energies

MDPI

Article

Time Series Forecasting with Multi-Headed Attention-Based Deep Learning for Residential Energy Consumption

Seok-Jun Bu and Sung-Bae Cho *

Department of Computer Science, Graduate School of Artificial Intelligence, Yonsei University, Seoul 03722, Korea; sjbuhan@yonsei.ac.kr

* Correspondence: sbcho@yonsei.ac.kr

Received: 7 June 2020; Accepted: 2 September 2020; Published: 10 September 2020

Abstract: Predicting residential energy consumption is tantamount to forecasting a multivariate time series. A specific window for several sensor signals can induce various features extracted to forecast the energy consumption by using a prediction model. However, it is still a challenging task because of irregular patterns inside including hidden correlations between power attributes. In order to extract the complicated irregular energy patterns and selectively learn the spatiotemporal features to reduce the translational variance between energy attributes, we propose a deep learning model based on the multi-headed attention with the convolutional recurrent neural network. It exploits the attention scores calculated with softmax and dot product operation in the network to model the transient and impulsive nature of energy demand. Experiments with the dataset of University of California, Irvine (UCI) household electric power consumption consisting of a total 2,075,259 time-series show that the proposed model reduces the prediction error by 31.01% compared to the state-of-the-art deep learning model. Especially, the multi-headed attention improves the prediction performance even more by up to 27.91% than the single-attention.

Keywords: convolutional recurrent neural network; multi-headed attention; time-series forecasting; energy consumption prediction

1. Introduction

According to the World Energy Outlook 2019, the International Energy Agency (IEA) pointed out that the energy demand will rise by 1.3% each year to 2040 with unrestrained planning by further efforts to improve efficiency [1]. The residential power consumption sector, which is a major factor that accounts for 27% of global electricity consumption [2], provides an ideal testbed for power demand prediction and analysis with a relatively limited and closed environment. The machine learning approach including the deep learning algorithms is convincing as a premise for power supply planning due to its non-linear learning capability [3].

Residential power consumption prediction is defined as a multivariate time series prediction problem [4]. As depicted in Figure 1, the differentiated features from sensor-level signals consisted of energy consumption attributes are extracted to predict the power consumption levels with the prediction model [5]. The process of predicting the future power demand from the historical power consumption with the other power attributes is essential in the field of energy management system (EMS) including the recently noted smart grid services. The main issue at the planning stage of the plan-do-check-act operation cycle [6] in traditional EMS is modeling the nature of energy consumption, which is limited to naive time-series models such as linear regression.

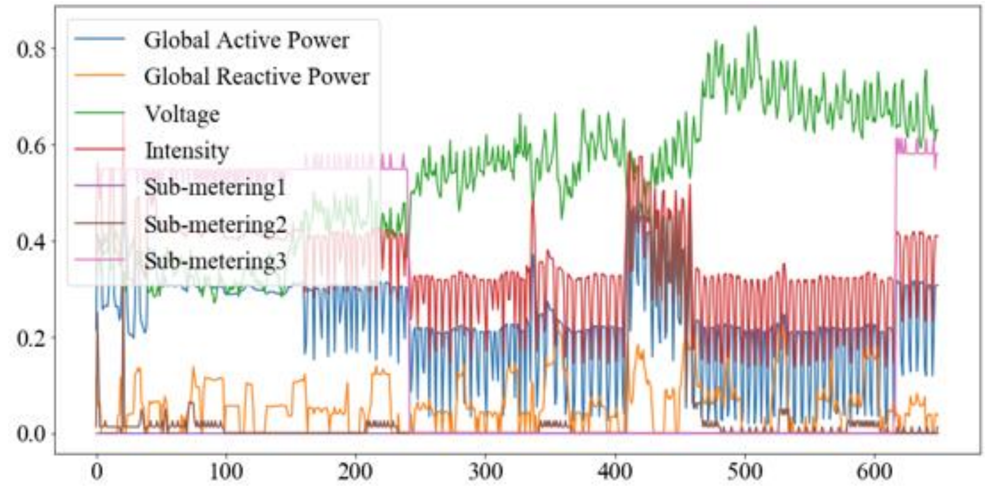


Figure 1. A typical problem formulation and the hyperparameters of the energy consumption prediction task with the resolution of the time and the time lag parameter.

In the domain of energy, machine learning aims at exploiting a database generated with the historical data of all the clients and predicting the future energy demand. Various regression techniques, including autoregressive integrated moving average [7] and support vector regressor [8], which are representative methods for the energy prediction, successfully model the long-term behavior of end-users and improve the prediction accuracy. Yet, in the prediction task under high-temporal resolution conditions [9] aimed at modeling the short-term behaviors that expose the behavioral patterns, a more complex and practical time series modeling method is required.

The power prediction models based on deep learning which achieved the highest performance so far [10], on the other hand, encounter two major hurdles: One is the multicollinearity between active power consumption patterns and other power attributes [11,12], and the other is the transient and impulsive nature of power consumption mainly occurring from the usage of electronic products. Although the convolution operation to learn the filter to extract the local correlation has been devised, the simultaneous or exclusive usage of power consuming facilities that have a critical effect on the active power cause the performance degradation of the deep learning models. Specifically, it is known that the state-of-the-art models that do not focus on modeling the multivariate and impulsive properties have their performance degradation in energy peaks.

The dataset of UCI household electric power consumption consisting of a total 2,075,259 time-series and seven variables is a representative example of the typical difficulty of modeling between the active power and the power consuming facilities. Figure 2 shows the correlation between the active, reactive power, voltage, intensity and three sub metering attributes. With the color of the graph is shown the active power level grouped into three stages, the existing deep learning-based models have the limitation to hardly concentrate the control facility to predict power consumption.

In this paper, we propose a deep learning model based on multi-headed attention to model the local connectivity between electric attributes and active power with the learnable feature extraction of convolution and weighting mechanism in time-series modeling [13,14]. The key idea is to extract the features from the multivariate electric attributes with the convolution operation, perform time-series modeling of the power spectrum with the gating operation, and effectively predict the transient and impulsive values of power consumption with multi-headed attention that performs probabilistic data localization based on the softmax operation. Taken together, our hypothesis of the proposed method is that the weighting function of attention can model the short-term patterns of time-series electricity data including the energy peaks. We will show that the multi-headed attention selectively extracts the power consumption patterns to cope with the challenge mentioned previously. To the best of our knowledge, this is the first attempt that multi-headed attention is incorporated to predict power consumption. The main findings of this research can be summarized as follows:

- The multi-headed attention works well for modeling the short-term patterns of time-series data, resulting in the best deep learning model for predicting the energy demand.
- The class activation map appropriately visualizes how the proposed method forecasts the energy demand from the time-series data.

