A Survey of Spatial-Temporal Predictive Models

Xutao Cao[1,2*]

[1] University of Illinois at Urbana-Champaign;
[2]Zhejiang University;

[*]Corresponding to: xutaoc2@illinois.edu

Abstract

This paper is a short survey into the field of Spatial-Temporal models. I mainly introduce three different types of spatial-temporal models. The first one is the recurrent networks, Convolutional LSTM. It is the first to conbine convolution operations with LSTM to model both spatial and temporal dependecies. And it achieved great performance on a precipitation nowcasting task at that time. Besides ConvLSTM, I also introduce some improved RNN-based spatial-temporal networks after it. The second is a convolutional neural network based model called 3D ConvNets. It is a modification of conventional 2D convolutional networks and has great performance on video prediction tasks. The third type of model I introduce is Graph-based model, Spatio-Temporal Graph Convolutional Networks (STGCN).And I will also introduce someIvariant of STGCN. I group the papers together by illustrating their methodology, presenting their experiments and results, analyzing related works ,and making a contrast between them.

1. Introduction

The spatial-temporal sequence tasks involves two elements: time and space. The term ”time” here refers to the sequence preceding and following. The space here also refers to the target on the image and the spatial information about the target’s movement and change. Additionally, it refers to the presence of GPS data or spatial data such as x, y, or latitude and longitude in tabular data. For example using the first n frames of a video to predict the next m frames of the video is a typical spatio-temporal topic nameed video prediction.

Actually, there are many other tasks in which spatial-temporal models behaves well. (1) In the area of precipitation nowcasting, radar maps can be constructed to a time sequence and apply spatial-temporal models to predict the future precipitation. (2) In the area of autonomous driving where spatial-temporal models can be used to predict the motion of pedestrians. (3) In agricultural, satellite images are used to classify the crop types in the fields, and these images can also be contructed as time sequence and therefore applying spatial-temporal models to the classification task.

There are various types of spatial-temporal models. In these survey, I will briefly review the most common deep networks that are used as building blocks for the spatial-temporal tasks: recurrent networks, convolution neural networks and graph-based networks.

2. RNN based Approaches

Here we take Convolutional LSTM Network[23] as an example.

LSTM[6] as a special RNN structure has proven stable and powerful for modeling long-range dependencies in various studies. But it contains too much redundancy for spatial data. Based on LSTM, Convolutional LSTM (ConvLSTM)[23] was proposed to model both temporal and spatial dependencies. In this paper, the goal of this model was to tackle the precipitation nowcasting problem, and to give precise and timely prediction of rainfall intensity in a local region over a relatively short period of time (e.g. 0-6 hours).

2.1. Methodology

The research uses radar maps as their input dataset. The observation at every timestamp is a 2D radar echo map. Observations are recorded periodically to get a sequence of tensors $\hat{\mathcal{X}}_1, \hat{\mathcal{X}}_2, \ldots, \hat{\mathcal{X}}_t$. The spatiotemporal sequence forecasting problem is to predict the most likely length-K sequence in the future given the previous J observations which include the current one: $\tilde{\mathcal{X}}_{t+1}, \ldots, \tilde{\mathcal{X}}_{t+K} = \arg\max_{\mathcal{X}_{t+1},\ldots,\mathcal{X}_{t+K}} p\left(\mathcal{X}_{t+1}, \ldots, \mathcal{X}_{t+K} \mid \hat{\mathcal{X}}_{t-J+1}, \ldots, \hat{\mathcal{X}}_t\right)$

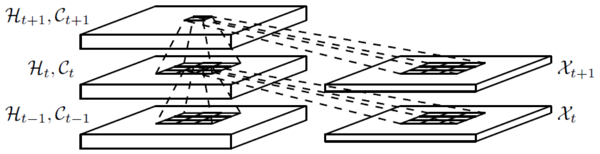
$\mathcal{H}_{t+1}, \mathcal{C}_{t+1}$
$\mathcal{H}_{t}, \mathcal{C}_{t}$
$\mathcal{H}_{t-1}, \mathcal{C}_{t-1}$
$\mathcal{X}_{t+1}$
$\mathcal{X}_{t}$

