文章编号:1671-6833(2024)01-0070-08

基于门控时空注意力的视频帧预测模型

李卫军[1,2], 张新勇[1], 高庾潇[1], 顾建来[1], 刘锦彤[1]

(1.北方民族大学 计算机科学与工程学院,宁夏 银川 750021;2.北方民族大学 图像图形智能处理国家民委重点实验室,宁夏 银川 750021)

摘 要:针对循环式视频帧预测架构存在精度低、训练缓慢,以及结构复杂和误差累积等问题,提出了一种基于门控时空注意力的视频帧预测模型。首先,通过空间编码器提取视频帧序列的高级语义信息,同时保留背景特征;其次,建立门控时空注意力机制,采用多尺度深度条形卷积和通道注意力来学习帧内及帧间的时空特征,并利用门控融合机制平衡时空注意力的特征学习能力;最后,由空间解码器将高级特征解码为预测的真实图像,并补充背景语义以完善细节。在 Moving MNIST、TaxiBJ、WeatherBench、KITTI 数据集上的实验结果显示,同多进多出模型 SimVP 相比,*MSE* 分别降低了 14.7%、6.7%、10.5%、18.5%,在消融扩展实验中,所提模型达到了较好的综合性能,具有预测精度高、计算量低和推理效率高等优势。

关键词:视频帧预测; 卷积神经网络; 注意力机制; 门控卷积; 编解码网络

中图分类号:TP391.41;TP183 **文献标志码**:A **doi**:10.13705/j.issn.1671-6833.2024.01.017

近年来,随着科技的飞速发展,智能设备得到了广泛的普及,由此产生了海量的无标签视频数据。智能预测与决策系统在生活中具有重要的地位,视频帧预测作为智能预测的关键技术,能够为决策系统提供支持,在气象预警[1]、交通流量[2]等领域具有广泛的应用前景。

目前,视频帧预测模型的多帧预测能力不足,其复杂的时空结构导致视频帧预测仍然是一项非常具有挑战性的任务。现有的视频帧预测方法可以分为两类,主要包括单进单出预测架构和多进多出预测架构。其中,单进单出预测架构是视频帧预测的主流结构。Srivastava 等[3]通过编码器将视频序列重建为固定长度的特征向量,并输入到长短期记忆网络(long short term memory,LSTM)中进行多帧预测。为提高 LSTM 的特征捕捉能力,Shi 等[4]采用卷积结构对 LSTM 的状态转移函数进行了扩展。为增强不同层次循环网络间的联系,Wang 等[5]通过在自底向上和自顶向下的方向上建立记忆流,使模型能够同时对短期变化和长期动态趋势进行建模。在此基础上,Wang 等[6]建立了一种基于因果 LSTM 的循环网络,由级联的双存储器和梯度高速单元组成,能够自适应地捕获短期和长期依赖关系。上述方法能够有效增强模型的特征学习能力,但随着预测长度的增加会存在误差累积的问题,导致预测精度迅速下降。

随着神经网络结构的快速发展,多进多出预测架构能够有效避免在长期预测中受到的误差累积影响。Liu 等[7]采用 3D 卷积自编码器学习体素流,并通过现有的流动像素值来合成未来视频帧。Aigner 等[8]提出一种基于时空三维卷积的生成式对抗网络(genertive adversarial network,GAN),该架构能够一次预测多个未来帧。Ye 等[9]分别对空间特征和时间特征进行建模,并采用对抗损失函数来提高预测清晰度。对抗网络和 3D 卷积的引入虽然能够有效提高预测性能,但也导致模型变得更加复杂。

为了平衡模型的综合性能,Gao 等[10]提出了一种简单的视频预测模型(simple video prediction,SimVP),通过采用简单的组成结构和训练策略,以

收稿日期:2023-08-07;**修订日期**:2023-09-21

基金项目:中央高校基本科研业务费专项资金(2021JCYJ12);国家自然科学基金资助项目(61962001);宁夏自然科学基金资助项目(2021AAC03215);北方民族大学研究生创新项目(YCX23147)

作者简介:李卫军(1979—),男,陕西渭南人,北方民族大学讲师,博士,主要从事本体论的构建与再利用、知识图谱构建、深度学习研究,E-mail:lwj@nmu.edu.cn。

引用本文:李卫军,张新勇,高庾潇,等. 基于门控时空注意力的视频帧预测模型[J]. 郑州大学学报(工学版),2024,45(1):70-77,121.(LI W J, ZHANG X Y, GAO Y X,et al. Video frame prediction model based on gated spatio-temporal attention[J]. Journal of Zhengzhou University (Engineering Science),2024,45(1):70-77,121.)

有效减少模型的参数量和训练时间。但 SimVP 仍然存在两个问题:①时空特征学习能力仍然不足;②难以平衡空间特征及时间特征的捕捉能力,导致对时间维度的信息学习不充分。受图像分割[11]领域最新进展的启发,本文提出了门控时空注意力。其中,空间注意力关注帧内空间位置下的相互关系,时间(通道)注意力[12]则关注帧间的变化趋势,并采用门控机制来融合获得的时间特征和空间信息。

1　相关工作

1.1　基于循环神经网络的单进单出预测架构

目前,基于循环神经网络的单进单出预测架构被广泛用于处理序列数据。Wang 等[13]利用相邻隐藏状态之间的差异信息对时空动力学中的非平稳和近似平稳特性进行建模。从预测编码的角度,Lotter 等[14]将真实信号和预测信号之间的差异信息作为网络参数的更新指标。此外,受偏微分方程(PDEs)的启发,Guen 等[15]提出了物理动力学网络(physical dynamics network,PhyDNet),采用双分支架构来分离视频中的物理动力学和未知因素。然而,该模型难以平衡长期和短期的预测性能。因此,Pan 等[16]提出了基于特征分离原理的泰勒网络(Taylor network,TaylorNet),该架构采用泰勒级数对视频序列进行建模,有效提高了模型的多帧预测能力。上述方法通常采用堆叠各种特征学习模块来提高预测效果,导致模型的计算量和参数量过大,这限制了模型的进一步广泛应用。

