Energy & Buildings 249 (2021) 111211

Contents lists available at ScienceDirect

Energy & Buildings

journal homepage: www.elsevier.com/locate/enb

Dual-stage attention-based long-short-term memory neural networks for energy demand prediction

Jieyang Peng [a,b], Andreas Kimmig [b], Jiahai Wang [a], Xiufeng Liu [c,*], Zhibin Niu [c,d,*], Jivka Ovtcharova [b]

[a] Advanced Manufacturing Technology Center, Tongji University, Shanghai 200092, PR China
[b] Karlsruhe Institute of Technology, Karlsruhe 76133, Germany
[c] Department of Technology, Management and Economics, Technical University of Denmark, Kgs. Lyngby 2800, Denmark
[d] College of Intelligence and Computing, Tianjin University, Tianjin 300072, PR China

ARTICLE INFO

Article history:
Received 16 March 2021
Revised 19 May 2021
Accepted 20 June 2021
Available online 24 June 2021

Keywords:
Energy demand forecasting
Energy consumption pattern recognition
Long short-term memory network
Attention mechanism
Word embedding

ABSTRACT

Forecasting energy demand of residential buildings plays an important role in the operation of smart cities, as it forms the basis for decision-making in the planning and operation of urban energy systems. Deep learning algorithms are commonly used to reliably predict potential energy usage since they can overcome the issue of dependency on long-distance data in energy forecasting relative to the standard regression model. However, there are still two problems to be solved for energy forecasting, including the encoding of categorical characteristics and adaptive extraction of the most relevant characteristics for the use in predictions. To address the problems, we proposed a sequential forecasting model for medium- and long-term energy demand forecasting based on an embedding mechanism and a two-stage attention-based long-term memory neural network. An empirical study was conducted on three years of daily electricity consumption data from the residential buildings of the Pudong district of Shanghai to evaluate the model. The results show that the model can effectively extract the key features that are highly correlated with energy consumption dynamics by employing long-term dependencies in time series. In addition, the hybrid model outperforms others in terms of long-term forecasting capability. This paper also discusses future research directions and the possibilities for applying deep learning techniques in the energy sector.

1. Introduction

Studies by the International Energy Agency have shown that the global electrical power consumption is growing faster than any other form of energy demand. Electric power's share of the total energy consumed has risen from 9% to 17% and will increase to more than 25% by 2050. In the new liberalized environment of smart cities, it is essential to understand and to predict the impact of natural variables on energy demand, in order to effectively manage power generation and supply [1]. An early understanding of power demand behavior is critical to the planning, analysis, and operation of energy systems and the ability of ensuring an uninterrupted, reliable, secure, and economical power supply [2].

For this purpose, recurrent neural networks (RNNs) have been applied extensively to energy demand forecasts. With RNNs, practical values for long-term dependencies are usually limited by the problem of disappearance or explosion of gradient. The most common solutions are long short-term memory (LSTM) and gated recurrent units (GRUs). The main idea is to create information paths through time whose derivatives are robust to the problems of disappearance and explosion of gradients. Unlike feed-forward artificial neural networks, LSTM performs the same task for each element of a sequence, with output depending on previous calculations. In addition, by introducing gate control, LSTM can solve the problems of disappearance and exploding gradients, making LSTM as one of the most popular deep learning networks in recent years.

Although LSTM can solve the problem of dependency on remote information during RNN training, there are several other challenges in forecasting energy consumption. First, the time series of energy consumption usually follow a periodic pattern (usually annually) [3], which changes over time and varies geographically. To learn the dynamic temporal correlations, the LSTM neural network usually compresses all necessary information from a source file into a fixed-length vector [4]. It is difficult to obtain all the necessary information (depending on the vector length) if the input sequence is too long, and this is often the case when predicting

* Corresponding author.
E-mail addresses: xiuli@dtu.dk (X. Liu), mind3str@gmail.com (Z. Niu).

https://doi.org/10.1016/j.enbuild.2021.111211
0378-7788/

Nomenclature

Nomenclature

LSTM	long short-term memory
DL	deep learning
ANN	artificial neural network
SVR	support vector regression
RFR	random forest regression
ABR	ada boost regression
ASLSTM	single attention LSTM
DALSTM	dual attention LSTM
EDA-LSTM	embedding dual attention LSTM
MAE	mean absolute error
MAPE	mean absolute percentage error
RMSE	root mean squared error
GDP	gross domestic product
LT	daily low temperature
HT	daily high temperature
DHT	daily high dew temperature
LH	daily low humidity
HH	daily high humidity
LW	daily low wind speed
HW	daily high wind speed
LA	daily low air pressure
HA	aily high air pressure
AP	air pollution index
WT	weather type
WD	wind direction
HI	Holiday index
GNP	gross national product

energy data. Thus, the traditional LSTM model cannot glean the necessary knowledge from data over a sufficiently long period of time.

Second, factors (features) influencing energy consumption are diverse in nature. For example, there are continuous numerical features such as temperature, air pressure, wind speed, etc. as well as categorical aspects such as weather type, wind direction, day of the week, and so forth. In energy consumption forecasting, these categorical attributes are usually covered directly in an single step, by enumeration in whole numbers or code. For example, Nan [5] directly used data as training characteristics. Since the different dimensions in the eigenvector are entirely independent of one another, the eigenvector cannot represent similarities between the meanings of category labels (or words) in the above-listed attributes, leading to poor generalization in the model.

In response to the above challenges, we employ an attention-based LSTM model to predict energy demand. First, based on the classical factors in the field of energy, we introduce the embedding mechanism to encode category data. The embedding mechanism preserves the relative closeness of the original samples in the semantic space for a category feature. Then, a temporal attention mechanism is used to adaptively extract the relevant driving time series related to the previous hidden state of the encoder, while a spatial attention layer is applied for the decoder to select the most relevant hidden states in the feature space at each moment. The experiments show that the attention-based LSTM can capture the subtle temporal consumption pattern persisting in energy load profiles and produce the best forecasts for the majority of cases. On the aggregation level, we also demonstrate that aggregating separate energy consumption patterns are generally more accurate than directly forecasting the aggregated loads.

The remainder of this paper is structured as follows. Section 2 provides the background of the load forecasting. Section 3 conducts an exploratory data analysis to explain the challenge of the long-term load forecasting problem. Section 4 introduces the structure and implementation of the hybrid LSTM framework. Section 5 introduces other benchmarks including three empirical predictors and discusses the experimental results. Section 6 concludes the paper and points out the future research directions.

2. Related work

Several studies have shown that the factors most affecting energy consumption tend to fall into one of the following three categories: weather conditions, economic indicators and regional differences (see Fig. 1). Among them, weather conditions are usually used for short- to mid-term energy demand forecasts, whereas economic factors are more relevant for long-term forecasts. Moreover, the demographic and lifestyles of various areas often contribute to a disparity in the consumption of energies.

Many utilities found that the weather condition influences short-term energy demand, including the aspects such as temperature, humidity, wind, and precipitation (in decreasing order of importance) [6]. Yan [7] used climate variables to present residential power consumption predictions of Hongkong, finding that the mean temperature strongly correlated with power consumption. Pardo [8] researched temperature and seasonal influences on Spainish energy demand. Sforna [9] identified an energy demand/temperature function by using the group method of data handling. Islam [10] developed novel artificial neural network-based weather-load and weather energy models to forecast power demand up to 24 months ahead. Wei [11] provided a deep learning model of natural gas load in consideration of temperature, dew point temperature and other weather factors. Wright investigated the forecasting of electrical demand at an electric utility within the province of Saskatchewan and proposed a multi-region load forecasting system based on weather related demand variables [12]. Taspinar [13] proposed multilayer perception of ANNs to forecast short-term natural gas consumption. Liu et al. proposed the periodic auto-regression with exogenous variables (PARX) model to predict electricity consumption [42] and generated smart meter data of electricity [43]. Meteorological data (moisture, atmospheric pressure, wind speed, and ambient temperature) in the last 4 years are used to construct a well-tuned algorithm. Zhang [14] presented an algorithm based on the ANN and weather conditions for forecasting electricity consumption in the high energy-consuming city. A better prediction result is gained, and the defect of falling into local hypo-strongpoint is overcomed. Additional research on the relationship between weather conditions and electric power system loads are available in the literature [15–18].