Figure 1. Inner structure of ConvLSTM

2.1.1 Structure of Convolutional LSTM

The primary disadvantage of Fully connected LSTMs for handling spatiotemporal data is their reliance on full connections for input-to-state and state-to-state transitions that do not encode spatial information. To overcome this problem, a distinguishing feature of ConvLSTM is that all the inputs $x_1, ..., x_t$, cell outputs $C_1, ..., C_t$, hidden states $H_1, ..., H_t$ and gates i_t, f_t, o_t of the ConvLSTM are 3D tensors whose last two dimensions are spatial dimensions (rows and columns). The ConvLSTM determines the future state of a certain cell by the inputs and past states of its local neighbors. This can be easily achieved by using a convolution operator in the state-to-state and input-to-state transitions. The key equations of ConvLSTM are shown below. where * denotes the convolution operator and ∘ denots the Hadamard product.

$$\begin{aligned}
i_t &= \sigma\left(W_{xi} * X_t + W_{hi} * H_{t-1} + W_{ci} \circ c_{t-1} + b_i\right) \\
f_t &= \sigma\left(W_{xf} * X_t + W_{hf} * H_{t-1} + W_{cf} \circ c_{t-1} + b_f\right) \\
c_t &= f_t \circ c_{t-1} + i_t \circ \tanh\left(W_{xc} * X_t + W_{hc} * H_{t-1} + b_c\right) \\
o_t &= \sigma\left(W_{xo} * X_t + W_{ho} * H_{t-1} + W_{co} \circ c_t + b_o\right) \\
h_t &= o_t \circ \tanh\left(c_t\right)
\end{aligned}$$

It is very similar to the original LSTM design except for the 3D tensor input and the convolution operations used between tensors.

Besides, as the author changed the input tensor to 3d tensors, extra padding is needed before covolution operation to maintain the number of rows and number of solumns same as input. In the paper, zero padding is performed on hidden states. And author also gives his explanation why zero padding is chosen here. He believes padding of hidden states on the voundary inputs can be viewed as using the state of the ourside world for calculation, and when zero padding is performed, we are setting the state of the outside world to zero and assume no prior knowledge about the outside. In this way, we can treat the boundary point differently, which is helpful in many cases.

2.1.2 The Encoding-Forcasting Structure

ConvLSTM can also be a building block for more complex structures. In this spatiotemporal sequence forcasting problem, they use the Encoding-Forecasting structure shown in the figure below. The encoding LSTM compresses the entire input sequence into a hidden state tensor, which is then unfolded to produce the final prediction. We can find that this structure is similar to the Encoding-Decoding structure in the seq2seq[14].

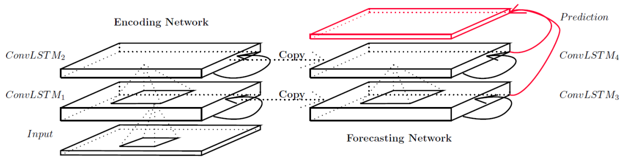

Figure 2. Encoding-forecasting ConvLSTM network for precipitation nowcasting

2.2. Experiments and Results

The author compared the ConvLSTM network with the Fully-Connected on a synthetic Moving-MNIST dataset to gain some basic understanding of the behavior of ConvLSTM. And he also build a new radar echo dataset and compare the model with the state-of-the-art ROVER algorithm[21] based on several commonly used precipitation nowcasting metrics. Through these experiments, the author comes to following conclusions: 1. ConvLSTM is better than FC-LSTM in handling spatial-temporal correlations. 2. Making the size of state-to-state convolution kernel bigger than 1 is essential for capturing the spatiotemporal motion patterns. 3. Deeper models can produce better results with fewer parameters. 4. ConvLSTM performs better than ROVER for precipitation nowcasting.

2.3. Conclusions

This is a pioneering paper in the spatial-temporal field. Compared with standard LSTMs, the ConvLSTM is able to model the spatiotemporal structures simultaneously by explicitly encoding the spatial information into tensors, overcoming the limitation of vector-variate representations in standard LSTM where the spatial information is lost.

2.4. Related Works

In [10], Shi pointed out that though ConvLSTM is proven to be better than the fully-connected recurrent structure in capturing spatial-temporal correlations, it is not optimal and leaves room for improvements. For rotational and scaling motion patterns, the local correlation structure of consecutive frames will vary according to spatial location and timestamp. Thus, it is inefficient to represent such location-variant relationships using convolution, which employs a location-invariant filter. So in his new paper, he proposed the Trajectory GRU (TrajGRU) which can actively learn the locatioin-variant structure for recurrent connections.

