Video Diffusion Models论文解读

目录

引言

最近一段时间，扩散模型在文图生成领域可谓名声大噪。实际上，扩散模型可以应用到各类AIGC任务上，除文图生成为代表的图片生成外，扩散模型还可以进行音频生成、时间序列生成、3D点云生成、文本生成。而这篇论文，即**Video Diffusion Models**，就将扩散模型用到了视频生成任务上，本文将对该论文展开讲解。该论文官方没有公开源码，但是会有一些基于该论文的相关开源工作，比如PaddleNLP的PPDiffusers，本文后续也会结合相关代码进行讲解。该论文一作作者为Jonathan Ho和Tim Salimans，来自于谷歌，论文目前已被NeurIPS 2022接收。

论文贡献

论文的主要贡献如下：

- 这是第一个使用扩散模型进行视频生成任务的论文工作，这里的视频生成任务包括无条件和有条件两种设定。
- 论文针对扩散模型中的UNet网络结构进行修改，使其适用于视频生成任务，提出了3D UNet，该架构使用到了space-only 3D卷积和时空分离注意力。这种时空分离注意力的UNet可以应用在可变序列长度上，因此可以在视频和图像建模目标上进行联合训练。
- 为了生成比训练时帧数更多的视频，论文还展示了如何使用梯度条件法进行重构指导采样，从而可以自回归地将模型扩展到更长的时间步长和更高的分辨率。

