基于空间自适应卷积 LSTM 的视频预测

吴哲夫 张令威 刘光宇 刘光灿*
(南京信息工程大学江苏省大数据分析技术重点实验室 江苏 南京 210044)

摘 要 在视频预测领域,传统的 CNN 与 LSTM 都不能充分表征视频中的时空特征。针对这一问题提出空间自适应卷积 LSTM 算法。受空间变换网络启发,在卷积 LSTM 内部的“input-to-state”计算过程中将传统卷积操作改为空间自适应卷积:利用额外卷积层获得自适应卷积所需的位置参数,令自适应卷积根据时空信息选择卷积位置,提升模型捕捉时空变换特征的性能;并针对雷达回波预测提出多分支编码预测的网络架构,根据降水类别训练 4 个不同的支路,以提升网络的预测性能。在合成数据集与真实数据集上的实验结果表明,该模型取得了有竞争力的结果,单独设计一个模块让网络显式地学习某种特征会使网络有更好的性能。

关键词 卷积 LSTM 空间变换网络 视频预测

中图分类号 TP183 **文献标志码** A DOI:10.3969/j.issn.1000-386x.2020.09.011

VIDEO PREDICTION BASED ON SPATIAL ADAPTIVE CONVLSTM

Wu Zhefu Zhang Lingwei Liu Guangyu Liu Guangcan*
(*Jiangsu Key Laboratory of Big Data Analysis Technology*, *Nanjing University of Information Science and Technology*, *Nanjing 210044*, *Jiangsu*, *China*)

Abstract In the field of video prediction, neither CNN nor LSTM can fully represent the spatiotemporal features in video. To solve this problem, we propose a spatial adaptive convolution LSTM algorithm. Inspired by the spatial transformation network, the traditional convolution operation is changed to spatial adaptive convolution in the “input-to-state” calculation process inside the convolutional LSTM. The additional convolution layer was used to obtain the position parameters needed for the adaptive convolution, so that the adaptive convolution could select the convolution position according to the spatiotemporal information, and improve the performance of the model to capture the spatiotemporal transformation features. A multi branch coding prediction network architecture was proposed for radar echo prediction, and four different branches were trained according to the precipitation category to improve the prediction performance of the network. The experimental results on the synthetic dataset and the real dataset show that the model has achieved competitive results. And designing a module to let the network learn certain features explicitly make the network have better performance.

Keywords ConvLSTM Spatial transformation network Video prediction

0 引 言

近年来,随着算法研究的深入和硬件的飞速发展,深度学习[1]在计算机视觉、自然语言处理、模式识别等诸多领域的应用愈加广泛。随着社会数据的体量不断增大,我们能够利用海量的历史信息进行预测。视频预测因其先天的数据量优势和无须人工标注的特点,逐渐成为深度学习的一个火热领域。

视频预测,即给定初始的若干帧图像信息,要求深度网络模型可以预测并输出后若干帧的图像信息。该技术多用于行为预测、气象预测、自动驾驶等领域。预

收稿日期:2019-07-12。国家自然科学基金优秀青年基金项目(61622305);国家自然科学基金青年科学基金项目(61502238);江苏省自然科学基金杰出青年基金项目(BK20160040)。**吴哲夫**,硕士生,主研领域:计算机视觉。**张令威**,硕士生。**刘光宇**,硕士生。**刘光灿**,教授。

测任务的关键在于同时捕捉给定视频的内容和动态。将卷积神经网络(CNN)[2]与循环神经网络(RNN)[3]结合,是近年来视频预测的主流方法。Lotter 等[4]提出了 PredNet,将图像预测误差在网络中前向传递,虽然学习视频表征能力较强,但测试时存在误差,因此只能实现单帧预测,预测时间短且不清晰;Kim 等[5]将 CNN 嵌入 RNN 模块中,提出了卷积 LSTM,提高了预测时间长度,但无法保持细节;Villegas 等[6]利用卷积 LSTM 的优势提出了 MCNet,将视频预测任务分为预测内容和预测动态信息两个子任务,将子任务的输出整合后进行编码最后输出预测视频,其细节效果比卷积 LSTM 略优,但依然没有解决复杂时空变化的预测问题。本文旨在解决视频预测中的复杂变化的预测问题。

本文基于卷积 LSTM,在经典的卷积操作之前加入空间变换网络[7]获得位置参数,用位置参数指导卷积位置,提高模型的精准度;提出多分支预测以解决气象雷达图预测的强降水预测问题。实验表明,本文模型能够更加高效地预测复杂动态,并提高针对强降水的预测性能。

1　卷积 LSTM

LSTM 的内部的计算为矩阵乘算,多用于处理时序数据如语音、语句,若直接将其用于图像处理,其覆盖整幅图片的全连接操作计算代价过高,且全连接操作忽视了图像的空间信息,因此无法保留空间特征。文献[8]提出了卷积 LSTM, 将 CNN 与长短时记忆网络(LSTM)结合,使模型不仅具有时序建模能力,而且能刻画局部空间特征。

