Convolutional LSTM Network: A Machine Learning Approach for Precipitation Nowcasting
卷积LSTM网络:利用机器学习预测短期降雨

施行健

香港科技大学

VALSE 2016/03/23

Content

- Quick Review of Recurrent Neural Network
- Introduction to Precipitation Nowcasting (短期降雨预报)
  - Goal of Precipitation Nowcasting
  - Classic Approaches
- Convolutional LSTM
  - Motivation
  - Formulation of the precipitation nowcasting problem
  - Model
  - Experiments

Content

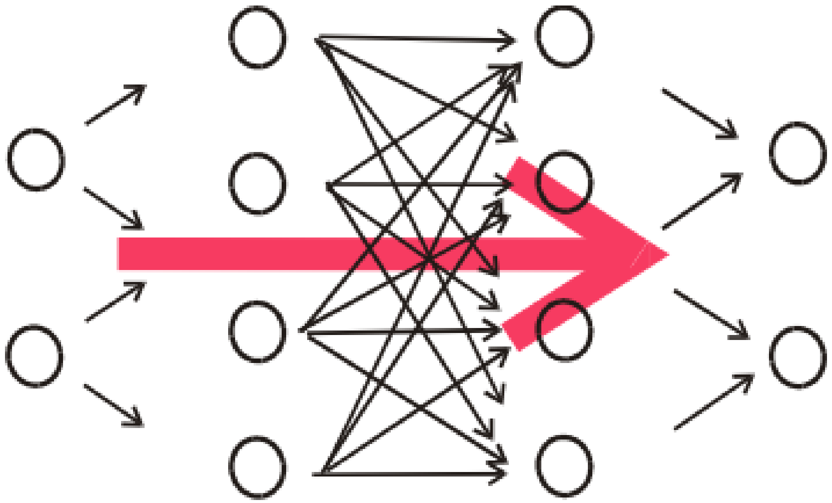
- Quick Review of Recurrent Neural Network
- Introduction to Precipitation Nowcasting (短期降雨预报)
  - Goal of Precipitation Nowcasting
  - Classic Approaches
- Convolutional LSTM
  - Motivation
  - Formulation of the precipitation nowcasting problem
  - Model
  - Experiments

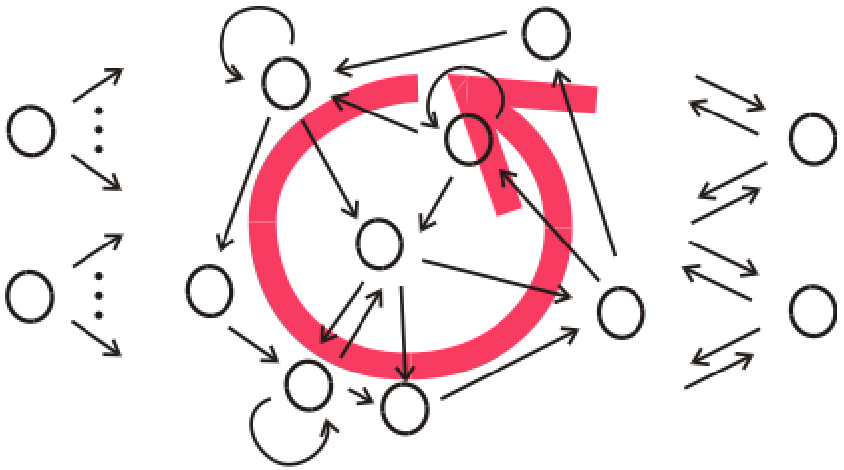
From FNN to RNN

Structural Generalization

Feedforward Neural Network is acyclic. There is no loop

Recurrent Neural Network can be arbitrary. Cycles are allowed in the network

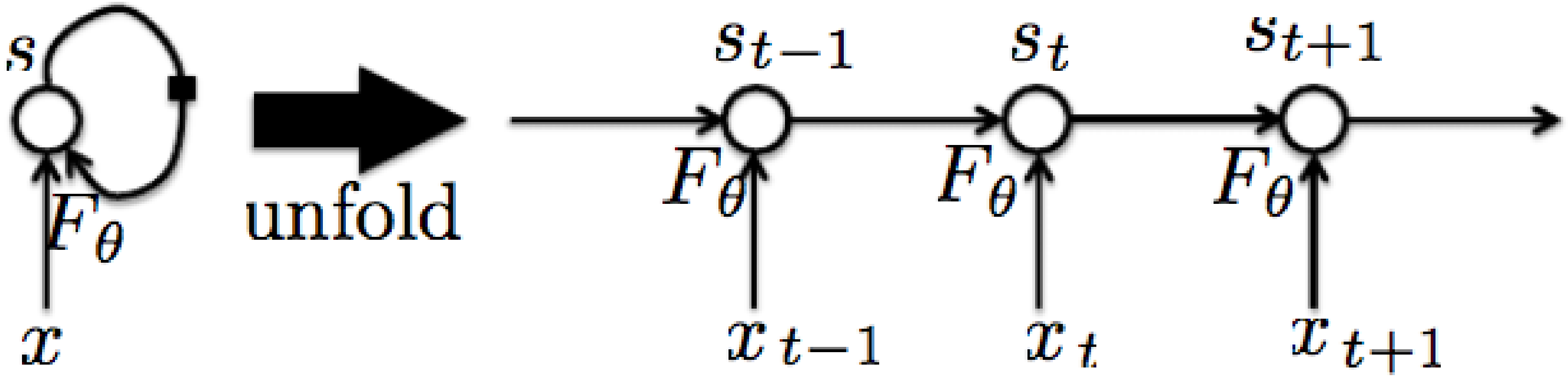




From FNN to RNN

- After unfolding the structure, recurrent neural network can be viewed as a type of feedforward neural network with shared transitional weights
- Example