方法

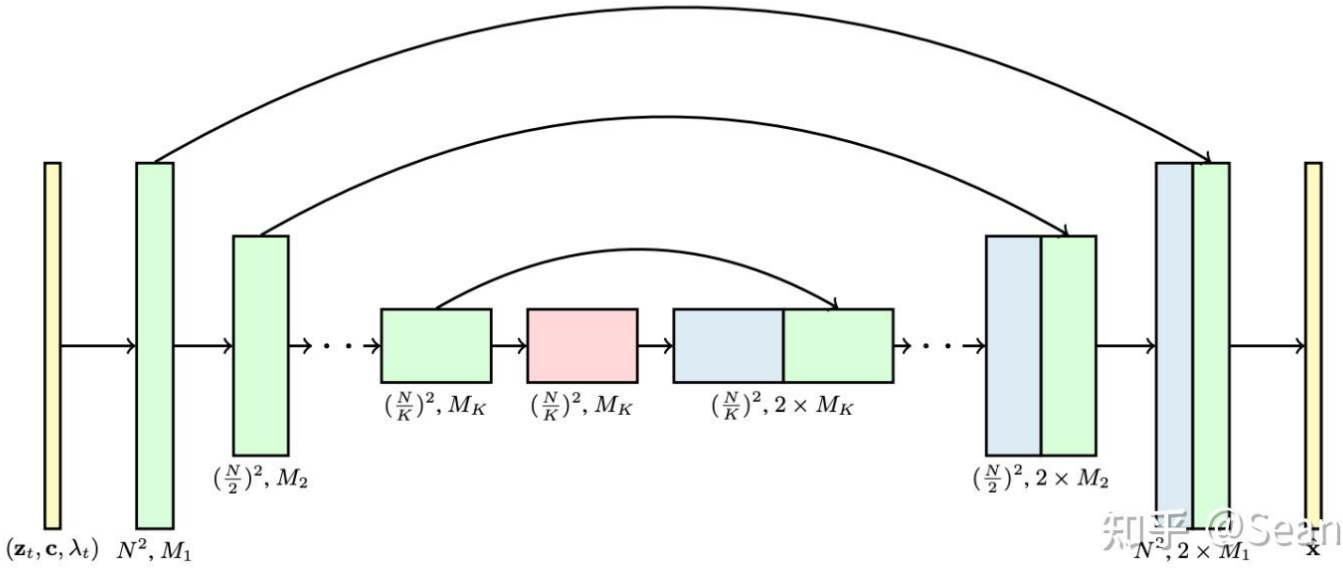
3D UNet

论文提出的视频生成方法基于标准的扩散模型（即高斯扩散模型），最核心的改进在于对作为backbone的UNet网络架构的修改，从而让扩散模型适应视频生成任务。

在图像生成任务上的扩散模型所用到的UNet网络架构通常来自PIXELCNN++改进版本，它先进空间下采样（spatial downsampling）然后进行空间上采样（spatial upsampling），其中在进行每一步空间上采样时都通过跳跃连接（skip connections）来同对应步的下采样过程的特征图进行联系。网络上采样和下采样的基本单元都是2D卷积残差块（2D convolutional residual blocks），只是在改进的版本中，为了能够引入文本这类的条件信息，每个2D卷积残差块后面还跟着一个注意块或者说空间注意块（spatial attention block）。

论文将这一图像扩散模型架构扩展至视频，Video Diffusion Models提出了3D UNet架构。具体来说，该架构将原UNet中的2D卷积替换成了space-only 3D卷积（space-only 3D convolution），举例来说，如果原来用的是3x3卷积，那么现在就要把它替换为1x3x3卷积（其中第一个维度对应视频帧，即时间维度，第二个和第三个维度对应帧高和帧宽，即空间维度，由于第一个维度是1所以对时间没有影响只对空间有影响）。随后的空间注

意块仍然保留，但只针对空间维度进行注意力操作，也就是把时间维度flatten为batch维度。在每个空间注意块之后，新插入一个时间注意块（temporal attention block），该时间注意块在第一个维度即时间维度上执行注意力，并将空间维度flatten为batch维度。论文在每个时间注意力块中使用相对位置嵌入（relative position embeddings），以便让网络能够不依赖具体的视频帧时间也能够区分视频帧的顺序。这种先进行空间注意力，再进行时间注意力的方式，可以称为时空分离注意力（factorized space-time attention）。3D UNet模型的整体架构如下图所示。



3D UNet模型整体架构

其中每个方块都代表一个四维的张量（即frames × height × width × channels），该图中每个方块的纵轴长度表示张量的长或宽大小（即height或width ），横轴长度表示张量的通道大小（即channels）。输入是噪音视频 $\mathbf{z}_t$、条件 $\mathbf{c}$以及log SNR λ_t，下采样或上采样中块与块之间的空间分辨率（即height × width）调整比率是2，使用通道乘子（channel multipliers）$M_1, M_2, \ldots, M_K$来指定通道的数目。模型通过卷积和时空分离注意力的方式来处理每一个块，在进行每一步空间上采样时都通过跳跃连接来同对应步的下采样过程的特征图进行联系。

这种时空分离注意力的方式有一个好处是可以对视频和图片生成进行联合建模训练。就是说可以在每个视频的最后一帧后面添加随机的多张图片，然后通过掩码的方式来将视频以及各图片进行隔离，从而让视频和图片生成能够联合训练起来，论文在实验部分也对这种联合建模训练的方式其进行了探索。