ConvLSTM 的结构与 FC-LSTM[9]相同,利用三个门限层来控制记忆的存取,内部结构如图 1 所示,公式如下:

$$\boldsymbol{i}_t=\sigma(W_{xi}*x_t+W_{hi}*\boldsymbol{H}_{t-1}+W_{ci}\circ\boldsymbol{C}_{t-1}+b_i) \quad (1)$$

$$\boldsymbol{f}_t=\sigma(W_{xf}*x_t+W_{hf}*\boldsymbol{H}_{t-1}+W_{cf}\circ\boldsymbol{C}_{t-1}+b_f) \quad (2)$$

$$\boldsymbol{C}_t=\boldsymbol{f}_t\circ\boldsymbol{C}_{t-1}+\boldsymbol{i}_t\circ\tanh(W_{xc}*x_t+W_{hc}*\boldsymbol{H}_{t-1}+b_c) \quad (3)$$

$$\boldsymbol{o}_t=\sigma(W_{xo}*x_t+W_{ho}*\boldsymbol{H}_{t-1}+W_{co}\circ\boldsymbol{C}_t+b_o) \quad (4)$$

$$\boldsymbol{H}_t=\boldsymbol{o}_t\circ\tanh(\boldsymbol{C}_t) \quad (5)$$

式中:σ 代表激活函数;W 代表各个门限层的权重;x_t 代表当前时间步 t 的输入图像;b_i 代表输入门对应的偏置;“$*$”表示卷积操作;“$\circ$”表示 Hadamard 乘积;$\boldsymbol{X}$、$\boldsymbol{C}$、$\boldsymbol{H}$、$\boldsymbol{i}$ $\boldsymbol{f}$、$\boldsymbol{o}$ 均为三维的张量,分别对应于图像的通道、空间的行、列信息。ConvLSTM 将传统的 FC-LSTM 中“input-to-state”和“state-to-state”的前馈神经网络操作替换成卷积操作,不仅可以使网络接收图像输入,而且能够捕捉空间局部特征,更好地针对图像进行时序预测。

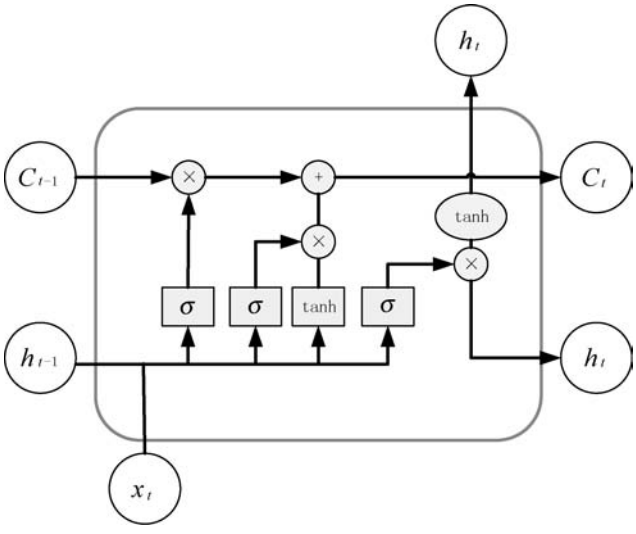

图 1　卷积 LSTM 结构

ConvLSTM 有一些变体,如 ConvGRU[10]等,多数变体通过改变门限层结构,使网络适应不同任务需要。

2　模型设计

在 LSTM 中加入卷积操作使其可以接受图像级的输入,但并没有触碰到视频预测的痛点,单纯地利用卷积操作并不能充分表征图像序列的空间变换信息。卷积网络对旋转、缩放等变化的表征能力不强,不能满足视频预测的性能需求,普通的 CNN 由于池化层的加入使之具有一定的平移不变性,并通过数据增强使网络能够隐式地获得一定的旋转、缩放不变性。但文献[11]提出,与其让网络隐式地学习到某种能力,不如为网络设计一个显式的处理模块,专门处理以上的各种变换。基于以上思想,本文提出空间自适应卷积 LSTM 网络模型。

2.1　网络结构

本文网络结构(图 2)与经典视频预测网络结构相似,即编码器-预测器的结构,网络堆叠了三层隐藏层,即空间自适应卷积 LSTM 层,隐藏层之间插入降采样层或上采样层。本文中的采样层为一次卷积操作,使网络有针对性地分别对低级局部细节动态和高级全局动态信息进行表征。网络输出端置于网络底层,因此高级时空特征能够由上至下指导低级局部时空特征的校准与更新,并利用低层的状态信息提升对细节的预测性能。

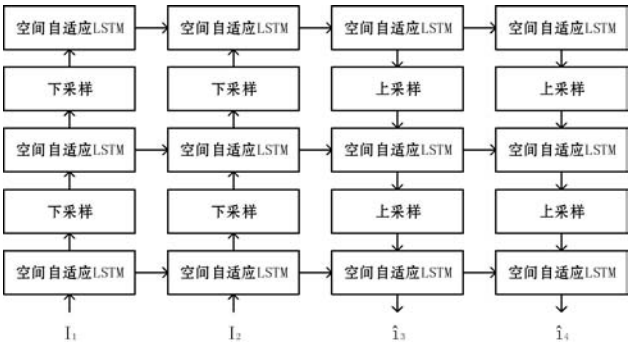

图 2 自编码预测结构

此外,可以根据预测对象优化网络结构:在本文对气象雷达回波图进行预测时,会有针对性地训练 4 个模型,4 个模型的结构完全相同,根据各个数据的降水类型决定每个数据的输入分支。