$$\boldsymbol{s}_t = F_\theta(\boldsymbol{s}_{t-1}, \boldsymbol{x}_t) \qquad \mathbf{x}_t = \sigma(\mathbf{W}_{rec}\mathbf{x}_{t-1} + \mathbf{W}_{in}\mathbf{u}_t + \mathbf{b})$$



- Back propagation through time (BPTT)
  - Unfold the RNN to FNN
  - Use backpropagation, can use SGD and any of its variants

[Goodfellow et.al, 2016] Deep Learning (http://www.deeplearningbook.org/)

Vanishing Gradient & Exploding Gradient

- Since the network is so deep, long term information in the gradient will contain a product of a large number of Jacobian.

This determinant will go to infinity or zero.

$$\mathbf{x}_t = \mathbf{W}_{rec}\sigma(\mathbf{x}_{t-1}) + \mathbf{W}_{in}\mathbf{u}_t + \mathbf{b}$$

$$\mathcal{E}_t = \mathcal{L}(\mathbf{x}_t)$$

$$\mathcal{E} = \textstyle\sum_{1\leq t\leq T}\mathcal{E}_t$$

$$\frac{\partial \mathcal{E}}{\partial \theta} = \sum_{1\leq t\leq T} \frac{\partial \mathcal{E}_t}{\partial \theta}$$

$$\frac{\partial \mathcal{E}_t}{\partial \theta} = \sum_{1\leq k\leq t} \left(\frac{\partial \mathcal{E}_t}{\partial \mathbf{x}_t} \frac{\partial \mathbf{x}_t}{\partial \mathbf{x}_k} \frac{\partial^+ \mathbf{x}_k}{\partial \theta} \right)$$

$$\frac{\partial \mathbf{x}_t}{\partial \mathbf{x}_k} = \prod_{t\geq i>k} \frac{\partial \mathbf{x}_i}{\partial \mathbf{x}_{i-1}} = \prod_{t\geq i>k} \mathbf{W}_{rec}^T diag(\sigma'(\mathbf{x}_{i-1}))$$

[Pascanul et.al, ICML2013] On the difficulty of training recurrent neural networks

Constant error carousel to avoid vanishing gradient

- Long term part of the gradient will contain a sum of product of Jacobians. It will not vanish if one of them does not vanish.

$$s_{t+1,i} = (1 - \frac{1}{\tau_i})s_{t,i} + \frac{1}{\tau_i}\sigma(b_i + W s_t + U x_t)$$

Make det(Jacobian) → 1

- Long-Short Term Memory

$$i_t = \sigma(W_{xi}x_t + W_{hi}h_{t-1} + W_{ci} \circ c_{t-1} + b_i)$$
$$f_t = \sigma(W_{xf}x_t + W_{hf}h_{t-1} + W_{cf} \circ c_{t-1} + b_f)$$
$$c_t = f_t \circ c_{t-1} + i_t \circ \tanh(W_{xc}x_t + W_{hc}h_{t-1} + b_c)$$
$$o_t = \sigma(W_{xo}x_t + W_{ho}h_{t-1} + W_{co} \circ c_t + b_o)$$
$$h_t = o_t \circ \tanh(c_t)$$

Cell is the constant error carousel

Long-term information will be considered if we initialize the bias of forget gate to a large value