Within the socio-economic field, Kermanshahi [19] found that unlike short-term power demand forecasting, long-term power demand forecasting is mainly affected by economic indicators rather than weather conditions. Canyurt [20] modeled Turkish future energy demand based on the gross domestic product (GDP) and import and export figures. Egelioglu [21] used multiple linear regression analysis to study the influence of economic variables on the annual power consumption of northern Cyprus. Harris [22] found that price played a major role in explaining consumers' energy conservation behavior. Lakhani [23] used residential price of power by per capita income and estimated the long-term

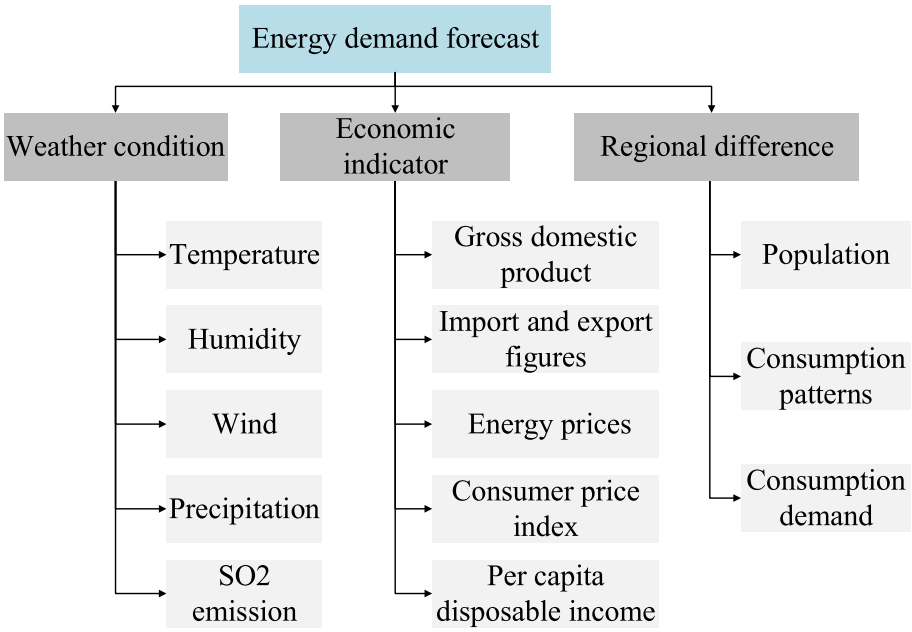

Fig. 1. Major factors affecting energy demand forecasting.

elasticity of demand to forecast energy demand in Maryland. Gautam [24] found that energy demand in the residential sector is responsive to price changes over the long term. Liu [25] used GDP, real power price, and population to forecast power consumption of Singapore. In the study of Gent, gross national product (GNP) and the price of electricity were found to relate best to the growth in electricity, while the effect of the weather was not considered [26]. Minato's study concerns the forecasting process of a region's energy demands, addressed not only by time series analysis but also by economic and social indexes [27]. Leung evaluated the accuracy of electricity consumption forecasts using various commonly employed methods based on average historical growth rates or the historical relationships between electricity consumption and key economic and demographic variables [28].

Additionally, regional differences are also crucial factors in energy demand forecasting. Rajanuu [29] expressed energy consumption patterns as a function of local weather and population. Wei collected historical datasets for four representative cities in three climate zones and developed a hybrid model to forecast daily natural gas consumption for each city.

3. Exploratory analysis and problem identification

The datasets for the empirical study are introduced in this section. The main challenge of the long term household load forecasting problem is the diversity and volatility. Therefore, we explore the consumption patterns from the perspective of clustering.

3.1. Dataset

The raw dataset used in this research is the electricity consumption data of the Pudong district of Shanghai from July 2015 to June 2018. The focus of this study is the Pudong region in the eastern Shanghai, with an area of 1,210 square kilometers and 5,501 million inhabitants. As weather is a factor closely related to electricity consumption, weather data from the Pudong district was also collected. These datasets are the key to enabling further studies of many smart grid technologies in a real-world context.

3.2. Exploratory analysis

According to the method proposed in [1,31], we can explore electricity consumption patterns by a visual analysis method. The dimensionality of the time series is first reduced to a 2D space using the t-SNE algorithm, then time series are visualized as a point in the scatter plot view. The time series of the electricity consumption with similar patterns are tightly bundled. The analysis is based on the daily electricity data of all the households of the Pudong district in 2017. The input of t-SNE is the daily electricity consumption data of each household, the dimension of the embedded space is set to 2 for the visual analysis purpose. The perplexity value is set to 30. The learning rate is set to 1,000 to avoid falling into a local optimum. The number of iterations and gradient norm are 200 and 1e-7, respectively. Furthermore, the principal components method of analysis is used to initialize the data and improve global stability. When PCA is conducted to initialize the raw data, the 0–1 normalization method is adopted to scale the data between 0 and 1, and the number of components is 365. The results of the cluster analysis are shown in Fig. 2.

As shown in Fig. 2, each point in the clustering map represents a household, and the distance between two points represents the difference in electricity consumption patterns. The horizontal axis of each of the six subgraphs corresponds to 365 days (1 year), and the vertical axis represents the electricity consumption of each day. The electricity consumption patterns can be classified into six categories.

- Bimodal pattern is the one closely related to the changes of weather temperature, and has been subdivided into summer and winter patterns. The two energy consumption patterns can be explained by the use of energy for heating in winter and for cooling in summer.
- High energy consumption and energy savings correspond to two extremes of electricity consumption. There were far fewer households with this pattern than with the bimodal pattern. High energy consumption pattern is characterised by constant high consumption throughout the year and low fluctuations within a fixed value range. In contrast to the electricity con-

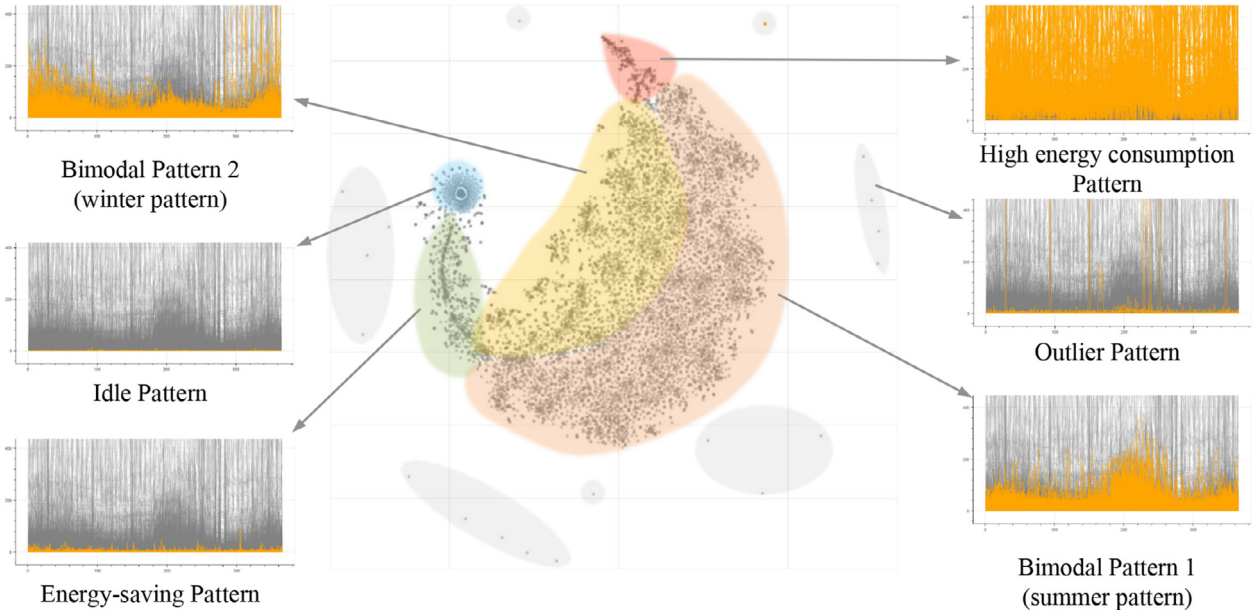

Fig. 2. Clustering analysis of the households in the Pudong district in 2017. The five typical patterns by different colors: blue = idle, green = energy savings, pink = high energy consumption, gray = outlier, yellow = bimodal 1, and orange = bimodal 2.

sumption with a bimodal pattern, the consumption with this pattern is above the average daily consumption on most days. This pattern occurs due to several reasons, for example a household may have low efficienct appliance or has a large home.