2.2 空间自适应卷积 LSTM

2.2.1 卷积操作的等价形式

在卷积 LSTM 中,卷积操作的对象是当前时间步的输入和上一时间步的状态变量,通过多层卷积操作提取输入和状态的空间特征,以决定在每个空间位置上的状态变量和输入信息的取舍。

卷积计算是将输入图片的目标位置及其周围若干固定位置的像素信息映射到输出图像的对应位置。以 3×3 卷积操作为例,其实质为输入到输出的映射,输出的每个位置的像素值都与输入的对应位置周围的 9 个点有关,分别找到所有目标位置对应输入的位置后,再对同一位置的不同通道给予不同的权重后求和,最后将不同位置的加权结果求和,得到输出(如图 3 所示),计算过程如下:

$$y_{i,j} = \sum_{l}^{L} W_l \cdot x_{p_{i,i,j},q_{l,i,j}} \tag{6}$$

式中:L 代表输出的每一点与输入相关的连接数,对应于传统卷积操作的卷积核尺寸,3×3 卷积操作中 $L=9$;$p_{l,i,j}$和 $q_{l,i,j}$表示与输出位置为(i,j)的第 l 个连接的位置参数,本例中 $p_{1,i,j}=i-1$,$q_{1,i,j}=j-1$,$p_{2,i,j}=i$,$q_{2,i,j}=j-1$,$\cdots$,$p_{9,i,j}=i+1$,$q_{9,i,j}=j+1$。

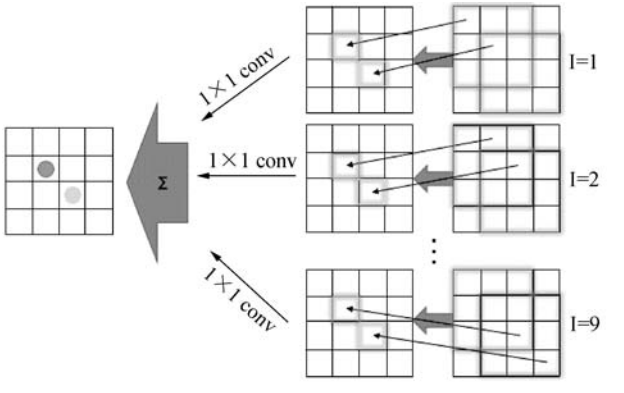

图 3 普通 3×3 卷积

在面对复杂的时空变化时,当前时间步的某类信息所在的位置不一定与上一时间步状态变量的对应类信息位置相同,用尺寸固定、参数固定的卷积核进行卷积操作难以进行精确的空间信息的取舍。基于这一情况,本文提出不固定卷积核尺寸。改变“input-to-state”的卷积方式,令卷积操作中的每一个卷积空间位置都能够随时间自适应改变(见图 4),以提高模型对时空相关性的捕捉能力。

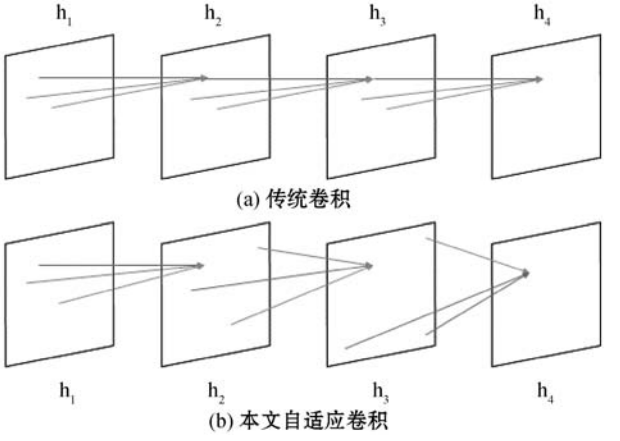

图 4 两种卷积操作

2.2.2 引入位置参数

本文受式(6)和空间变换网络启发,引入空间自适应卷积操作。首先确定卷积连接数 L,其含义与式(6)中的 L 相同,用位置参数 U_t和 V_t表示输入中所有与输出相关的位置,根据位置参数寻找对应位置的输入。随后将输出图像中的每个位置都与输入图像中若干个位置对应起来,用新的卷积公式实现自适应卷积,具体公式如下:

$$\boldsymbol{f}_t = \sigma\left(\sum_{l=1}^{L} W_{fx}^{l} * \mathrm{trans}(x_t, \boldsymbol{U}_{t,l}, \boldsymbol{V}_{t,l}) + W_{fh} * h_{t-1} + b_f\right) \tag{7}$$

$$\boldsymbol{i}_t = \sigma\left(\sum_{l=1}^{L} W_{ix}^{l} * \mathrm{trans}(x_t, \boldsymbol{U}_{t,l}, \boldsymbol{V}_{t,l}) + W_{ih} * h_{t-1} + b_i\right) \tag{8}$$