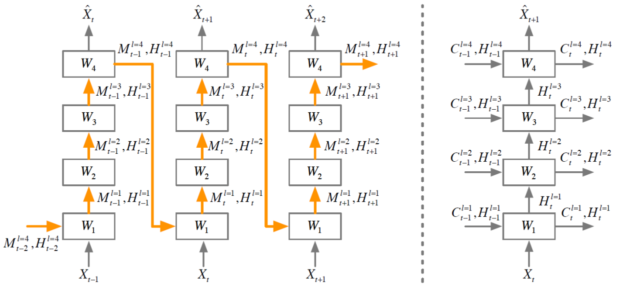
$\hat{X}_t$
$\hat{X}_{t+1}$
$\hat{X}_{t+2}$
$\hat{X}_{t+1}$
$M_{t-1}^{l=4}, H_{t-1}^{l=4}$
$M_t^{l=4}, H_t^{l=4}$
$M_{t+1}^{l=4}, H_{t+1}^{l=4}$
$C_{t-1}^{l=4}, H_{t-1}^{l=4}$
$C_t^{l=4}, H_t^{l=4}$
W_4
$M_{t-1}^{l=3}, H_{t-1}^{l=3}$
$M_t^{l=3}, H_t^{l=3}$
$M_{t+1}^{l=3}, H_{t+1}^{l=3}$
$H_t^{l=3}$
$C_{t-1}^{l=3}, H_{t-1}^{l=3}$
$C_t^{l=3}, H_t^{l=3}$
W_3
$M_{t-1}^{l=2}, H_{t-1}^{l=2}$
$M_t^{l=2}, H_t^{l=2}$
$M_{t+1}^{l=2}, H_{t+1}^{l=2}$
$H_t^{l=2}$
$C_{t-1}^{l=2}, H_{t-1}^{l=2}$
$C_t^{l=2}, H_t^{l=2}$
W_2
$M_{t-1}^{l=1}, H_{t-1}^{l=1}$
$M_t^{l=1}, H_t^{l=1}$
$M_{t+1}^{l=1}, H_{t+1}^{l=1}$
$H_t^{l=1}$
$C_{t-1}^{l=1}, H_{t-1}^{l=1}$
$C_t^{l=1}, H_t^{l=1}$
W_1
$M_{t-2}^{l=4}, H_{t-2}^{l=4}$
X_{t-1}
X_t
X_{t+1}
X_t

Figure 3. Left: The ConvLSTM network with a spatiotemporal memory flow. Right: The conventional ConvLSTM architecture. The orange arrows denote the memory flow derection for all memory cells.

In [19], the author presented a new structure, predictive recurrent neural network (PredRNN). In this paper, he demonstrated how a four-layer ConvLSTM encoder-decoder network works by feeding input frames into the first layer and generating the future video sequence in the fourth. Spatial representations are encoded layer by layer in this process, with hidden states delivered from bottom to top. However, the memory cells in these four layers are completely independent of one another and are only updated in the time domain. Under these circumstances, the bottom layer would completely disregard what the top layer had memorized during the previous time step. The new model PredRNN proposed in this paper is an improvement for ConvLSTM, which models spatial deformations and temporal variations simutaneously and achieved state-of-the-art performance on three video prediction dataset s including both synthetic and natural video sequences.

And in 2018, a improvement of PredRNN was proposed by the same research team. In [17], they proposed PredRNN++ towards a resolution of the spatiotemporal predictive learning dilemma between deep-in-time structures and vanishing gradientsThey designed the causal LSTM with a cascaded dual memory structure to enhance its ability to model short-term dynamics. To circumvent the vanishing gradient problem, they proposed a gradient highway unit that connected distant previous inputs to distant future predictions. Additionally, they obtained state-of-the-art prediction results on a synthetic moving digits dataset and a real video dataset.

In [20], Wang et al. believe that the high-order non-stationarity of video dynamics has been overlooked in the preceding work, which uses relatively simple temporal transition methods, either controlled by recurrent gate structures or implemented via feed-forward network recursion. They proposed Memory In Memory (MIM) networks and corresponding recurrent blocks for mitigating the non-stationary learning difficulty by utilizing high-order differencing. As a result, they achieved state-of-the-art prediction performance on four datasets: a synthetic dataset of flying digits, a traffic flow prediction dataset, a weather forecasting dataset, and a human pose video dataset.