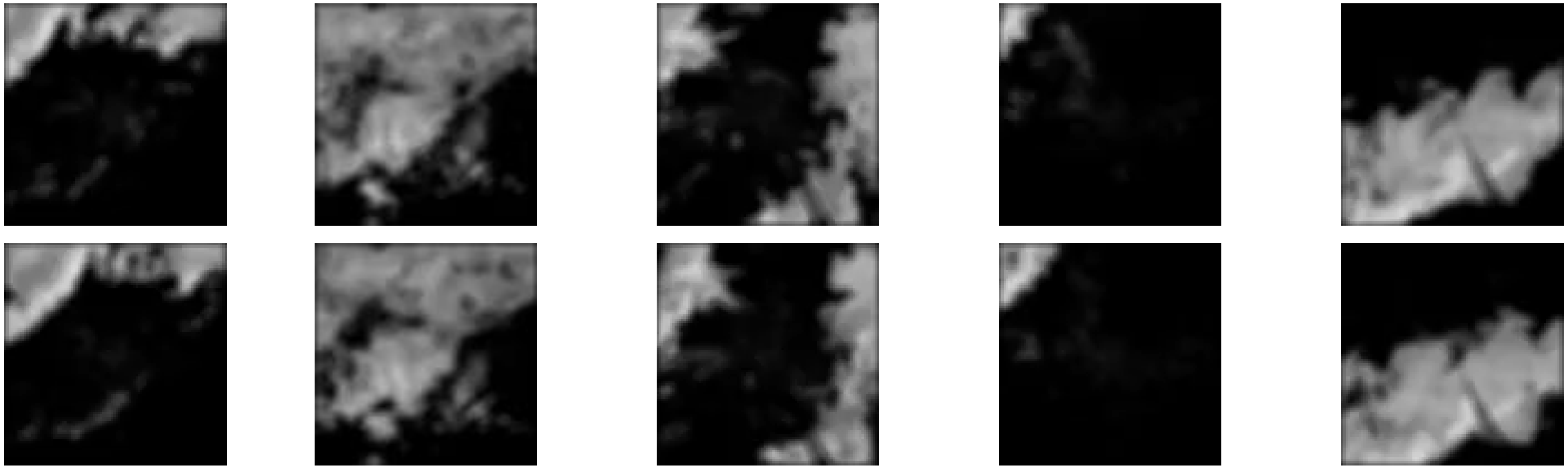
[Jozefowicz et.al, ICML2015] An Empirical Exploration of Recurrent Network Architectures

Content

- Quick Review of Recurrent Neural Network
- Introduction to Precipitation Nowcasting (短期降雨预报)
  - Goal of Precipitation Nowcasting
  - Classic Approaches
- Convolutional LSTM
  - Motivation
  - Formulation of the precipitation nowcasting problem
  - Model
  - Experiments

Goal of Precipitation Nowcasting

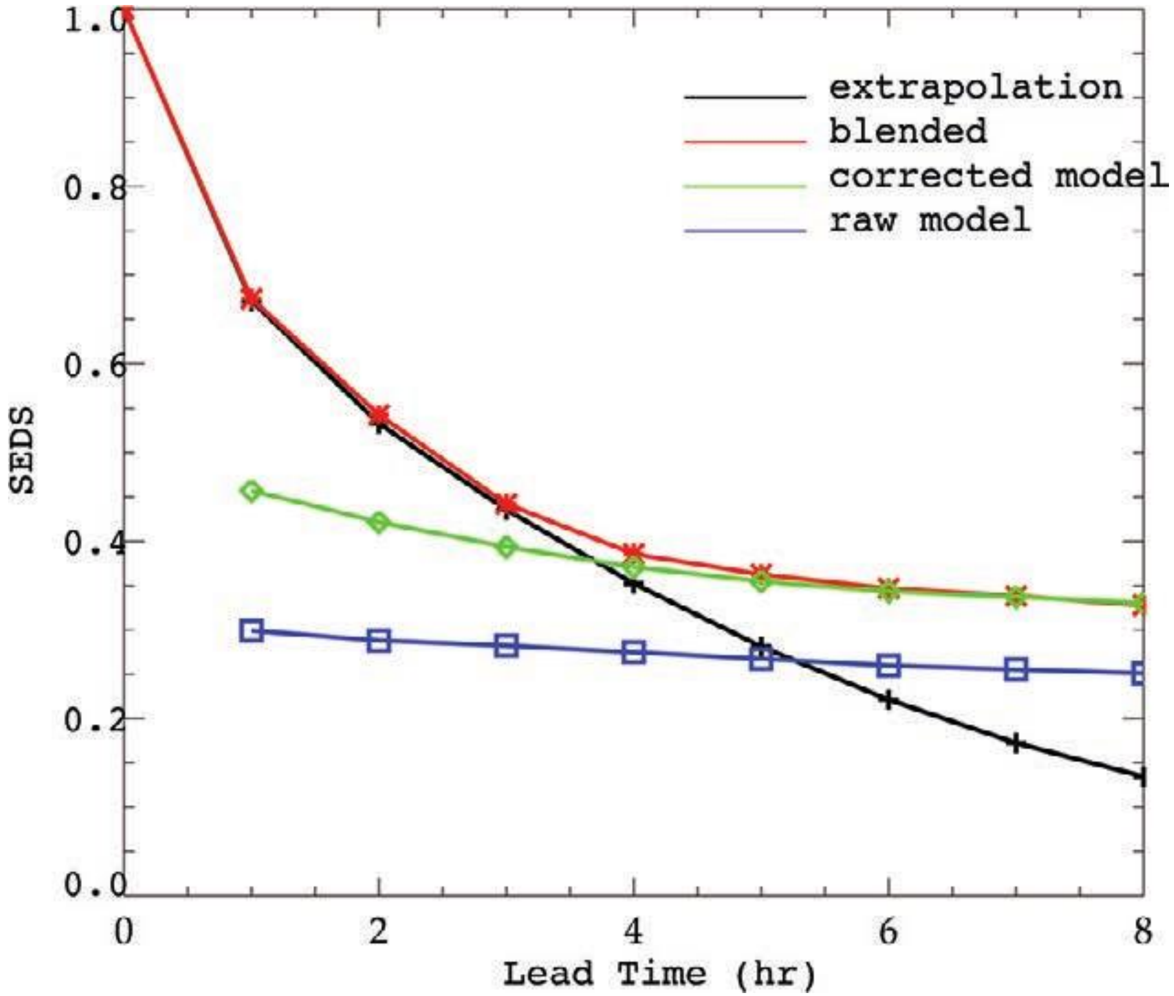
- Give precise and timely prediction of rainfall intensity in a local region over a relatively short period of time (e.g., 0-6 hours)
  - High resolution & Accurate timing
  - High dimensional spatiotemporal data