1.2　基于卷积神经网络的多进多出预测架构

近年来,基于卷积神经网络的多进多出预测架构开始被应用在视频帧预测领域中。Sun 等[17]提出了一种新的 U-net 预测架构,能够对神经网络不同层次中的多个时间和空间尺度进行统一建模。受 Transformer 在计算机视觉领域成功应用的启发,Ning 等[18]提出了一种基于局部时空块扩展的 Transformer 预测架构,通过将二维卷积融合到多头注意力中以捕捉序列中的长期依赖关系。此外,Tan 等[19]提出了一种轻量型时空预测学习框架,采用膨胀卷积构建时空注意力来增强模型的特征捕捉能力。多进多出预测架构通常构建各种模块来增强空间特征的获取能力,但对时间特征的学习仍然不足。

本文受 SimVP 框架的启发,构建了基于门控时空注意力的视频帧预测模型。通过多尺度深度条形卷积和通道注意力来捕捉复杂的时空运动趋势,同时采用门控机制来平衡模型的时空特征学习能力,有效地增强了模型的时空动力学建模能力。

2　本文算法

2.1　问题描述

定义一个 $\boldsymbol{X}=\{\boldsymbol{x}_{t+1},\boldsymbol{x}_{t+2},\cdots,\boldsymbol{x}_{t+\mathrm{m}}\}$ 表示长度为 m 的输入视频帧序列,$\boldsymbol{Y}=\{\boldsymbol{y}_{t+1},\boldsymbol{y}_{t+2},\cdots,\boldsymbol{y}_{t+\mathrm{n}}\}$ 表示待预测的未来 n 帧真实序列,$\boldsymbol{Y}'=\{\boldsymbol{y}'_{t+1},\boldsymbol{y}'_{t+2},\cdots,\boldsymbol{y}'_{t+n}\}$ 表示模型预测的未来 n 帧视频序列,其中 $\boldsymbol{x}_t$,$\boldsymbol{y}_t$ 和 $\boldsymbol{y}'_t$ 分别表示第 t 时刻的原始帧、真实帧和预测帧。模型训练的目的就是通过输入的视频序列 $\boldsymbol{X}$ 来预测未来的视频序列 $\boldsymbol{Y}'$,同时对模型的可学习参数 Θ 进行优化,使真实序列 $\boldsymbol{Y}$ 和预测序列 $\boldsymbol{Y}'$ 之间的差异最小:

$$\Theta^{*}=\operatorname{argmin} L(F_{\Theta}(\boldsymbol{X}),\boldsymbol{Y})。\tag{1}$$

式中:Θ^{*} 为模型的最佳参数;F_{Θ} 为神经网络模型;L 为评估差异的 MSE 损失函数。

2.2　网络结构

目前,在未来帧预测任务中领先的方法是 SimVP 架构,本文方法采用了类似的设计思想。如图 1 所示,模型主要由空间编码器、时空预测模块和空间解码器组成。空间编码器通过多层 2D 卷积来实现特征提取和下采样操作,该模块能够将输入的帧序列编码到低维潜在空间。时空预测模块主要由多个堆叠的门控时空网络(MST)构成,MST 通过对输入的低维特征信息进行时空动力学建模,以学习视频序列中的时间趋势和空间相关性。此外,MST 之间共享参数,这有效地减少了模型的参数量。空间解码器由 2D 卷积和上采样操作组成,通过将时空预测模块的输出作为解码器的输入,以实现低维信息向真实预测帧的转换,并且得到的预测序列可继续作为模型的输入进行后续的长期预测。

2.3　空间编码器

如图 1 所示,综合考虑模型的计算量和参数量,空间编码器采用了多层纯卷积结构,主要由 Conv2d、GroupNorm、SiLU 组成。由于需要充分捕捉视频帧的空间特征,并避免在下采样过程中造成过多的信息损耗,本文在编码器和解码器之间建立了残差连接,最大限度保留视频帧的背景语义 $\boldsymbol{B}_{\mathrm{bn}}$。空间编码器提取视频序列高级特征信息的过程可以表示为

$$\boldsymbol{Z}_{\mathrm{en}},\boldsymbol{B}_{\mathrm{bn}}=\sigma(\mathrm{Norm2d}(\mathrm{Conv2d}(\boldsymbol{X}_{\mathrm{n}})))。\tag{2}$$

式中:σ 为激活函数 SiLU;Norm2d 为组归一化层;$\boldsymbol{X}_{\mathrm{n}}$ 为输入序列;Conv2d 为 2D 卷积运算符;$\boldsymbol{Z}_{\mathrm{en}}$ 为获取的低维信息。通过将 2D 卷积的步长(*step*)设置为 2 实现下采样,而设置为 1 则进行卷积操作。