重构指导采样

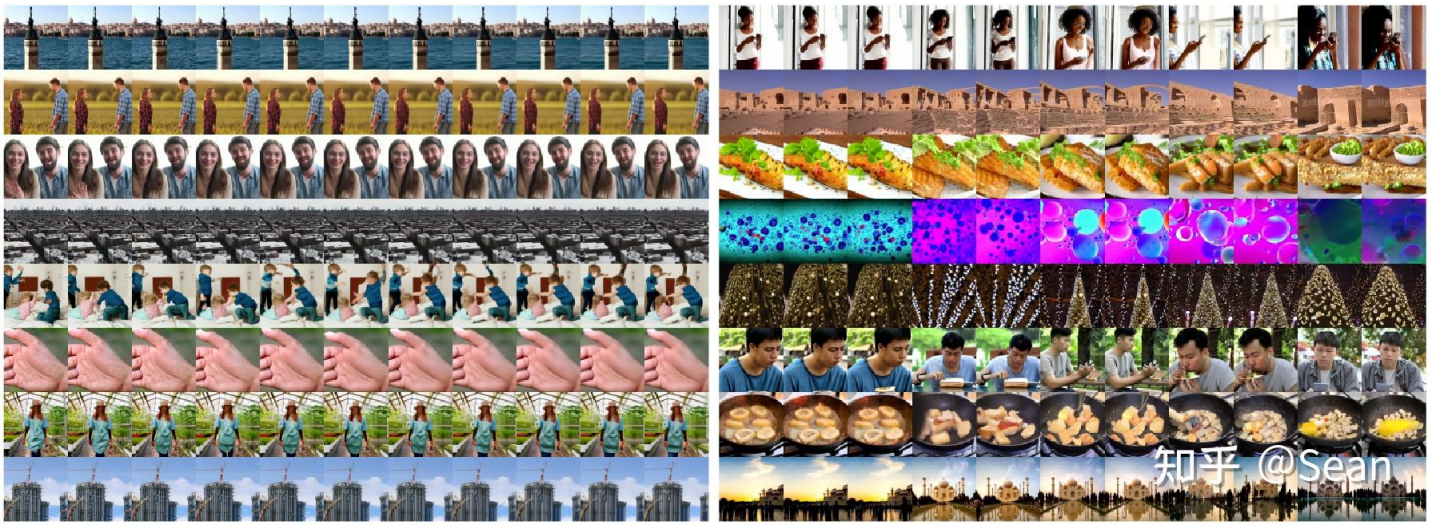
论文的另一个主要创新是为无条件扩散模型提供了一种条件生成的方法。这种条件生成方法称为梯度条件法（gradient conditioning method），它修改了扩散模型的采样过程，使用基于梯度优化的方式来改善去噪数据的条件损失（conditioning loss），从而可以让生成的视频通过自回归地方式扩展至更长的时间步和更高的分辨率。由于梯度条件法中所使用的附加梯度项可以解释为一种额外的指导，而这种指导其实基于模型对条件数据的重建，因此论文将该

方法称为重建引导采样（reconstruction-guided sampling），或简单地称为重建指导（reconstruction guidance）。

在训练的阶段，由于计算资源限制，往往只能在一个固定的帧数下面训练一个视频，而且这个帧数往往很短（比如论文中的16帧）。但是在采样（推理）阶段，我们可以先生成一个16帧的视频 $\mathbf{x}^{\mathbf{a}} \sim p_\theta(\mathbf{x})$，然后在这个基础上拓展得到第二个视频 $\mathbf{x}^{\mathrm{b}} \sim p_\theta\left(\mathbf{x}^{\mathrm{b}} \mid \mathbf{x}^{\mathrm{a}}\right)$，这样一来就可以通过自回归的方式拓展采样的视频到任意长度。但是这种采样方式需要我们显式地训练一个条件生成模型 $p_\theta\left(\mathbf{x}^{\mathrm{b}} \mid \mathbf{x}^{\mathrm{a}}\right)$，或者通过插值的方式从无条件生成模型 $p_\theta(\mathbf{x})$近似得到。之前的工作通常都是通过“替换法”来进行自回归视频拓展，简单来说替换法就是对两条件样本进行联合训练 $p_\theta\left(\mathbf{x}=\left[\mathbf{x}^{\mathrm{a}}, \mathbf{x}^{\mathrm{b}}\right]\right)$，但是在扩散模型前向的具体过程中 $\mathbf{x}^{\mathrm{b}}$对应的部分保持正常的迭代更替，而 $\mathbf{x}^{\mathrm{a}}$对应的部分被替换为一个固定的 $q\left(\mathbf{z}_s^{\mathrm{a}} \mid \mathbf{x}^{\mathrm{a}}\right)$或者 $q\left(\mathbf{z}_s^{\mathrm{a}} \mid \mathbf{x}^{\mathrm{a}}, \mathbf{z}_t^{\mathrm{a}}\right)$，也就是始终需要参考 $\mathbf{x}^{\mathrm{a}}$。 论文认为这种方式下$\mathbf{x}^{\mathrm{b}}$对应的部分的更替仅仅是 $\hat{\mathbf{x}}_\theta^{\mathrm{b}}\left(\mathbf{z}_t\right) \approx \mathbb{E}_q\left[\mathbf{x}^b \mid \mathbf{z}_t\right]$，而真正理想的更替应该是 $\mathbb{E}_q\left[\mathbf{x}^b \mid \mathbf{z}_t, \mathbf{x}^a\right]$， 这样才能够和上一个视频有更好的一致性。论文将上述更替方案改写为 $\mathbb{E}_q\left[\mathbf{x}^b \mid \mathbf{z}_t, \mathbf{x}^a\right]=\mathbb{E}_q\left[\mathbf{x}^b \mid \mathbf{z}_t\right]+\left(\sigma_t^2 / \alpha_t\right) \nabla_{\mathbf{z}_t^b} \log q\left(\mathbf{x}^a \mid \mathbf{z}_t\right)$，它使用基于梯度优化的方式来改善去噪数据的条件损失，提出的条件重构采样，其公式如下：