Classic Approaches

- Numerical Weather Prediction (NWP) based Methods
  - Build a model with several physical equations. Simulation.
  - More accurate in the longer term
  - The first 1-2 hours of model forecasts may not be available
- Extrapolation based Methods
  - Optical flow estimation + Extrapolation (Semi-Lagrangian Extrapolation)
  - More accurate in the first 1-2 hours
  - [27th Conference of Severe Local Storm] ROVER by HKO
- Hybrid Method
  - For the first several hours of now-casting, we use extrapolation based methods, while using NWP for longer term prediction

Classic Approaches



Black: Extrapolation
Red: Hybrid
Green: Corrected NWP
Blue: NWP

[Bulletin of American Meteorological Society 2014] Use of NWP for Nowcasting Convective Precipitation: Recent Progress and Challenges

Content

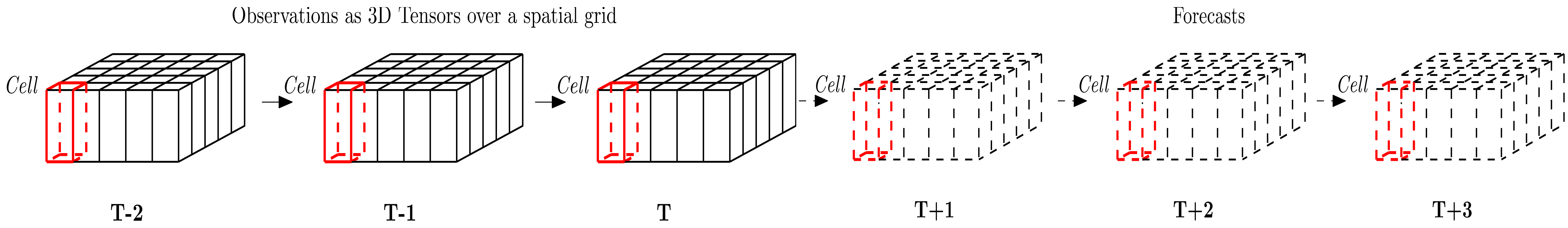
- Quick Review of Recurrent Neural Network
- Introduction to Precipitation Nowcasting (短期降雨预报)
  - Goal of Precipitation Nowcasting
  - Classic Approaches
- Convolutional LSTM
  - Motivation
  - Formulation of the precipitation nowcasting problem
  - Model
  - Experiments

Motivation

- The limitation of optical flow based methods
  - Flow estimation step and Radar echo extrapolation step are separated, accumulative error
  - Hard to estimate the parameters
- A machine learning based, end-to-end approach for this problem
- Machine learning based approach is not trivial
  - Multi-step prediction (size of the search space grows exponentially)
  - Spatiotemporal data (take advantage of the spatiotemporal correlation within the data)

Formulation of the precipitation nowcasting problem

- Periodic observations taken from a dynamic system over a spatial MXN grid → sequence of tensors



- Predict the most likely length-K sequence in the future given the previous J observations

$$\tilde{\mathcal{X}}_{t+1}, \ldots, \tilde{\mathcal{X}}_{t+K} = \underset{\mathcal{X}_{t+1}, \ldots, \mathcal{X}_{t+K}}{\arg\max}\; p(\mathcal{X}_{t+1}, \ldots, \mathcal{X}_{t+K} \mid \hat{\mathcal{X}}_{t-J+1}, \hat{\mathcal{X}}_{t-J+2}, \ldots, \hat{\mathcal{X}}_t)$$

How to perform multi-step prediction?

