A Novel Video Prediction Algorithm based on Robust Spatiotemporal Convolutional LongShort Term Memory

Wael Saideni, David Helbert, Fabien Courrèges, Jean Pierre Cances

A Novel Video Prediction Algorithm based on Robust Spatiotemporal Convolutional Long Short Term Memory (Robust-ST-ConvLSTM)

Wael Saideni, David Helbert, Fabien Courreges, and Jean Pierre Cances

XLIM Research Institute, UMR CNRS 7252
✉wael.saideni@xlim.fr
✉david.helbert@univ-poitiers.fr
✉fabien.courreges@unilim.fr
✉cances@ensil.unilim.fr

Abstract. Recently, video prediction algorithms based on neural networks have become a promising research direction. Therefore, a new recurrent video prediction algorithm called "Robust Spatiotemporal Convolutional Long Short-Term Memory" (Robust-ST-ConvLSTM) is proposed in this paper. Robust-ST-ConvLSTM proposes a new internal mechanism that is able to regulate efficiently the flow of spatiotemporal information from video signals based on higher order Convolutional-LSTM. The spatiotemporal information is carried through the entire network to optimize and control the prediction potential of the ConvLSTM cell. In addition, in traditional ConvLSTM units, cell states, that carry relevant information throughout the processing of the input sequence, are updated using only one previous hidden state, which holds information on previous data unit already seen by the network. However, our Robust-ST-ConvLSTM unit will rely on N previous hidden states, that provide temporal context for the motion in video scenes, in the cell state updating process. Experimental results further suggest that the proposed architecture can improve the state-of-the-art video prediction methods significantly on two challenging datasets, including the standard Moving MNIST dataset, and the commonly used video prediction KTH dataset, as human motion dataset.

Keywords: Video prediction, deep learning, neural networks, computer vision, ConvLSTM, memory flow, hidden states

1 Introduction

Video prediction, one of the emerging fields of computer vision, is facing several challenges [1–5]. Actually, it has gained significant interests due to its broad-ranging realistic forecasting applications, such as traffic flow prediction and video surveillance.

The great progress made by deep learning in a wide range of applications and research fields, motivated authors to explore deep leaning architectures to predict future video frames. The main advantage of deep learning models is their

potential to learn adequate features from high-dimensional data, such as videos, in an end-to-end manner without hand-designed features [6]. However, despite the significant progress in deep learning architectures, video prediction is still considered as a big challenge especially in terms of output visual quality and long-term prediction. Therefore, our Robust Spatiotemporal Convolutional Long Short-Term Memory (Robust-ST-ConvLSTM) algorithm is proposed as a long-term prediction algorithm that outperforms the state-of-the-art approaches in terms of quality performances. Our algorithm is based on a modified version of ConvLSTM cell. Obviously, ConvLSTM is not very efficient in handling long sequences. Indeed, ConvLSTM based algorithms focus on stochastic features of the data rather than its spatial distortion. Also, a temporal information encoding in ConvLSTM unit [7] is based on 1^{st}-order Markovian architecture. Thus, making long-range temporal correlations hard to extract. In addition, the vanishing gradient problem often occurs in training 1^{st}-order RNN based predictive algorithms [8].
Bearing all these drawbacks in mind, we propose our Robust-ST-ConvLSTM algorithm for video prediction. With the following properties, we hope our algorithm will pave the way for the application of recurrent neural network on real-wold datasets:

- Spatial and temporal data are taken into consideration jointly.
- The new spatiotemporal memory (STM) cell transfers low-level and semantic aspects of the dynamic scene which are the key of generating future frames.
- The Robust-ST-ConvLSTM new internal mechanism offers new cell state and hidden state transition functions to efficiently regulate the flow of spatiotemporal information from the input videos.
- The algorithm aims to rely on N previous hidden states, that provide temporal context for the motion in video scenes, to update one cell state at every timestep.

The remainder of this paper is organized as follows: The related works on video prediction are discussed in Section 2. In Section 3, our proposed Robust-ST-ConvLSTM algorithm is presented. Section 4 provides the experimental results. Section 5 concludes the paper.