$$\tilde{\mathbf{x}}_\theta^b\left(\mathbf{z}_t\right)=\hat{\mathbf{x}}_\theta^b\left(\mathbf{z}_t\right)-\frac{w_r \alpha_t}{2} \nabla_{\mathbf{z}_t^b}\left\|\mathbf{x}^a-\hat{\mathbf{x}}_\theta^a\left(\mathbf{z}_t\right)\right\|_2^2 \quad (1)$$

论文发现，在确保生成样本与条件信息的一致性方面，梯度条件法比之前的方法更加有效。下图左为利用梯度条件法通过自回归方式拓展得到的视频帧，图右为利用基线的替换（replacement）法通过自回归方式拓展得到的视频帧。可以看到，使用梯度条件法采样的视频比基线方法具有更好的时间一致性。



梯度条件法（左）和基线"替换"法（右）在自回归扩展下生成的视频帧。与基线方法相比，使用梯度条件法采样的视频获得了更好的时间一致性。

实验

论文在无条视频生成（Unconditional video modeling）、视频预测（Video prediction）、文本条件视频生成（Text-conditioned video generation）三个任务上进行了实验和评估。

无条件视频生成

无条件视频生成使用Soomro数据集。论文对该数据集的16帧短视频片段进行建模，并下采样到64x64的空间分辨率。实验对比结果如下，相比之前的模型取得了新的SOTA结果：

Method	Resolution	FID↓	IS↑
MoCoGAN [52]	16x64x64	26998 ± 33	12.42
TGAN-F [26]	16x64x64	8942.63 ± 3.72	13.62
TGAN-ODE [18]	16x64x64	26512 ± 27	15.2
TGAN-F [26]	16x128x128	7817 ± 10	22.91 ± .19
VideoGPT [62]	16x128x128		24.69 ± 0.30
TGAN-v2 [41]	16x64x64	3431 ± 19	26.60 ± 0.47
TGAN-v2 [41]	16x128x128	3497 ± 26	28.87 ± 0.47
DVD-GAN [14]	16x128x128		32.97 ± 1.7
Video Diffusion (ours)	16x64x64	**295 ± 3**	**57 ± 0.62**
real data	16x64x64		6[illegible]

视频预测

视频预测任务是指给定视频的第一帧预测剩余的帧，它是一个有条件的视频生成任务。论文使用前面所说的重构指导采样来进行该任务。论文在BAIR Robot Pushing和Kinetics-600两个数据集上进行实验，实验结果如下，同样达到了新的SOTA水平：

Method	FVD↓
DVD-GAN [14]	109.8
VideoGPT [62]	103.3
TrIVD-GAN-FP [33]	103.3
Transframer [35]	100
CCVS [31]	99
VideoTransformer [59]	94
FitVid [4]	93.6
NUWA [61]	86.9
Video Diffusion (ours)	
ancestral sampler, 512 steps	68.19
Langevin sampler, 256 steps	**66.92**