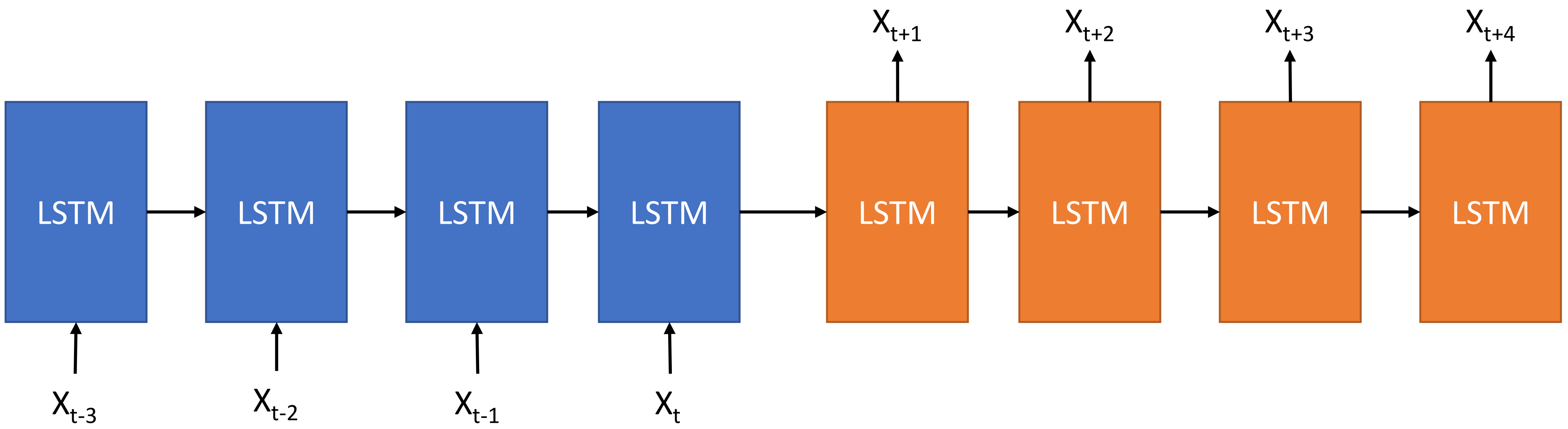
- Encoding-Forecasting Structure

$$
\begin{aligned}
\tilde{\mathcal{X}}_{t+1}, \ldots, \tilde{\mathcal{X}}_{t+K} &= \underset{\mathcal{X}_{t+1}, \ldots, \mathcal{X}_{t+K}}{\arg\max}\; p(\mathcal{X}_{t+1}, \ldots, \mathcal{X}_{t+K} \mid \hat{\mathcal{X}}_{t-J+1}, \hat{\mathcal{X}}_{t-J+2}, \ldots, \hat{\mathcal{X}}_t) \\
&\approx \underset{\mathcal{X}_{t+1}, \ldots, \mathcal{X}_{t+K}}{\arg\max}\; p(\mathcal{X}_{t+1}, \ldots, \mathcal{X}_{t+K} \mid f_{encoding}(\hat{\mathcal{X}}_{t-J+1}, \hat{\mathcal{X}}_{t-J+2}, \ldots, \hat{\mathcal{X}}_t)) \\
&\approx g_{forecasting}(f_{encoding}(\hat{\mathcal{X}}_{t-J+1}, \hat{\mathcal{X}}_{t-J+2}, \ldots, \hat{\mathcal{X}}_t))
\end{aligned}
$$

  - Using Recurrent Neural Network to encoding and forecasting
  - [NIPS2014] Sequence to sequence learning with neural networks
  - [ICML2015] Unsupervised learning of video representations using LSTMs

How to deal with spatiotemporal data?

- A pure Encoding-Forecasting structure is not enough
- We are dealing with spatiotemporal data!



NOT ENOUGH!!!

How to deal with spatiotemporal data?

- We need to design a specific network structure for spatiotemporal data
- What's the characteristics of the spatiotemporal data we are dealing with?
  - Strong correlation between local neighbors, i.e, neighbors tend to act similarly
- Fully-connected LSTM (FC-LSTM) → Convolutional LSTM (ConvLSTM)
  - Regularize the network by specifying the structure
  - Use convolution instead of fully-connection in state-to-state transition!

Comparison between FC-LSTM & ConvLSTM

FC-LSTM

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

Input & state at a timestamp are 1D vectors. Dimensions of the state can be permuted without affecting the overall structure.

