OS1A-S4

VQ-VDM: ベクトル量子化変分オートエンコーダと拡散モデルを用いた動画生成モデル

電気通信大学　梶 凌太　柳井 啓司

はじめに

近年画像生成は目覚ましい発展を遂げており、その次のモダリティとして動画生成に対しても注目が集まっている

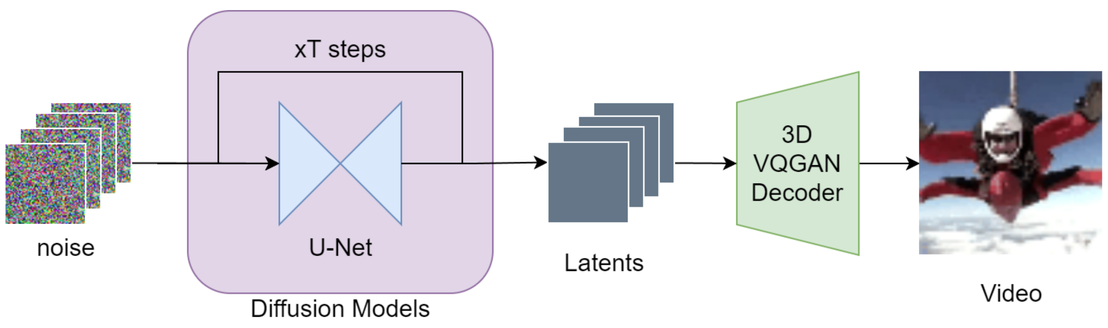
しかしながら、**動画生成**は画像生成と比較して
・計算量の多さ
・時間方向の考慮
などから困難なタスクとなっている