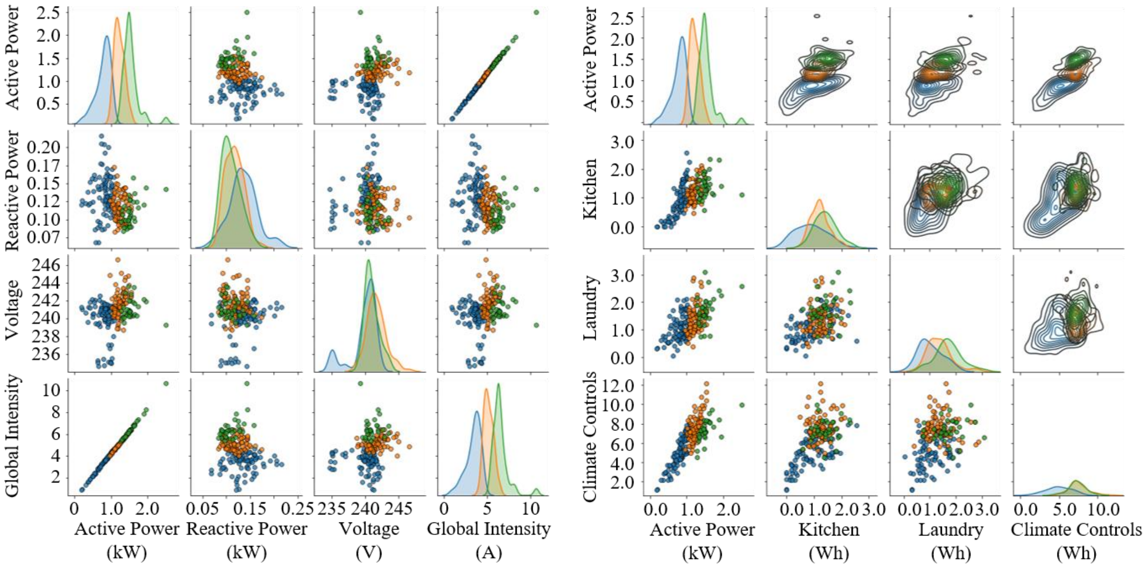


Figure 2. Results of multivariate correlation of the energy attributes by the consumption level.

The remainder of this paper is organized as follows. In Section 2, we review the previous power consumption models based on deep learning and clarify the contributions of this paper by discussing the differences between them. In Section 3 we illustrate how the electric attributes are selectively extracted by the deep learning model with multi-headed attention. The performance of the model is evaluated in Section 4 through various experiments, including the visualization of multi-headed attention vectors and comparison with recent models based on deep learning. Finally, Section 5 concludes the paper with some discussion of future directions.

2. Related Works

In this section, we review the relevant models based on deep learning for forecasting energy consumption. According to the similarities of the fields and the techniques used, we present the traditional signal processing methods as well as the power prediction studies based on deep learning. Table 1 summarizes the significant studies of the last five years on predicting power consumption in terms of the feature extraction and time series modeling. Most of the methods before the inception of deep learning focused on the issue of time series modeling based on the symbolic-dynamic approach [15,16]. Due to the limitations that slight shifts along the time axis causing a large distance between the two time series, Lin et al. extracted and modeled the rotation-invariant symbols and constructed the bag-of-patterns [17].

Meanwhile, the superiority claims by the power demand models based on machine learning encountered a major hurdle: the methods are evaluated for short-term forecasting horizons and do not consider medium and long-term ones [7]. In order to build time-invariant features and perform the non-linear mapping for predicting the power consumption, Tso and Yau presented a neural network and compared the performance with existing prediction methods based on the rules and symbols [18]. Among the power demand forecasting methods based on machine learning such as the autoregressive integrated moving average (ARIMA) [7] and decision tree [8], the neural network achieved the best performance, and its non-linear mapping capability attracted much attention [19]. In particular, combining the approach of machine learning algorithms like the ensemble of recurrent neural network and support vector regressor [20] improved the accuracy and the stability of power demand prediction.

Table 1. Related works on the residential energy consumption prediction based on deep learning.

Time-Scale	Author	Domain/Data	Forecasting Method	Prediction MSE (Time-Scale)
Short (30-min or 1-h)	Mocanu [21]	UCI Household Power	Factored Conditional Restrict Boltzmann Machine	0.6211 (1-h)
	Marino [22]	UCI Household Power	Sequence-to-sequence LSTM	0.6420 (1-h)
	Li [20]	New South Wales and Victorian Australia	RNN-SVR Ensemble	0.4059 (1-h)
	Kong [23]	Smart Grid Smart City (SGSC) Australia	RNN, LSTM	0.2903 (1-h)
	Li [24]	Applied Building Energy	Stacked Autoencoder, Extreme Learning Machine	40.6747 (30-min)
	Makridakis [7]	M3-competition Time-series Data	ARIMA, NN, GP, k-NN, SVR	0.3252 (1-h)
	Rahman [25]	Salt Lake City Public Safety Building	LSTM	0.2903 (1-h)
	Gonzalez-Briones [8]	Shoe Store Power Data Spain	LR, DT, RF, k-NN, SVR	0.4234 (1-h)
	Shah [26]	Nord Pool Electricity Data	Spline Function-based, Polynomial Regression, ARIMA	0.2428 (1-h)
	Fan [27]	Hong Kong Educational Building Operational Data	NN, RNN, LSTM	118.2 (24-h)
	Kim [10]	UCI Household Power	CNN-LSTM	0.2803 (1-h)
	Wang [28]	Irish Customer Behavior Trials, Low Carbon London	Quantile Loss-guided LSTM	0.1552 (1-h)
	Kim [29]	UCI Household Power	CNN-LSTM with Particle Swarm Optimization-based Architecture Optimization	0.3258 (30-min)
Medium (1-month)	Shi [30]	Smart Metering Electricity Customer Behavior Trial (CBT)	Pooling-based RNN	0.4505 (1-month)
	Bouktif [31]	French National Energy Consumption	LSTM with Genetic Algorithm-based Time Lag Optimization	270.4 (1-month)
Long (1-year)	Guo [32]	China Jiangsu Province Power	LSTM, Quantile Regression	594.8 (1-year)

The deep learning models, including the long short-term memory (LSTM) which can learn temporal gating functions and the convolutional neural network (CNN) which can extract local correlation between power spectrums, are making remarkable achievements in the field of energy consumption forecasting and energy pattern classification [10]. Moreover, designed as a probabilistic approach using CNN and LSTM layers as building blocks, autoencoder (AE) [24] and adversarial learning models like generative adversarial network (GAN) [33] indirectly demonstrate the popularity and possibility of deep learning for power consumption prediction.

