirection}}$.

↳ Flatten, Global feature 반영 문제 해결을 위한 시도.

# Diffussm Block

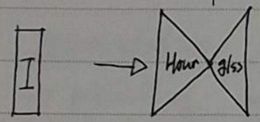
구성요소

① Gated bidirectional SSM

② hourglass Architecture (with m/p) → flattened input → m/p 적용

→ Sequence.

모래시계 → expanding ⟷ Contracting. 교차 반복.




$J \times D$

$m = L/J$ → 고해상도 (long length) / → 저해상도 (low ~~log~~ length).

m → 비율

flatten. → m/p based 기반.