2 Related Works

Video prediction algorithms have used various deep learning architectures to enhance the quality performance of the predicted frames and to fasten the process. Deep learning has been extensively used to analyze the frames and extract their features exploited in spatiotemporal predictive learning.
Recent deep learning approaches can be categorized into three classes: recurrent neural approaches, convolutional networks based algorithms and generative networks.
Recurrent neural networks (RNN) have demonstrated a significant success in

recent video prediction related works [9–25]. ConvLSTM [7] is considered as a crucial branch in predicting future frame. A two-stream architecture based on adversarial training to model deterministic dynamics is proposed by Zhang et al [12]. It enables to update hidden states along a z-order curve. Wang et al. [26] proposed PredRNN as a sequence of recurrent blocks defining an additional global memory cell in order to ameliorate the prediction ability of the network. However, the proposed memory cell transfers long-term and short-term data at the same time which can restrict the predictive performances of the network. Therefore, a pair of memory cell is introduced in [27] and explicitly decoupled to deal with different variations. Also, reverse scheduled sampling strategy was added to learn temporal dynamics and reduce the training discrepancy between the encoding and the prediction structures.
Convolutional networks, considered as feed-forward neural networks, are also commonly used in future prediction problems. A multi-model is defined in [28] to model dynamic patterns and learn image representation by combining temporal and spatial sub-networks. In [29], Deep Voxel Flow (DVF) is trained to synthesize future frames by flowing pixel values directly from input frames. It can predict the in-between frames (interpolation) and the future frames (extrapolation) of the input video. Another interesting convolutional networks for video prediction are 3D convolutions based models to capture temporal consistency [30–33].
Generative networks are used to synthesize new frames by learning a probability distribution from the input data. Generative Adversarial Networks (GAN) [34] are commonly used in video prediction architectures. Y.-H. Kwon et al. [35] proposed a retrospective cycle GAN based algorithm to predict video frames. In [36], it is confirmed that conditional Generative Adversarial Networks (cGAN) can ensure the spatiotemporal coherence between the input videos and the generated frames. Designing a network by dividing the video data into content part and motion part is discussed in [37]. The content part detects the objects in the sequence and the motion part captures their movements. This video prediction framework introduces a new adversarial learning scheme.

3 The proposed Robust Spatiotemporal ConvLSTM architecture

Our algorithm is based on Robust Spatio-Temporal Convolutional Long Short Term Memory (Robust-ST-ConvLSTM) cell that is an extended version of ConvLSTM cell.

3.1 Convolutional Long Short Term Memory (ConvLSTM)

ConvLSTM is considered as a Long Short Term Memory (LSTM) [38] network applied on high dimensional data. In fact, LSTM is a powerful network commonly

used to solve time series problems thanks to its ability to avoid long-term dependency problems and remember information for long periods of time. It main structure enables to connect previous information to the future function. However, LSTM is inadequate to process high dimensional data since it requires 1D vectors as input. Therefore, ConvLSTM was proposed to extract spatial features for the prediction mode. Different from LSTM unit, ConvLSTM cell structure is based on 3D tensors, including the inputs X_t, the cell states C_t and the hidden states H_t.

3.2 The proposed Robust Spatiotemporal ConvLSTM algorithm

Robust Spatiotemporal ConvLSTM (Robust-ST-ConvLSTM) algorithm shows a new internal mechanism that is able to regulate efficiently the flow of spatiotemporal information from video signals based on higher order Convolutional-LSTM. The proposed algorithm decides the cell state C_t from N previous hidden states $(H_{t-2}, ..., H_{t-N})$. N will be fixed by the user depending on the application, the reconstruction quality required and the computational resources available. The proposed Robust-ST-ConvLSTM requires also to implement a memory flow to hold the spatiotemporal information in order to optimize and control the prediction abilities of ConvLSTM. Indeed, the memory flow will be a second cell state to handle spatiotemporal data since the cell state C_t handles temporal data and will not be eliminated. Robust-ST-ConvLSTM uses a stack of ConvLSTM units to learn spatial correlations and temporal dynamics from the input scene. These features will be used to predict the future frames. Thus, a novel transition function is defined based on spatiotemporal memory flow to support previous hidden states.
The process of updating temporal cell states C_t, in ConvLSTM, is activated from one timestep to another. However, successive frames have temporal correlations and very close spatial data distribution. Hence, these properties can be exploited to make better predictions. Therefore, Robust-ST-ConvLSTM, considered as a higher-order ConvLSTM based on memory flow, will exploit the global motion changes of the consecutive frames and the spatiotemporal memory information to forecast future frames. An horizontal diagram flow can represent the memory state updating process for the original stacked ConvLSTM. We suggest here to upgrade the previous model by updating the memory state horizontally (cell state C_t) and also vertically (spatiotemporal memory state STM_t) as shown in Figure 1. This process will enhance the way spatiotemporal information is handled from the input to the output and allow to connect all the recurrent units of the entire network.
From a mathematical perspective, the new robust spatiotemporal unit, illus-