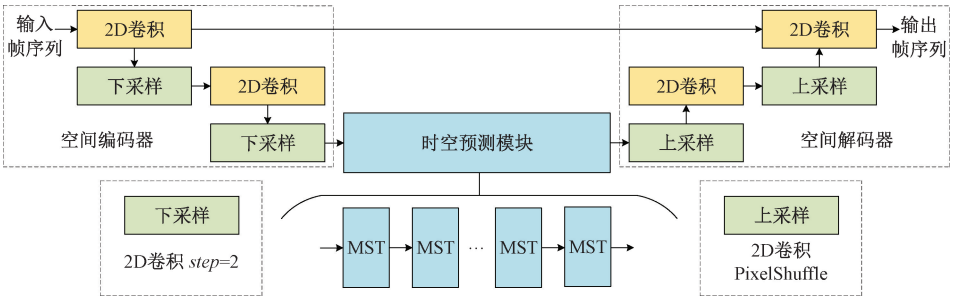

图 1 网络架构图

Figure 1 Network architecture diagram

2.4 时空预测模块

时空预测模块位于整个模型的中间部分。同空间编码器和空间解码器对单帧图像进行操作不同,预测模块处理沿时间维度堆叠形成的视频帧序列。由于视频帧预测是一种像素密集型任务,预测输出和输入的视频帧分辨率相同,因此,预测模块即要高效提取时空特征,又要尽可能避免预测过程中增大感受野导致的细节缺失。因此,本文提出了一种新的门控时空网络(MST),如图 2 所示。MST 是一种基于 Transformer 的变体,由归一化层(Batch Norm)、门控时空注意力层和全连接层组成。其中,门控时空注意力层主要包括空间注意力、时间注意力和门控融合机制 3 个部分,空间注意力能够学习帧内的多尺度特征信息,而时间注意力能够捕捉帧间的时间变化趋势。此外,门控融合机制能够有效地融合空间信息和时间特征,使模型能够采取相同的重视程度来学习序列中的空间相关性和时间趋势。门控时空注意力对视频序列中每个时空位置下的运动强度进行合理的权重分配,这有效平衡了时间特征及空间信息的捕捉能力,同时能够有效提高模型的时空预测建模能力。

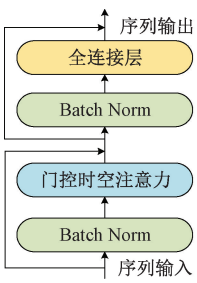

图 2 MST 网络结构图

Figure 2 MST network structure diagram

2.4.1 时空注意力

为了有效捕捉空间相关性和时间依赖关系,注意力机制需要分解为空间注意力和时间注意力,以充分学习帧内和帧间的相互作用。由于传统空间注意力的特征捕捉能力不足,并忽略了多尺度感受域的重要性,因此,本文采用多尺度深度条形卷积来构建空间注意力,同时使用大卷积核来增强模型的特征捕捉能力。如图 3 所示,空间注意力获取特征信息的过程主要包括 2 个阶段:首先建立基于大卷积核的多尺度深度条形卷积 $\mathrm{Cdw}_{1\times k}$ 和 $\mathrm{Cdw}_{k\times 1}$,以提取视频序列 $\boldsymbol{Z}_{\mathrm{i}}$ 中的多尺度特征信息;然后通过大小为 1×1 的卷积核 $\mathrm{Conv2d}_{1\times 1}$ 来聚合捕捉到的多尺度信息 $\boldsymbol{Z}_{\mathrm{m}}$。空间注意力捕捉多尺度特征信息的过程可以表示为

$$\boldsymbol{Z}_{\mathrm{m}} = \sum_{k\in\{7,11,21\}} \mathrm{Cdw}_{k\times 1}(\mathrm{Cdw}_{1\times k}(\boldsymbol{Z}_{\mathrm{i}}))\text{;} \quad (3)$$

$$\boldsymbol{Z}_{\mathrm{h}} = \mathrm{Conv2d}_{1\times 1}(\boldsymbol{Z}_{\mathrm{m}})\text{。} \quad (4)$$

式中:k 为卷积核大小,$k\in\{7,11,21\}$ 代表 k 分别取 7、11 和 21;$\boldsymbol{Z}_{\mathrm{h}}$ 为聚合后的多尺度信息。

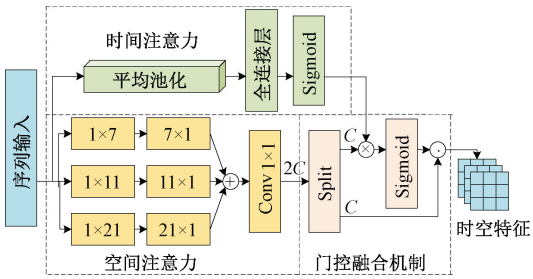

图 3 门控时空注意力网路结构

Figure 3 Structure of the gated spatio-temporal attention network

空间注意力能够有效捕捉帧内的空间相关性,但难以完整学习帧间的时间变化趋势。因此,本文采用通道注意力作为时间注意力,利用通道间的相互关系获取时间权重 $\boldsymbol{S}_{\mathrm{a}}$。该过程可以表示为

$$\boldsymbol{S}_{\mathrm{a}} = \mathrm{FC}(\mathrm{Avgpool}(\boldsymbol{Z}_{\mathrm{i}}))\text{。} \quad (5)$$

式中:$\boldsymbol{Z}_{\mathrm{i}}$ 为原始输入信息;Avgpool 为全局平均池化;FC 为全连接层。

2.4.2 门控融合机制

为了使模型对空间特征和时间特征采取相同的重视程度,本文提出了门控融合机制对空间注意力和时间注意力进行深度融合。如图 3 所示,门控融合过程可以分为 3 个阶段:首先,通过拆分操作 split

将通道数为 $2C$ 的多尺度空间信息 $\boldsymbol{Z}_{\mathrm{h}}$ 拆分为通道数为 C 的空间特征 $\boldsymbol{G}_{\mathrm{s}}$ 和 $\boldsymbol{Z}_{\mathrm{t}}$;其次,将空间信息 $\boldsymbol{Z}_{\mathrm{t}}$ 同时间权重 $\boldsymbol{S}_{\mathrm{a}}$ 相乘,并通过激活函数 Sigmoid 将其映射至[0,1]以获得时空权重;最后,将空间特征 $\boldsymbol{G}_{\mathrm{s}}$ 乘以时空权重以获得多尺度时空特征 $\boldsymbol{Z}''_{\mathrm{i}}$。整个注意力的融合过程可以表示为

$$\boldsymbol{G}_{\mathrm{s}},\boldsymbol{Z}_{\mathrm{t}}=\mathrm{split}(\boldsymbol{Z}_{\mathrm{h}}); \tag{6}$$

$$\boldsymbol{Z}''_{\mathrm{i}}=\sigma(\boldsymbol{S}_{\mathrm{a}}\otimes\boldsymbol{Z}_{\mathrm{t}})\odot(\boldsymbol{G}_{\mathrm{s}})。 \tag{7}$$

