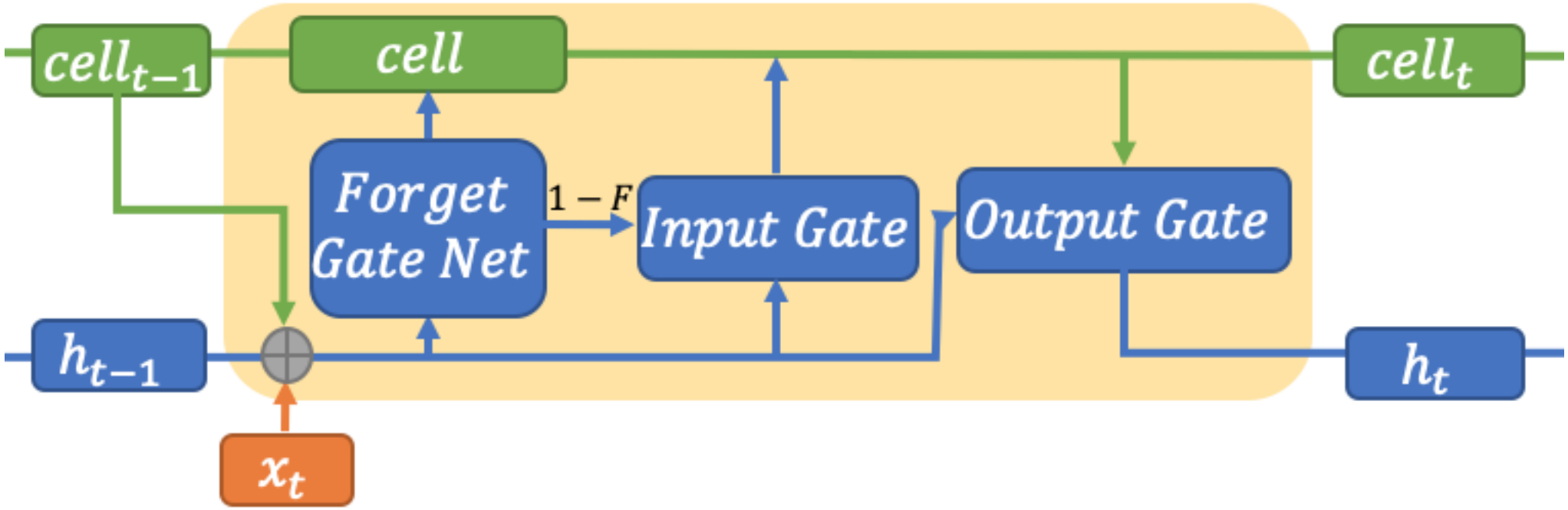
MA-LSTM: A Multi-attention Based LSTM for Complex Pattern Extraction

Jingjie Guo
2Sun Yat-sen University
1Shenzhen Institutes of Advanced Technology,
Chinese Academy of Sciences.
Email: jj.guo@siat.ac.cn

