## 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

introduction  Diffusion to the test of the set of the space model (Diffusion of the set	Diffusion models without Attention 17/16	
Diffusion to the total the total to pace model (Diffusion of the state	introduction	
U-Net, transformer 2 2/2.  Tepresentation Compression  Depresentation Compression  Depresentation Compression  Depresentation Compression  Depresentation Compression  Depresentation Compression  Chade-off) Atructural integrity  When they I seem the series of the serie	Diffusion -> HUZED HY SET. > Diffusion State Space model (	Diffu 45/4
Depresentation Compression  Depresentation Compression  Depresentation Compression  Depresentation Compression  That is information  Depresentation Compression  That is information  That is informat	Httl et 2 Attention > 4 Labe 4 pace model backbone = 2 24/21	
Depresentation Compression  Depresentation Compression  Depresentation Compression  Depresentation Compression  That is information  Depresentation Compression  That is information  That is informat	U-Net, transformer 87 2/2.	
Depresentation Compression  Depresentation Compression  Depresentation  Depresentation  Charles in the printing of the content of the printing of the partial detail of (down-sample)  Attention offset.  Charlest p May 35th  Degolishan  Cone of attention approximation method)  Diffu SS/N  Deform the CARA. And is a few for the partial sequence of the state	gratial int	ormation
That sole resolution   the patial detail of (down-sample)  Attention offet.   o Generate (122) (Up-gample)  Channot + mysts  Legolotion   the solution   the following pace  Cone of attention approximation method)  Diffussin  Plated state space model (45/M) back bone the solution of the	1) representation Compression	
Thurston office desolution   the potial detail to (down-sample)  Attention office   o Generate . 122 (Up-gample)  Charact -> Haysot.  Jesulation   p sub-guadratic space  (one of Attention Approximation method)  Diffussin   p (LRA, Autio > 423)  -> gated state space model (49/M) back bone 24%  Sequence model (49/M) back bone 24%  Sequence model (49/M) back bone 24%  VAL (Latent model) 21 encoder 45.	- patchiffing. The pale of the fuery	
Thurston office desolution   the potial detail to (down-sample)  Attention office   o Generate . 122 (Up-gample)  Charact -> Haysot.  Jesulation   p sub-guadratic space  (one of Attention Approximation method)  Diffussin   p (LRA, Autio > 423)  -> gated state space model (49/M) back bone 24%  Sequence model (49/M) back bone 24%  Sequence model (49/M) back bone 24%  VAL (Latent model) 21 encoder 45.	(trade-off) structural integrity-	
Attention offity.  Charact & Milyson   2egolation  Cone of Attention Approximation method)  (one of Attention Approximation method)  Diffu SSM  Fratel Flate Space model (49M) back bone 24%  Sequence model (49M) back bone 24%  Sequence model (49 m) back bone 24%  Vet (Latent model) 2 encoder 4%.	HLAKI ABY > EV	4/72 In of
Attention offity.  Charact & Milyson   2egolation  Cone of Attention Approximation method)  (one of Attention Approximation method)  Diffu SSM  Fratel Flate Space model (49M) back bone 24%  Sequence model (49M) back bone 24%  Sequence model (49 m) back bone 24%  Vet (Latent model) 2 encoder 4%.	-> but - gralo resolution   2/21 -> Sportial detail to (down	-gample)
Chappet - 7 Highton  2egolotion  (one of Attention Approximation method)  (one of Attention Approximation method)  (one of Attention Approximation method)  - D (LRA, Autio > 425)  - D fated Flate apace model (45M) back bone 225  Sequence model (45M) back bone 225  Sequence model (45M) back bone 225  VAZ (Latent model) -   encoder 45  VAZ (Latent model) -   encoder 45	Attention offety. O Generate . 124 (UP-9a)	mple)
(one of Attention Approximation method)  (a) Diffu SSM  Diffu Dif		
(one of attention Approximation method)  (a) Diffu SSM		
Sequence model (49/M) back bone \$25  Sequence model (49/M) back bone \$25  Sequence model (49/M) back bone \$25  Degreence model (49/M) back bone \$25  VAZ (Latert model) =   encoder \$25  VAZ (Latert model) =   encoder \$25		
Sequence model (49/M) back bone \$25  Sequence model (49/M) back bone \$25  Sequence model (49/M) back bone \$25  Degreence model (49/M) back bone \$25  VAZ (Latert model) =   encoder \$25  VAZ (Latert model) =   encoder \$25	( LRA, ANDIO > ARE)	
Sequence model st of of open of the open of the sequence of the first of the open of the sequence of the seque	- + fatel state space model (69/h) back bone \$15	
De hourglass architecture VAZ (Latert moder) =   encoder &		
De hourglass architecture VAZ (Latert moder) =   encoder &	Sequence model "Sto that - 124 \$24.	
SE Greate 12 12)	1 hourglass architecture	
SE Greate 12 12)	Was 1 to 1 to 1 to 1 de	
P? (depresentation) = flatter = 14 421.	Se guence 72 421) VHE ( Latert model ) 2   lu Coder 40	
1 (september) = flatter of 421	42 (delegately) · Ny	
	flatter =19 42	

State space models (55m) input sequence of scalars = h, ur, uz, ... h\_ output = y, yr, ... . YL Upotion =>  $P|_{X} = \overline{A} d_{K-1} + \overline{B}u_{K}$   $\int_{X}^{X} = \overline{C} d_{K}$   $\int_{X}^{X} \int_{X}^{X} \int$ 3/28: long convolutional that (with linear equation) 57 : Off (fast fourier Transform)

D Sub-guaratic Alforithm

D (N24) > 54% Alforithm O(LlogL) D Continuous - time 44te-space only Itzent > Stable & effective to 22 75 四州, 老和树 > A.B.C = Continuous & szynky discrete # 母語是 往 (S4D) 1 Bidirectional SSM layer As > KNN ~ Bidirection Lp Flather, Global feature God & Em = 542 915 45

