Proceedings of the 4th International Conference on Signal Processing and Machine Learning
DOI: 10.54254/2755-2721/53/20241288

Exploring the development and application of LSTM variants

Nuo Chen

Southwest Jiaotong University, 999 Xian Road, Sichuan, China

chennuo2022@my.swjtu.edu.cn

Abstract. Long Short-Term Memory (LSTM) is receiving increasing attention as the development of deep learning technology. The gate structure of LSTM enhances long-term memory, forming its superior capacity to complete tasks that challenge traditional RNN. However, considering the wide variety of applications, a comprehensive understanding of the development and application of the model, which is vital for future research, is comparatively lacking. Therefore, this paper is produced with the hope of offering an overview of the development of LSTM. It shows the process of development from RNN to LSTM and explains the aim and necessity of LSTM's birth. After that it introduces the structure of LSTM, analyses its advantages over RNN, and discusses the application of some popular LSTM variants, such as peephole LSTM, bidirectional LSTM, and GRU. Hopefully, this work can provide a more profound knowledge of LSTM's benefits and potential, identifying worthwhile avenues or fields of future research.

Keywords: Deep Learning, Neural Networks, Long Short-Term Memory (LSTM).

1. Introduction

As an essential branch of Recurrent Neural Networks (RNN), Long Short-Term Memory (LSTM) is widely adopted in tasks like NLP, stock forecasting, image processing, etc. First introduced by Hochreiter et al., LSTM is regarded as a remedy to the problem of gradient vanishing and gradient exploding in traditional RNNs [1]. The hidden layers of the model pass on cell state and hidden state, which preserve short-term and long-term information separately. The novel structure reduces the effect of large time lags, significantly improving LSTM's stability and accuracy compared with traditional RNNs. The model is being explored and optimized. Many branches and variants, such as bidirectional LSTM or GRU, were designed by researchers during the development of LSTM, broadening its application continuously. This paper aims to briefly introduce the theory of LSTM, show its structure, analyze its advantages over RNN, and discuss the application of some popular LSTM variants. By showing the application and importance of LSTM, it is hoped that this paper can provide a deeper understanding of its advantages and potential, revealing worthwhile directions or fields of future research.

2. Introduction of RNN

The Recurrent Neural Network is a popular model in various fields like time series prediction, machine translation, and voice detection. RNN has a similar structure as Muti-layer Perceptron (MLP), which is one of the most basic and widely applied deep learning algorithms. The main distinction between RNN

DOI: 10.54254/2755-2721/53/20241288

and MLP is that in an RNN model, the hidden layers at different moments are connected to not only other layers but also itself.

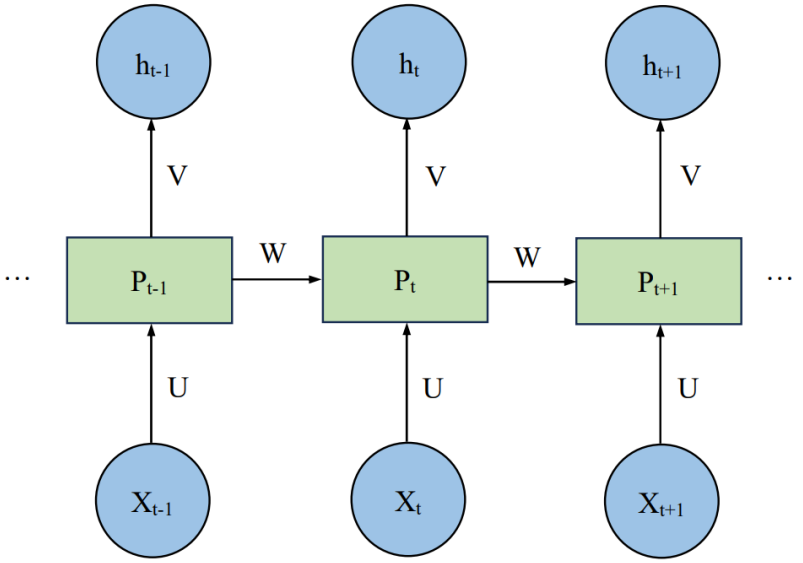

Figure 1. RNN Structure [2].