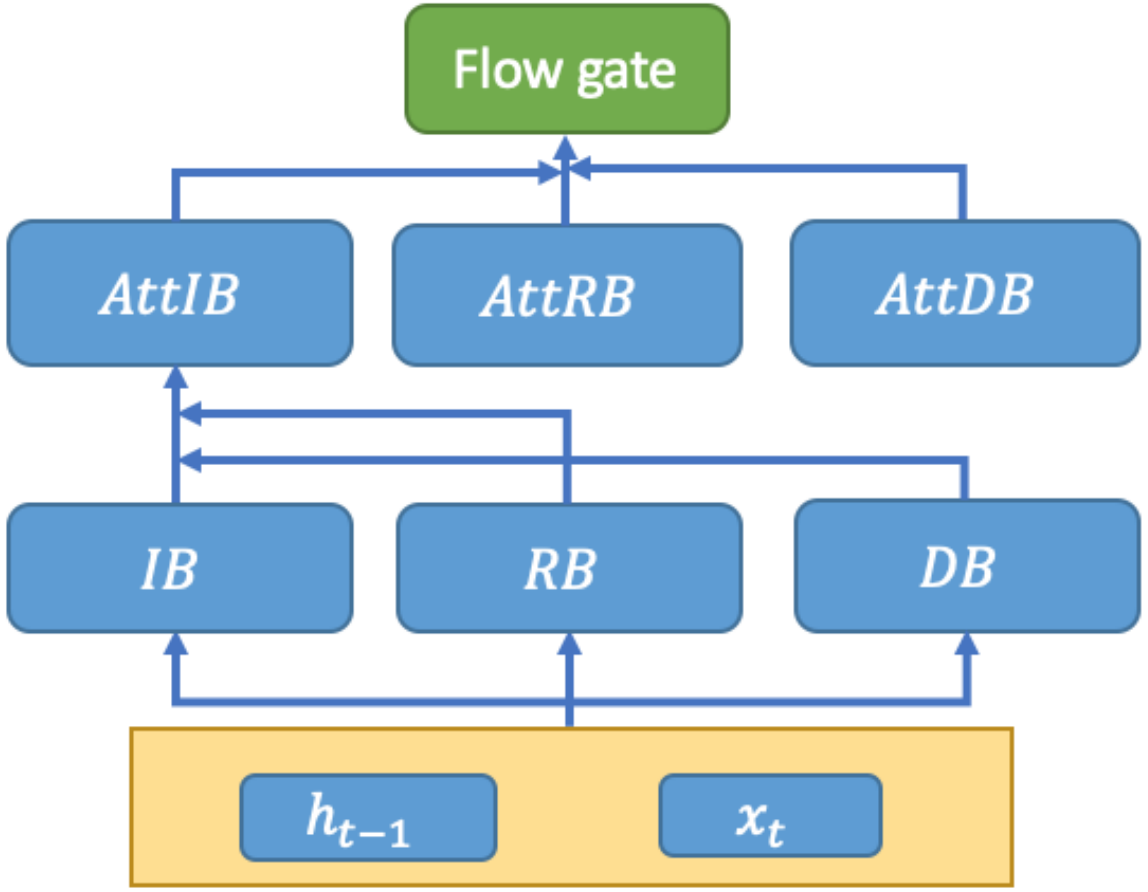
Abstract

Recurrent neural network (RNN) is a class of artificial neural networks where connections between nodes form a directed graph along a temporal sequence. This allows it to exhibit temporal dynamic behavior, which makes it applicable to tasks such as handwriting recognition or speech recognition. However, the RNN relies heavily on the automatic learning ability to update parameters which concentrate on the data flow but seldom considers the feature extraction capability of the gate mechanism. In this paper, we propose a novel architecture to build the forget gate which is generated by multiple bases. Instead of using the traditional single-layer fully-connected network, we use a Multiple Attention (MA) based network to generate the forget gate which refines the optimization space of gate function and improve the granularity of the recurrent neural network to approximate the map in the ground truth. Credit to the MA structure on the gate mechanism. Our model has a better feature extraction capability than other known models. MA-LSTM is an alternative module which can directly replace the recurrent neural network and has achieved good performance in many areas that people are concerned about.

Model:


$cell_{t-1}$
cell
$cell_t$
Forget Gate Net
$1-F$
Input Gate
Output Gate
h_{t-1}
h_t
x_t

Forget Gate Net:



- ***random base:***
  Random base means that we have no limit on the generation of the base which retains the excellent characteristics of traditional LSTM.

$$rb_t = \sigma(W_{rb}x_t + U_{rb}h_{t-1} + b_{rb}) \tag{8}$$

- ***increasing base:***

$$rb_t = cum(softmax(\sigma(W_{rb}x_t + U_{rb}h_{t-1} + b_{rb}))) \tag{9}$$

- ***decreasing base:***

$$db_t = 1 - cum(softmax(\sigma(W_{db}x_t + U_{db}h_{t-1} + b_{db}))) \tag{10}$$

tips:

- cum(): cumulative sum
  Example: $cum([1, 2, 3]) = [1, 3, 6]$

Experiment:

Traffic prediction:
Problem statement: Given the data until time interval t, the traffic volume prediction problem aims to predict the inflow and outflow traffic volume at time interval t + 1.