式中:σ 为激活函数 Sigmoid;⊙为哈达玛积(Hadamard product);⊗为克罗内克积(Kronecker)。

2.5　空间解码器

如图 1 所示,空间解码器由 Conv2d、GroupNorm、PixelShuffle 组成,通过将预测模块输出的预测信息输入到空间解码器中,能够将低维预测信息 $\boldsymbol{Z}_{\mathrm{c}}$ 解码为图像序列 $\boldsymbol{Y}'$,同时补充背景语义 $\boldsymbol{B}_{\mathrm{bn}}$。空间解码器输出预测图像序列的过程可以表示为

$$\boldsymbol{Y}'=\sigma(\mathrm{Norm2d}(\mathrm{Conv2d}(\boldsymbol{Z}_{\mathrm{c}},\boldsymbol{B}_{\mathrm{bn}})))。 \tag{8}$$

式中:σ 为激活函数 SiLU;Conv2d 为 2D 卷积,通过像素重组层(PixelShuffle)实现上采样操作,否则进行步长为 1 的卷积操作。

3　实验结果及分析

3.1　实验环境及模型参数

本文采用的软件运行平台为 Windows10 专业版 64 位,深度学习环境软件配置为 Python3.8 和 PyTorch1.10。硬件配置为 NVIDIA TITAN V 显卡,采用 CUDA10.2,使用 Adam 优化器、OneCycle[20] 及余弦退火学习率调整策略来训练模型。

该模型的超参数主要包括学习率、训练次数、*drop_path*、批处理大小、MST 单元数等。其中,在 Moving MNIST、TaxiBJ、WeatherBench 和 KITTI 数据集上,学习率分别设置为 0.001 0、0.000 5、0.005 0、0.005 0,训练次数分别为 600、50、50、100,而 *drop_path* 分别设置为 0、0.2、0.2、0.2,批处理大小统一设置为 16,MST 单元数分别设置为 8、8、8、6。

本文采用 *MSE* 损失函数来对模型进行训练,并通过均方误差(*MSE*)、平均绝对误差(*MAE*)、结构相似指数(*SSIM*)和均方根误差(*RMSE*)来评估预测图像的质量。

3.2　实验评估

本文在 Moving MNIST[3] 数据集上进行根据 10 个条件帧来预测 10 个未来帧的实验,并同先进的循环式模型和多进多出预测方法对比来评估模型的时空预测学习能力。如表 1 所示,尽管没有采用循环式设计,本文方法在 Moving MNIST 数据集上依然获得了较高的预测精度,同 SimVP 相比,*MSE* 和 *MAE* 分别降低了 14.7%、8.9%,同时参数量和计算量也有所下降。虽然推理效率有所降低,但时空特征学习能力更强,这显著地减少了模型的训练次数,同时训练时间缩短了近 61 h。同最先进的循环式模型 TaylorNet 相比,本文模型虽然计算量有所增加,但 *MSE* 和 *MAE* 也分别降低了 8.6%、3.7%,同时推理效率提高了 12%,并显著地缩短了训练时间。可以看出,本文方法有效解决了循环式架构预测精度低、推理效率低和训练时间长等问题。此外,同最先进的多进多出模型 SimVP+gSTA 相比,*MSE* 和 *MAE* 也下降了 9.0%、7.0%,在相同的训练次数下,本文方法获得了更高的预测精度和推理效率。

表 1　在 Moving MNIST 数据集上的实验结果

Table 1　Experimental results on the Moving MNIST dataset

方法	*MSE*	*MAE*	*SSIM*	参数量/M	计算量/GFlops	帧率/(帧·s^{-1})	训练次数	训练时间/h
ConvLSTM[4]	103.3	182.9	0.707	33.78	127.01	153	—	—
PredRNN[5]	56.8	126.1	0.867	23.83	116.00	124	—	—
PredRNN++[6]	46.5	106.8	0.898	38.58	171.73	95	—	—
MIM[13]	44.2	101.1	0.910	38.00	179.18	84	—	—
MAU[21]	27.6	80.3	0.937	4.50	17.82	168	—	—
PhyDNet[15]	24.4	70.3	0.947	3.10	15.33	181	2 000	≈242
SimVP[10]	23.8	68.9	0.948	57.90	19.43	333	2 000	≈101
SimVP+gSTA[19]	22.3	67.5	0.951	46.81	16.52	245	600	≈38
TaylorNet[16]	22.2	65.2	0.955	3.31	15.72	225	1 000	≈129
本文方法	20.3	62.8	0.955	46.93	16.53	252	600	≈40

图 4 所示为 Moving MNIST 数据集的预测结果,其中,误差特征图为真实帧和预测帧之间差值的绝对值。可以看出,随着预测长度的增加,在 $t=10$ 时,TaylorNet 由于受到误差累积的影响,产生了最密集的误差图。SimVP 虽然解决了误差累积的问题,但特征学习能力仍然不足,其误差主要集中在图像细节。而本文方法避免了误差累积的影响,同时具有高效的特征学习能力,获得了最佳的预测图像。