Figure 1 shows an unrolled graph of an RNN model, where U, V, and W are the weight matrix between layers. the value of P_t at moment t is affected by both the current input X_t and P_{t-1}.

Assume that the activation functions of the hidden and the output layer are $f(.)$and$g(.)$ respectively, then P_t is

$$P_t = f(UX_t + WP_{t-1}) \tag{1}$$

and h_t is

$$h_t = g(VP_{t-1}) \tag{2}$$

With the connection between hidden layers, an RNN model gains the ability to form the memory of a time series, which is lacking in regular BPNN models. This characteristic means that RNN is significantly superior in tasks where the inputs naturally appear as sequences, such as processing and predicting a paragraph of text or a segment of audio.

One drawback of RNN is that vanishing gradient and exploding gradient may occur during the backpropagation. This is mainly due to the high time dependency of the matrix W between hidden layers. In other words, W remains constant and does not change with different P_t. As a consequence, RNN tends to be weak in recognizing patterns or relativities with large time lags. Therefore, new methods and improvements are necessary for better precision and stability.

3. Theory and Applications of LSTM

To prevent gradient vanishing and exploding of standard RNN, Hochreiter et al. introduce LSTM, which can bridge larger scales of time lags and maintain the errors [1]. The structure was later enhanced by Gers et al. by adding a “forget gate” that allows the cells to reset when necessary [2]. The forget gate effectively avoids the saturation of the outputs or even breakdown, therefore it has been included in most LSTM designs ever since.

DOI: 10.54254/2755-2721/53/20241288

Staudemeyer et al. offer a specified analysis of the theory of LSTM and its variants [3]. Since this paper aims to give a review of LSTM's development and application, the explanation of the theory will be simplified.

Figure 2 demonstrates the unit of a classic LSTM model with multiple gate structures, which is called a cell.

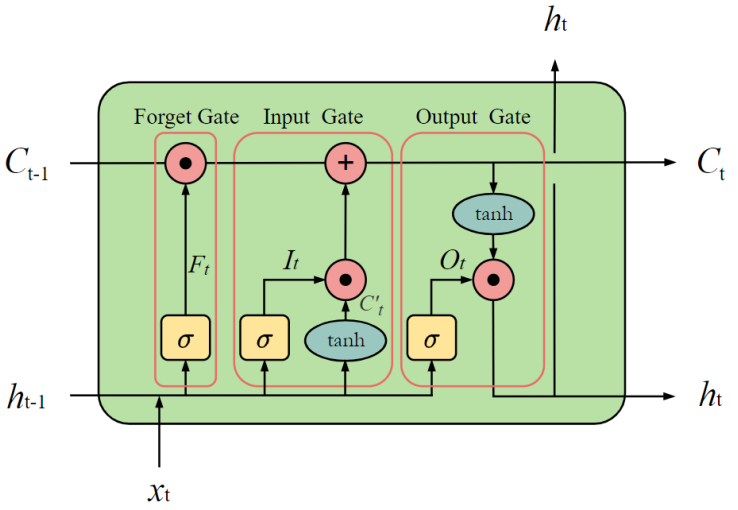

Figure 2. LSTM With a Forget Gate [2].

Notice that, unlike the single constant matrix W in the standard RNN model, the cell state C_t and the hidden state h_t here are both changed and passed on to the next cell. This allows the LSTM model to preserve information on short-term and long-term separatelSigma standsfor the Sigmoid function and tanh stands for the Tanh function. The forget gate gives a portion by which the data should be forgotten:

$$F_t = \sigma(W_F x_t + U_F h_{t-1} + b_F) \tag{3}$$

Where W_F and U_F stands for the weights while b_F stands for the bias.