3. CNN-Based Approaches

Convolutional layers are the fundamental building blocks of deep learning architectures for visual reasoning, as they efficiently model the spatial structure of images using Convolutional Neural Networks (CNNs). However, their performance in the field of spatial-temporal sequence prediction is limited by intra- and inter-frame dependencies. Convolution operations account for short-range intra-frame dependencies as a result of their limited receptive fields, which are determined by the kernel size. This is a well-resolved issue, which many authors addressed by 1. stacking additional convolutional layers, 2. increasing the kernel size, 3. linearly combining multiple scales, as in the reconstruction of a Laplacian pyramid, 4. using dilated convolutions to capture long-range spatial dependencies, and 5. enlarging the receptive fields or resampling.

The above-mentioned vanilla CNNs still lack explicit inter-frame modeling capabilities. To accurately model the variability between frames in a spatio-temporal sequence, 3D convolutions offer a promising alternative to recurrent modeling. The above-mentioned vanilla CNNs still lack explicit inter-frame modeling capabilities. To accurately model the variability between frames in a spatio-temporal sequence, 3D convolutions offer a promising alternative to recurrent modeling.

Here I'd like to introduce one of most pioneering 3D convolution paper, [15] published in ICCV 2015.

3.1. Methodology

The author noted that 3D ConvNets are well-suited for learning spatiotemporal features. In comparison to 2D ConvNets, 3D ConvNets are more capable of modeling temporal information due to their 3D convolution and pooling operations. Convolution and pooling operations are performed spatiotemporally in 3D ConvNets, whereas they are performed spatially only in 2D ConvNets. 2D convolution applied on an image will output an image, 2D convolution applied on multiple images (treating them as different channels) also results in an image. As a result, 2D ConvNets immediately lose temporal information about the input signal following each convolution operation. Only three-dimensional convolution preserves the temporal information contained in the input signals, resulting in a volume output. The same phenomenon occurs with 2D and 3D polling.

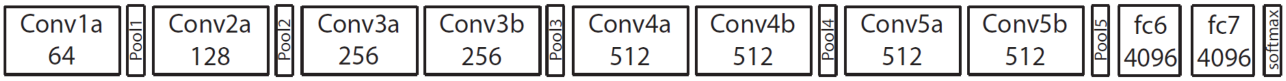

Figure 4. **C3D architecture.** C3D net has 8 convolution, 5max-pooling, and 2 fully connected layers, followed by a softmax output later. All 3D convolution kernels are $3 \times 3 \times 3$ with stride 1 in both spatial and temporal dimensions. Number of filters are denoted in each box. The 3D pooling laters are denoted from pool1 to pool5. All pooling kernels are $2 \times 2 \times 2$, except for pool1 is $1 \times 2 \times 2$. Each fully connnected layer has 4096 output units.

3.1.1 Model Architecture

After experimenting with various network architectures and exploring different kernel temporal depth, the author came to the conclusion that $3 \times 3 \times 3$ is the best kernel choice for 3D ConvNets. And 3D ConvNets are consistently better than 2D ConvNets for spatio-temporal tasks like video classification. They designed their 3D ConvNets (C3D) to have 8 convolutional layers, 5 pooling layers, followed by two fully connected layers, and a softmax output layer. The network structure is presented in figure 4. All of 3D convolution filters are $3 \times 3 \times 3$ with stride $1 \times 1 \times 1$. All 3D pooling layers are $2 \times 2 \times 2$ with stride $2 \times 2 \times 2$ except for pool1 which has kernel size of $1 \times 2 \times 2$ and stride $1 \times 2 \times 2$ with the intention of preserving the temporal information in the early phase. Each fully connected layer has 4096 output units.

3.1.2 What does C3D learn

Additionally, the author used deconvolution to deduce what C3D is learning internally from a motion video dataset. They observe that C3D initially focuses on appearance and then on salient motion in subsequent frames. In the first example, it focuses on the entire person before tracking the pole vault performance throughout the frame. Similarly, in the second example, it focuses first on the eyes and then on the motion occurring around the eyes as the makeup is applied. Thus, C3D is distinguished from conventional 2D ConvNets in that it attends selectively to both motion and appearance.