ConvLSTM

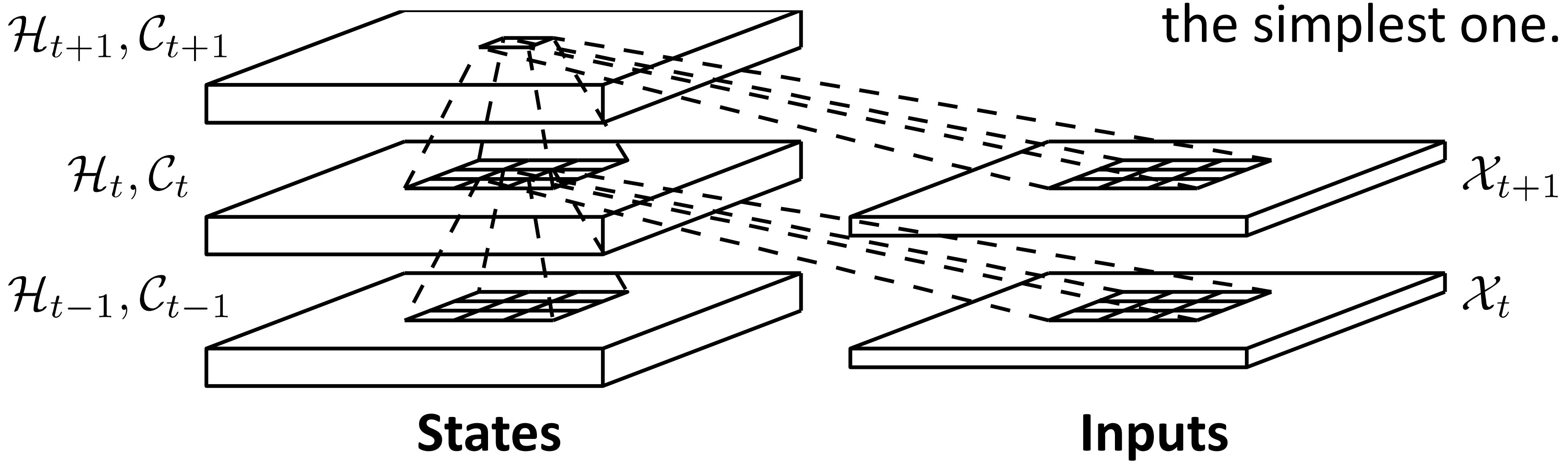
$$
\begin{aligned}
i_t &= \sigma(W_{xi} * \mathcal{X}_t + W_{hi} * \mathcal{H}_{t-1} + W_{ci} \circ \mathcal{C}_{t-1} + b_i) \\
f_t &= \sigma(W_{xf} * \mathcal{X}_t + W_{hf} * \mathcal{H}_{t-1} + W_{cf} \circ \mathcal{C}_{t-1} + b_f) \\
\mathcal{C}_t &= f_t \circ \mathcal{C}_{t-1} + i_t \circ \tanh(W_{xc} * \mathcal{X}_t + W_{hc} * \mathcal{H}_{t-1} + b_c) \\
o_t &= \sigma(W_{xo} * \mathcal{X}_t + W_{ho} * \mathcal{H}_{t-1} + W_{co} \circ \mathcal{C}_t + b_o) \\
\mathcal{H}_t &= o_t \circ \tanh(\mathcal{C}_t)
\end{aligned}
$$

Input & state at a timestamp are 3D tensors. Convolution is used for both input-to-state and state-to-state connection.

Use Hadamard product to keep the constant error carousel (CEC) property of cells

Convolutional LSTM

Using ‘state of the outside world’ for boundary grids. Zero padding is used to indicate ‘total ignorance’ of the outside. In fact, other padding strategies (learn the padding) can be used, we just choose the simplest one.

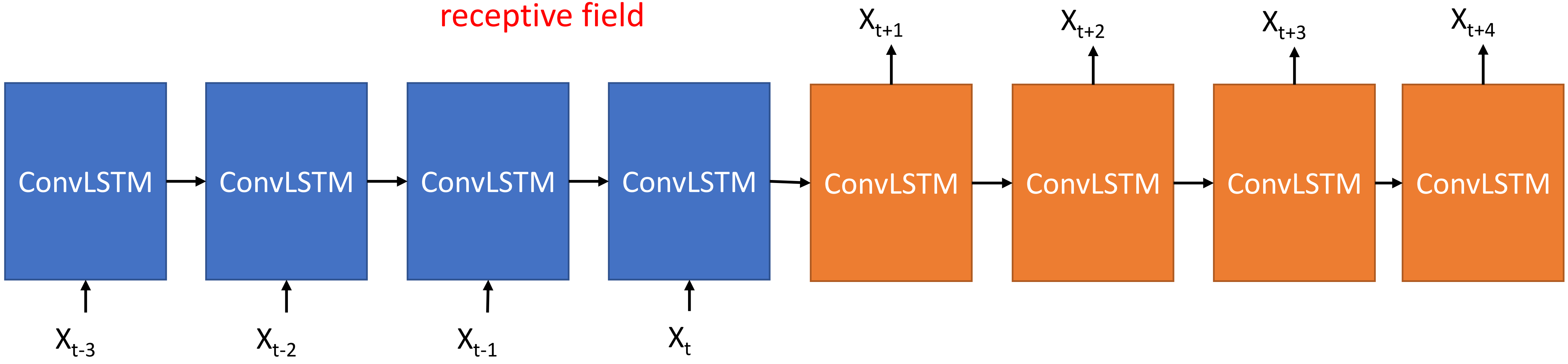


FC-LSTM can be viewed as a special case of ConvLSTM with all features standing on a single cell.

For convolutional recurrence, 1X1 kernel and larger kernels are totally different!
Later states → Larger receptive field

Convolutional LSTM

With the help of convolutional recurrence, the final state has large receptive field