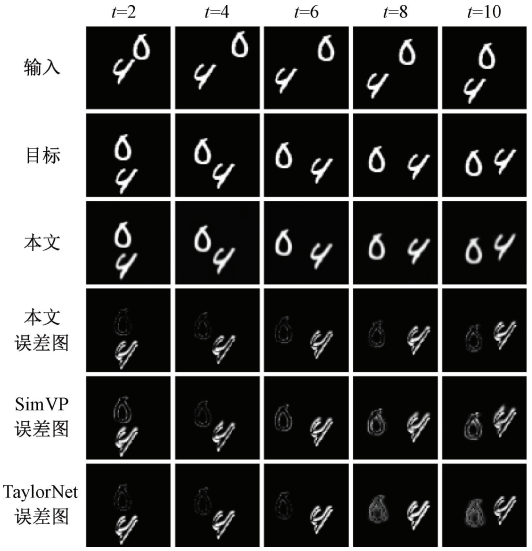

图 4 Moving MNIST 数据集预测结果

Figure 4 Moving MNIST dataset prediction results

本文在 TaxiBJ[22] 数据集上同经典的基线模型和最新的先进方法对比来评估模型的交通流预测性能,如表 2 所示。可以看出,本文方法获得了较高的预测精度,同最先进的循环式模型 PredRNN 相比,*MSE* 和 *MAE* 分别降低了 4.1%、2.6%,同时计算量减少了 39.8 GFlops。因此基于端对端的多进多出预测架构显著优于循环式单进单出预测架构,能够有效增强模型的预测性能,并减少计算量。而同最先进的多进多出模型 TAU 相比,*MSE* 也降低了 1.3%,并且计算量仅略微增加。此外,SimVP 是近期提出的一种简单的多进多出纯卷积网络,该模型构造简单,具有较高的综合性能,本文方法同 SimVP 相比,在 *MSE* 和 *MAE* 上也分别降低了 6.7%、3.2%,同时能够显著减少计算量。

表 2 在 TaxiBJ 数据集上的实验结果

Table 2 Experimental results on the TaxiBJ dataset

方法	*MSE*/10^{-2}	*MAE*	*SSIM*	计算量/GFlops
PhyDNet[15]	36.2	15.5	0.982	5.6
ConvLSTM[4]	33.5	15.3	0.983	20.7
PredRNN++[6]	33.4	15.3	0.983	63.0
SimVP[10]	32.8	15.4	0.983	3.6
MAU[21]	32.6	15.2	0.983	6.0
SimVP+gSTA[19]	32.4	15.0	0.984	2.6
PredRNN[5]	31.9	15.3	0.983	42.4
TAU[23]	31.0	14.9	0.984	2.5
本文方法	30.6	14.9	0.984	2.6

图 5 所示为 TaxiBJ 数据集的预测结果,可以看出,随着预测长度的增加,在 $t=4$ 时,循环式模型受到误差累积的影响,导致 MAU 的预测效果迅速下降,SimVP 虽获得了不错的预测效果,但对时间趋势的捕捉能力仍然不足。本文方法能够有效地平衡时间及空间特征的学习能力,取得了最佳的预测效果,具有很好的交通流预测性能。

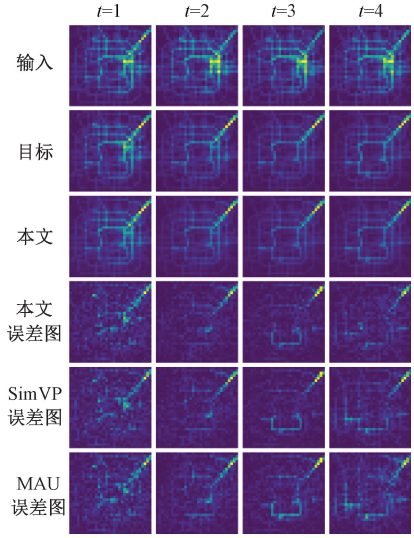

图 5 TaxiBJ 数据集预测结果

Figure 5 TaxiBJ dataset prediction results

气候预测是时空预测学习的另一项基本任务,本文在 WeatherBench[24] 数据集上同时空预测学习方法进行了对比试验。如表 3 所示,循环式时空预测学习方法虽取得了一定效果,但复杂的结构也导致计算量过大,而本文方法采用多进多出预测架构实现了更好的综合性能。其中,同最先进的循环式模型 MAU 相比,*MSE* 降低了 11%,并且计算量减小了 32.6 GFlops。而同最先进的多进多出模型 SimVP+gSTA 相比,在 *MAE* 上也降低了 0.9%。此外,同 SimVP 模型相比,*MSE* 和 *MAE* 分别降低了 10.5%、7.5%。

表 3 在 WeatherBench 数据集上的实验结果

Table 3 Experimental results on the WeatherBench dataset

方法	*MSE*	*MAE*	*RMSE*	计算量/GFlops
MIM[13]	1.784	0.871	1.336	109
ConvLSTM[4]	1.521	0.794	1.233	136
PredRNN++[6]	1.634	0.788	1.278	413
PredRNN[5]	1.331	0.724	1.154	278
MAU[21]	1.251	0.703	1.119	39.6
SimVP[10]	1.238	0.703	1.113	8.0
TAU[23]	1.162	0.670	1.078	6.7
SimVP+gSTA[19]	1.105	0.656	1.051	7.0
本文方法	1.108	0.650	1.055	7.0

图 6 所示为 WeatherBench 数据集预测结果。可以看出,随着预测长度的增加,在 $t=12$ 时,SimVP 模型难以完整地预测图像细节,MAU 由于预测机制的原因,在长期预测中精度会迅速下降。而本文方