trated in Figure 2, can be defined as:

$$\begin{aligned}
I_t &= \sigma(W_i * X_t + f(H^l_{t-1}, ..., H^l_{t-N})) \\
F_t &= \sigma(W_f * X_t + f(H^l_{t-1}, ..., H^l_{t-N})) \\
\hat{C}_t &= tanh(W_{\hat{c}} * X_t + f(H^l_{t-1}, ..., H^l_{t-N})) \\
C^l_t &= F_t \circ C^l_{t-1} + I_t \circ \hat{C}_t \\
I'_t &= \sigma(W'_i * X_t + M'_i * STM^{l-1}_t) \\
F'_t &= \sigma(W'_f * X_t + M'_f * STM^{l-1}_t) \\
\hat{C}'_t &= tanh(W'_{\hat{c}} * X_t + M'_{\hat{c}} * STM^{l-1}_t) \\
STM^l_t &= F'_t \circ STM^{l-1}_t + I'_t \circ \hat{C}'_t \\
O_t &= \sigma(W_{ox} * X_t + f(H^l_{t-1}, ..., H^l_{t-N}) \\
&\quad + W_{oc} * C^l_t + W_{ostm} * STM^l_t) \\
H^l_t &= O_t * tanh(W_{1\times1} * [C^l_t, STM^l_t])
\end{aligned} \tag{1}$$

Where σ is the sigmoid activation function, $*$ and $\circ$ represent the convolution operator and the Hadamard product respectively. Same as ConvLSTM structure, I_t and I'_t denote the input gates, F_t and F'_t symbolize the forget gates, $\hat{C}_t$ and $\hat{C}'_t$ represent the potential cell states, O_t denotes the output gate. X_t represents the input at the time step t. H^l_t symbolizes the hidden state of the lth layer at the time step t. C^l_t is the memory state of the lth layer at the time step t. STM^l_t represents the spatiotemporal memory of the lth layer at the time step t. f denotes the function combining N previous hidden states.
The design of the function f must satisfy the following conditions:

- Hidden states have a spatial structure that should be preserved
- To capture the context of the previous frames (timesteps), the size of the filters controlling the previous hidden states structure should increase over timesteps.
- Computational complexity does not have to explode

In order to implement f, our approach is inspired from recursive least squares filters used in signal processing [39]. Indeed, the idea is to focus on returning the mean value of all elements in the input tensor that handle the previous hidden states. In Robust-ST-ConvLSTM, combining multiple preceding hidden states generates a feedback signal. Then, the state of the N-order Robust-St-ConvLSTM is recursively updated with the following function f :

$$f(H^l_{t-1}, ..., H^l_{t-N}) = \frac{1}{N} \sum_{n=1}^{N} \alpha^n_n W_{hn} H^l_{t-n} \tag{2}$$

where α denotes the forgetting factor. The parameter α $(0 < \alpha < 1)$ gives more weight to recent hidden states.
Unlike ConvLSTM based architectures, Robust Spatiotemporal unit depends on

the previous hidden states from the previous timesteps at the same layer and the spatiotemporal memory state. Precisely, the first layer in a stacked ConvLSTM model at time step t receives the spatiotemporal memory of the last layer in the stacked model of the previous time step as illustrated in Figure 1 ($STM_t^1 = STM_{t-1}^L$ with L is the number of stacked layers).
Consequently, the main structure of ConvLSTM has been modified by adding a second gated structure for the spatiotemporal memory STM_t^1. Yet, the final hidden state H_t^l depends on the fusion of the spatiotemporal memory state STM_t^l and the temporal memory state C_t^l.
The spatiotemporal memory is implemented to reduce the loss of spatiotemporal information in multidimensional data from the top layer to the bottom layer of the network. Moreover, previous hidden states, used as input, are implemented to enlarge the visibility of the neural units about the context of the ongoing events at different timesteps.
In comparison with standard ConvLSTM model, our proposed approach increases the number of parameters, especially with the addition of a second gated structure. However, it prevent an unnecessarily expenditure of ConvLSTM model (by adding some hyperparameters) to obtain the same performances.