For the input gate

$$I_t = \sigma(W_I x_t + U_I h_{t-1} + b_I) \tag{4}$$

$$C'_t = tanh(W_{C'} x_t + U_{C'} h_{t-1} + b_{C'}) \tag{5}$$

$$C_t = F_t \cdot C_{t-1} + I_t \cdot C'_t \tag{6}$$

For the output gate

$$O_t = \sigma(W_O x_t + U_O h_{t-1} + b_O) \tag{7}$$

$$h_t = o_t \cdot tanh(C_t) \tag{8}$$

In the fields of time series prediction, LSTM shows a higher level of accuracy and flexibility compared with models such as ARIMA and Prophet. Kong et al. apply LSTM in short-term household energy consumption prediction research on a subset of the SGSC dataset [4]. 70% of the 92-day period is chosen to be the training set, 20% as a testing set, and 20% as a validation set. For horizontal comparisons, besides an LSTM-based framework with 2 hidden layers and 20 nodes in each, they also connect a BPNN of the same size, a KNN algorithm with 20 neighbors, an ELM with an exhaustive search strategy, and an IS-HF with 1 hidden layer with 6 neurons [4]. The prediction results show that LSTM is outstanding in both individual household prediction and aggregating forecasts. Kong et al. suggest that the advantages of LSTM will be more obvious with the increasing in inconsistency of the data [4].

DOI: 10.54254/2755-2721/53/20241288

LSTM's capacity to complete tasks like machine translation and speech recognition is also widely recognized. Muhammad et al. adapted the Word2Vec and the LSTM model for the sentiment analysis of Indonesian hotel comments and reached an accuracy of 85.96%, showing LSTM's high-level performance in various NLP tasks [5]. Moreover, its diverse variants also expand the fields of application.

4. Variants of LSTM

On the base of classic LSTM, researchers explore many LSTM variants with different structures, characteristics, and application scenarios. In this part, some of the popular LSTM variants are introduced with cases of application.

4.1. Peephole LSTM

In a classic LSTM model, the gates cannot receive information from the cell state, which leads to possible information loss if the output of a cell is turned off. Peephole LSTM adds a peephole on each of the gates so that what the gates in this model receive is not only the hidden state at the last moment and the current input but also the cell state (Yang et al.) [6]. Yang et al. obtained the spring wind speed data every hour for one month at Rong Cheng Bureau port Xi Hu, processed it by the wavelet decomposition method, and made prediction by building a peephole LSTM model. Among all the data gained, those in the first 25 days are used for training and the left is for verifying [6]. The result is that the predicted data has a mean absolute percentage error (MAPE) of 0.0421, which proves that peephole LSTM is a suitable model for such time series prediction tasks [6].

4.2. Bidirectional LSTM

A single LSTM can only save information from the past to the future. To identify the data forward and backward at the same time, a bidirectional LSTM can be formed by combining two LSTM models in a contract direction. Gupta et al. adopted LSTM and bidirectional LSTM for poisonous comment detection [7]. The comments gained from the Kaggle dataset are first pre-processed and embedded, then split into the training set and testing set. An LSTM model and a bidirectional LSTM model are built to complete the multi-label classification task. It finally turns out that the bidirectional LSTM model gives significantly higher validation accuracy (98.07%) than standard LSTM [7]. This case offers a good example of the advantages of bidirectional LSTM in text-related problems since understanding the whole context is considerably vital in these problems.

4.3. Graph LSTM(G-LSTM)

In Graph LSTM, the propagation is not linear but through the graph formed by nodes, hence the structure provides more flexibility than chain LSTM. Graph LSTM reaches more than simply classifying sentiment but deeper opinion mining. It is applied in short-text sentiment classification by Wan [8]. The features of short text, such as irregularity and interactivity, are analyzed, which are challenges for normal linear LSTM. The dataset is randomly generated from the annotation corpus, embedded, and classified by LSTM, SVM, and Graph LSTM. The result is that Graph LSTM has the highest accuracy 85% while LSTM has 75% and SVM has 79% [8].