Final Structure

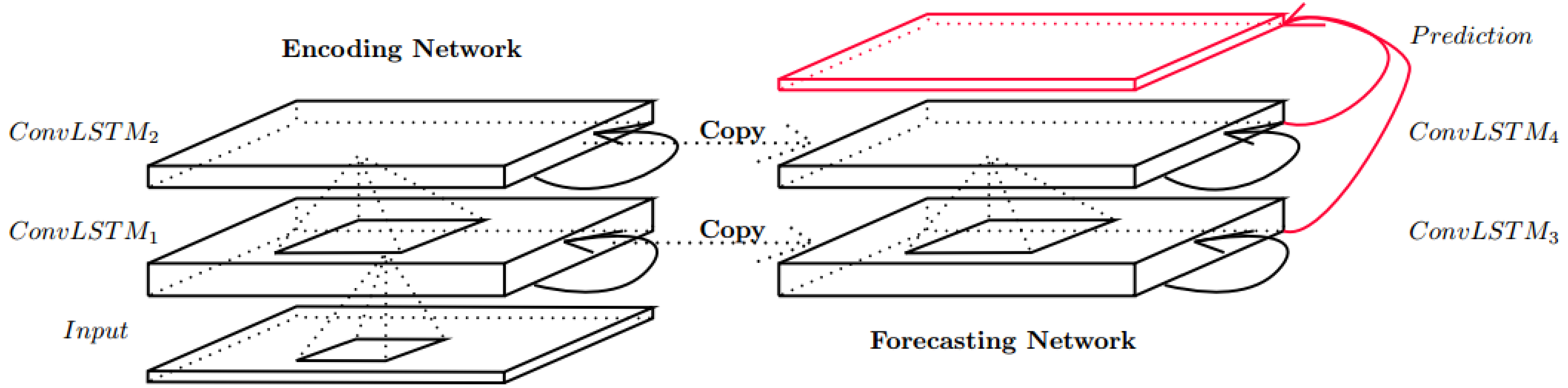


Figure 3: Encoding-forecasting ConvLSTM network for precipitation nowcasting

Cross Entropy Loss + BPTT + RMSProp + Early-stopping

Experiments

- Experiments on a synthetic Moving-MNIST dataset
  - Gain some basic understanding of the model
  - Test the effectiveness of ConvLSTM on synthetic data.
- Experiments on the real-life HKO Radar Echo dataset
  - Test if the proposed approach is effective for our precipitation nowcasting problem.

Moving-MNIST

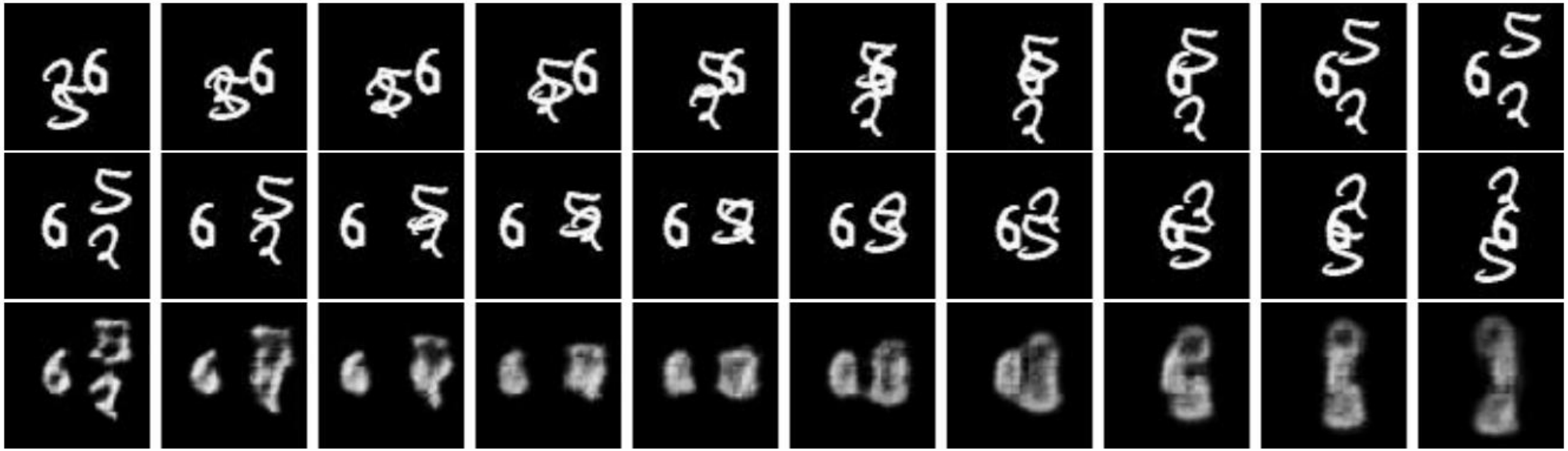
- 2 characters bouncing inside a 64X64 box
- 10000 training sequences + 2000 validation sequence + 3000 testing sequences, 10 frames for input & 10 frames to predict
- Different parameters of the model.

Model	Number of parameters	Cross entropy
FC-LSTM-2048-2048	142,667,776	4832.49
ConvLSTM(5x5)-5x5-256	13,524,496	3887.94
ConvLSTM(5x5)-5x5-128-5x5-128	10,042,896	3733.56
ConvLSTM(5x5)-5x5-128-5x5-64-5x5-64	**7,585,296**	**3670.85**
ConvLSTM(9x9)-1x1-128-1x1-128	11,550,224	4782.84
ConvLSTM(9x9)-1x1-128-1x1-64-1x1-64	8,830,480	4231.50

Convolutional state-to-state transition is important! Kernel Size >1 is important!