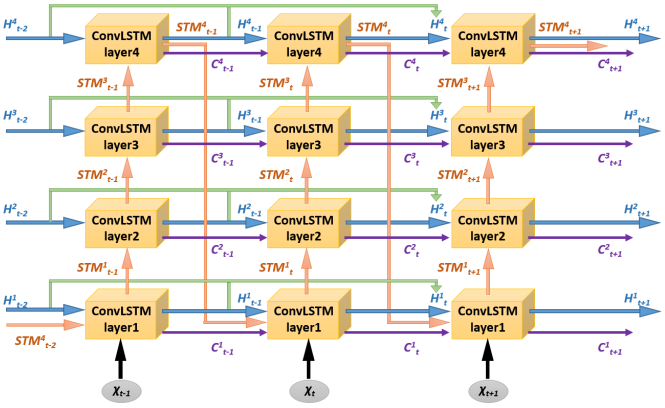

Fig. 1. The main structure of Robust Spatiotemporal LSTM

4 Experiments

4.1 Datasets and performance metrics

Robust-ST-ConvLSTM architecture is tested on two motion video datasets: KTH [40] for human motion and Moving MNIST [41]. To compare its performances with the state-of-art approaches, frames quality evaluation metrics are used. Those metrics are Peak Signal-to-Noise Ratio (PSNR) and Structural Similarity Index (SSIM) [42].

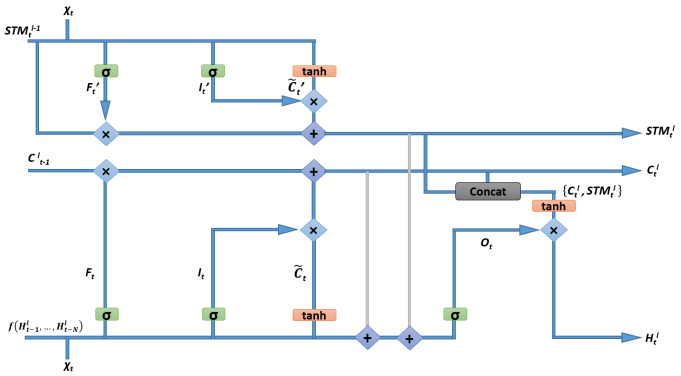

Fig. 2. Robust Spatiotemporal Unit

4.2 Implementation details of Robust-ST-ConvLSTM

Python 3.6 is used to implement the proposed architecture. For its ability to store and process multidimensional data, Pytorch 1.4.0 is used to develop the deep learning framework.
Adam optimizer [43] is used as an optimization algorithm to minimize the loss function with a learning rate of 0.0001. We choose a mini-batch of 2 sequences at each training iteration. We put an end to the training process after 100.000 iterations. As illustrated in Figure 1, our proposed architecture is composed of 4 stacked ConvLSTM layers for each timestep. Three hidden states are used to enhance the prediction process and our model becomes a 3rd-order Robust-ST-ConvLSTM. The dimensions of the hidden state depend on the input frames.
We train our implementation on the RTX 2060 GPU. It takes about a week to train the entire network on KTH dataset and about 4 days on Moving MNIST.

4.3 Comparisons with the State-of-the-Art Methods

Quantitative results of the proposed algorithm and state-of-the-art networks on KTH dataset are illustrated in Table 1. Table 1 summarizes the comparisons with previous methods on PSNR and SSIM. The corresponding frame-wise comparisons are presented in Figure 3 and Figure 4. It can be observed that our proposed Robust-ST-ConvLSTM for video prediction outperforms the others. it increases the average PSNR and SSIM over the same number of predicted frames by 26% and 21.31%, respectively, in comparison with standard ConvLSTM based method. Also, Robust-ST-ConvLSTM performs favorably against PredRNN-v2017 [26] and PredRNN-v2021 [27]. It performs better than PredRNN-v2021 by 1.72% and 2.77% in terms of PSNR and SSIM, respectively. The efficiency of our proposed approach in forecasting future frames in a video is proved by the qualitative results. Figure 5 plots the generated frames of different methods compared with the ground truth. Robust-ST-ConvLSTM provides clearer and