$$\tilde{\boldsymbol{C}}_t = \tanh\left(\sum_{l=1}^{L} W_{cx}^{l} * \mathrm{trans}(x_t, \boldsymbol{U}_{t,l}, \boldsymbol{V}_{t,l}) + W_{ch} * h_{t-1} + b_c\right) \tag{9}$$

$$\boldsymbol{C}_t = \boldsymbol{f}_t \circ \boldsymbol{C}_{t-1} + \boldsymbol{i}_t \circ \tilde{\boldsymbol{C}}_t \tag{10}$$

$$\boldsymbol{o}_t = \sigma\left(\sum_{l=1}^{L} W_{ox}^{l} * \mathrm{trans}(x_t, \boldsymbol{V}_{t,l}, \boldsymbol{U}_{t,l}) + W_{oh} * h_{t-1} + b_o\right) \tag{11}$$

$$\mathrm{h}_t = \boldsymbol{o}_t \circ \tanh(\boldsymbol{C}_t) \tag{12}$$

式中:$\boldsymbol{U}_{t,l}$和 $\boldsymbol{V}_{t,l}$分别表示第 l 个连接位置的横、纵坐标;W_{fh}、W_{ih}、W_{ch}、W_{oh}为各个门限层的权重,通过训练学习获得权重参数,其尺寸为 $C\times1\times1$,C 为输入图像的

通道数,文中每个门限层的权重均有 L 个,故参数量为 $C\times L$(对应于传统卷积核的尺寸 $C\times W\times H$)。

本文的位置参数不能直接确定,需要通过深度网络训练获得。位置参数(i,j)本身是离散的,无法通过反向传播求导以学习位置参数。为使位置参数可导,本文引入双线性插值法[12]。令输出特征图某一位置(i,j)对应到输入特征图的卷积位置为(u,v),若 u、v 为小数,则利用双线性插值法求得小数位置的像素值,再将该像素值作为自适应卷积的输入。像素值的计算方法以 warp 函数表示,若 $Y=\text{warp}(X,U,V)$,则有:

$$Y_{c,i,j}=\sum_{h=1}^{H}\sum_{w=1}^{W}X_{c,m,n}\max(0,1-|i+V_{i,j}-h|)\max(o,1-|j+U_{i,j}-w|) \tag{13}$$

2.2.3 位置参数的学习

为学习位置参数,本文为网络设计一个显式的处理模块,将当前时间步的输入和上一时间步的隐状态通道联结后对其进行卷积操作,其结果的尺寸为$(2\times L)\times w\times h$,公式如下:

$$\boldsymbol{U}_t,\boldsymbol{V}_t=\gamma(x_t,h_{t-1}) \tag{14}$$

式中:x_t表示当前时间步的输入,h_{t-1}表示上一时间步的隐状态,两者空间尺寸均为 $w\times h$;将 x_t与 h_{t-1}通道级联后做一次普通卷积操作,以 γ 表示,该卷积的输出尺寸为结果为$(2\times L)\times w\times h$,将其沿通道维度拆分,获得 2 个尺寸为 $L\times w\times h$ 的张量,用 $\boldsymbol{U}_t$和 $\boldsymbol{V}_t$表示,其空间尺寸为 $w\times h$,通道数为 L。

空间自适应卷积的输出特征图内位置(i,j)的结果来源于输入特征图中位置$(\boldsymbol{V}_{i,j},\boldsymbol{U}_{i,j})$的权重求和,若相关连接数为 L 个,则第 l 个相关连接的位置为$(\boldsymbol{V}_{l,i,j},\boldsymbol{U}_{l,i,j})$。

传统的卷积 LSTM 中,直接将当前时间步的图片或上层卷积 RNN 的输出作为当前时间步的输入。而本文空间自适应卷积 LSTM 结构在输入图像之前,通过 γ 卷积操作获得自适应卷积层的输出与输入之间的拓扑链接(即位置参数),利用拓扑链接对当前 LSTM 的输入作空间变换,使其与隐状态中的信息对齐,以此实现精准的记忆保存和图像序列预测。

2.3 损失函数

在进行普通视频预测时,我们的损失函数采用 L2 损失函数:

$$L=\frac{1}{N}\times\sum_{n}^{N}\sum_{i}^{W}\sum_{j}^{H}(y_{n,i,j}-\hat{y}_{n,i,j})^2 \tag{15}$$

式中:y 和 $\hat{y}$ 分别表示实际视频帧和预测视频帧;N 表示视频帧数;W 和 H 分别为视频的宽和高。损失函数就是遍历所有像素点求预测图像序列与 GroundTruth 的差平方的平均值(MSE)。

预测雷达降水回波图时,为更精准预测高降水区,本文为不同降水等级设定不同权重,根据权重比决定不同降水程度的误差对损失的影响。本文将像素值在[0,30)区间的权重设定为 0.3,像素值在[30,50)的权重设定为 0.3,像素值在[50,80]的权重设定为 0.4。在计算损失时,先根据 GroundTruth 判断当前像素位置的损失权重,最后按照权重比计算损失:

$$\omega_{n,i,j}=\begin{cases}0.3 & y_{n,i,j}\in[0,30)\\ 0.3 & y_{n,i,j}\in[30,50)\\ 0.4 & y_{n,i,j}\in[50,80]\end{cases} \tag{16}$$

$$L=\frac{1}{N}\times\sum_{n}^{N}\sum_{i}^{W}\sum_{j}^{H}\omega_{n,i,j}(y_{n,i,j}-\hat{y}_{n,i,j})^2 \tag{17}$$