As the results are shown in Table I, we can see that our model performance better than not only the traditional methods but also some models with state-of-the art results. The model with MA-LSTM can better extract the characteristics of spatiotemporal data and provide more accurate prediction capabilities.

TABLE I
STAOC AND BASELINE MODELS PERFORMANCE

Data	Method	Start		End	
		rmse	mape	rmse	mape
Taxi	ARIMA	36.53	22.21%	27.25	20.91%
	LR	28.51	19.94%	24.38	20.07%
	MLP	26.67	18.43%	22.08	18.31%
	XGBoost	26.07	19.35%	21.72	18.70%
	LinUOTD	28.48	19.91%	24.39	20.03%
	ConvLSTM	28.13	20.50%	23.67	20.70%
	DeepSD	26.35	18.12%	21.95	18.15%
	ST-ResNet	26.23	21.13%	21.63	21.09%
	DMVST-Net	25.74	17.38%	20.51	17.14%
	STDN	24.10	16.30%	19.05	16.25%
	STAOC(ours)	**23.56**	**15.63%**	**18.43**	**15.56%**
Bike	ARIMA	11.53	26.35%	11.25	25.79%
	LR	10.92	25.29%	10.33	24.58%
	MLP	9.83	23.12%	9.12	22.40%
	XGBoost	9.57	23.52%	8.94	22.54%
	LinUOTD	11.04	25.22%	10.44	24.44%
	ConvLSTM	10.40	25.10%	9.22	23.20%
	DeepSD	9.69	23.62%	9.08	22.36%
	ST-ResNet	9.80	25.06%	8.85	22.98%
	DMVST-Net	9.14	22.20%	8.50	21.56%
	STDN	8.85	21.84%	8.15	20.87%
	STAOC(ours)	**8.54**	**21.25%**	**8.12**	**20.70%**

Expertiment:

Handwritten recognition:
Handwritten recognition is to recognize numbers in the images. In other words, it is to map an m*m grid data to a vector of length 10, with a value ranging from 0-1 which represent the probabilities of 1-10.

As the results are shown in Table II, we can see that the MA-LSTM performs better than LSTM, GRU and GIFG on the Handwriting Recognition problem. It means that MA-LSTM can extract feature better than many RNN architecture network and recognize more diverse patterns in the computer vision. It provides a new method for many currently difficult CV problems and can easily improve the performance with the model using the traditional RNN architecture.

TABLE II
MA-LSTM AND BASELINE MODELS PERFORMANCE

Dataset	Method	Tess Loss	Test Accuracy
Mnist	LSTM	0.07298	98.56%
	GIFG	0.07043	98.28%
	GRU	0.07739	98.45%
	MA-LSTM(ours)	**0.06722**	**98.93%**

Experiment:

Language Model:
Typical word-level language models are specified as a product of conditional probabilities using the chain rule.

As the results are shown in Table III, we can see that MA-LSTM performs better than others in one layer RNN model. It means that MA-LSTM can work well in NLP problem. However, when I design multiple layers RNN, the performance of MA-LSTM will turn poor.

TABLE III
MA-LSTM AND BASELINE MODELS PERFORMANCE

Dataset	Method	Tess Loss	Test Perplexity
PTB	LSTM	4.50	90.47
	GRU	4.52	91.02
	On-LSTM	4.48	88.52
	MA-LSTM(ours)	**4.47**	**87.07**