sharper prediction than other approaches. Details are predicted accurately because of the memory flow which enhances the long term prediction ability of the ConvLSTM cell.
Furthermore, the qualitative evaluation of our Robust-St-ConvLSTM and the state-of-the-art algorithms on Moving MNIST dataset by predicting 10 frames based on the features of the previous 10 input frames is illustrated in Table 2. As presented in Table 2, our proposed model outperforms the state-of-the-art approaches in both metrics, thus confirming the previous observations on KTH dataset. Our model increases the average PNSR over the 10 predicted frames by 3.15% by comparing it with PredRNN-v2021. In terms of SSIM, our Robust-ST-ConvLSTM outperforms PredRNN-v2021 by 0.22%. Also, compared with the standard ConvLSTM based model, our proposed algorithm has better PSNR ($\geq$14.59%) and SSIM ($\geq$26.95%) performances.

In this research work, various values of α have been tested randomly ($0 < \alpha < 1$) and the optimum value was selected for the comparison ($\alpha = 0.9$). This means that determining the optimal value of α could be an interesting research direction. Previous observations about the value of α and the number of hidden states confirm that a trade-off should be done between quality performances and computational cost, in future research work, to enhance the quality performances of the predicted images without training a computationally expensive algorithm.

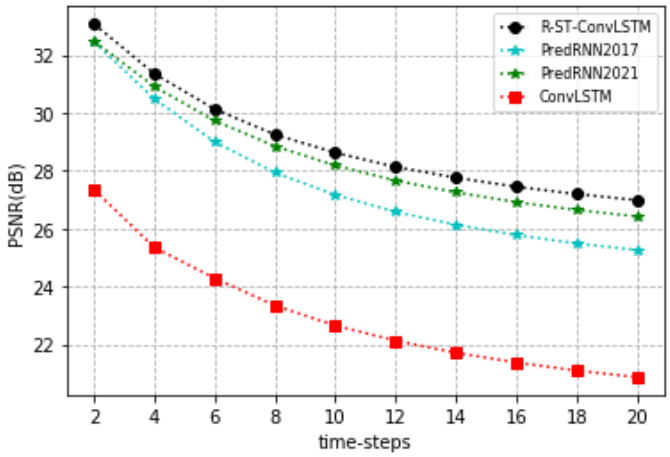

Fig. 3. Frame-wise PSNR comparisons of different models on KTH dataset after 100 000 iterations

5 Conclusion

In this paper, we present a new recurrent neural network model for predicting future video frames named "Robust Spatiotemporal ConvLSTM" (Robust-ST-

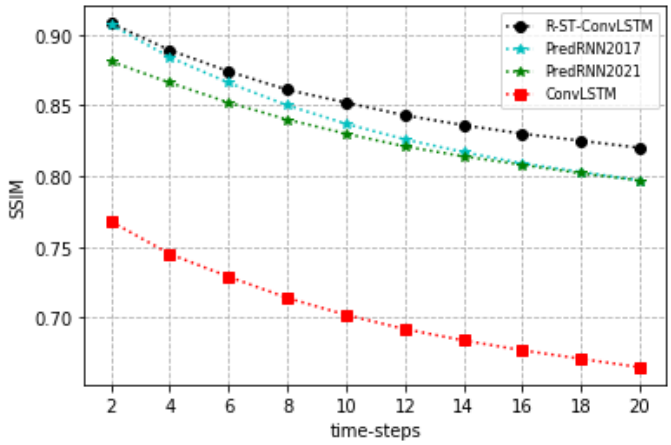

Fig. 4. Frame-wise SSIM comparisons of different models on KTH dataset after 100 000 iterations