3.2. Experiments and Results

The author experimented the C3D model on the UCF101 dataset[13], the ASLAN dataset consists of 3,631 videos from 432 action classes and 2 benchmarks YUPENN[4] and Maryland[11]. They concluded that C3D is capable of modeling both appearance and motion simultaneously and outperforms 2D ConvNet features on a variety of spatio-temporal tasks, including video analysis. They demonstrated that when combined with a linear classifier, C3D features can outperform or approach current best methods for video analysis. Last but not least, C3D features they proposed are efficient, compact and extremely simple to use.

3.3. Related Works

In [2], the author noted that one disadvantage of 3D models is that they have significantly more parameters than 2D ConvNets due to the addition of the kernel dimension, which makes them more difficult to train. Additionally, they appear to preclude the benefits of ImageNet pre-training, and thus previous work has defined and trained relatively shallow custom architectures from scratch. The author introduces a new two-stream inflated 3D ConvNet (I3D) based on 2D ConvNet inflation: the filters and pooling kernels of extremely deep image classification ConvNets are expanded into 3D, enabling the learning of seamless spatio-temporal feature extractors from video while leveraging successful ImageNet architecture designs and even parameters.

In [18], The success of 3D CNNs prompted the author to propose a new model for spatiotemporal predictive learning that incorporates both recurrent modeling (for temporal dependency) and feed-forward 3D-Conv modeling (for local dynamics). This model is referred to as the Eidetic 3D LSTM (E3D-LSTM). Additionally, they introduce an eidetic 3D memory to: a) memorize local appearance and motion in a small spatiotemporal volume, and b) recall the long-range historical context through the acquisition of the ability to attend to previous memory states. In many cases, spatiotemporal predictive modeling is highly dependent on temporally adjacent appearances and ongoing short-term motions. With a short time convolution window, all the information is encapsulated in the eidetic 3D memory cell and used in recurrent transitions. Experiments conducted in the paper demonstrate that the E3D-LSTM model outperforms state-of-the-art methods on video prediction and early activity recognition tasks.

4. Graph Models

In 2018 IJCAI, Yu etal.[24] proposed a novel deep learning frame work. Spatio-Temporal Graph Convolutional Networks (STGCN), to tackle the time series prediction problem in traffic domain which is also a typical spatio-temporal problem. Rather than using standard convolutional and recurrent units, they formulate the problem in terms of graphs and construct the model using fully convolutional structures, which enables significantly faster training with fewer parameters.

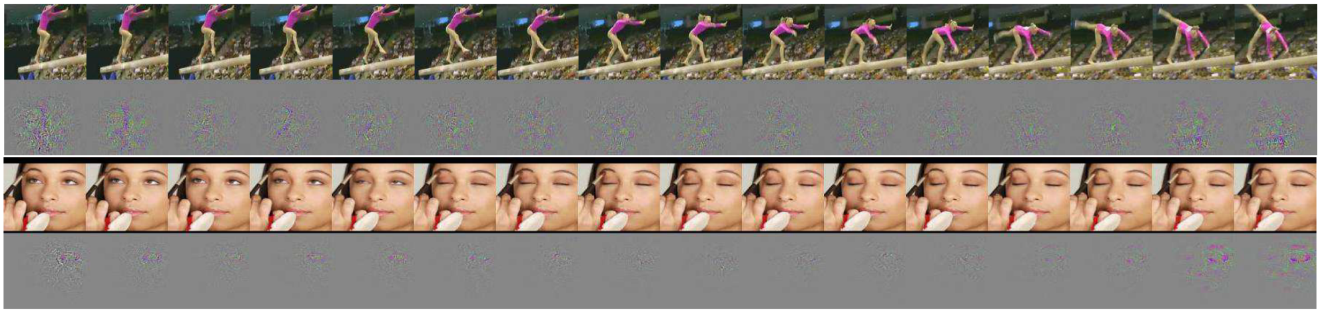

Figure 5. **Visualization of C3D model, using the method from [25]** Interestingly, C3D captures appearance for the first fewframes but thereafter only attends to salient motion.

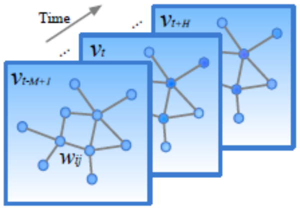

Figure 6. Gragh-structured traffic data. Each v_t indicates a frame of current raffic status at time step t, which is recorded in a graph-structured data matrix.