4.4. GRU

GRU has a similar structure to LSTM, but it does not contain the output gate. The simpler structure enhances GRU's flexibility and efficiency in implementation. Fu et al. predict traffic flow using ARIMA, LSTM, and GRU models and compare the prediction accuracies [9]. For the data collected from the PeMS dataset, GRU's MAE is 5% lower than LSTM and 10% lower than ARIMA [9]. Although the simple structure, fast speed, and high accuracy of GRU make it more and more popular, Yang et al. suggest that GRU is comparatively weaker when handling short text and large datasets [10]. The data loss of GRU under these circumstances is likely to be larger than that of LSTM. Concludingly, GRU

DOI: 10.54254/2755-2721/53/20241288

can be seen as a simpler solution instantly, while LSTM has more potential with the growth of computing power.

4.5. CNN-LSTM

Like LSTM, CNN is also a popular model in the field of deep learning. With the superior capacity to mine information and detect potential features, CNN is often utilized in natural language processing, object detection, and other tasks. To take advantage of both models, the combination of LSTM and CNN is a hotspot of research in recent years. A 3D CNN-LSTM model is constructed by Akilan et al. to realize a novel method for a video-based foreground-background (FG-BG) segment [11]. A Conv-LSTM2D model is first built and then some of the Vanilla layers are replaced by 3D conv layers. Finally, it forms a 3D CNN-LSTM model with 42 layers and 221,367 trainable parameters. The model is trained by 16 benchmark datasets chosen from the change detection 2014 benchmark database, with an Adadelta optimizer. In the sanity test, the CNN-LSTM model is 4% overall better than the original model and has 9 FPS higher inferencing speed [11]. Although the model is relatively weak when facing moving camera scenarios, it performs better than conventional methods in most other circumstances. It should also be emphasized that the model has more potential to handle many computer-vision (CV) based problems other than the FG-BG segment, which is left for future research and experiments.

5. Conclusion

This paper introduces the development and applications of LSTM. The gate structure of the LSTM cells promotes its ability to remember long-term memory and avoid the possibility of gradient vanishing and exploding in the traditional RNN model. The model is utilized in time series prediction, nature language processing, object detection, and many other edging deep learning tasks. On the base of LSTM, different variants, such as peephole LSTM and bidirectional LSTM also spring, expanding the fields of application. Moreover, the combination of LSTM and other models is also a worthy direction for research, which may produce novel methods to solve many conventional problems. Finally, due to the limitation in personal academic ability and research angle, this paper only aims to provide an overview of LSTM's development and application. A detailed analysis of LSTM's mathematical theory is considered in future work, while more complicated variants can also be introduced to help form a more generalized and deep understanding.

References

[1] Hochreiter S and Schmidhuber J 1997 J. Neural Comput. 9 8 1735-1780
[2] Gers F A Schmidhuber J and Cummins F 2000 J. Neural Comput. 12 10 2451-2471
[3] Staudemeyer R C and Morris E R 2019 J. Prep ArXiv.1909 09586
[4] Kong W, Dong Z Y, Jia Y, Hill D J, Xu Y and Zhang Y 2019 J. IEEE. T. Smart Grid. 10 1 841–851
[5] Muhammad P F, Kusumaningrum R and Wibowo A 2021 J. Proc. Comput. Sci. 179 728–735
[6] Yang T, Wang H, Aziz S, Jiang H and Peng J 2018 C. IEEE. 364-369
[7] Gupta A, Nayyar A, Arora S and Jain R 2021 C. Inter. Conf. on Advanced Informatics for Computing Research (Singapore) 1393100–112
[8] Wan Y 2019 C. 2019 Int. Conf. on Artificial Intelligence and Advanced Manufacturing (AIAM) (Dublin, Ireland). 35-38
[9] Fu R, Zhang Z and Li L 2016 J. Conf. of Chinese Association of Automation (YAC) (Wuhan, China). 324-328
[10] Yang S, Yu X and Zhou Y 2020 C.Int. Workshop on Electronic Communication and Artificial Intelligence (IWECAI) (Shanghai, China). 98-101
[11] Akilan T, Wu Q J, Safaei A, Huo J and Yang Y 2020 J. IEEE Transactions on Intelligent Transportation Systems. 21 3 959–971