- The energy saving pattern has approximately the same shape as the bimodal pattern, but the consumption with this pattern is much lower than the consumption with the bimodal pattern. The energy-saving consumption pattern may be explained by the fact that these households live in new apartments with energy efficient appliance, or that they are low-income families who are very cautious about consuming too much energy.
- The outlier pattern cannot be explained by existing knowledge; it may be caused by irregular living habits or electricity theft. Although it is difficult to determine the reasons, the pattern provides users with a tracking tip.
- It is worth noting that the idle pattern also accounts for a considerable proportion, due to a large number of vacant apartments in Shanghai. After zooming in with the exploration tool, at a fixed date in each month, some slight variations in electricity consumption can be observed for some households, meaning that these homes were unoccupied, but regularly inspected. Whilst, the homes with zero consumption were unoccupied throughout the year, e.g., new apartments.

According to the above analysis, the electricity consumption data of the Pudong district have several patterns. Each pattern represents a specific temporal correlation. Such temporal correlations exist widely in the consumption pattern because they are based on residents' behaviors that are hard to learn, leading traditional regression methods to perform poorly. In this regard, a learning algorithm with the capability of abstracting previous observations as some hidden knowledge of residents' behaviors and establishing a correlation between the abstraction and energy demand is the key to better forecasting performance. In the case of residential load forecasting, the attention-based LSTM network can be used to generalize residents states from the patterns of the input consumption profiles, maintain the memory of the states, and finally make a prediction based on the learned information.

4. Methods

This section describes in detail the motivation, the modeling process, and the main parameters of the forecasting model.

4.1. Framework overview

The embedding dual-stage attention-based long short-term memory neural network (EDA-LSTM) model is mainly composed of feature engineering and deep neural network modules. Fig. 3 illustrates the framework. In the feature engineering module, we use a normalization method to scale the value of each numerical attribute between 0 and 1. We also use word embedding to map categorical attributes to fixed-length feature vectors. The preprocessed data are transferred to the prediction model to predict future energy consumption. The prediction model integrates the attention mechanism into the traditional encoder-decoder architecture of the LSTM. Finally, the prediction accuracy is analyzed and compared with other algorithms, and the application of the attention mechanism to the prediction of energy demand is also investigated.

4.2. Feature engineering

We extract the following 15 features for the energy consumption forecasting process.

The numerical attributes are: (1) daily low temperature (LT), (2) daily high temperature (HT), (3) daily low dew temperature (LDT), (4) daily high dew temperature (DHT), (5) daily low humidity (LH), (6) daily high humidity (HH), (7) daily low wind speed (LW), (8) daily high wind speed (HW), (9) daily low air pressure (LA), (10) daily high air pressure (HA), and (11) air pollution index (AP).

The categorical attributes are (12) weather (WT), (13) wind direction (WD), (14) weekday (Wd), and (15) holiday index (HI). Among them, weather includes light rain, clouds, sleet, and 11 other typical weather conditions. Holiday index is the number of vacation days, usually ranging from 1 to 8 (a workday is represented by 0). However, the Spring Festival is the most important

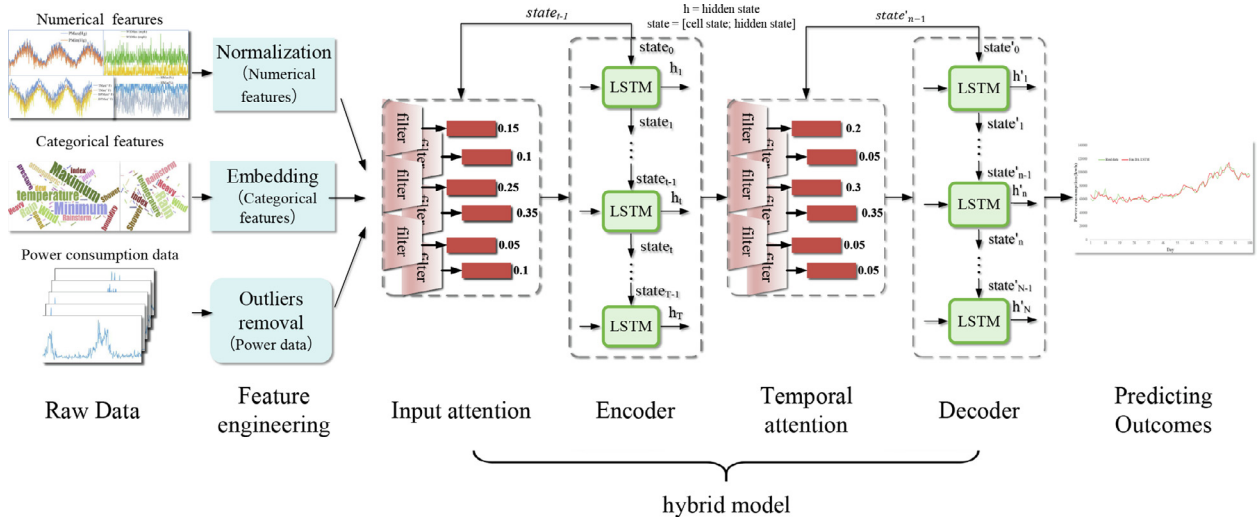

Fig. 3. Data processing flow. The process can roughly be divided into three steps: data pre-processing, hybrid modeling, and results comparison.

holiday in China (most people leave from the city to their hometowns during the Spring Festival), so it is coded separately as 9.

Normalization. Since the value range of the raw data varies, the value range of all numeric characteristics needs to be normalized so that each characteristics contributes approximately proportionally to the forecasting model. Another reason for the use of feature scaling is that the gradient descent converges much faster with feature scaling than without [33]. The general equation for a min–max of [0, 1] is as follows:

$$x_{scale} = \frac{x - x_{min}}{x_{max} - x_{min}} \tag{1}$$

where x is the original value, x_{min} and x_{max} are the minimum and maximum values of a specific feature, respectively, and x_{scale} is the normalized value.

Word embedding. In general, categorical attributes are usually encoded into fixed-length vectors using the one-hot encoding approach. However, such mapping is completely uninformed; "similar" categories are not placed closer together in the vector spaces. Therefore, word embedding is introduced in the present research to meaningfully represent categories in the transformed space. The process of embedding is shown in Fig. 4.

In the case of weather attribute, the input sequence has 11 different weather types, including sun, clouds and cloudiness. After being input, all weather features are converted into a numerical dictionary. The "embedding lookup" module based on the TensorFlow platform is used to map all embedding functions. During the training process, the batch size is set to 128, the number of weather features to 11, and the input time step to 15. After training, the weather feature is a three-dimensional tensor of 15*128*5.