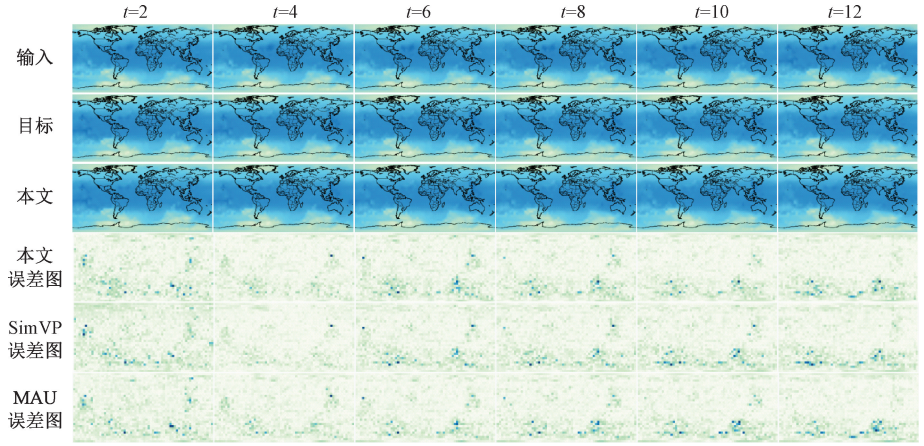

图 6　WeatherBench 数据集预测结果

Figure 6　WeatherBench dataset prediction results

法获得了最稀疏的误差图,高效的特征提取能力能够学习到更多的图像细节,并且不受误差累积的影响,在全球气候预测任务中表现出极佳的性能。

复杂的真实世界往往包含了不同运动对象的各种非线性时空运动,这导致时空预测学习更加具有挑战性。为了评估模型的泛化能力和适应性,本文在 KITTI[14] 数据集上进行训练,并在 CalTech Pedestrian 数据集[14] 上进行最终测试。其中,模型在 KITTI 和 Caltech Pedestrian 上采用了相同的参数设置,统一进行通过 10 个条件帧来预测 1 个未来帧的对比实验。

如表 4 所示,本文方法在真实数据集 KITTI 上获得了较高的预测精度,同基线模型 SimVP 相比,*MSE* 和 *MAE* 分别降低了 18.5%、12.3%。而同最先进的循环式模型 ConvLSTM 相比,本文方法在 *MSE* 和 *MAE* 上也分别降低了 6.4%、6.4%,同时计算量更小。此外,同最先进的多进多出模型 SimVP+gSTA 相比,虽然 *MSE* 略微有所上升,但 *MAE* 降低了 1.7%,并且计算量减少了 45.6 GFlops。可以看出,多进多出预测架构在预测精度上显著优于循环式预测架构,而本文方法通过较少的计算量达到了和 SimVP+gSTA 模型同样先进的预测性能,并且显著优于其他时空预测学习方法,具有很好的自动驾驶预测能力。

表 4　在 KITTI 数据集上的实验结果

Table 4　Experimental results on the KITTI dataset

方法	*MSE*	*MAE*	*SSIM*	计算量/GFlops
PhyDNet[15]	312.2	2 754.8	0.862	40.4
MAU[21]	177.8	1 800.4	0.918	172.0
SimVP[10]	160.2	1 690.8	0.934	60.6
PredNet[14]	159.8	1 568.9	0.929	42.8
ConvLSTM[4]	139.6	1 583.3	0.935	595.0
TAU[23]	131.1	1 507.8	0.946	92.5
SimVP+gSTA[19]	129.7	1 507.7	0.945	96.3
本文方法	130.6	1 482.3	0.945	50.7

3.3　消融扩展实验

为分析门控时空注意力每个局部模块对最终预测性能的影响,本文在 TaxiBJ 数据集上进行了消融实验。表 5 所示为消融实验结果,其中“No/MST”表示用 1×1 卷积替换门控时空注意力层,“No/Sat-3×3”和“No/Sat-7×7”分别是将空间注意力的多尺度深度卷积替换成 3×3 卷积和 7×7 卷积,“No/Tat”表示没有设置时间注意力,“No/Mk”表示不采用门控融合机制平衡注意力。而“MST-4”、“MST-6”和“MST-10”则表示 MST 的数量分别设置为 4、6 和 10。

表 5　在 TaxiBJ 数据集上的消融实验结果

Table 5　Ablation experimental results on the TaxiBJ dataset

方法	$MSE/10^{-2}$	*MAE*	参数量/M	计算量/GFlops
No/MST	34.53	15.50	7.91	2.1
No/Sat-3×3	31.76	15.09	9.89	2.6
No/Sat-7×7	30.94	14.98	9.97	2.6
No/Tat	31.17	15.17	9.97	2.6
No/Mk	31.11	15.05	9.56	2.5
MST-4	31.31	15.19	4.57	1.2
MST-6	30.75	15.08	7.30	1.9
MST-10	30.58	14.81	12.76	3.3
本文方法	30.60	14.91	10.03	2.6

如表 5 所示,采用门控时空注意力层使得 *MSE* 和 *MAE* 分别降低了 11.4% 和 3.8%。同 3×3 卷积和 7×7 卷积相比,使用多尺度深度条形卷积能够增强模型的感受野和捕捉多尺度特征的能力,使得 *MSE* 分别降低了 3.7%、1.1%。通过时间注意力学习帧间的相互作用,使 *MSE* 也降低了

1.8%。而门控机制深度融合了两种注意力,*MSE* 降低了 1.6%。可以看出,模型中的每个模块都能够有效提高最终的预测精度。此外,设置过多的 MST 单元带来的效果提升并不明显,同时导致了模型的参数量和计算量增大。因此,本文将 MST 数量设置为 8,并同上述 3 个模块进行集成获得了最佳的时空预测性能。

本文在 TaxiBJ 数据集上进行了卷积扩展实验如表 6 所示。其中,Dw 为本文采用的多尺度深度条形卷积,Dc 代表使用多尺度膨胀卷积,Mm 代表采用多尺度 2D 卷积,并在最终测试阶段通过重参数融合法[25]压缩模型,Mc 为使用多尺度 2D 卷积,其中 7×7 卷积被 3 个 3×3 卷积所代替。同 Dc 和 Mc 相比,Dw 在预测性能、参数量及推理效率方面具有显著优势,而 Mm 由于采用了重参数融合法,获得了最佳的推理效率,但本文方法获得了更高的预测精度,同时具有很好的推理效率。