However, electricity demand forecasting is a difficult task due to the characteristics that demand time-series exhibit. The characteristics include the non-constant mean and variance, calendar effects, multiple periodicities, high volatility, jumps, et cetera. Mocanu et al. introduced a stochastic pretraining stage into the power consumption prediction with a neural network and evaluated it under various temporal conditions [21]. Restricted Boltzmann machine (RBM) is a representative neural network model of unsupervised learning that aims to minimize the divergence of the Kullback–Leibler divergence between the layers [34]. The stacked RBM is suitable for learning the prior distribution from the power consumption data and has been contributed significantly to improving the prediction performance. Marino et al. and Kong et al. attempted to predict the power consumption by using the recurrent neural network (RNN) with LSTM designed for time series modeling, and verified that the feasibility of a neural network with memory cell could significantly improve the prediction performance [22,23].

To account for the different characteristics of the demand-series, recently, researchers suggested optimization methods for forecasting models. The optimization of deep learning architecture or loss function has been performed for the remarkable performance improvement in various domains. Li et al. introduced the autoencoder before the time-series modeling to extract multivariate features [24], and the RNN with pooling operation to selectively update the gradient [30]. A method of optimizing the time lag parameters, a critical factor for prediction performance, by genetic algorithm (GA) was also introduced [31]. Furthermore, the optimization of the power prediction model with particle swarm optimization (PSO), which has faster convergence and larger exploration of searching the space than GA, attracted the attention by outperforming the performance of existing deep learning models [29]. In addition, studies have proceeded to optimize or develop the loss functions, such as adaptation of quantile loss function into neural networks [27,32].

From the relevant studies on deep learning applications to predicting energy consumption, it is obvious that this area serves as a competitive platform for various deep learning techniques. In this paper, we present another model based on the multi-headed attention on top of the most superior model based on CNN-LSTM (convolutional neural network-long short-term memory).

3. Convolutional Recurrent Neural Network with Multi-Headed Attention

In this section, we describe the architecture of the CNN-LSTM network with multi-headed attention that extracts the spatiotemporal features and models the power consumption from the power spectrum. Two major components of the conventional power prediction model are adopted and modified with end-to-end neural network architecture. Here, the multi-headed attention is implemented with softmax and dot product operation to model the transient and impulsive values of electricity demand.

3.1. Structure Overview

The main objective of the method is to forecast the electricity demand using CNN-LSTM network. We incorporate the multi-headed attention mechanism and formulate it with the function $\phi(\cdot)$ that extracts the spatiotemporal features and predicts the future energy demand. Since there is a complex non-linear mapping expressed by stacking the multiple layers, we adopt a direct forecast strategy [35] that avoids the accumulation of bias in the recursive strategy. The direct forecast strategy is formulated with direct model ϕ_h and its parameter Θ_h:

$$\hat{y}_t = \phi_h\left(X^{\omega}_{t-h}; \Theta_h\right) + e_{t,h} \tag{1}$$

where $X^{\omega}_{t-h} = [R_{t-h}, \ldots, R_{t-h-\omega}]$ with time lag ω, $\Theta_h = [\psi_h, \theta_h]$ is the model parameter for horizon h where ψ_h is a set of hyperparameters and θ_h is a set of parameters, and $e_{t,h}$ is the forecast error of the model ϕ_h. The electricity demand forecaster predicts h -step ahead, and we verify the model under various conditions of {1-min, 10-min, ..., 1-day, 1-week} in Section 4. The input element R_t of the sequence of the power attribute vector is composed of global active power, reactive power, voltage, intensity, and three sub-meterings.

It is well known that the convolutional recurrent neural networks have the advantages represented by data-driven filter learning focused on extracting spatiotemporal features in the field of signal processing, including predicting the power consumption [10,36]. Figure 3 illustrates the overall architecture of the CNN-LSTM with the multi-headed attention for predicting power consumption. The proposed model for predicting the future power consumption from the input power data consists of two major stages.