3 实　验

3.1 手写体视频

3.1.1 数据集

本文手写体视频实验数据来源于 MNIST 手写体数据集[13]。MNIST 手写数据集有 60 000 幅图片,取其中 50 000 幅作为训练素材,另 10 000 幅图片作为测试素材。训练集为 50 000 幅训练集素材生成的 80 000 个长度为 20 帧的图像序列;测试集为测试集素材生成的 20 000 个长度为 20 帧的图像序列。由素材生成数据集的方式为:从 0 ~ 9 中随机选取 3 个数字,再从 MNIST 素材中随机选取对应的 3 幅数字图片,设定好随机旋转角度范围、平移速度、缩放尺寸倍率等超参数,根据超参数结合帧生成算法生成 20 帧的手写体视频,其中前 10 帧作为输入,后 10 帧作为 GroundTruth。

3.1.2 模型参数

本文的硬件配置为 4 块 TESLA K80 GPU,优化器采用 Adam Optimizer,学习率为 10^{-4},动量(momentum)设置为 0.5,默认 batch size 设置为 8。激活函数采用 LeakyReLU。本实验有三层自适应卷积 LSTM,由下至上每层的门限层卷积核个数各有 64 个、192 个、192 个。

3.1.3 实验结果及分析

量化评估如表 1 和表 2 所示,训练迭代 4 个 epoch,即训练 32 万次图像序列。误差计算方式为预测的 10 个图像序列分别与测试集的后 10 幅正确图片的 MSE。

表 1　预测结果的误差对比

预测方案	预测对象	
	MNIST	雷达回波图
PredNet	0.035 84	0.091 06
卷积 LSTM	0.002 63	0.006 88
自适应($L=9$)	0.001 93	0.007 46
自适应($L=17$)	0.001 65	0.005 88

表 2 MNIST 视频预测逐帧的结构相似性评估

预测方案	帧次									
	11	12	13	14	15	16	17	18	19	20
PredNet	0.978 1	0.821 0	0.631 1	0.413 1	0.396 4	0.341 3	0.267 9	0.246 1	0.294 1	0.206 0
卷积 LSTM	0.935 8	0.878 2	0.808 4	0.753 9	0.725 6	0.714 9	0.692 5	0.655 9	0.610 0	0.551 0
自适应(L=9)	0.966 8	0.934 1	0.898 7	0.855 5	0.823 3	0.820 1	0.810 3	0.785 0	0.755 1	0.707 8
自适应(L=17)	0.977 4	0.948 6	0.928 4	0.897 8	0.862 2	0.853 6	0.835 1	0.796 4	0.763 0	0.723 1

为分析预测序列的差异,本文额外计算结构相似性(SSIM),由表 2 可知,在用 PredNet 进行多帧预测时,由于缺少 GroundTruth 来计算误差,因此无法在预测时进行误差前向传播,具体表现为从预测的第二帧开始迅速模糊,最终的多帧预测效果很差。本文提出的空间自适应卷积 LSTM 相较于传统的卷积 LSTM 和 PredNet,预测结构相似度以及长期预测的清晰度都有可观的提升。

MNIST 实验效果如图 5 – 图 8 所示,由于版面限制,使用泛用性最好的卷积 LSTM 对比。从上至下分别为 Ground Truth、经典的卷积 LSTM 预测序列、空间自适应卷积 LSTM(L = 9)预测序列,以及空间自适应卷积 LSTM(L = 17)预测序列(L 代表自适应卷积的相关连接数),由左到右为从预测的序列中抽出的第 2 帧、第 5 帧、第 10 帧的实验结果。可以看出,经典的卷积 LSTM 处理较明显的旋转、缩放等复杂变换时,图像开始变得模糊,而 9 链接和 17 链接的自适应卷积 LSTM,都能很好地预测到旋转缩放等复杂变换,且 17 链接能够相对更好地保持清晰度,同时对动态变化预测得更加精准。推断其原因是 17 链接的输出到输入的映射连接数更多,因此能够在不过拟合的情况下用更多的参数来更加精细地表征视频序列中的时空变化。实验过程中,17 链接的迭代速度也略慢于 9 链接的速度,相对于性能的提升,这种计算代价是可以接受的。