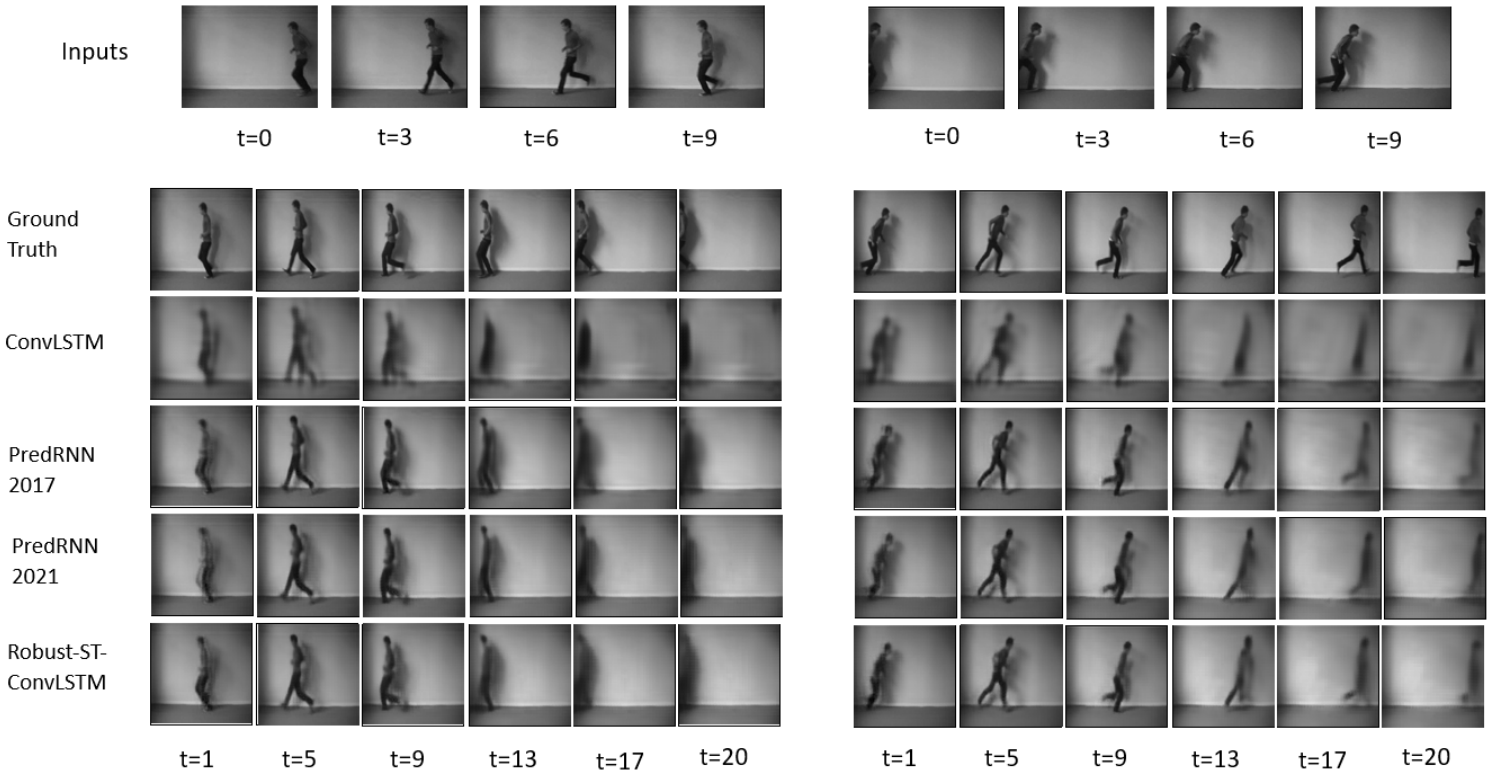

Fig. 5. Prediction examples on the KTH data set, where we predict 20 frames into the future based on the past 10 frames

Table 1. Quantitative evaluation of different algorithms on the KTH dataset. The metrics are averaged over the 20 predicted frames.

Model	PSNR(dB)	SSIM
ConvLSTM (Shi et al., 2015)	23.009	0.704
PredRNN (Wang et al., 2017)	27.624	0.839
PredRNN (Wang et al., 2021)	28.502	0.831
Robust-ST-ConvLSTM	**28.992**	**0.854**

Table 2. Quantitative evaluation of different algorithms on the MNIST dataset. The metrics are averaged over the 10 predicted frames.