Moving-MNIST

- Out-of-domain test
  - How the model performs for out-of-domain samples? Generate dataset with 3 characters.

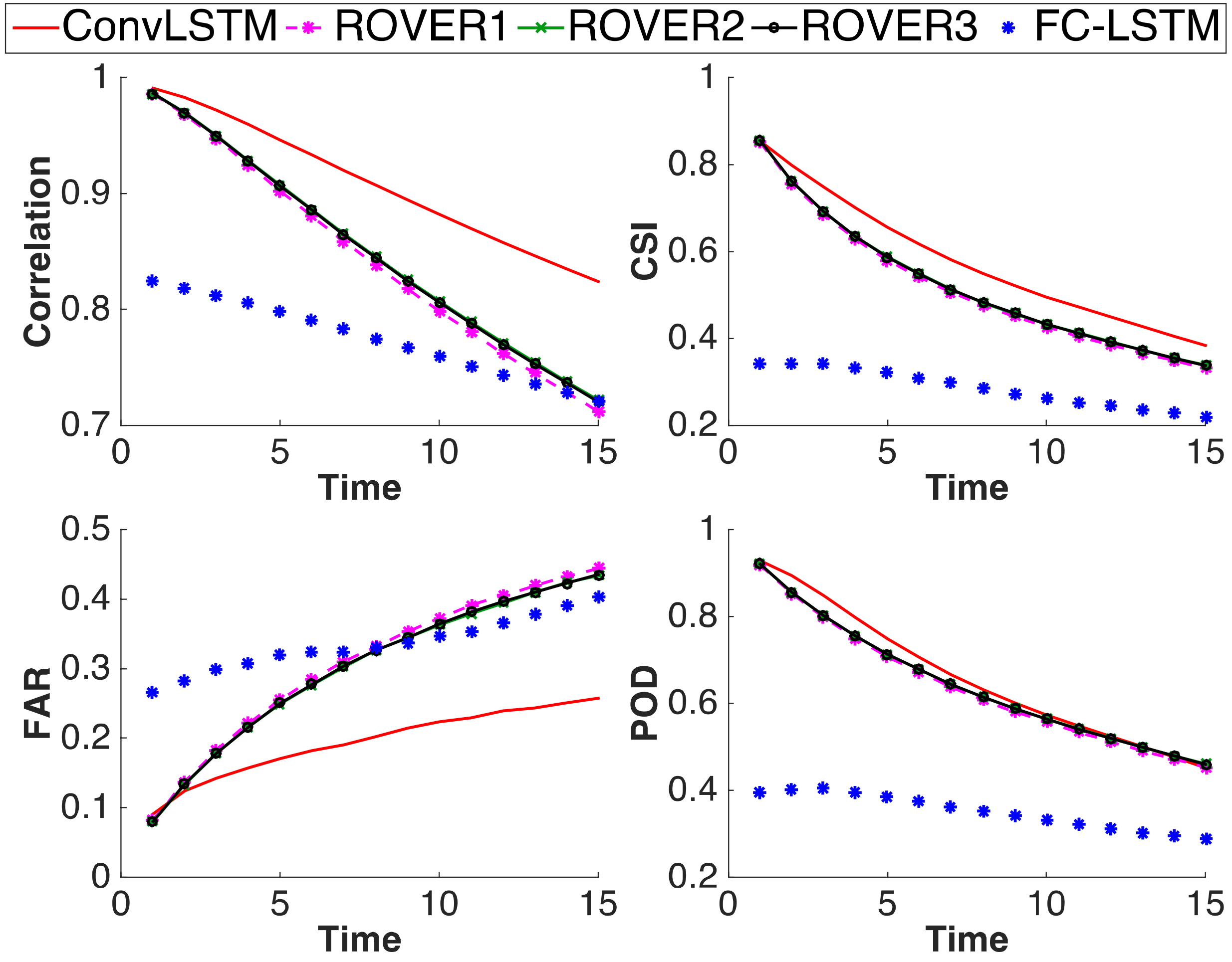


HKO Radar Echo

- Slice several separated training & testing sequences. 5 frames for input & 15 frames to predict. The size of each frame is 100X100.

Model	Rainfall-MSE	CSI	FAR	POD	Correlation
ConvLSTM(3x3)-3x3-64-3x3-64	**1.420**	**0.577**	**0.195**	**0.660**	**0.908**
Rover1	1.712	0.516	0.308	0.636	0.843
Rover2	1.684	0.522	0.301	0.642	0.850
Rover3	1.685	0.522	0.301	0.642	0.849
FC-LSTM-2000-2000	1.865	0.286	0.335	0.351	0.774

HKO Radar Echo



Discussion

- ConvLSTM for other spatiotemporal problems like human action recognition and object tracking
  - [Ballas, ICLR2016] Delving deeper into convolutional networks for learning video representations
  - [Ondru´ska, AAAI2016] Deep Tracking: Seeing Beyond Seeing Using Recurrent Neural Networks
- Imposing structure in recurrent connection.