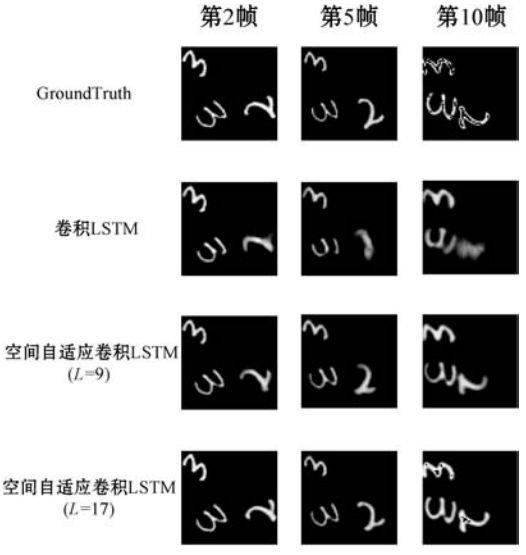

图 5 MNIST 实验结果对比 1

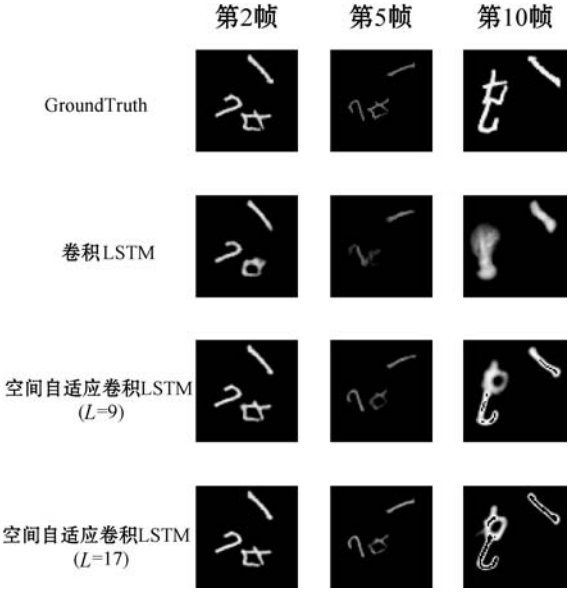

图 6 MNIST 实验结果对比 2

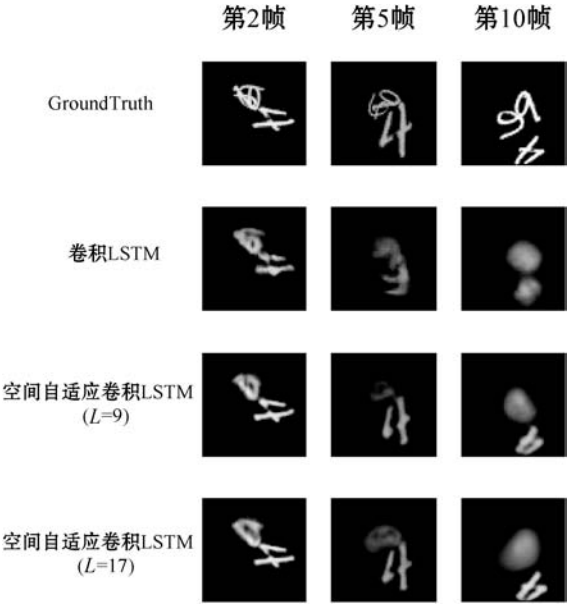

图 7 MNIST 实验结果对比 3

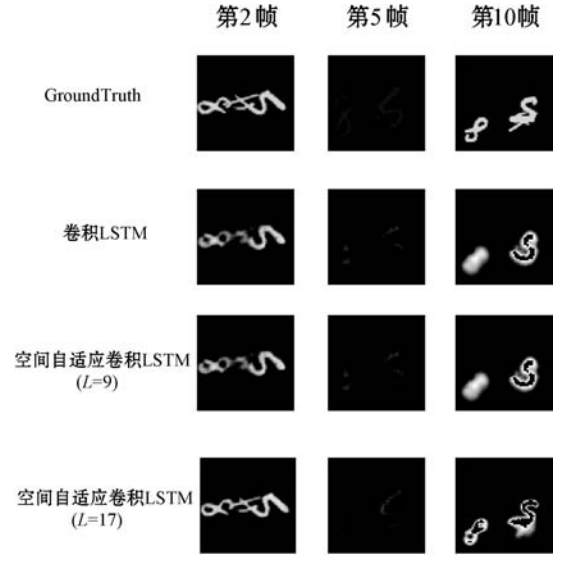

图 8 MNIST 实验结果对比 4

3.2　雷达降水回波图

3.2.1　数据集

为获取预测模块的实际应用中的泛用性,本文额外使用雷达回波图来进行气象预测。气象雷达图数据集来自四川自贡气象局,共 6 万组回波数据。每组回波记录有 61 幅图片,雷达回波图为 501 ×501 的单通道灰度图像,初始缺省值均为 255,为方便观测预测效果,输入网络前将所有雷达回波图的缺省值更改为 0。本文取前 31 幅图片作为输入,后 30 幅作为 Ground-Truth。在针对降水雷达回波图的预测时,由于硬件性能限制,故先将 501 ×501 的灰度图像降采样为 64 ×64 的单通道灰度图像,然后针对 64 ×64 的图像序列进行预测。

3.2.2　模型框架

针对雷达回波图进行训练时,本文将所有雷达回波数据分为 4 部分:当某个雷达回波图片段中第 11 帧和第 31 帧中白点个数均大于 5 万个时,将此片段分至Ⅰ类;第 11 帧少于 5 万,第 31 帧大于 5 万,将此片段分至Ⅱ类;第 11 帧多于 5 万,第 31 帧小于 5 万,分至Ⅲ类,第 11 帧与第 31 帧均少于 5 万个白点,分至Ⅳ类。在训练时,向网络中输送数据之前首先对图片序列进行分类,再根据类别送入 4 个不同模型中的其中一个模型,不同模型处理不同的气象变化趋势,以此提高模型对不同气象类别的精准预测能力。多分支预测结构如图 9 所示。