4.3. Deep network structure

The EDA-LSTM model proposed in this paper consists of two modules: an input attention layer and an attention-based coding-decoding layer.

Input attention layer. As mentioned above, after feature engineering, the raw data are converted into 15 dimensional feature vectors, x_t^k, where t represents the time, and k represents different attributes.

At time step t, given the $15-th$ input sequence $x^k = (x^1, x^2, \ldots, x^{15})^T \in R^T$, an input attention layer can be constructed through a deterministic attention model by referring to the state vector $state_{(t-1)}$ in the encoder LSTM unit with:

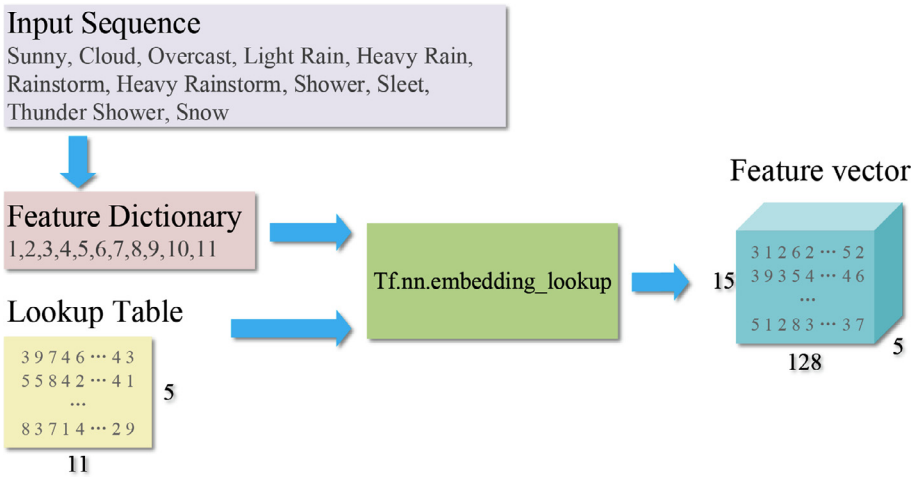

Fig. 4. Embedding training process in a deep learning framework.

$$\begin{aligned} e_t^k &= f(state_{t-1}, x^k) \\ &= v_e^T tanh(W_e[h_{t-1}; c_{t-1}] + U_e x^k + b_e) \\ &where k \in (1,15), t \in (1,10) \end{aligned} \tag{2}$$

In Eq. (2), e_t^k is the weight vector in the attention mechanism, the state vector $state_{(t-1)}$ includes the hidden state $h_{(t-1)}$ and cell-state $c_{(t-1)}$ of the encoder LSTM unit (refer to Fig. 3), b_e is the bias terms, $v_e^T \in R^{(Tm)}, W_e \in R^T, U_e \in R^{(TT)}$ is the weight matrix to be trained.

Then, the weight vector e_t^k is normalized through the SoftMax function and multiplied by the initial input sequence x^k. The normalized equation is as follows:

$$\alpha_t^k = softmax(e_t^k) = \frac{exp(e_t^k)}{\sum_{k=1}^{15} exp(e_t^k)} \tag{3}$$

where α_t^k is the normalized attention weight measuring the importance of the k-th input feature at time t. A SoftMax function is applied to e_t^k to ensure that the sum of all attention weights is 1. The spatial attention mechanism is a feed-forward network that can be trained jointly with other components of the RNN. With these attention weights, the initial input sequence can be adaptively extracted with:

$$\tilde{x_t} = (\alpha_t^1 x_t^1, \alpha_t^2 x_t^2, \ldots, \alpha_t^{15} x_t^{15})^\tau \tag{4}$$

$\tilde{x_t}$ is the new sequence of temporal features processed with the input attention. Compared with the original input $x_t, \tilde{x_t}$ pays more attention to the local connection between the original features and the output, and at the same time suppresses useless information.

Attention-based encoding–decoding layer. The encoder-decoder model is a popular framework in deep learning. Its limitation is that encoders and decoders are only connected by a vector C of a fixed length. However, the vector C cannot completely represent the information for the entire sequence. Also, the information transferred from the previous input to the network is overwritten by later input. The attention-based encoding–decoding model can learn the meaning of each element from the input sequence and provide the decoder with information about the hidden state of each encoder. This allows the model to focus on useful parts of an input sequence.

For time series forecasting, given the input sequence $\tilde{x}_t = (\tilde{x}_1, \tilde{x}_2, \ldots, \tilde{x}_T)$, the encoder is applied to learn the mapping from $\tilde{x}_t$ to h_t (at time step t) with:

$$h_t = LSTM(h_{t-1}, \tilde{x}_t) \tag{5}$$

where $h_t \in R^m$ is the hidden state of the encoder at time t, m is the size of the hidden state, and $LSTM(x)$ is the non-linear activation function for capturing long-term dependencies. The structure of an single LSTM cell is shown in Fig. 5.

Each LSTM unit has a memory cell with the cell-state c_t at time t. Three sigmoid gates control access to the memory cell: forget gate f_t, input gate i_t, and output gate o_t. The following equations describes the computational process of an single LSTM:

$$\begin{aligned} f_t &= \delta\left(W_f\left[h_{t-1}; \tilde{x_t}\right] + b_f\right) \\ i_t &= \delta\left(W_i\left[h_{t-1}; \tilde{x_t}\right] + b_i\right) \\ o_t &= \delta\left(W_o\left[h_{t-1}; \tilde{x_t}\right] + b_o\right) \\ \tilde{c_t} &= tanh\left(W_c\left[h_{t-1}; \tilde{x_t}\right] + b_c\right) \\ c_t &= f_t \times c_{t-1} + i_t \times \tilde{c_t} \\ h_t &= o_t \times tanh(c_t) \end{aligned} \tag{6}$$

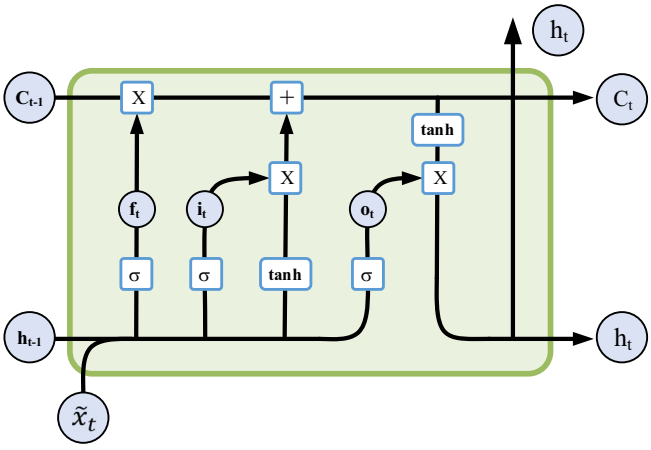

Fig. 5. Structure of an single LSTM cell.

where $\tilde{x}_t$ is the input from the Input-attention layer at time t, h_t is the hidden state at time t.i_t, f_t and o_t represent the input gate, forget gate, and output gate, respectively; c_t is the cell state; $\tilde{c}_t$ is the candidate cell state; and $W_f, W_i, W_o, W_c \in R^{m(m+n)}$ and $b_f, b_i, b_o, b_c \in R^m$ are the parameters to be learnt. The key reason for using an LSTM unit is that the cell state sums activities over time, which can overcome the problem of vanishing gradients and capture long-term dependencies of time series in better.