Model	PSNR(dB)	SSIM
ConvLSTM (Shi et al., 2015)	28.380	0.705
PredRNN (Wang et al., 2017)	30.569	0.869
PredRNN (Wang et al., 2021)	31.525	0.893
Robust-ST-ConvLSTM	**32.520**	**0.895**

ConvLSTM). It is based on a new robust spatiotemporal unit, an extension architecture of ConvLSTM. Our approach learns extra information from the memory flow that handle the spatiotemporal information to significantly improve the long-term frame prediction. We further improve the temporal context for the motion in videos by opting for a higher order ConvLSTM approach to enable cell states update from previous hidden states. Qualitative and quantitative results demonstrate the superiority of our algorithm dealing with video prediction, showing state-of-the-art performance in KTH and Moving MNIST datasets. This architecture inspires us to further explore recurrent structures to optimize the computational cost of the algorithm and generate accurate predictions in future research work.

Acknowledgments

This work was supported in part by the sensors generation project of Nouvelle Aquitaine region (2018-1R50214).

References

1. K. M. Kitani, B. D. Ziebart, J. A. Bagnell, and M. Hebert, “Activity forecasting,” in ECCV, 2012.
2. C. Vondrick, H. Pirsiavash, and A. Torralba, “Anticipating visual representations from unlabeled video,” in CVPR, 2016.
3. K. Zeng, W. B. Shen, D. Huang, M. Sun, and J. C. Niebles, “Visual forecasting by imitating dynamics in natural sequences,” in ICCV, 2017.
4. A. Bhattacharyya, M. Fritz, and B. Schiele, “Long-term on-board prediction of people in traffic scenes under uncertainty,” in CVPR, 2018.

5. A. Hu, F. Cotter, N. Mohan, C. Gurau, and A. Kendall, "Probabilistic future prediction for video scene understanding," arXiv:2003.06409, 2020.
6. Y. LeCun, Y. Bengio, and G. E. Hinton, "Deep learning," Nature, vol. 521, no. 7553, 2015.
7. S. Xingjian et al., "Convolutional LSTM network: A machine learning approach for precipitation nowcasting," in Proc. Adv. Neural Inf. Process. Syst., 2015, pp. 802–810.
8. R. Soltani and H. Jiang, "Higher order recurrent neural networks," arXiv preprint arXiv:1605.00064, 2016
9. W. Lotter, G. Kreiman, and D. D. Cox, "Unsupervised learning of visual structure using predictive generative networks," arXiv:1511.06380, 2015.
10. N. Wichers, R. Villegas, D. Erhan, and H. Lee, "Hierarchical long-term video prediction without supervision," in ICML, ser. Proceedings of Machine Learning Research, vol. 80, 2018.
11. R. Villegas, J. Yang, Y. Zou, S. Sohn, X. Lin, and H. Lee, "Learning to generate long-term future via hierarchical prediction," in ICML, 2017.
12. J. Zhang, Y. Wang, M. Long, W. Jianmin, and P. S. Yu, "Z-order recurrent neural networks for video prediction," in ICME, July 2019.
13. M. Ranzato, A. Szlam, J. Bruna, M. Mathieu, R. Collobert, and S. Chopra, "Video (language) modeling: a baseline for generative models of natural videos," arXiv:1412.6604, 2014.
14. N. Srivastava, E. Mansimov, and R. Salakhutdinov, "Unsupervised Learning of Video Representations using LSTMs," in ICML, 2015.
15. W. Lotter, G. Kreiman, and D. Cox, "Deep Predictive Coding Networks for Video Prediction and Unsupervised Learning," in ICLR (Poster), 2017.
16. W. Byeon, Q. Wang, R. K. Srivastava, and P. Koumoutsakos, "Contextvp: Fully context-aware video prediction," in CVPR (Workshops), 2018.
17. V. Patraucean, A. Handa, and R. Cipolla, "Spatio-temporal video autoencoder with differentiable memory," (ICLR) Workshop, 2015.
18. C. Lu, M. Hirsch, and B. Scholkopf, "Flexible Spatio-Temporal ¨ Networks for Video Prediction," in CVPR, 2017.
19. E. L. Denton and V. Birodkar, "Unsupervised learning of disentangled representations from video," in NeurIPS, 2017.
20. J. Oh, X. Guo, H. Lee, R. L. Lewis, and S. P. Singh, "ActionConditional video prediction using deep networks in atari games," in NeurIPS, 2015.
21. E. Denton and R. Fergus, "Stochastic video generation with a learned prior," in ICML, ser. Proceedings of Machine Learning Research, J. G. Dy and A. Krause, Eds., vol. 80, 2018.
22. S. shahabeddin Nabavi, M. Rochan, and Y. Wang, "Future semantic segmentation with convolutional LSTM," in BMVC, 2018.
23. S. Vora, R. Mahjourian, S. Pirk, and A. Angelova, "Future segmentation using 3d structure," arXiv:1811.11358, 2018.
24. A. Terwilliger, G. Brazil, and X. Liu, "Recurrent flow-guided semantic forecasting," in WACV, 2019.
25. X. Shi, Z. Chen, H. Wang, D. Yeung, W. Wong, and W. Woo, "Convolutional LSTM network: A machine learning approach for precipitation nowcasting," in NeurIPS, 2015.
26. Y. Wang, M. Long, J. Wang, Z. Gao, and S. Y. Philip, "PredRNN: Recurrent neural networks for predictive learning using spatiotemporal lstms," in NeurIPS, 2017, pp. 879–888