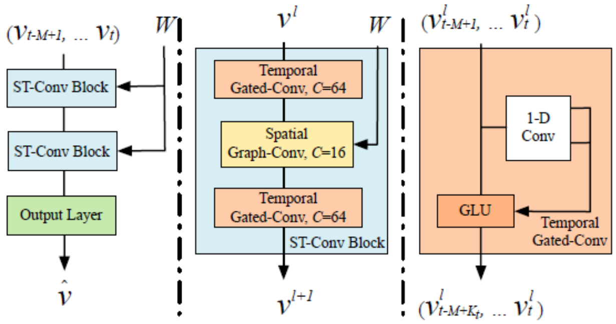
$(v_{t-M+1}, \ldots v_t)$
W
ST-Conv Block
ST-Conv Block
Output Layer
$\hat{v}$
v^l
W
Temporal Gated-Conv, C=64
Spatial Graph-Conv, C=16
Temporal Gated-Conv, C=64
ST-Conv Block
v^{l+1}
$(v^l_{t-M+1}, \ldots v^l_t)$
1-D Conv
GLU
Temporal Gated-Conv
$(v^l_{t-M+K_t}, \ldots v^l_t)$

Figure 7. Architecture of spatio-temporal graph convolutional networks.

4.1. Methodology

Traffic forecast is to predict the most likely traffic measurements (e.g. speed or traffic flow) in the next H time steps given the previous M traffic observations as,

$$\hat{v}_{t+1}, \ldots, \hat{v}_{t+H} = \underset{v_{t+1}, \ldots, v_{t+H}}{\arg\max} \log P\left(v_{t+1}, \ldots, v_{t+H} \mid v_{t-M+1}, \ldots, v_t\right)$$

They define the traffic network on a graph and focus on structured traffic time series. The observation v_t is not independent but linked by pairwise connection in graph. Therefore, the data point v_t can be regarded as a graph signal that is defined on an undirected graph (or directed one) $\mathcal{G}$ with weights w_{ij} as shown in Figure 6 At the t-th time step, in graph $\mathcal{G}_t = \left(\mathcal{V}_t, \mathcal{E}, W\right), \mathcal{V}_t$ is a finite set of vertices, corresponding to the observations from n monitor stations in a traffic network; $\mathcal{E}$ is a set of edges, indicating the connectedness between stations; while $W \in \mathbb{R}^{n \times n}$ denotes the weighted adjacency matrix of $\mathcal{G}_t$.

4.1.1 Convolutions on Graphs

There are two main ways of generalizing CNNs to structured data forms. One is to expand the spatial definition of a convolution [9] and the other is to manipute in the spectral domain with graph Fourier transforms [1]. In this paper, the author introduce the notion of graph convolution operator " $*\mathcal{G}$ " based on the conception of spectral graph convolution, as the multiplication of a signal $x \in \mathbb{R}^n$ with a kernel Θ,

$$\Theta *_{\mathcal{G}} x = \Theta(L)x = \Theta\left(U\Lambda U^T\right)x = U\Theta(\Lambda)U^T x$$

where graph Fourier basis $U \in \mathbb{R}^{n \times n}$ is the matrix of eigenvectors of the normalized graph Laplacian $L = I_n - D^{-\frac{1}{2}} W D^{-\frac{1}{2}} = U\Lambda U^T \in \mathbb{R}^{n \times n}$ (I_n is an identity matrix, $D \in \mathbb{R}^{n \times n}$ is the diagonal degree matrix with $D_{ii} = \Sigma_j W_{ij}$; $\Lambda \in \mathbb{R}^{n \times n}$ is the diagonal matrix of eigenvalues of L, and filter $\Theta(\Lambda)$ is also a diagonal matrix. By this definition, a graph signal x is filtered by a kernel Θ with multiplication between Θ and graph Fourier transform $U^T x$ [1].

4.1.2 Network Architecture

As shown in Figure 7, The framework STGCN consists of two spatio-temporal convolutional blocks (ST-Conv blocks) and a fully-connected output layer in the end. Each ST-Conv block contains two temporal gated convolution layers and one spatial graph convolution layer in the middle. The residual connection and bottleneck strategy are applied inside each block. The input $v_{t-M+1}, \ldots, v_t$ is uniformly processed by ST-Conv blocks to explore spatial and temporal dependencies coherently. Comprehensive features are integrated by an output layer to generate the final prediction $\hat{v}$.