To predict the output y_n, we use another LSTM neural network to decode the encoded input. Following the encoder with spatial attention, a temporal attention mechanism is used in the decoder to adaptively select relevant hidden encoder states across all time steps. Specifically, the new attention weight of each hidden encoder state at time t is calculated based on the decoder state $state'_{n-1}$ and hidden state of the encoder LSTM unit h_t with:

$$\begin{aligned} l_n^t &= f(state'_{n-1}, h_t) \\ &= v_l^\tau tanh(W_l[h'_{n-1}; c'_{n-1}] + U_l h_t + b_l) \\ &t \in (1,T), n \in (1,N) \\ \beta_n^t &= softmax\left(l_n^t\right) = \frac{exp(l_n^t)}{\sum_{t=1}^{10} exp(l_n^t))} \end{aligned} \tag{7}$$

where $state'_{n-1}$ includes the hidden state h'_{t-1} and cell-state c'_{t-1} of the decoder LSTM unit. v_l^τ, W_l,and U_l are the weight matrixes; and b_l is the bias terms. The attention weight β_n^t represents the importance of the $t-th$ hidden encoder state for forecasting. Since each hidden encoder state h_t is mapped to a temporal component of the input, the attention mechanism computes the vector C_n as a weighted sum of all the hidden encoder states, $(h_1, h_2, \ldots, h_{10})$.

$$C_n = \sum_{t=1}^{T} \beta_n^t h_t \tag{8}$$

Generally, C_n is directly used as the input of the decoder in the traditional encoder-decoder model. In the energy field, considering that the energy consumption at time t is usually related to the consumption at time $t-1, C_n, y_{n-1}$ and h'_{n-1} are performed as nonlinear transformations based on neural network, where y_{n-1} is the predicted value at time $t-1, h'_{n-1}$ is the hidden state in the decoder, and the result C'_n is the input of the decoder for decoding.

Similar to the encoder process, LSTM computes the mapping from an input dataset C'_n to an output series h'_t with three different gates. The following equation describes the mapping process:

$$h'_t = LSTM(h'_{t-1}, C'_n) \tag{9}$$

In the above equation, h'_t can be updated as:

$$\begin{aligned} f'_t &= \delta\left(W'_f[h'_{t-1}; C'_n] + b'_f\right) \\ i'_t &= \delta(W'_i[h'_{t-1}; C'_n] + b'_i) \\ o'_t &= \delta(W'_o[h'_{t-1}; C'_n] + b'_o) \\ \tilde{c}'_t &= tanh(W'_c[h'_{t-1}; C'_n] + b'_c) \\ c'_t &= f'_t \times c'_{t-1} + i'_t \times \tilde{c}'_t \\ h'_t &= o_t \times tanh(c'_t) \end{aligned} \tag{10}$$

where $[h'_{t-1}; C'_n]$ is a concatenation of the hidden state h'_{t-1} and the decoder input C'_n. W'_f, W'_i, W'_o, W'_c and b'_f, b'_i, b'_o, b'_c are the parameters to be trained; and $\delta(x)$ and $tanh(x)$ are a logistic sigmoid function and hyperbolic function, respectively.

$$\begin{aligned} y_n &= F(y_1, \ldots, y_{n-1}, x_1, \ldots, x_T) \\ &= v_y^T(W_y[h'_n; C'_n] + b_w) + b_v \end{aligned} \tag{11}$$

where $[h'_n; C'_n]$ is a concatenation of the hidden decoder state and connection vector. The parameters, W_y and b_w, map the concatenation to the size of the hidden decoder states. The linear function with weights v_y^T and b_v produces the final forecasting result.

5. Evaluation and case study

Since the temporal prediction in this paper is based on time series prediction, this section examines the effectiveness and robustness of the clustering-based LSTM. In addition, to investigate the superiority of the proposed model in electricity consumption forecasting, the results from the EDA-LSTM and other transition models (i.e., machine learning models, standard LSTM, single-attention LSTM, and dual-attention LSTM) are obtained. The results are compared and discussed below.

5.1. Experimental settings

Datasets. The following experiments use the daily electricity consumption data of three years, approximately 10,000 households, from the Pudong district of Shanghai. The first 800 days of data are for training, the next 100 days for model validation, and the last 100 days as a test set for model performance evaluation.. As discussed earlier, weather is the main factor affecting short-term electricity consumption. Therefore, 15 weather parameters are selected as the features for model training.

Comparison with classic approaches. In previous studies, many comparisons were conducted between standard LSTM and other energy forecasting algorithms. For example, Fan [38] used conventional recurring units (i.e., GRU and LSTM) for model development and compared the performances for energy forecasting. There were two benchmark models presented in the work [39] to estimate medium to the long-term wind and photo-voltaic electricity consumption, the persistence model based on LSTM, and the support vector regression (SVR) model. Wei [40] proposed a hybrid algorithm based on LSTM to predict future gas consumption and compared it with SVR and back-propagation neural network algorithms. Three machine learning algorithms were developed for the comparison in that research: SVR, random forest regression, and Ada-Boost regression. In general, deep learning algorithms work better for large datasets than machine learning methods. Therefore, in this paper, we will focus on the comparison against the standard LSTM model and its variants.

Parameters and evaluation metrics. An Adam optimizer [34] was used to train our model. According to the experiments in [35], the batch size and epochs were set to 128 and 5,000, respectively. To avoid gradient explosion, we set the maximum gradient to 2; the dropout rate of each LSTM function was set to 0.3 to avoid overfitting.

Since the proposed EDA-LSTM model is smooth and differentiable, it can be trained via a back-propagation algorithm [36]. To quantify the performance of the models, we used the following three metrics: mean absolute error (MAE), mean absolute percentage error (MAPE), and root mean squared error (RMSE) [37]. They are defined as follows:

$$\begin{aligned} MAE &= \frac{1}{N}\sum_{i=1}^{N}|y_i^* - y_i| \\ MAPE &= \frac{1}{N}\sum_{i=1}^{N}\left|\frac{y_i^* - y_i}{y_i}\right| \times 100\% \\ RMSE &= \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i^* - y_i)^2} \end{aligned} \tag{12}$$

where N is the length of the time series, y_i is the actual data, and y_i^* is the forecasted value.

5.2. Results and analysis

The derivation process of our model first adds spatial attention to the standard LSTM, then adds temporal attention to generate a dual attention LSTM, and finally integrates the embedding mechanism to create the proposed embedding/dual attention LSTM.

Forecast of urban long-term electricity demand. Table I shows the average forecast errors for the three machine learning models and four LSTM models. The forecasting of each LSTM model was repeated 10 times to obtain the average of the forecast errors. The predicted results of the four LSTM models and real value curves are shown in Fig. 6.

In Fig. 6, the yellow, green, red, and blue lines represent the forecast results of the four LSTM models, while the black line is real data. All the results are from 16 March 2018 to 23 June 2018, when the energy usage rose as the temperature increased. As shown in Fig. 6, the standard LSTM model can only predict the trend towards increasing electricity consumption, but it does not perform well in detail, while the hybrid (EDA-LSTM) model shows good accuracy in both stable and fluctuating periods.

According to Table 1, the deep learning algorithm performs better than the machine learning algorithms in general, due to its ability to learn high-level features from data in an incremental manner. For the four deep learning algorithms, the MAPE of the standard LSTM model is 5.91%, while the hybrid (EDA-LSTM) model is 4.57%. Compared to the standard LSTM model, the MAPE value of the hybrid model decreases by 25.9%. The advantage of the hybrid model can be explained by the reasons as below.

There are three innovative mechanisms in the hybrid model: embedding, spatial attention, and temporal attention. The main task of the embedding layer is to identify the distribution of categorical characteristics. This provides information on the relationship between categorical characteristics, which are, otherwise, very difficult to understand. As the results above show, the embedding mechanism improves the hybrid model only slightly, as only four of the fifteen features are categorical. If more categorical features were collected, embedding would further improve the hybrid model.