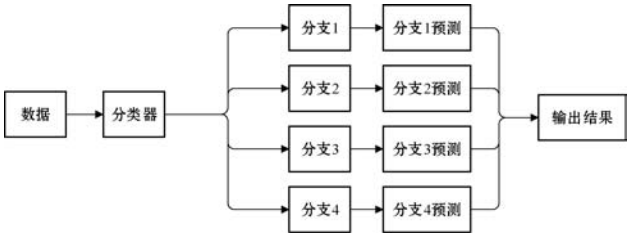

图 9　多分支预测结构

3.2.3　实验结果及分析

雷达回波实验效果如图 10 和图 11 所示,实验迭代次数为 8 万次,在预测的 30 帧视频中,每 6 帧中抽出 1 帧作为实验结果对比,共抽出 5 帧。可以看到,即使是将回波图压缩至 64 ×64 大小,预测结果依然有所区别,传统的卷积 LSTM 在预测后期图像时部分细节会丢失。与之相比,本文的自适应卷积 LSTM 和多分支网络结合的方法能够在一定程度上改善预测结果,尤其在降水量较高地区(图 10、图 11 中的偏白地区),采用空间自适应卷积模型能够更好地预测对应范围内的时空变化,推测其原因是用了带权重判定的损失函数。其他灰色区域的预测也能够更好地拟合 GroundTruth 的轮廓。

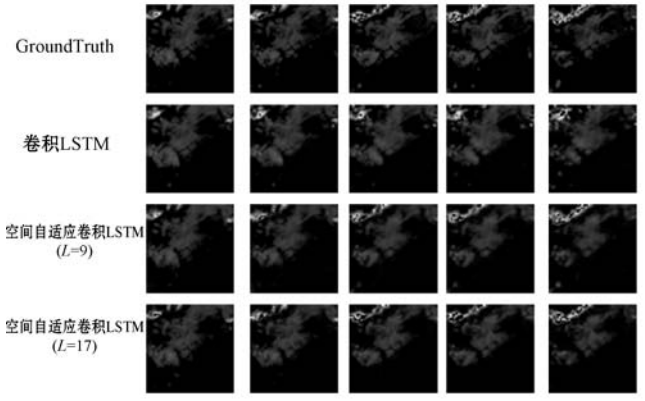

图 10　雷达回波实验结果 1

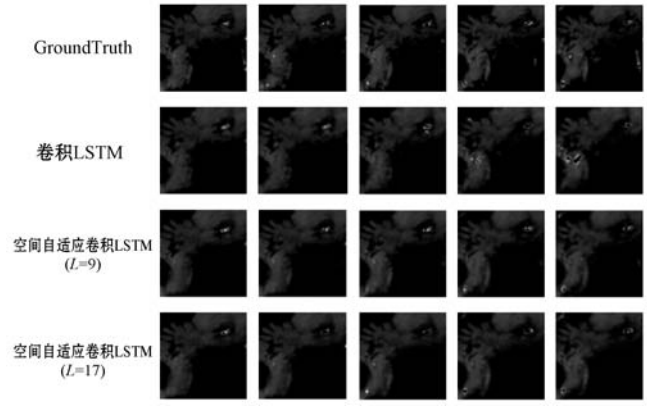

图 11　雷达回波实验结果 2

预测降水雷达回波图主要预测目标是强降水区域,为清楚地表示实验结果,本文对有代表性的强降水序列进行 gray2RGB 处理,如图 12 所示。可以看出,本文的自适应 LSTM 与多分支编码预测网络架构在处理强降水序列时有更加精准的结果。

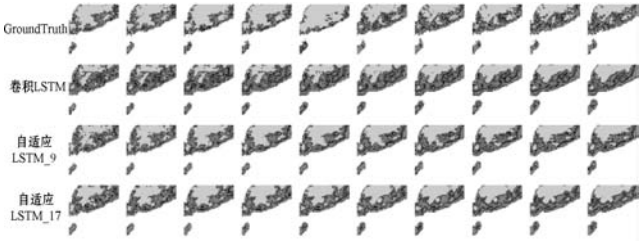

图 12　雷达回波实验结果 RGB 化

4　结　语

本文对基于深度学习的视频预测进行研究,基于传统卷积 LSTM 改变其“input-to-state”的计算过程,在其中添加空间转换层以显式学习时空变化特征。通过手写体视频片段的预测结果评测模型性能。实验证明,在某些情况下,单独设计一个模块让网络显式地学习某种特征会使网络有更好的泛化性能。本文的空间自适应卷积 LSTM 相较于传统的卷积 LSTM 确实有可

(下转第 110 页)

参 考 文 献

[1] Wang J, Song Q, Jiang Z, et al. A novel InSAR based off-road positive and negative obstacle detection technique for unmanned ground vehicle[C]//2016 IEEE International Geoscience and Remote Sensing Symposium(IGARSS). IEEE, 2016:1174 - 1177.