表 6 卷积扩展实验对比结果

Table 6 Convolution extension experiment comparison results

方法	$MSE/10^{-2}$	*MAE*	参数量/M	计算量/GFlops	帧率/(帧·s^{-1})
Dw	30.60	14.91	10.03	2.61	1 020
Dc	30.92	15.03	10.07	2.62	949
Mm	31.30	14.83	10.72	2.79	1 131
Mc	31.07	14.89	11.01	2.86	705

为了探究不同预测架构对收敛性能的影响,本文在 Moving MNIST 数据集上进行了扩展实验。图 7 所示为不同模型收敛速度的对比结果。可以看出,同单进单出预测架构 PhyDNet 相比,多进多出预测策略在收敛性能方面具有显著优势。其中,本文方法实现了比 SimVP 更快的收敛速度,获得了较好的收敛效果。这表明,在每次训练中,模型能够捕捉到更多的时空动态趋势,这将会有效缩短模型的整体训练时间。

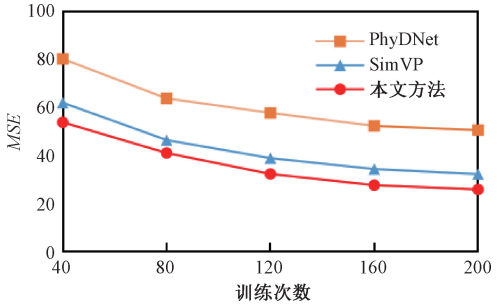

图 7 收敛性能实验结果

Figure 7 Convergence performance experimental results

4 应用前景展望

随着计算机视觉和深度学习技术的不断发展,视频预测技术将会具有更加广泛的应用前景。在交通领域中,视频预测技术可用于交通流监测、交通事故预测和城市规划,通过分析实时的视频流,交通系统可以更好地调度交通信号、减少拥堵,有效提高交通系统的效率。在气象领域中,视频预测技术可用于监测自然灾害,通过分析卫星和地面摄像头的视频数据,能够提前发现灾害迹象并发出预警提示,有效减少损失。视频预测技术的发展将会产生很多新的应用领域,在医疗领域中,视频预测技术将可以用于远程患者的监测、手术中的实时病情分析,医生可以利用视频预测技术来提高手术的准确性和安全性。视频预测技术将在多个领域引领创新和变革,将会有助于提高效率和安全性,并有潜力挖掘出更多的应用场景,为未来创造更多的可能性。

5 结论

本文提出了门控时空注意力来生成帧内和帧间相互关系的时空权重,以充分学习视频序列中空间维度和时间维度下有意义的时空信息,并采用门控融合机制平衡空间及时间注意力的特征捕捉能力,在 Moving MNIST、TaxiBJ、WeatherBench、KITTI 数据集上的实验结果均优于对比算法。此外,现有方法并未充分考虑帧内的多尺度信息交互作用对预测精度的影响,在今后的工作中,将研究如何更加高效地捕捉帧内及帧间的信息交互关系,同时保持模型结构简单、参数量低和推理效率高等优势。

参考文献:

[1] DAI K, LI X T, YE Y M, et al. MSTCGAN: multiscale time conditional generative adversarial network for long-term satellite image sequence prediction [J]. IEEE Transactions on Geoscience and Remote Sensing, 2022, 60: 1-16.

[2] TAN C, LI S Y, GAO Z Y, et al. OpenSTL: a comprehensive benchmark of spatio-temporal predictive learning [EB/OL]. (2023-06-20)[2023-07-20]. https://arxiv.org/abs/2306.11249.

[3] SRIVASTAVA N, MANSIMOV E, SALAKHUTDINOV R. Unsupervised learning of video representations using LSTMs[EB/OL]. (2016-01-04)[2023-07-20]. https://arxiv.org/abs/1502.04681.

[4] SHI X J, CHEN Z R, WANG H, et al. Convolutional LSTM network: a machine learning approach for precipi-

tation nowcasting[EB/OL]. (2015-09-19)[2023-07-20]. https://arxiv.org/abs/1506.04214.

[5] WANG Y B, LONG M S, WANG J M, et al. PredRNN: recurrent neural networks for predictive learning using spatiotemporal LSTMs[C]//NIPS'17: Proceedings of the 31st International Conference on Neural Information Processing Systems. Cham: Springer, 2017: 879-888.

[6] WANG Y B, GAO Z F, LONG M S, et al. PredRNN++: towards a resolution of the deep-in-time dilemma in spatiotemporal predictive learning[EB/OL]. (2018-11-19)[2023-07-20]. https://arxiv.org/abs/1804.06300.

[7] LIU Z W, YEH R A, TANG X O, et al. Video frame synthesis using deep voxel flow[C]//2017 IEEE International Conference on Computer Vision (ICCV). Piscataway: IEEE, 2017: 4473-4481.

[8] AIGNER S, KÖRNER M. FutureGAN: anticipating the future frames of video sequences using spatio-temporal 3d convolutions in progressively growing GANs[EB/OL]. (2018-11-26)[2023-07-20]. https://arxiv.org/abs/1810.01325.

[9] YE X, BILODEAU G A. VPTR: efficient transformers for video prediction[C]//2022 26th International Conference on Pattern Recognition (ICPR). Piscataway: IEEE, 2022: 3492-3499.

[10] GAO Z Y, TAN C, WU L R, et al. SimVP: simpler yet better video prediction[C]//2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE, 2022: 3160-3170.

[11] GUO M H, LU C Z, HOU Q B, et al. SegNeXt: rethinking convolutional attention design for semantic segmentation[EB/OL]. (2022-09-18)[2023-07-20]. https://arxiv.org/abs/2209.08575.

[12] HU J, SHEN L, SUN G. Squeeze-and-excitation networks[C]//2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2018: 7132-7141.