The input attention mechanism adaptively selects the respective driving series, i.e., at time step t, the input characteristics relevant for the predicted value are searched, and weighted appropriately during operation. In summer, for example, the influ-

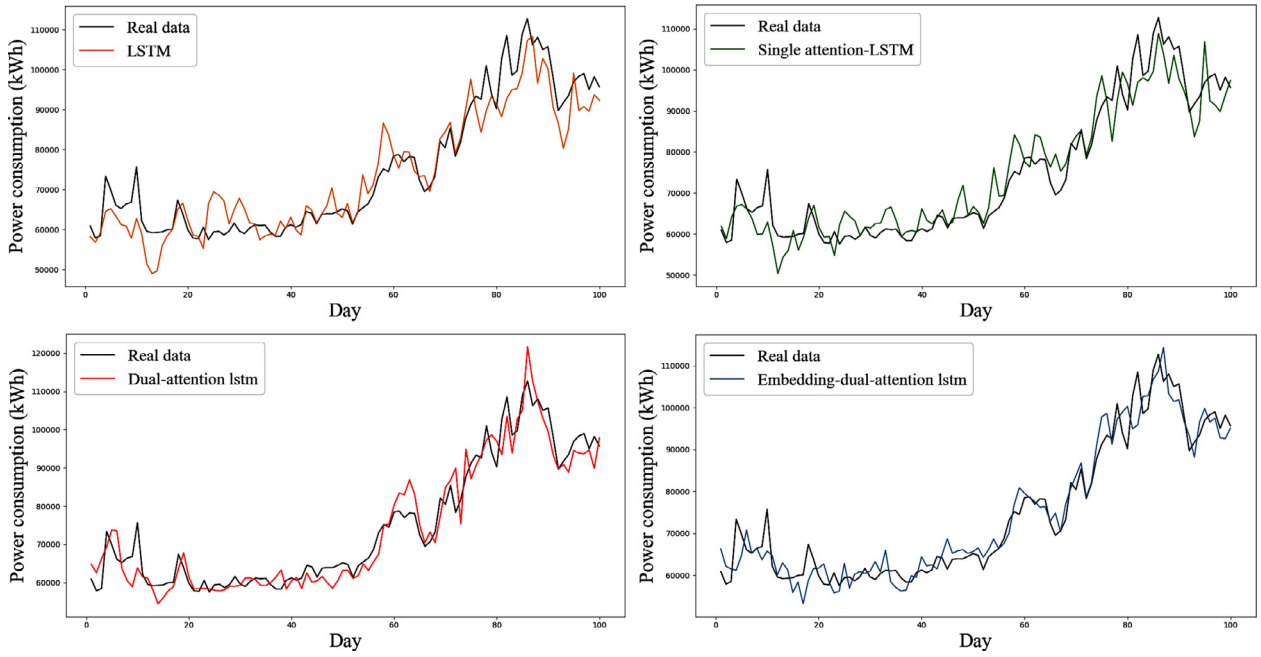

Fig. 6. Prediction results from the four deep learning models.

Table 1
Average forecasting errors for different regression models. In this paper, four traditional machine learning regression models and four variants of the LSTM algorithm are selected for the comparison. The results show that the EDA-LSTM model has the lowest error rate, which canbe more suitable for long-term prediction.

Forecast period	Indicators	Support Vector Regression	RandomForest Regression	AdaBoost Regression	LSTM	Single Attention LSTM	Dual Attention LSTM	EDA-LSTM
30 days	MAE	22295.55	24447.46	28071.37	3853.58	3609.19	3284.69	2660.89
	RMSE	24781.69	27538.85	31184.33	4873.49	4625.17	4154.61	3148.23
	MAPE	17.96%	19.38%	22.29%	5.64%	5.34%	4.18%	4.39%
60 days	MAE	17385.83	18636.19	21078.68	3831.58	3438.71	3719.63	2879.47
	RMSE	21239.38	23137.55	26134.22	5181.99	4385.36	4178.83	3776.15
	MAPE	14.47%	15.18%	17.12%	6.01%	5.81%	4.44%	4.42%
100 days	MAE	14267.98	13641.09	14895.94	4253.34	4045.75	3351.99	3111.55
	RMSE	17936.49	18579.77	20811.23	5645.73	5149.94	4216.06	4015.04
	MAPE	14.02%	12.48%	13.34%	8.09%	5.91%	4.89%	4.57%

ence of temperature on electricity consumption is higher than that any other characteristic.

The temporal attention mechanism can capture long-term information for coded inputs, helping the model to automatically determine how characteristics affects energy consumption during the whole time period. More precisely, the time step with the most significant impact on future consumption receives the most attention. For example, the months of February 2015 and February 2016 deserves more attention in terms of forecasting consumption for 1 February 2017 because they are in the same season.

In addition, in the four deep learning algorithms, the prediction error accumulates continuously with the extension of the prediction period, which is reflected by the MAE value of the LSTM model (see Table 1). However, the cumulative effect of the error is the weakest in the EDA-LSTM model, meaning that the EDA-LSTM model is more resistant to the cumulative effect of the error than the ordinary LSTM model (see Fig. 7). The scatter plot in Fig. 7 depicts the distribution of MAE values for the four deep learning algorithms for the prediction of 100 days. The four lines in the figure correspond to the linear fitting of the MAE values of the four deep learning models. Among them, the MAE line of the EDA-LSTM model is the stablest, reflecting its stability in long-term power demand forecasts.

Forecast of different energy consumption patterns. To study the prediction effectiveness of different prediction algorithms on different patterns, the data from 2016 is used as a training set to predict the electricity consumption for the whole year of 2017. In Fig. 8, the horizontal axis represents 365 days of 2017. It can be seen that the peak electricity consumption is in July and August for both bimodal patterns. The difference between the two patterns is that there is another peak in December and January for the winter pattern. In addition, the pattern of high energy consumption also has the bimodal characteristics. The prediction results of the LSTM algorithm in different patterns are shown in Fig. 8, and the prediction accuracy of all methods are shown in Table 2.

According to the MAPE in Table 2, the prediction results of the bimodal model are the best. The MAPEs for the summer and winter models are 4.17% and 4.23%, respectively. The reason is that customers with these two patterns possess relatively stable consumption behavior, which is mainly influenced by temperature. For example, the peak consumption of customers in the summer pattern is concentrated in July and August, with a peak of about 180 kWh per day. For the rest of the year, their daily consumption remains relatively low and stable (about 80 kWh per day). For customers with the winter pattern, their peak electricity consumption

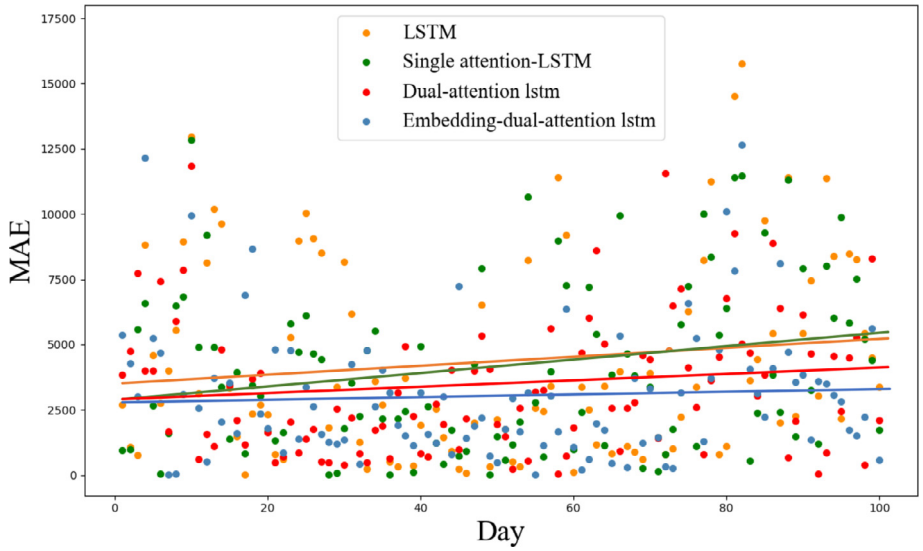

Fig. 7. MAE distribution of the forecasting models.