4.2. Experiment and Result

The team verify their model on two real-world traffic datasets BJER4 and PeMSD7 collected by Beijing Municipal Traffic Commission and California Department of Transportation, respectively. The standards for evaluating different models are Mean Absolute Error (MAE), Mean Absolute Percent Error (MAPE) and Mean Square Error (RMSE). And the model proposed in the paper has achieved very good results in short, medium and long-term predictions with much shorter training time. The author concluded that the experiments show that their model out performs other state-of-the-art methods on two real-world datasets.

4.3. Related Works

Numerous studies in deep learning have been motivated by graph convolution in spatio-temporal tasks. Seo et al. [3] developed the graph convolutional recurrent network (GCRN) to extract both spatial and dynamic variation from structured data sequences. The primary objective of this study is to identify the optimal combinations of recurrent networks and graph convolution in a variety of contexts. Li et al. [7] successfully used gated recurrent units (GRU) with graph convolution for long-term traffic forecasting using the aforementioned principles. In contrast to these works, we construct our model entirely from convolutional structures; the ST-Conv block is specifically designed to process structured data uniformly using a residual connection and bottleneck strategy; and our model also makes use of more efficient graph convolution kernels.

Following that, A STGCN[5] employs two attention layers to capture the dynamics of spatial and temporal dependencies. Graph WaveNet[22] creates a self-adaptive matrix that takes into account the variations in influence between nodes and their neighbors. It utilizes temporal correlations to exponentially expand the receptive field. In [12] Spatial-Temporal Synchronous Graph Convolutional Networks (STSGCN) is proposed for the same task. This model is able to effectively capture the complex localized spatial-temporal correlations through an elaborately designed spatial-temporal synchronous modeling mechanism.

5. Conclusions

The first type of method I introduce in this article is ConvLSTM. This is a pioneering paper in the field of spatial-temporal models and has affacted many researchers after it. After the publication of [23], Researchers began to focus on improving the memory structure of LSTM and the stacking architectures between layers. In this context, many variants of ConvLSTM were proposed, like the PredRNN, TrajGRU, PredRNN++ mentioned above and they all achieved state-of-the-art performance. Nowadays as the publishment of Attention mechanism [16] some teams even tried to combine the ConvLSTM with attention (Self-attention ConvLSTM [8]) to further improve the ConvLSTM. However, recurrent nets still suffers from the high computational cost and has space for performance improvement. Up to now, the recurrent network based model is still the mainstream method for spatial-temporal task.

The second type of method I introduce in this article is 3D ConvNets (C3D). In comparison to 2D Con vNets, 3D ConvNets are more capable of modeling temporal information due to their 3D convolution and pooling operations. In the video prediction task, C3D tends to model both the intra-frame dependencies and the inter-frame dependencies abd us proven tobe efficient compact and extremely simple to use. However, the C3D also suffers from high computational cost and some new models like I3D, E3D-LSTM are published as improvements of C3D.

The third type of method I introduce in this article is Graph-based models, STGCN. Graph models are fairly new topics in the field of deep learning, and are proven to be efficient in the field of spatial-temporal prediction. Graph-based models extend the field of spatio-temporal tasks, as the frame no longer need to be rectangular and can be anything that could be abstracted as an graph. As these graph may not contain so many pixels like the video frames in video prediction task or the radar maps in precipitaion nowcasting task, the computational cost could come down. This is one promising feature of Graph-based models. However, the performance of these kind model still has much space to be improved.

Although researches on spatial-temporal tasks are not so popular like other CV tasks (object detection, pattern recognition, etc). It still has a hugh number of real life applications. So I believe there will be better and better spatial-temporal models in the near future.

References

[1] Joan Bruna, Wojciech Zaremba, Arthur Szlam, and Yann LeCun. Spectral networks and locally connected networks on graphs. *arXiv preprint arXiv:1312.6203*, 2013. 5

[2] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017. 4

[3] Quanjun Chen, Xuan Song, Harutoshi Yamada, and Ryosuke Shibasaki. Learning deep representation from big and heterogeneous data for traffic accident inference. In *Thirtieth AAAI conference on artificial intelligence*, 2016. 6

[4] Konstantinos G Derpanis, Matthieu Lecce, Kostas Daniilidis, and Richard P Wildes. Dynamic scene understanding: The role of orientation features in space and time in scene classification. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1306–1313. IEEE, 2012. 4

