# Sales Forecaster: A Bis-Attention LSTM Encoder-Decoder for Local Store Sales Forecasting

**Ziye Zhou[1], Zhiying Chen[1], XuYUAN[2]**

HJYCloud Technology, School of Software, Dalian University of Technology, Shenzhen, Guangdong, China

**Keyword:** Time Series, Forecasting, Attention, Neural Network, Sales

**Abstract:** Time series forecasting is deeply and broadly studied throughout the years with various applications in multiple industries, such as finance, weather, and environment. Deep neural networks combined with attention mechanism is capable of capturing long-term patterns while effectively incorporating information from other variables. This paper applies and modifies a few of the deep neural networks based on attention mechanism for a real-world problem, local store sales forecasting, in a theoretical manner. Apart from the effect of long-term dependencies, this model is able to take dynamic spatial correlation into account, which is widely encountered in local store sales forecasting.

## 1. Introduction

The wide application of deep neural networks along with attention mechanism [1, 2, 3] has provided a possible solution for multivariate time series forecasting. Long- and Short-term Time-series Network (LSTNet) [1] is proved to be a reliable model to perform multivariate time series forecasting. Based on this, the dual-stage attention based recurrent neural network (DA-RNN) [2] is considered as the state-of-the-art method in time series prediction.

This paper considers the problem of store sales forecasting in a time series framework whose target series is sales and tries to model the relationship among historical sales on both itself and other stores, driving series and non-time-series data. There existed only a few research [4] in the intersection of machine learning and sales prediction before because it is a complicated problem that the effects among a multitude of variables are tricky to extract and there lacks appropriate models for this problem. Traditional models either fail to capture the spatio-temporal interactions among different segments of exogenous data, or are unable to model the relation between driving series and target data, let alone the extremely long-term temporal effects. Fortunately, the wide applications of end-to-end deep learning models provide us a potential way to this problem because of the incorporation of spatio-temporal feature extraction for exogenous data, temporal dynamical modeling of target series, and the inclusion of external factors.

Our model proposed to store sales forecasting has two parts, a linear component whose inputs are external factors and historical sales and a non-linear component, responsible for extracting the spatial and temporal effects among target sales, driving series and historical sales through an encoder-decoder architecture.

## 2. Related Work

Deep neural networks, especially Recurrent Neural Networks (RNNs) and Convolutional Neural Networks (CNNs), are two neural networks that are investigated in the context of time series forecasting and proved to be naturally suitable for such problems. LSTNet [1] is a typical combination of both CNNs and RNNs as it leverages CNN to capture short-term dependencies and RNN along with Recurrent-skip (LSTNet-S) [1] or Attention mechanism (LSTNet-A) [1] to model long-term dependencies. In terms of the very long-term dependencies in time series, a memory component and attention mechanism [5] are more effective compared to aforementioned models.

  DOI: 10.25236/icemeet.2021.032

The tremendous breakthroughs in natural language processing, encoder-decoder networks for a sequence to sequence [6] have become successful in time series forecasting as well. The incorporation of attention mechanism [7] alleviates the issue that the performance of this architecture will deteriorate rapidly as the length of input sequences increase such as Dual-stage Attention-based Recurrent Neural Network (DA-RNN) [2]. However, these models do not consider the spatial correlations among different segments of exogenous data which is common in store sales forecasting. A Multi-level Attention Network to predict the readings of a Geo-sensor (GeoMAN) [3] is proposed to model the dynamic spatio-temporal dependencies.

The nearest deep model proposed before to our model is DA-RNN, which inspires us to apply a dual stage attention mechanism, encoder input mechanism and decoder temporal mechanism. However, they fail to consider the geospatial effect when providing the forecasts. GeoMAN offers a potential way for us to incorporate the geospatial effect via the weighted sum of attention weights and geospatial similarity. External factors from GeoMAN are included in the linear component of model rather than acting as an input for the decoder in GeoMAN. Another drawback of DA-RNN and GeoMAN is that they do not consider the autoregressive part for final forecasting to alleviate the insensitiveness of predictions in neural network. The proposal of linear and non-linear component refers to the introduction of wide and deep model [8] which divides a joint training neural network into wide and deep part.