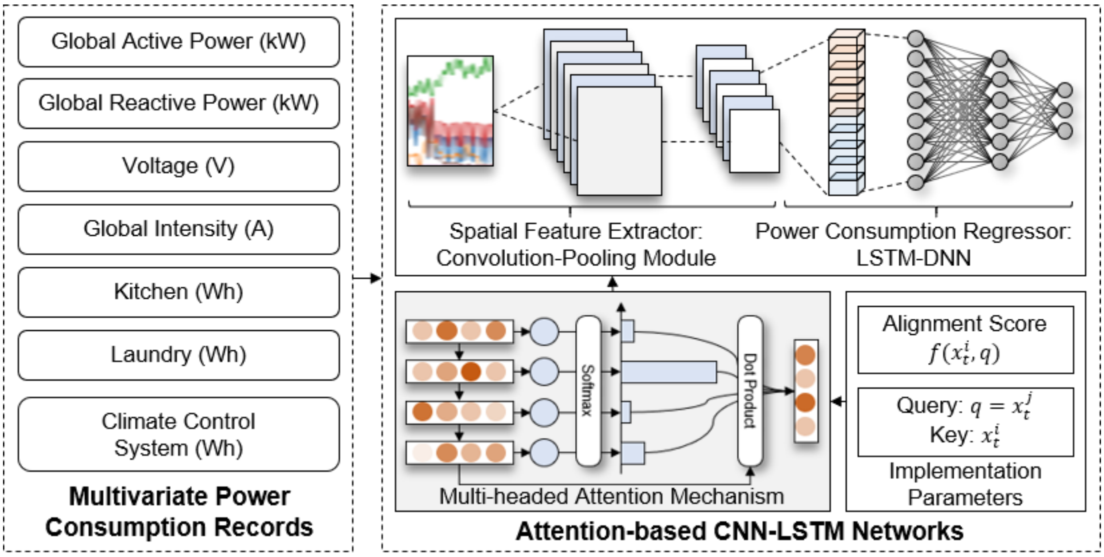

Global Active Power (kW)
Global Reactive Power (kW)
Voltage (V)
Global Intensity (A)
Kitchen (Wh)
Laundry (Wh)
Climate Control System (Wh)
Multivariate Power Consumption Records
Spatial Feature Extractor: Convolution-Pooling Module
Power Consumption Regressor: LSTM-DNN
Softmax
Dot Product
Multi-headed Attention Mechanism
Alignment Score $f(x_t^i, q)$
Query: $q = x_t^j$
Key: x_t^i
Implementation Parameters
Attention-based CNN-LSTM Networks

Figure 3. Overall structure of the proposed CNN-LSTM with multi-headed attention.

First, the data preprocessing stage defines the hyperparameters required for training the CNN-LSTM model. Min–max normalization per each power attribute and the sliding-window, one of the most common preprocessing methods, are performed before defining the hyperparameter. The sliding-window is defined by the length of the input signal ω called time lag with the stride parameter τ that determines the amount of overlapping time steps. The power consumption data are sampled with each period, and the proposed method is verified under the temporal resolution R_t of 1, 15, 30, 45, and 60 min, 1 day, and 1 week, respectively.

Second, the modeling stage adjusts the weights of CNN-LSTM model according to the hyperparameters defined. To extract spatiotemporal features from the power spectrum, we construct a convolution-pooling operation that models the hidden correlations between power attributes [37], and the gating operation applied to a recurrent memory cell that models the temporal relations from time-series data [38]. Meanwhile, the multi-headed attention designed to interpret the activation value as a probability and construct a correlation matrix with itself is implemented as a layer of the CNN-LSTM model to be placed between each convolutional and recurrent layer [39]. In addition, the class activation map (CAM) is put in the last convolutional layer for the analysis of the network outputs, which can localize the receptive field by summing the weights of CNN's top-most feature maps [40].

3.2. Convolutional Recurrent Neural Networks

The major hurdle in modeling the power consumption with neural regressor lies in extracting the spatiotemporal features from the limited consumption samples [41]. We construct the CNN and LSTM for learning the features from the time-series power consumption data. The two deep learning models are combined in a sequential manner, while maintaining the complementary relations from spatiotemporal features.

The convolution $\phi_c(\cdot)$ and the pooling operation in CNNs, which have been successfully applied in the field of the signal processing, are suitable to model the sequence of power consumption and extract the features using local connectivity between windowed signals. The convolution operation is known to reduce the translational variance between features [42] and preserves the spatial relationship between power attributes by learning filters to extract the hidden correlations. Given the t-th time step, the sequence of the convolutional operation is applied using $m \times 1$ sized filter W with the a-th node in the l-th layer and τ-th element in sequence of the power attributes R_τ:

$$\phi_c^l(X_t) = \left[\sum_{a=0}^{m-1} W_a^{l-1} R_\tau \right]_{\tau=1}^{t-\omega} \tag{2}$$

Because the dimension of the output vector that has been distorted and copied by the convolution operation $\phi_c(\cdot)$ is increased by the number of convolution filters, the summary statistic from nearby node activations is extracted from $\phi_p(\cdot)$ by a max-pooling operation. Pooling refers to a dimension reduction process used in CNN in order to impose the capacity bottleneck and facilitate faster computation [43]. The max-pooling operation has effects on feature selection and dimensionality reduction under $k \times 1$ sized area with pooling stride. The proposed 1D convolution-pooling operation aims to extract the spatial features from the power spectrum per attributes and deliver the series of encoded vectors to the following LSTM. The spatial features extracted by convolution-pooling function ϕ_c contain the time-series information of the window size ω according to the sliding-window preprocessing. The key idea of LSTM is adapting the gating operation which is composed of input gate, forget gate, and output gate o_t and producing the encoded vector $\phi_L(\cdot)$ with the cell state c_t and the hidden value h_t at the time step t:

$$\phi_L(\cdot) = h_t = o_t \circ tanh(c_t) \tag{3}$$

where $\circ$ denotes the element-wise product and b denotes bias term. After the spatiotemporal features are extracted by CNN-LSTM, the typical multi-layer perceptron (MLP) is used to complete the regression function $\phi(\cdot)$ with activation function σ and weight matrix W^l:

$$\phi^l(X_t) = \sigma\Big(W^l \phi_L^{l-1}\Big(\phi_c^{l-2}(X_t)\Big) + b^l\Big) \tag{4}$$

where a linear activation function is used in the last layer of MLP so that the output scalar value $\hat{y}$ is interpreted as a power prediction. The CNN-LSTM regressor is updated by the backpropagation algorithm with gradient descent optimization, by minimizing the loss function represented by mean squared error (MSE) where n denotes the number of observations:

$$\mathrm{L}_{MSE} = \frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \tag{5}$$

3.3. Multi-Headed Attention

Attention is used to compute an alignment score between elements from two sources [44,45]. Intuitively, the attention mechanism is formulated as an operation to calculate the similarity between query and key, and to extract the value related to the query as a weighted sum. Given the time step t of the window $X_t^\omega = [R_t, \ldots, R_{t-\omega}]$ and the spatiotemporal feature vector representation of a query q, attention computes the alignment score by a compatibility function $f(R_t, Q)$ which measures the correlation between R_t and Q. The alignment score vector $A_t = [f(R_\tau, Q)]_{\tau=t}^{t-\omega}$ consists of a series of correlations between the elements of query and key measured by the compatibility function:

$$A_t = [f(R_\tau, Q)]_{\tau=t}^{t-\omega} \tag{6}$$

The compatibility function is interpreted as a probability distribution $p(z \middle| X,\ Q)$ by a softmax operation, with the indicator variable z as defined as follows:

$$p(z|R_t, Q) = softmax(A_t) \tag{7}$$

The correlation between the query Q and the key expressed on the scale of [0,1] can be expressed as a random variable defined in the following equations as the attention score s, and can be written as the expectation of the energy consumption sampled according to its importance:

$$p(z = t|Q) = \frac{exp(f(R_t, Q))}{\sum_{\tau=t}^{t-\omega} exp(f(R_\tau, Q))} \tag{8}$$

$$S_t = \sum\nolimits_{\tau=t}^{t-\omega} p(z = \tau|R_t, Q) = E_{t\sim p(z|R_t,Q)}(R_t) \tag{9}$$