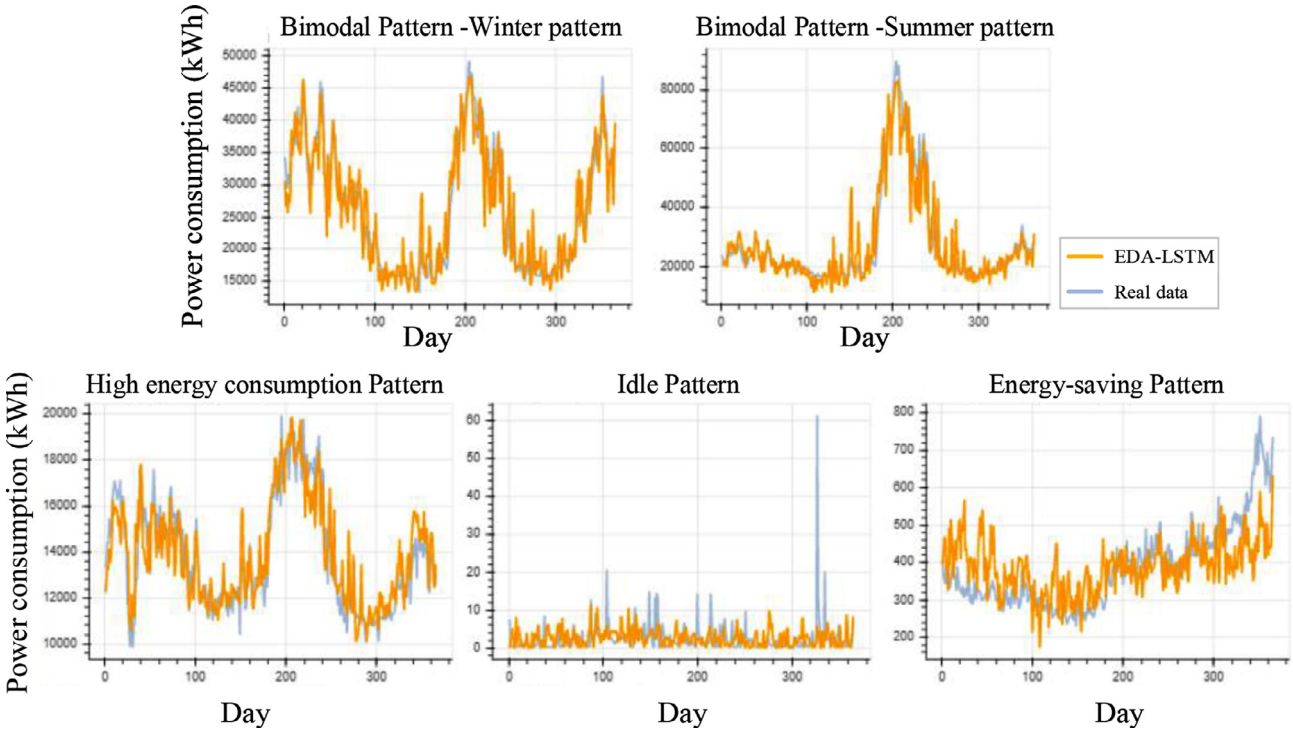

Fig. 8. Prediction results of EDA-LSTM model in different patterns.

usually occurs between December and January, with peaks around 200 kWh per day. Meanwhile, these customers still maintain high daily consumption (around 180 kWh) in the middle of July and August.

The customers in high power consumption mode are characterized by constant high consumption throughout the year, and usually with more than 1000 kWh of daily consumption. Their dailyconsumption fluctuates dramatically, thus reflecting a stronger discrete distribution, as shown in Fig. 2. From the perspective of time-series forecasting, the electricity consumption of these customers will be influenced by many other factors besides the climate factors, such as the enriching life and the preference for high-power appliances, etc, resulting in relatively low prediction accuracy.

The estimation errors of the idle pattern and energy-saving pattern are the highest. This result can be attributed to the fact that although the electricity consumption behaviors of these customers are relatively stable (daily electricity consumption usually ranges from 0 to 10 kWh), their consumption behaviors are almost completely unaffected by weather temperature, which means that their electricity consumption data are highly randomized. Besides, from the prediction perspective, there is a large number of zeros in the energy data (labels) of the two patterns of customers, while most of the data with zero consumption correspond to different

Table 2
Comparison of accuracy of different algorithms in different patterns.

		MAE	RMSE	MAPE
Bimodal pattern winter	Support Vector	3724.6	5033.2	23.3%
	Random Forest	3349.4	5143.1	12.1%
	AdaBoost	3666.2	5364.7	16.4%
	LSTM	3197.4	4910.9	4.6%
	EDA-LSTM	3119.5	4177.7	4.23%
Bimodal pattern summer	Support Vector	5700.0	7912.2	22.7%
	Random Forest	3874.0	6389.3	12.5%
	AdaBoost	6274.8	8839.4	12.3%
	LSTM	5871.8	3874.0	4.3%
	EDA-LSTM	3718.3	5871.8	4.17%
High energy consumption pattern	Support Vector	1759.9	2049.2	21.9%
	Random Forest	1829.7	2131.4	15.7%
	AdaBoost	1667.8	1878.2	13.0%
	LSTM	1511.9	1752.1	9.7%
	EDA-LSTM	1483.1	1740.2	7.6%
Low energy consumption pattern	Support Vector	380.2	404.2	50.3%
	Random Forest	382.5	405.2	51.9%
	AdaBoost	389.5	416.1	51.6%
	LSTM	374.5	398.9	51.2%
	EDA-LSTM	342.7	367.2	50.1%
Idle pattern	Support Vector	32.6	39.7	86.4%
	Random Forest	33.7	40.9	90.7%
	AdaBoost	33.3	40.5	87.9%
	LSTM	33.5	40.7	86.1%
	EDA-LSTM	33.3	40.5	86.1%

features (weather data), which is disastrous for the training process of LSTM. Fortunately, the energy consumption of the two patterns accounts for a tiny proportion of the total. For example, the energy consumption of the idle pattern and the energy-saving pattern only account for 0.0033% and 0.54% of total energy consumption in 2017, respectively. Therefore, the impact of the two patterns on the total prediction error is limited.

We also found that aggregating the forecast results of all patterns yields a better forecast for the aggregation level, compared with the conventional strategy of directly forecasting the aggregated load. This is because the temporal relationship of individual patterns is more distinct and the signal-to-noise ratio is higher.

6. Conclusions and future work

Considering the unsatisfactory accuracy of the standard LSTM forecasting model and the difficulty of encoding categorical characteristics in the energy domain, this paper proposes a hybrid energy demand forecasting model based on a dual attention mechanism to predict future energy demand. Compared to the standard LSTM model, the proposed forecasting model offers clear advantages in terms of predictive accuracy and robustness of mid- to long-term forecasts; the model can thus be applied to more complex time series data when forecasting energy consumption.

In addition, we also found that the temporal attention mechanism performed better with sequence data than other models, meaning that it may be more suitable for data with sequencial correlations. The spatial attention mechanism can automatically select features with the most significant impact on the results, so it is appropriate for processing larger numbers of input features. Furthermore, when there are categorical attributes in the input features, the embedding mechanism can improve the reliability of the algorithm.

Finally, through cluster analysis, we found that the total electricity consumption in Pudong can be divided into several specific patterns. Each pattern can be described with a specific temporal relationship. The ratio of the number of customers with different patterns was related to the local climate, economy, and social customs. This also is significant for further research on electricity consumption issues.