## 3. Problem Formulation

Our problem is to model target variable $S^i$ given both historical sales $S^j_{t<T}$ and driving series $F_{t<T}$, where T is our targeted time stamp. Formally, given a set of multivariate driving series $\mathrm{F} = \{F^1, F^2, \cdots, F^D\}$, where $F^i \in \mathbb{R}$, D is the number of features, historical sales for itself $S^i = \{s^i_1, s^i_2, \cdots, s^i_{T-1}\}$, and historical sales for other local stores $S^{j\neq i} = \{s^j_1, s^j_2, \cdots, s^j_{T-1}\}$ where $\mathrm{j} \neq \mathrm{i} \in \mathbb{N}$, we aim to predict the target variable in a rolling forecasting fashion, that is, $S^i = \{s^i_T, s^i_{T+1}, \cdots, S^i_{T+k}\}$, where k is the predictive horizon.

## 4. Methodology

A novel input attention mechanism that can adaptively select the relevant driving series is proposed and another attention mechanism incorporating with the spatial similarity supple the encoder architecture. A temporal attention mechanism is used to learn a soft alignment between encoder hidden states and target sales across all time stamps in the decoder. Finally, a linear regression of AR components and external features is included in the final prediction.

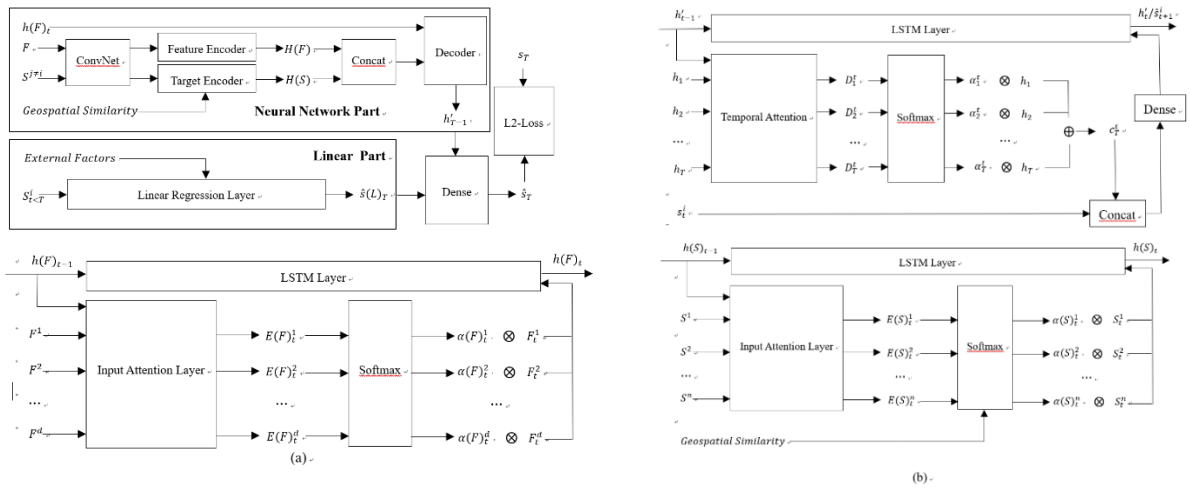


Figure 1. Graphical illustration of Bis-Attention LSTM Encoder-Decoder Sales Forecaster. (a) and (b) are Feature Encoder and Target Encoder. The upper left part is the general framework and the upper right part is the decoder.

### 4.1 Encoders

The encoder is made up of two encoders for different purposes. The first encoder is composed of a CNN and an RNN along with attention mechanism that encodes the inputs into a vector representation. Besides, the second encoder includes a matrix to measure the geospatial similarity which is learned from the network of stores as a prior knowledge for attention weights.