Method	FVD↓	IS↑
Video Transformer [59]	170 ± 5	
DVD-GAN-FP [14]	69.1 ± 0.78	
Video VQ-VAE [57]	64.3 ± 2.04	
CCVS [31]	55 ± 1	
TrIVD-GAN-FP [33]	25.74 ± 0.66	12.54
Transframer [35]	25.4	
Video Diffusion (ours)		
ancestral, 256 steps	18.6	15.39
Langevin, 128 steps	**16.2 ± 0.34**	**15.64**

文本条件视频生成

在该任务中，论文使用了1千万的带标题视频（captioned videos）作为训练数据。论文使用BERT-large获取标题的词嵌入表征作为扩散模型的条件输入。论文对视频-图片联合训练（joint video-image training）、无分类器指导（classifier-free guidance）、自回归视频拓展（autoregressive video extension for longer sequences）进行了探索。

对于视频-图片联合训练，下表展示了文本条件下，16x64x64视频的实验结果，每个视频有额外的0、4或8个独立图像帧进行联合训练。可以看到，随着添加更多独立的图像帧，采样视频和图像的样本质量明显改善。

Image frames	FVD↓	FID-avg↓	IS-avg↑	FID-first↓	IS-first↑
0	202.28/205.42	37.52/37.40	7.91/7.58	41.14/40.87	9.23/8.74
4	68.11/70.74	18.62/18.42	9.02/8.53	22.54/22.19	10.58/9.91
8	57.84/60.72	15.57/15.44	9.32/8.82	19.25/18.98	10.81/10.12

对于无分类器指导，在文本条件视频生成任务中增加指导可以提高每帧图片的保真度（fidelity），增强条件信号的效果。下表展示了不同指导权重（guidance weight）下的指标，可以看到更高指导权重下IS指标上的表现有着明显的提升，而FID指标表现上随着指导权重的增加先提升后下降，这个结果也和之前文本条件图片生成的结果一致。

Frameskip	Guidance weight	FVD↓	FID-avg↓	IS-avg↑	FID-first↓	IS-first↑
1	1.0	41.65/43.70	12.49/12.39	10.80/10.07	16.42/16.19	12.17/11.22
	2.0	50.19/48.79	10.53/10.47	13.22/12.10	13.91/13.75	14.81/13.46
	5.0	163.74/160.21	13.54/13.52	14.80/13.46	17.07/16.95	16.40/14.75
4	1.0	56.71/60.30	11.03/10.93	9.40/8.90	16.21/15.96	11.39/10.61
	2.0	54.28/51.95	9.39/9.36	11.53/10.75	14.21/14.04	13.31/12.63
	5.0	185.89/176.82	11.82/11.78	13.73/12.59	16.59/16.44	16.24/14.62

对于自回归视频拓展，论文对比了重构指导采样(reconstruction guidance)和替换法(replacement)生成视频的效果，并发现这种方法的确实优于替换法，对比结果如下表所示：

Guidance weight	Conditioning method	FVD↓	FID-avg↓	IS-avg↑	FID-first↓	IS-first↑
2.0	reconstruction guidance	136.22/134.55	13.77/13.62	10.30/9.66	16.34/16.46	14.67/13.37
	replacement	451.45/436.16	25.95/25.52	7.00/6.75	16.33/16.46	14.67/13.34
5.0	reconstruction guidance	133.92/133.04	13.59/13.58	10.31/9.65	16.28/16.53	15.09/13.72
	replacement	456.24/441.93	26.05/25.69	7.04/6.78	16.30/16.54	15.11/13.69

效果展示

来自文本条件视频扩散模型的样本，以“烟花”为条件

参考

https://arxiv.org/pdf/2204.03458.pdf

https://video-diffusion.github.io/

https://github.com/PaddlePaddle/PaddleNLP/tree/develop/ppdiffusers