[2] 刘家银,唐振民,王安东,等. 基于多激光雷达与组合特征的非结构化环境负障碍物检测[J]. 机器人,2017,39(5):638 - 651.

[3] Larson J, Trivedi M. Lidar based off-road negative obstacle detection and analysis[C]//2011 14th International IEEE Conference on Intelligent Transportation Systems (ITSC). IEEE,2011:192 - 197.

[4] Karunasekera H, Zhang H, Xi T, et al. Stereo vision based negative obstacle detection[C]//2017 13th IEEE International Conference on Control and Automation(ICCA). IEEE, 2017: 834 - 838.

[5] 吴一全,孟天亮,吴诗婳. 图像阈值分割方法研究进展 20 年(1994—2014)[J]. 数据采集与处理,2015,30(1):1 - 23.

[6] ElBayoumi HarbSM, Isa N A M,Salamah S A. Improved image magnification algorithm based on Otsu thresholding[J]. Computers and Electrical Engineering,2015,46:338 - 355.

[7] Xu H, Wang Y, Wu Y, et al. Infrared and multi-type images fusion algorithm based on contrast pyramid transform [J]. Infrared Physics and Technology,2016,78:133 - 146.

[8] 齐继阳,李金燕,陆震云,等. 改进的 Otsu 法在焊接图像分割中的应用[J]. 焊接学报,2016,37(10):97 - 100,135.

[9] 袁小翠,吴禄慎,陈华伟. 基于 Otsu 方法的钢轨图像分割[J]. 光学精密工程,2016,24(7):1772 - 1781.

[10] 马天兵,刘健,杜菲,等. 基于改进 Otsu 方法的振动图像分割研究[J]. 电光与控制,2019,26(2):15 - 19,35.

[11] Fan J L, Lei B. A modified valley-emphasis method for automatic thresholding[J]. Pattern Recognition Letters,2012, 33(6):703 - 708.

[12] 申铉京,张赫,陈海鹏,等. 快速递归多阈值分割算法[J]. 吉林大学学报(工学版),2016,46(2):528 - 534.

[13] 周榆丰. 天车吊运系统中运动目标识别与匹配方法研究[D]. 唐山:河北联合大学,2014.

[14] 韩青松,贾振红,杨杰,等. 基于改进的 OTSU 算法的遥感图像阈值分割[J]. 激光杂志,2010,31(6):33 - 34.

[15] 崔长彩,王克贤,黄国钦,等. 单层钎焊金刚石砂轮表面磨粒全场快速测量[J]. 中国机械工程,2019,30(14):1639 - 1645.

(上接第 67 页)

观的性能提升,且捕捉复杂时空变化特征的能力更强,更能胜任像素级视频预测的任务。若针对任务内容对网络结构进行改进,会获得更加可观的性能提升。

此外,本文提出的网络结构依然具有改进的空间,在面对像素级预测任务时,可以加入注意力机制,在每次提取特征时都可以对不同通道加入不同的权重,以提高预测深度的效果。

参 考 文 献

[1] Rosenblatt F. The perceptron: a probabilistic model for information storage and organization in the brain[J]. Psychological review, 1958, 65(6): 386.

[2] Lecun Y, Bottou L, Bengio Y, et al. Gradient-based learning applied to document recognition[J]. Proceedings of the IEEE, 1998, 86(11): 2278 - 2324.

[3] Zaremba W, Sutskever I, Vinyals O. Recurrent neural network regularization[EB]. arXiv:1409.2329, 2014.

[4] Lotter W, Kreiman G, Cox D. Deep predictive coding networks for video prediction and unsupervised learning[EB]. arXiv:1605.08104, 2016.

[5] Kim S, Hong S, Joh M, et al. DeepRain: ConvLSTM network for precipitation prediction using multichannel radar data[EB]. arXiv:1711.02316, 2017.

[6] Villegas R, Yang J, Hong S, et al. Decomposing motion and content for natural video sequence prediction[EB]. arXiv: 1706.08033, 2017.

[7] Jaderberg M, Simonyan K, Zisserman A. Spatial transformer networks[C]//Advances in neural information processing systems. 2015: 2017 - 2025.

[8] Hochreiter S, Schmidhuber J. Long short-term memory[J]. Neural Computation, 1997, 9(8): 1735 - 1780.

[9] Gers F A, Schmidhuber J. Recurrent nets that time and count[C]//Proceedings of the IEEE-INNS-ENNS International Joint Conference on Neural Networks. IJCNN 2000. Neural Computing: New Challenges and Perspectives for the New Millennium, 2000.

[10] Huang B, Huang H, Lu H. Convolutional gated recurrent units fusion for video action recognition[C]//International Conference on Neural Information Processing, 2017.

[11] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need[C]//Advances in Neural Information Processing Systems, 2017.

[12] Li X, Orchard M T. New edge-directed interpolation[J]. IEEE transactions on image processing, 2001, 10(10): 1521 - 1527.

[13] Srivastava N, Mansimov E, Salakhudinov R. Unsupervised learning of video representations using lstms[C]//International Conference on Machine Learning, 2015.