[5] Shengnan Guo, Youfang Lin, Ning Feng, Chao Song, and Huaiyu Wan. Attention based spatial-temporal graph convolutional networks for traffic flow forecasting. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 922–929, 2019. 6

[6] Sepp Hochreiter and Jürgen Schmidhuber. Long Short-Term Memory. *Neural Computation*, 9(8):1735–1780, 11 1997. 1

[7] Yaguang Li, Rose Yu, Cyrus Shahabi, and Yan Liu. Diffusion convolutional recurrent neural network: Data-driven traffic forecasting. *arXiv preprint arXiv:1707.01926*, 2017. 6

[8] Zhihui Lin, Maomao Li, Zhuobin Zheng, Yangyang Cheng, and Chun Yuan. Self-attention convlstm for spatiotemporal prediction. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 11531–11538, 2020. 6

[9] Mathias Niepert, Mohamed Ahmed, and Konstantin Kutzkov. Learning convolutional neural networks for graphs. In *International conference on machine learning*, pages 2014–2023. PMLR, 2016. 5

[10] Xingjian Shi, Zhihan Gao, Leonard Lausen, Hao Wang, Dit-Yan Yeung, Wai-kin Wong, and Wang-chun WOO. Deep learning for precipitation nowcasting: A benchmark and a new model. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. 2

[11] Nitesh Shroff, Pavan Turaga, and Rama Chellappa. Moving vistas: Exploiting motion for describing scenes. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 1911–1918. IEEE, 2010. 4

[12] Chao Song, Youfang Lin, Shengnan Guo, and Huaiyu Wan. Spatial-temporal synchronous graph convolutional networks: A new framework for spatial-temporal network data forecasting. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 914–921, 2020. 6

[13] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. Ucf101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402*, 2012. 4

[14] Ilya Sutskever, Oriol Vinyals, and Quoc V Le. Sequence to sequence learning with neural networks. In *Advances in neural information processing systems*, pages 3104–3112, 2014. 2

[15] Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. Learning spatiotemporal features with 3d convolutional networks. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, December 2015. 3

[16] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008, 2017. 6

[17] Yunbo Wang, Zhifeng Gao, Mingsheng Long, Jianmin Wang, and Philip S Yu. PredRNN++: Towards a resolution of the deep-in-time dilemma in spatiotemporal predictive learning. In Jennifer Dy and Andreas Krause, editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 5123–5132. PMLR, 10–15 Jul 2018. 3

[18] Yunbo Wang, Lu Jiang, Ming-Hsuan Yang, Li-Jia Li, Mingsheng Long, and Li Fei-Fei. Eidetic 3d LSTM: A model for video prediction and beyond. In *International Conference on Learning Representations*, 2019. 4

[19] Yunbo Wang, Mingsheng Long, Jianmin Wang, Zhifeng Gao, and Philip S Yu. Predrnn: Recurrent neural networks for predictive learning using spatiotemporal lstms. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, pages 879–888, 2017. 3

[20] Yunbo Wang, Jianjin Zhang, Hongyu Zhu, Mingsheng Long, Jianmin Wang, and Philip S. Yu. Memory in memory: A predictive neural network for learning higher-order non-stationarity from spatiotemporal dynamics. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019. 3

[21] WC Woo and WK Wong. Application of optical flow techniques to rainfall nowcasting. In *the 27th Conference on Severe Local Storms*, 2014. 2

[22] Zonghan Wu, Shirui Pan, Guodong Long, Jing Jiang, and Chengqi Zhang. Graph wavenet for deep spatial-temporal graph modeling. *arXiv preprint arXiv:1906.00121*, 2019. 6

[23] SHI Xingjian, Zhourong Chen, Hao Wang, Dit-Yan Yeung, Wai-Kin Wong, and Wang-chun Woo. Convolutional lstm network: A machine learning approach for precipitation nowcasting. In *Advances in neural information processing systems*, pages 802–810, 2015. 1, 6

[24] Bing Yu, Haoteng Yin, and Zhanxing Zhu. Spatio-temporal graph convolutional networks: A deep learning framework for traffic forecasting. *arXiv preprint arXiv:1709.04875*, 2017. 4

[25] Matthew D Zeiler and Rob Fergus. Visualizing and understanding convolutional networks. In *European conference on computer vision*, pages 818–833. Springer, 2014. 5