#### 4.1.1. RNN with input attention

The encoder maps driving series into the hidden state with $\mathrm{h}(F)_t^{rec} = LSTM(h(F)_{t-1}^{rec}, F_t)$, where $h(F)_t^{rec}$ is the hidden state of the encoder at time t, LSTM is an activation function.

Given the $d$-th driving series denoted as $F^d$, an input attention mechanism for each time stamp is presented as the following equation by referring to the previous hidden state and cell state in the encoder, $\mathrm{E}(F)_t^d = v_d' \tanh(W_d[h(F)_{t-1}^{rec.}, c(F)_{t-1}^{rec}] + U_d F^d + b_d$ and $\alpha(F)_t^d = e^{E(F)_t^d} / \sum_i e^{E(F)_t^i}$, where $v_d \in \mathbb{R}^{T\times 1}$, $W_d \in \mathbb{R}^{T\times 2M}$, $U_d \in \mathbb{R}^{T\times T}$, and $b_d \in \mathbb{R}^{T\times 1}$ are learnable parameters.

The vector representation for modified driving series according to attention weight can be further obtained, $\hat{F}^d = \{\alpha(F)_1^d f_1^d, \alpha(F)_2^d f_2^d, \cdots, \alpha(F)_t^d f_t^d\}$. Then the updated hidden state at time step t is $h(F)_t^{rec} = LSTM(h(F)_{t-1}^{rec}, \hat{F}_t)$.

#### 4.1.2. Incorporation of geospatial similarity

The second encoder includes a geospatial similarity when applying different weights to certain sales series rather than treating all sales series equally.

We can incorporate the geospatial similarity in a linear weighted manner for target sales $S^i$: $\alpha(S)_t^n = e^{(1-y)E(S)_t^n + yCorr(n,i)} / \sum_j e^{(1-y)E(S)_t^j + yCorr(j,i)}$, where $\alpha(S)_t^n$ denotes the modified attention weight for the sales series $S^n$. $Corr(n,i)$ is the indicator to measure the geospatial similarity between two different local stores. $\gamma\in[0,1]$ is tunable hyperparameter as a trade-off determining how importance geospatial similarity accounts for the modified attention weight.

Hidden state in this encoder for sales in other stores' can be updated in the same manner as the first one, $\hat{S}^n = \{\alpha(S)_1^n s_1^n, \alpha(S)_2^n s_2^n, \cdots, \alpha(S)_t^n s_t^n$. Then the updated hidden state for other sales data at time $t$ is $h(S)_t^{rec} = LSTM(h(S)_{t-1}^{rec}, \hat{S}_t$.

### 4.2. Decoders

The attention weight of each encoded hidden state in encoder $[h(F)_t^{rec}; h(S)_t^{rec}]$ at time $t$ is calculated based on the previous decoder hidden state $h_{t-1}^{\prime rec}$ and the cell state $c_{t-1}^{\prime rec}$ in the LSTM units by $D_i^t = v_i' \, tanh(W_i[h_{t-1}^{\prime rec}; c_{t-1}^{\prime rec}] + U_i[h(F)_i^{rec}; h(S)_i^{rec}] + b_i$ where $h_{t-1}^{\prime rec}$, $c_{t-1}^{\prime rec} \in \mathbb{R}^p$. $p$ is the number of decoder hidden units. $[h(F)_t^{rec}; h(S)_t^{rec}] \in \mathbb{R}^{2p}$ is the concatenation of the previous hidden and cell state in the decoder at time $t$. $v_i \in \mathbb{R}^{m\times 1}$, $W_i \in \mathbb{R}^{m\times 2p}$, $U_i \in \mathbb{R}^{m\times 2m}$ and $b_i \in \mathbb{R}^{m\times 1}$ are learnable parameters. The final output of the decoder is a context vector which represented as $c_T^t = \sum_{i=1}^T \alpha_i^t \, [h(F)_i^{rec}; h(S)_i^{rec}]$, where $\alpha_i^t$ represents the attention weight of the i-th encoder hidden state. A simple dense layer mapping context vector and target sales to $\tilde{s}_t^i = w_i[s_t^i; c_t] + b_i$, where $[s_t^i; c_t] \in \mathbb{R}^{(2m+1)\times 1}$ is the concatenation of context vector and target sales at time $t$. $w_i \in \mathbb{R}^{(2m+1)\times(2m+1)}$, and $b_i \in \mathbb{R}^{(2M+1)\times 1}$ are learnable parameters. The final output can be regarded as $h_t^{\prime rec} = LSTM(h_t^{\prime rec}, \tilde{s}_t^i)$.