The compatibility function f is commonly used as an additive or multiplicative operation. The function f is implemented as the multiplicative (dot-product) operation that guarantees memory-efficient and fast convergence when considering the characteristics of the power consumption data of huge instances, with the spatiotemporal feature encoding function $\phi_{cp}(\cdot)$ defined in Section 3.2:

$$f(R_t, Q) = \phi_{cp}^{l}(R_t) \cdot \phi_{cp}^{l}(Q) \tag{10}$$

The attention mechanism is implemented as the deep learning layers; the attentive layer is placed between the modeling steps of the power spectrum and the time series. The attention layers are suitable for modeling the sudden increase in the usage of the power facilities which were difficult to predict by conventional deep models for predicting the power consumption. Self-attention is a special case of the attention: it replaces query Q with a source signal X_t. The single-attention mechanism intuitively performs dot-product on itself encoded by the convolution-pooling operation, and as the effect of obtaining the covariance between spatiotemporal feature vectors.

The proposed multi-headed attention, on the other hand, is an extension of the attention mechanism, which holds multiple attention in a single window and performs better than single-headed attention [45]. Figure 4 shows the multi-headed attention. Instead of computing a single scalar score $f(X_t,\ X_t)$ for each time step, we define the output of compatibility function f with the vector of the same length as X_t. Z_k denoting the alignment score from the compatibility function. Since we have expanded the dimension of the attention vector into $[f(R_\tau, Q)]_\tau$, we can formalize the importance weight vector P_{kt} for the element of the power attribute or encoded feature k in each step t:

$$P_{kt} = p(Z_k = i \middle| R_t, Q) \tag{11}$$

$$S_t = \left[\sum\nolimits_{\tau=t}^{t-\omega} P_{k\tau} R_t^{k\tau}\right]_{k=1}^{d_e} = \left[E_{t\sim p(Z_k|R_t,Q)}\left(R_t^{kt}\right)\right]_{k=1}^{d_e} \tag{12}$$

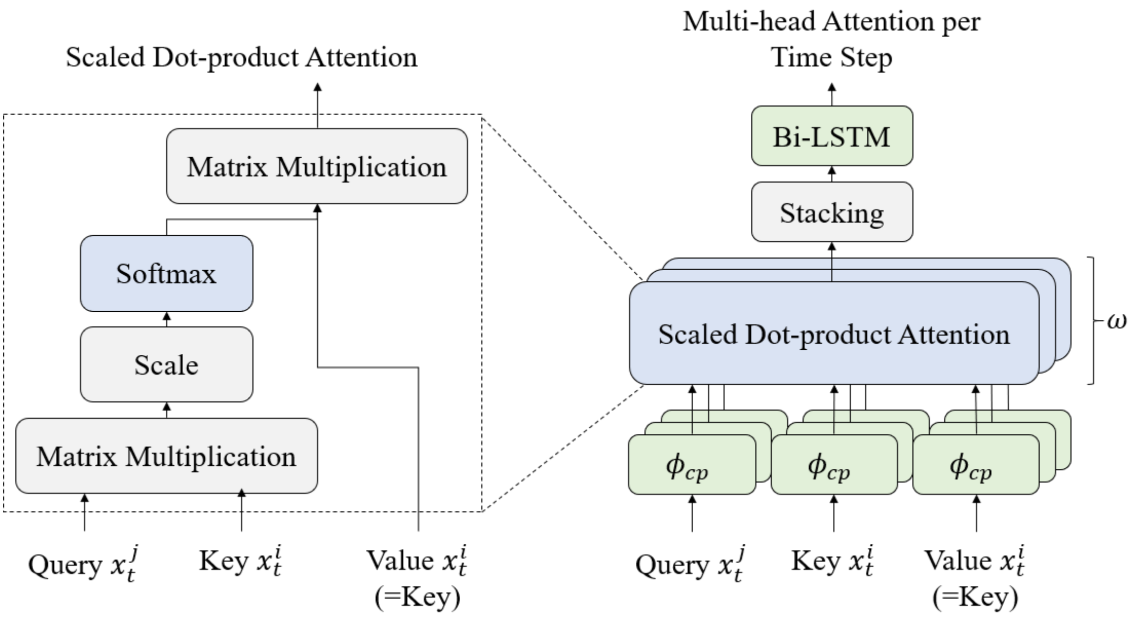


Figure 4. The operations of the multi-headed attention.

4. Experimental Results

In this section, we present how the CNN-LSTM with multi-headed attention predicts the power consumption and evaluate the performance with 10-fold cross-validation in terms of prediction error, which is followed by quantitative comparison with the relevant deep learning models.

4.1. Dataset and Implementation

We validate the proposed CNN-LSTM with multi-headed attention on the dataset of UCI household electric power consumption [46]. As shown in Table 2, the data were collected as approximately 2.07 million multi-channel sensors recording the household power consumption from December 2006 to November 2010, and the attributes include global active power (GAP), global reactive power (GRP), voltage, intensity, and additional three sub meterings. The data are normalized and processed in a sliding-window with time lag parameter ω. The prediction model receives the seven attributes under the time resolution condition and produces the GAP of the next time step.

Table 2. The attributes and statistics of the UCI household electric power consumption dataset.