Even the most sophisticated models cannot explain or predict reality in its entirety because the factors in real phenomena are infinite [41]. Among the three factors influencing energy demand, this study only forecast the electricity consumption of the Pudong district based on weather conditions, without considering regional differences and economic indices. Therefore, the next step is to collect more consumption data and economic indicators over a more extended period to analyze the impact of economic changes on energy demand.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgment

The research is partially supported by National Key R&D Program of China (No.2017YFE0101400), National Natural Science Foundation of China (No.61802278), European Union's Horizon 2020 research and innovation program under the Marie Sklodowska-Curie grant agreement (No.754462) and the Flexible Energy Denmark project (FED) funded by Innovationsfonden (8090-00069B). This article is also sponsored by the China scholarship council, and the Flexibility for Smart Urban Energy Systems project (FlexSUS) (91352) funded by the European Union's Horizon 2020 research and innovation programme under grant agreement No 775970.

References

[1] Z. Niu, J. Wu, X. Liu, L. Huang, P.S. Nielsen, Understanding energy demand behaviors through spatio-temporal smart meter data analysis, Energy 226 (2021) 120493.
[2] T. Zhang, R. Ramakrishnan, M. Livny, BIRCH: an efficient data clustering method for very large databases, ACM sigmod record 25(2), 1996, pp. 103–114.

[3] Y. Liang, S. Ke, J. Zhang, X. Yi, Y. Zheng, Geoman: Multi-level attention networks for geo-sensory time series prediction, IJCAI (2018).
[4] D. Bahdanau, K. Cho, Y. Bengio, Neural machine translation by jointly learning to align and translate, Computer Science (2014).
[5] N. Weia, C. Lia, X. Peng, et al., Daily natural gas consumption forecasting via the application of a novel hybrid model, Applied Energy 250 (2019) 358–368.
[6] P. Robinson, Modeling utility load and temperature relationships for use with long-lead forecasts, Journal of Applied Meteorology and Climatology 1997 36:591–8.
[7] Y. Yan, Climate and residential electricity consumption in Hong Kong, Energy 23 (1) (1998) 17–20.
[8] A. Pardo, V. Meneu, E. Valor, Temperature and seasonality influences on Spanish electricity load, Energy Econ 24 (2002) 55–70.
[9] M. Sforna, Searching for the electric load-weather temperature function by using the group method of data handling, Electric Power Systems Research 32 (1995) 1–9.
[10] S. Islam, S. Al-Alawi, K. Ellithy, Forecasting monthly electric load and energy for a fast-growing utility using an artificial neural network, Electric Power Systems Research 34 (1995) 1–9.
[11] N. Wei, J. Liu, F. Zeng, Daily Natural gas load forecasting based on a hybrid deep learning model, Energies 12 (2) (2019) 218.
[12] C. Wright, C.W. Chan, P. Laforge, Towards developing a decision support system for electricity load forecast, Decision Support Systems (2012).
[13] F. Taspinar, N. Celebi, N. Tutkun, Forecasting of daily natural gas consumption on regional basis in Turkey using various computational methods, Energy and Buildings 56 (2013) 23–31.
[14] P. Zhang, H. Wang, Fuzzy wavelet neural networks for city electric energy consumption forecasting, Energy Procedia 17 (2012) 1332–1338.
[15] A.D.P. Lotufo, C.R. Minussi, Electric power systems load forecasting: A survey. In *PowerTech Budapest 99*, Abstract Records (1999) 36.
[16] B. Soldo, Forecasting natural gas consumption, Applied Energy 92 (2012) 26–37.
[17] A. Tascikaraoglu, O. Erdinc, M. Uzunoglu, A. Karakas, An adaptive load dispatching and forecasting strategy for a virtual power plant including renewable energy conversion units, Applied Energy 119 (2014) 445–453.
[18] F. Magoulès, H.X. Zhao, *Data mining and machine learning in building energy*, in: *analysis*, John Wiley & Sons, 2016.
[19] B. Kermanshahi, H. Iwamiya, Up to year 2020 load forecasting using neural nets, Electrical Power Energy System 24 (2000) 789.
[20] O.E. Canyurt, H. Ceylan, H.K. Ozturk, A. Hepbasli, Energy demand estimation based on two-different genetic algorithm approaches, Energy Sources 26 (14) (2004) 1313.
[21] F. Egelioglu, A.A. Mohamad, H. Guven, Economic variables and electricity consumption in Northern Cyprus, Energy 26 (2001) 355.
[22] J.L. Harris, L.M. Liu, Dynamic structural analysis and forecasting of residential electricity consumption, International Journal of Forecasting 9 (4) (1993) 437–455.
[23] H.G. Lakhani, B. Bumb, Forecasting demand for electricity in Maryland: an econometric approach, Technological Forecasting and Social Change 11 (1978) 237.
[24] T.K. Gautam, K.P. Paudel, Estimating sectoral demands for electricity using the pooled mean group method, Applied Energy 231 (2018) 54–67.
[25] Liu X. Q., Ang B. W., Goh T. N., Forecasting of electricity consumption: a comparison between an econometric model and a neural network model. In *[Proceedings] 1991 IEEE International Joint Conference on Neural Networks 1991, pp.* 1254-1259..
[26] M.R. Gent, Electric supply and demand in the United States: Next 10 years, IEEE Power Engineering Review 12 (4) (1992) 8–13.
[27] Y. Minato, Y. Yokoi, Development of a forecasting method of a region's electric power demand (1), IEEJ Transactions on Power and Energy 116 (2) (1996) 147–154.
[28] P. Leung, W. Miklius, Accuracy of electric power consumption forecasts generated by alternative methods: the case of Hawaii, Energy Sources 16 (3) (1994) 289–299.
[29] M. Ranjan, V.K. Jain, Modelling of electrical energy consumption in Delhi, Energy 24 (1999) 351.
[31] X. Liu, Z. Niu, L. Yang, J. Wu, D. Cheng, X. Wang, VAP: a visual analysis tool for energy consumption spatio-temporal pattern discovery, in: Proceedings of the International Conference on Extending Database Technology, 2020, pp. 579–582.
[33] S. Ioffe, C. Szegedy, Batch normalization: accelerating deep network training by reducing internal covariate shift, International Conference on Machine Learning (2015), https://doi.org/JMLR.org.
[34] D. Kingma, J. Ba, Adam: A method for stochastic optimization (2014), arXiv: 1412.6980.
[35] Y. Qin, D. Song, H. Chen, W. Cheng, G. Jiang, G. Gottrell, A Dual-Stage Attention-Based Recurrent Neural Network for Time Series Forecasting (2017), https://doi.org/arXiv preprint arXiv:1704.02971.
[36] D.E. Rumelhart, G.E. Hinton, R.J. Williams, Learning representations by back-propagating errors, nature 323 (6088) (1986) 533–536.
[37] I.P. Panapakidis, A.S. Dagoumas, Day-ahead natural gas demand forecasting based on the combination of wavelet transform and ANFIS/genetic algorithm/neural network model, Energy 118 (2017) 231.
[38] C. Fan, J. Wang, W. Gang, S. Li, Assessment of deep recurrent neural network-based strategies for short-term building energy forecastings, Applied Energy 236 (2019) 700–710.
[39] S. Han, Y.H. Qiao, J. Yan, Y. Liu, L. Li, Mid-to-long term wind and photovoltaic power generation forecasting based on copula function and long short term memory network, Applied Energy 239 (2019) 181–191.
[40] N. Wei, C. Lia, X. Peng, Y. Li, F. Zeng, Daily natural gas consumption forecasting via the application of a novel hybrid model, Applied Energy 250 (2019) 358–368.
[41] S. Makridakis, S.C. Wheelwright, Forecasting Methods for Management, 5th ed., Wiley, New York, 1989.
[42] X. Liu, P.S. Nielsen, Scalable prediction-based online anomaly detection for smart meter data, Information Systems 77 (2018) 34–47.
[43] Iftikhar N., Liu X., Danalachi S., Nordbjerg F. E., Vollesen J. H., October. A scalable smart meter data generator using spark. In OTM Confederated International Conferences "On the Move to Meaningful Internet Systems" (2017) pp. 21-36, Springer, Cham.