### 4.3. A Linear Part of Autoregression and External Factors

The linear part of this model is a simple linear regression for both historical observations in a fixed window and external factors. Assumed that predictions of this linear part are $\tilde{s}(L)_t^i$, and the coefficients of this linear regression are both $W_i^L \in \mathbb{R}$ and $b_i^L \in \mathbb{R}$. The linear regression model is formulated as $\tilde{s}(L)_t^i = \sum_{k=1}^w w_i^L [s_{t-k}^i; E^i] + b_i^L$.

### 4.4. Joint Training of the Linear and Non-Linear Components

The model's prediction is generated via $\hat{s}_t^i = W_i^D h_{t-1}^{\prime rec} + W_i^L \tilde{s}(L)_t^i + b_i$, where $\hat{s}_t^i$ is the final prediction for target sales series at tim $t$. $W_i^D \in \mathbb{R}^{1\times p}$, $W_i^L \in \mathbb{R}^{1\times 1}$ and $b_i \in \mathbb{R}^{1\times 1}$ are learnable

parameters.

## 5. Conclusions

This research introduces a sales forecaster customized to our research problem and provides a promising tool for store sales prediction, greatly enriching the techniques for commercial prediction. The contribution for our research is that we propose a model specifically for the problem of sales forecasting considering multi-dimensional features including driving series for sales, other stores' sales, stores' geospatial similarities, and external factors. To the best of our knowledge, it is the first deep learning model for considering so many features in multivariate time series forecasting.

However, this research still has some limitations as well. It does not include any practical dataset and comparison between existing models and our proposed model. But these limitations will become our future research direction, and we are aiming to apply our proposed model to practical datasets and evaluate its performance. Besides, another direction is related to the estimation of geospatial similarity where we can inject some information learned from Graph Convolutional Networks.

## Reference


[1]. G. Lai, et al. "Modeling Long- and Short-Term Temporal Patterns with Deep Neural Networks", presented at the 41st International ACM SIGIR Conference on Research & Development in Information, 2018, pp.95-104.

[2]. Y. Qin, et al. "A Dual-Stage Attention-Based Recurrent Neural Network for Time Series Prediction", presented at Proceedings of the 26th International Joint Conference on Artificial Intelligence, 2017, pp.2627-2633.

[3] Y. Liang, et al. "GeoMAN: Multi-Level Attention Networks for Geo-Sensory Time Series Prediction", presented at Proceedings of the 27th International Joint Conference on Artificial Intelligence, 2018. pp.3428-3434.

[4]. Y. Kaneko, and K. Yada. "A Deep Learning Approach for the Prediction of Retail Store Sales", presented at 2016 IEEE 16th International Conference on Data Mining Workshops, Barcelona, Spain, 2016.

[5]. Y. Chang, et al. "A Memory-Network Based Solution for Multivariate Time-Series Forecasting", arXiv preprint arXiv: 1809.02105, 2018.

[6]. I. Sutskever, O. Vinyals, and Q. Le. "Sequence to Sequence Learning with Neural Networks", presented at 28th Conference on Neural Information Processing Systems, Montreal, Canada, 2014.

[7]. A. Vaswani, et al. "Attention is All You Need", presented at 31st Conference on Neural Information Processing Systems, 2017.

[8]. H. Cheng, et al. "Wide & Deep Learning for Recommender Systems", presented at Proceedings of the 1st Workshop on Deep Learning for Recommender Systems, 2016, pp.7-10.