Attributes	Date	Global Active Power (kW)	Global Reactive Power (kW)	Voltage (V)	Global Intensity (A)	Kitchen (Wh)	Laundry (Wh)	Climate Controls (Wh)
Average	-	1.0916	0.1237	240.8399	4.6278	1.1219	1.2985	6.4584
Std. Dev.	-	1.0573	0.1127	3.2400	4.4444	6.1530	5.8220	8.4372
Max	26/11/2010	11.1220	1.3900	254.1500	48.4000	88.0000	80.0000	31.0000
Min	16/12/2006	0.0760	0.0000	223.2000	0.2000	0.0000	0.0000	0.0000

The architecture of CNN-LSTM can be modified variously according to the number of stacked convolution-pooling and LSTM layers, as well as the number of convolutional filters, the kernel size and the number of the nodes in LSTMs. Given that typical deep learning models require an optimization process, it is essential to adjust and optimize the hyperparameters. The hyperparameters of the proposed model are determined by the intuition from the statistics of energy consumption as well as a through empirical study of iterative optimization summarized in Table 3. Figure 5 shows the overall architecture of the proposed model, where the time-distributed convolution-pooling layers, the self-attention, and LSTM layers are depicted. The spatiotemporal features from the power consumption data at each time step are exclusively extracted from time-distributed convolution and LSTM layers, respectively.

Table 3. Summary of the hyperparameters of the proposed model.

Operation	No. of Convolution Filters/Nodes	Kernel Size	Stride	Activation Function	No. of Parameters
TimeDistributed(Conv1D)	64	2 × 1	1	tanh	192
TimeDistributed(MaxPool1D)	-		2	tanh	0
TimeDistributed(Multi-Attention)	-		-	softmax	0
TimeDistributed(Conv1D)	64	2 × 1	1	tanh	8256
TimeDistributed(MaxPool1D)	-		2	tanh	0
TimeDistributed(Multi-Attention)	-		-	softmax	0
Dropout	0.5	-	-	-	0
LSTM	64	-	-	tanh	73,984
LSTM	64	-	-	tanh	33,024
Dropout	0.5	-	-	-	0
Dense	64			tanh	4160
Dense	64			tanh	4160
Dense	1			linear	65

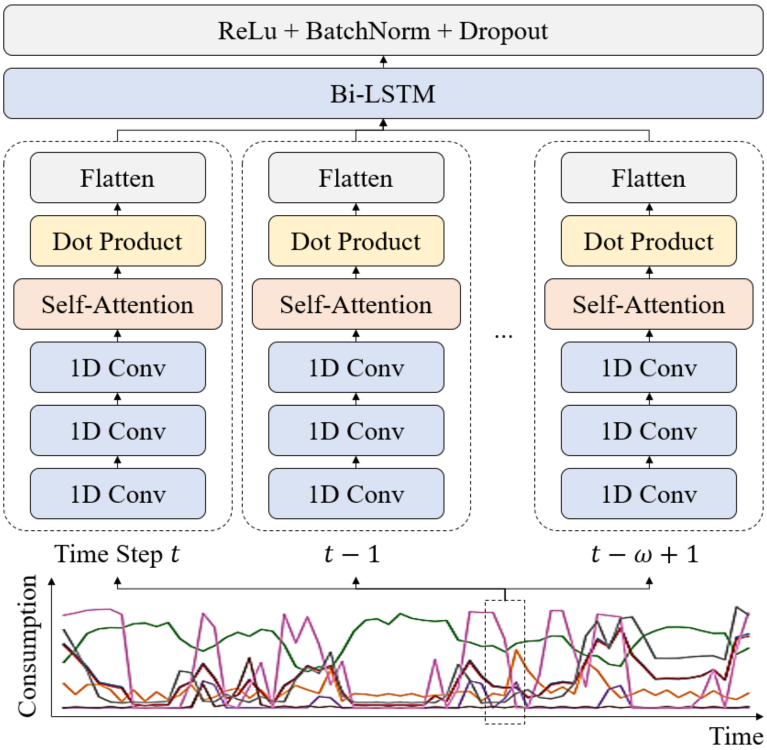


Figure 5. Implementation of the proposed CNN-LSTM with multi-headed attention.

4.2. Power Consumption Prediction Performance

Figure 6 compares the results of the prediction performance for the machine learning models including the convolution neural network and LSTM neural network. The proposed model designed to selectively model the spatiotemporal features has achieved the error reduction of 21.82%, compared to the conventional CNN-LSTM neural network. The evaluation is based on the mean squared error (MSE) for measuring the errors in Euclidean space:

$$MSE = \frac{1}{n}\sum\nolimits_{i=1}^{n}(y_t - \hat{y}_t)^2 = \frac{1}{n}\sum\nolimits_{i=1}^{n}(y_t - \phi(X_t))^2 \quad (13)$$

We further evaluate the proposed model in various time resolutions of 1, 15, 30, 45, and 60 min, 1 h, 1 day, and 1 week in Table 4. Each MSE is the result of the 10-fold cross validation. We compare machine learning methods for power demand forecasting published in the last two years. The prediction error is the highest at the unit time of 45 min and 1 h, and sometimes the performance degradation occurs due to the loss of short-term temporal features that might disturb the long-term temporal modeling. Considering that the end-user's long term behavior is reflected as a trend at low temporal resolution, the smoothing strategy is effective. It is known that the autoregressive integrated moving average (ARIMA) can model the overall trend of the time-series based on the moving average operation. As expected, the advantages of ARIMA emerge in the long period of 1W and 1D, and nonlinear mapping methods such as support vector regressor (SVR) and neural networks alleviate errors in a short period of 1H and 1M. The proposed method achieves the best performance in all temporal resolutions against the latest machine learning methods [7,8] and deep learning method [10].

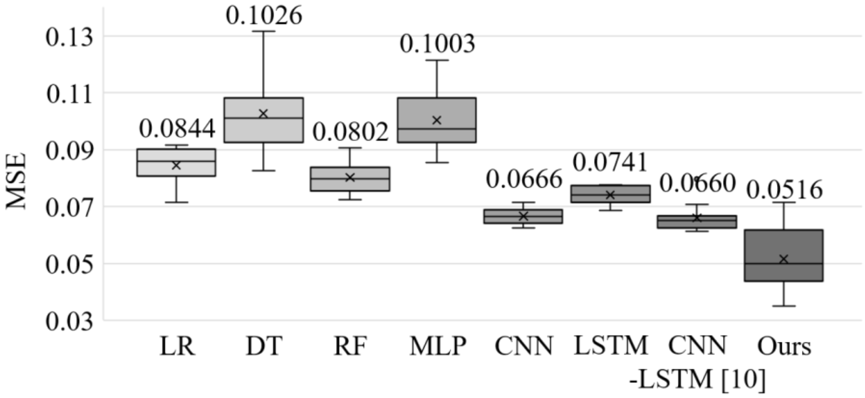


Figure 6. 10-fold cross validation of the methods including the deep learning models.