[13] WANG Y B, ZHANG J J, ZHU H Y, et al. Memory in memory: a predictive neural network for learning higher-order non-stationarity from spatiotemporal dynamics[C]// 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE, 2020: 9146-9154.

[14] LOTTER W, KREIMAN G, COX D. Deep predictive coding networks for video prediction and unsupervised learning[EB/OL]. (2017-05-01)[2023-07-20]. https://arxiv.org/abs/1605.08104.

[15] GUEN V L, THOME N. Disentangling physical dynamics from unknown factors for unsupervised video prediction [C]//2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE, 2020: 11471-11481.

[16] PAN T, JIANG Z Q, HAN J N, et al. Taylor saves for later: disentanglement for video prediction using Taylor representation[J]. Neurocomputing, 2022, 472: 166-174.

[17] SUN F, BAI C, SONG Y, et al. MMINR: multi-frame-to-multi-frame inference with noise resistance for precipitation nowcasting with radar[C]//2022 26th International Conference on Pattern Recognition (ICPR). Piscataway: IEEE, 2022: 97-103.

[18] NING S L, LAN M C, LI Y R, et al. MIMO is all you need: a strong multi-in-multi-out baseline for video prediction[EB/OL]. (2023-05-30)[2023-07-20]. https://arxiv.org/abs/2212.04655.

[19] TAN C, GAO Z Y, LI S Y, et al. SimVP: towards simple yet powerful spatiotemporal predictive learning[EB/OL]. (2023-04-26)[2023-07-20]. https://arxiv.org/abs/2211.12509.

[20] SMITH L N, TOPIN N. Super-convergence: very fast training of neural networks using large learning rates[EB/OL]. (2017-08-23)[2023-07-20]. https://arxiv.org/abs/1708.07120v1.

[21] CHANG Z, ZHANG X F, WANG S S, et al. MAU: a motion-aware unit for video prediction and beyond[C]// 35th Conference on Neural Information Processing Systems. Sydney: NeurIPS , 2021: 1-13.

[22] ZHANG J B, ZHENG Y, QI D K. Deep spatio-temporal residual networks for citywide crowd flows prediction[C]// Proceedings of the 31st AAAI Conference on Artificial Intelligence. New York: ACM, 2017: 1655-1661.

[23] TAN C, GAO Z Y, WU L R, et al. Temporal attention unit: towards efficient spatiotemporal predictive learning [C]//2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE, 2023: 18770-18782.

[24] RASP S, DUEBEN P D, SCHER S, et al. WeatherBench: a benchmark data set for data-driven weather forecasting[J]. Journal of Advances in Modeling Earth Systems, 2020, 12(11): 1-17.

[25] DING X H, ZHANG X Y, MA N N, et al. RepVGG: making VGG-style ConvNets great again[C]//2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE, 2021: 13728-13737.

(下转第 121 页)

Evaluation of Power Supply Reliability of Centralized Feeder Automation Distribution Network

CHEN Genyong[1], GAO Xiangyu[1], TAN Chao[2], FAN Xuguang[3]

(1. School of Electrical and Information Engineering, Zhengzhou University, Zhengzhou 450001, China; 2. Xinxiang Power Supply Company of State Grid Henan Electric Power Company, Xinxiang 453000, China; 3. Baofeng Power Supply Company of State Grid Henan Electric Power Company, Pingdingshan 467400, China)

Abstract: There were still some research gaps in reliability evaluation of distribution network considering centralized feeder automation, and most studies only focused on the impact of power failure. Considering the pre-arranged maintenance and capacity constraints, the influence of load transfer, and combined with the type and operation logic of feeder automation, according to the related technical indexes of feeder automation, the load nodes appearing in the process of power restoration were classified in detail. And the calculation formulas of expected power restoration time and power supply reliability indexes of different types of loads were derived. Combined with an example, the average outage duration *SAIDI* of feeder system was reduced by 0.95-1.08 h/(user·a) with different terminal configurations in the example, which showed that optimized the terminal configuration could effectively improve the power supply reliability of distribution network, which proved the accuracy and practicability of the evaluation method in this study. The influences of different terminal configurations on reliability were compared.

Keywords: centralized feeder automation; distribution network; power supply reliability; pre-arranged maintenance; load transfer

---

(上接第 77 页)

Video Frame Prediction Model Based on Gated Spatio-Temporal Attention

LI Weijun[1,2], ZHANG Xinyong[1], GAO Yuxiao[1], GU Jianlai[1], LIU Jintong[1]

(1. School of Computer Science and Engineering, North Minzu University, Yinchuan 750021, China; 2. The Key Laboratory of Images and Graphics Intelligent Processing of State Ethnic Affairs Commission, North Minzu University, Yinchuan 750021, China)

Abstract: A video frame prediction model based on gated spatio-temporal attention was proposed to address the issues of low accuracy, slow training, complex structure, and error accumulation in recurrent video frame prediction architectures. Firstly, high-level semantic information of the video frame sequence was extracted by a spatial encoder while preserving background features. Secondly, a gated spatio-temporal attention mechanism was established, utilizing multi-scale deep bar convolutions and channel attention to learn both intra-frame and inter-frame spatio-temporal features. A gate fusion mechanism was employed to balance the feature learning capability of spatio-temporal attention. Finally, a spatial decoder reconstructed the high-level features into predicted realistic images and complements background semantics to enhance the details. Experimental results on the Moving MNIST, TaxiBJ, WeatherBench, and KITTI datasets showed that compared to the multi-input multi-output model SimVP, the mean squared error (*MSE*) was reduced by 14.7%, 6.7%, 10.5%, and 18.5%, respectively. In ablation and expansion experiments, the proposed model achieved good overall performance, demonstrating advantages such as high prediction accuracy, low computational complexity, and efficient inference.

Keywords: video frame prediction; convolutional neural network; attention mechanism; gated convolution; codec network