27. Yunbo Wang, Haixu Wu, Jianjin Zhang, Zhifeng Gao, Jianmin Wang, Philip S. Yu, Mingsheng Long, "PredRNN: a recurrent neural network for spatiotemporal predictive learning". arXiv:2103.09504, 2021
28. J. Yan, G. Qin, R. Zhao, Y. Liang and Q. Xu, "Mixpred: video prediction beyond optical flow," in IEEE Access, vol. 7, pp. 185654-185665, 2019, doi: 10.1109/ACCESS.2019.2961383.
29. Z. Liu, R. A. Yeh, X. Tang, and Y. Liu, "Video frame synthesis using deep voxel Flow," in Proc. IEEE Int. Conf. Comput. Vis. (CVPR), Oct. 2017, pp. 44634471.
30. Y. Wang, L. Jiang, M.-H. Yang, L.-J. Li, M. Long, and L. Fei-Fei, "Eidetic 3d LSTM: a model for video prediction and beyond," in ICLR, 2019.
31. C. Vondrick, H. Pirsiavash, and A. Torralba, "Generating videos with scene dynamics," in NeurIPS, 2016.
32. S. Tulyakov, M.-Y. Liu, X. Yang, and J. Kautz, "MoCoGAN: de- composing motion and content for video generation," in CVPR, June 2018.
33. S. Aigner and M. K¨orner, "Futuregan: Anticipating the future frames of video sequences using spatio-temporal 3d convolutions in progressively growing autoencoder gans," arXiv:1810.01325, 2018.
34. I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in Proc. Advances Neural Information Processing Systems Conf., 2014, pp. 2672–2680.
35. Y.-H. Kwon and M.-G. Park, "Predicting future frames using retrospective cycle gan," in CVPR, 2019.
36. Sergiu Oprea, Pablo Martinez-Gonzalez, Alberto Garcia-Garcia, John Alejandro Castro-Vargas, Sergio Orts-Escolano, Jose Garcia-Rodriguez, and Antonis Argyros. A review on deep learning techniques for video prediction. arXiv preprint arXiv:2004.05214, 2020
37. S. Tulyakov, M.-Y. Liu, X. Yang, and J. Kautz, "MoCoGAN: Decomposing motion and content for video generation," in CVPR, June 2018
38. S. Hochreiter and J. Schmidhuber, "Long short-term memory," Neural Computation, vol. 9, no. 8, pp. 1735– 1780, 1997.
39. J. Cances and V. Meghdadi, "Joint channel estimation and data demodulation algorithms for fast time varying band limited frequency selective Rayleigh fading channels: A comparison study," Annales Des Telecommun., vol. 55, no. 56, pp. 226–237, May/Jun. 2000.
40. Christian Schuldt, Ivan Laptev, and Barbara Caputo, "Recognizing human actions: a local svm approach," in ICPR, 2004, pp. 32–36 Vol.3.
41. N. Srivastava, E. Mansimov, and R. Salakhutdinov, "Unsupervised Learning of Video Representations using LSTMs," in ICML, 2015.
42. A. Horé and D. Ziou, "Image quality metrics: PSNR vs. SSIM," 2010 20th International Conference on Pattern Recognition, Istanbul, 2010, pp. 2366-2369, doi: 10.1109/ICPR.2010.579.
43. D. Kingma and J. Ba. Adam: a method for stochastic optimization. In ICLR, 2015