Table 4. Comparison of MSE according to time resolution and modeling method (LR: linear regression; ARIMA: autoregressive integrated moving average; DT: decision tree; RF: random forest; SVR: support vector regression; MLP: multilayered perceptron; CNN: convolution neural network; and LSTM: long short-term memory).

Resolution	LR	ARIMA [7]	DT	RF	SVR [8]	MLP	CNN	LSTM	CNN-LSTM [10]	Ours
1M	0.0844	0.0838	0.1026	0.0802	0.0797	0.1003	0.0666	0.0741	0.0660	0.0516
15M	0.2515	0.2428	0.4234	0.3972	0.3228	0.2370	0.2123	0.2208	0.2085	0.1838
30M	0.3050	0.2992	0.5199	0.3951	0.3619	0.2813	0.2684	0.2773	0.2592	0.2366
45M	0.3321	0.3431	0.4804	0.4315	0.4247	0.3208	0.3183	0.3220	0.3133	0.2838
1H	0.3398	0.3252	0.5259	0.4344	0.4059	0.3072	0.2865	0.2903	0.2803	0.2662
1D	0.1083	0.0980	0.1891	0.1190	0.1311	0.1134	0.1069	0.1129	0.1013	0.0969
1W	0.0624	0.0616	0.0706	0.0617	0.0620	0.0441	0.0333	0.0387	0.0328	0.0305

We confirm the effect of changes in time lag parameter ω. Figure 7 shows the MSE performance by iterative evaluations according to the temporal resolution and the time lag. As in the previous experimental results, it is observed that the prediction error considerably increases and yields

0.3848 MSE in the condition of 45 to 60 min, and the prediction error similarly increases regardless of the resolution with a short time lag.

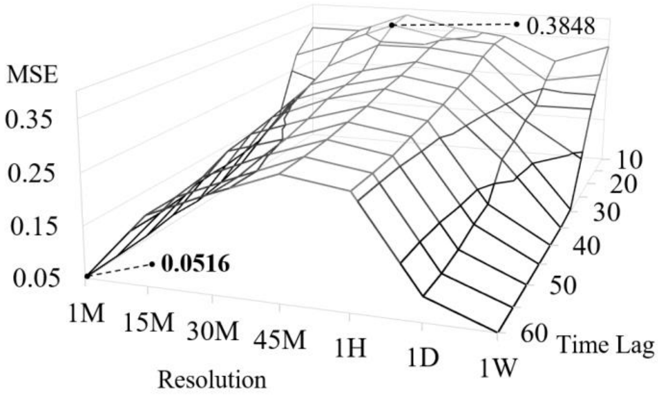


Figure 7. Comparison of MSEs by the time lag parameter ω.