目的

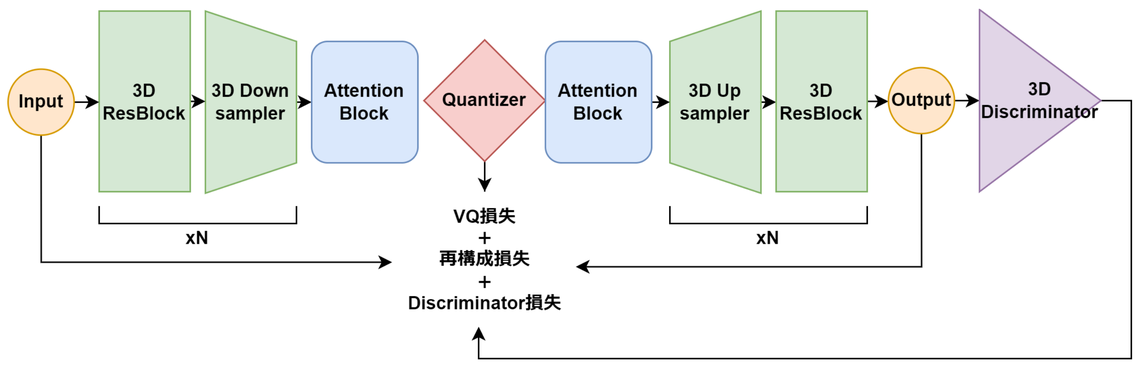
2-Stage手法による**軽量**な動画生成手法の提案

手法

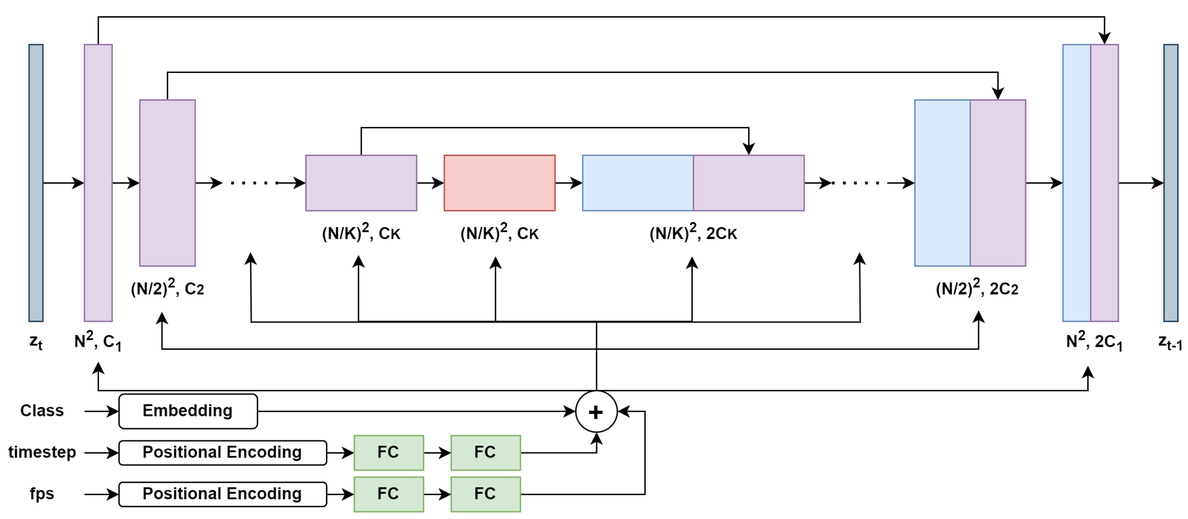
本研究は3D VQGANによる動画圧縮とDiffusion Modelsによる動画生成の2段階で構成される



3D VQGANは3D畳み込みを用いて実装され、再構成損失・VQ損失・Discriminator損失で学習される



Diffusion Modelsの各ブロックは3D Residual block・Spatial Attention・Temporal Attentionから構成される
クラス条件付けのための埋込みとFPSのエンコーディングを各Residual blockで条件付けする



Classifier-freeガイダンスを用いるため、0.5の確率でクラス情報を落として学習を行う

実験概要

動画データセットUCF-101とSky Time-lapseで学習・評価
・UCF-101は101種類のアクションが含まれた動画13320本
・Sky-timeplaseは空のタイムラプス動画約5000本

学習・評価時には16フレーム128x128解像度の動画を生成
評価指標にはISとFVD、およびKVDを用いた

生成速度の比較としてDiffusionベース手法のVDMと比較した

実験

1. UCF-101を用いた実験

最先端手法であるTATS**には劣るが、それ以外の手法を上回る**

手法	ベース	解像度	IS(↑)	FVD(↓)	会議
VideoGPT	自己回帰	128x128	24.69	-	arXiv 2021
DVD-GAN	GAN	128x128	27.38	-	arXiv 2019
TGANv2	GAN	128x128	28.87	1209	IJCV 2020
DIGAN	GAN	128x128	32.70	577	ICLR 2021
CogVideo*	自己回帰	160x160	50.46	626	arXiv 2022
VDM	Diffusion	64x64	57.00	-	NIPS 2022
TATS	自己回帰	128x128	79.28	332	ECCV 2022
Ours	Diffusion	128x128	**64.13**	**425**	-

＊CogVideoは学習済みの大規模画像生成モデルの重み利用＋540万本の動画による事前学習を行っている

2. Sky Time-lapseを用いた実験

提案手法は全ての手法を上回るFVDを達成した

手法	ベース	解像度	FVD(↓)	KVD(↓)	会議
MoCoGAN-HD	GAN	128x128	183.6	13.9	ICLR 2021
DIGAN	GAN	128x128	114.6	6.8	ICLR 2022
TATS	自己回帰	128x128	132.6	5.7	ECCV 2022
Ours	Diffusion	128x128	**109.4**	**5.9**	-

3. 生成速度の比較

提案手法は128x128の解像度、VDMは64x64の解像度で生成し100ステップあたりで比較

手法	解像度	100 step time [s]
VDM	16x64x64	35.26±2.43
Ours	16x128x128	**3.95±0.01**

提案手法は空間方向に4倍大きい動画を生成しているが、提案手法は約9倍高速に動作している

おわりに

3D VQGANとDiffusion Modelsを用いた動画生成モデルを提案
3D VQGANにより動画を256倍小さい潜在変数へ圧縮
Diffusion Modelsが潜在変数に対して動作するため学習・生成が高速
生成品質は一部データセットで最先端のFVDを達成
従来手法のVDMより高解像度な動画を約9倍高速に生成

今後の展望
Stable Diffusionなどの大規模画像生成モデルの事前知識を利用し生成品質を高める