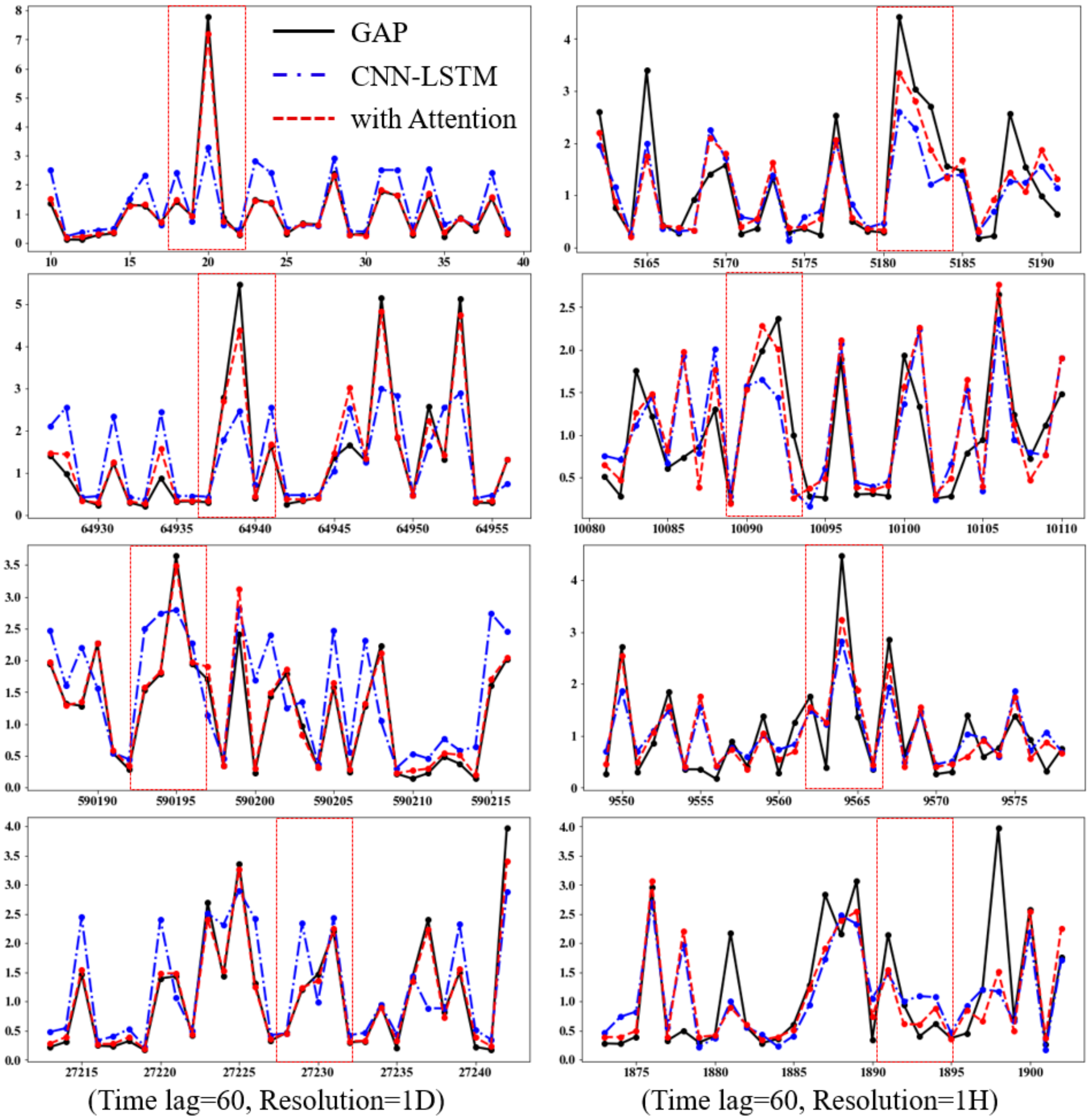
4.3. Effects of Multi-Headed Attention

For a thorough comparison, in addition to the neural networks proposed in previous works for energy prediction, we implement four additional neural networks and compared the prediction errors in Table 5. The scalability of the multi-headed attention can be evaluated by improving the different neural network architectures. It can be seen that the multi-headed attention improves the performance for power prediction in all cases. Interestingly enough, the single-attention significantly improves the performance of 2D-CNN, implying that a filter of extracting temporal features within a 2D convolution filter is appropriately learned from the local connectivity of time steps.

Table 5. Effect of various attention mechanisms on different neural network architectures (MSE).

Attention Type	CNN (1D)	CNN (2D)	LSTM	CNN-LSTM
None	0.0724	0.0666	0.0741	0.0676
Single-attention	0.0701	0.0542	0.0706	0.0660
Multi-headed Attention	0.0688	0.0538	0.0683	0.0516

We compare the prediction performance of the proposed model with that of a competitive CNN-LSTM model by plotting the ground-truth and prediction values in Figure 8. The prediction values in the red line are quite similar to the actual power consumption values in the black line, which shows the superiority of the prediction in the transient and impulsive cases mentioned in Section 1.



(Time lag=60, Resolution=1D)

(Time lag=60, Resolution=1H)

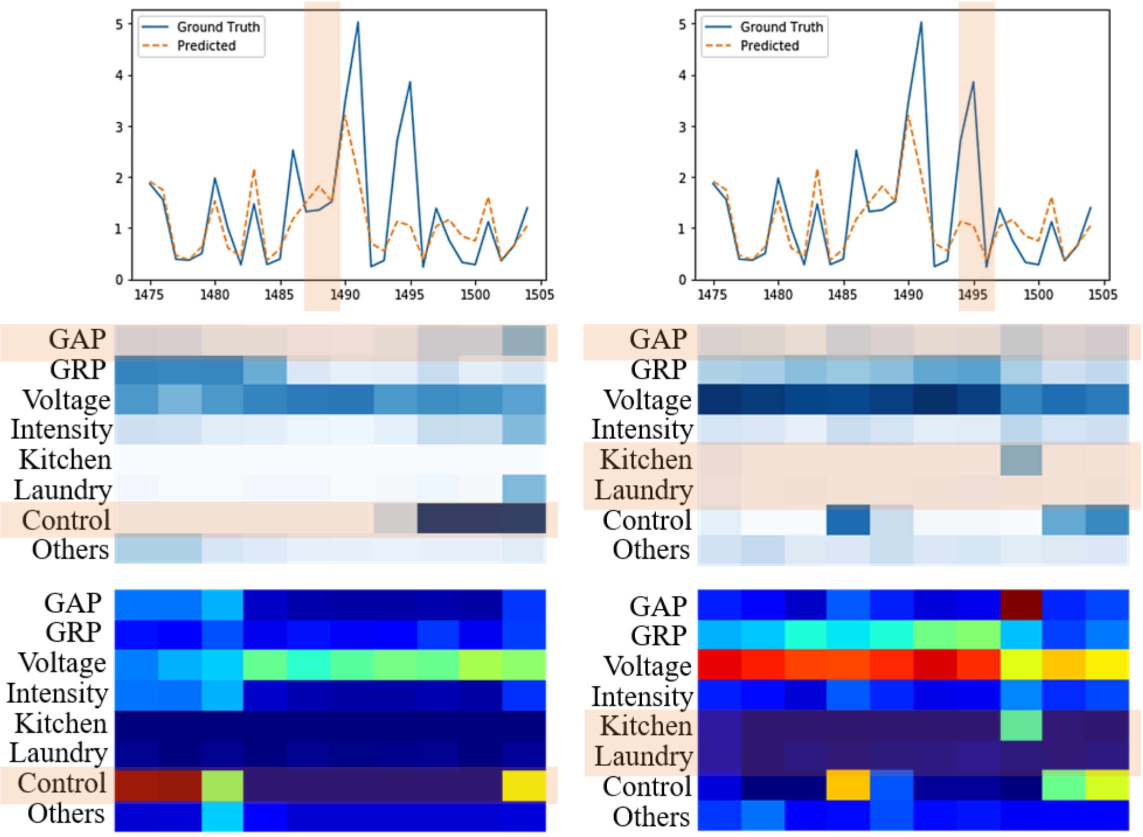
Figure 8. Verification of robustness to transient, impulsive characteristics of power consumption data.

4.4. Discussion

Meanwhile, we discover the two patterns of prediction failure mainly occurred in the case of Figure 9. The attention vector is uniformly distributed over all entire time steps in the window. Considering that the attention score is expressed as a probability distribution at the scale of (0,1) by the softmax operation, we conclude that the output of the deep learning models diverges and fails to extract spatiotemporal features.

Figure 9a shows the heat map of the convolutional filter for the two prediction failures that can be analyzed as the delay in weighting the control facility occurred in preparation for the sudden increase of the GAP. In Figure 9b, on the other hand, a sudden decrease in GAP without reason is observed, and it can be confirmed that it is difficult to model only by the historical power consumption data.

Taken together, in terms of the power prediction performance and the effect of the attention mechanism, we have evaluated the performance and the robustness in the transient and impulsive signals by quantitative and qualitative experiments. The general power prediction, however, still requires additional mechanism to cope with the aleatoric uncertainty [47] caused by the distribution of the power consumption data. This problem can be handled by extending the proposed model with generative deep learning and adopting the unsupervised learning.



(a) Prediction Failure on Increasing Watt-hour

(b) Prediction Failure on Decreasing Watt-hour

Figure 9. The gradients in the convolutional layers depicted by class activation map.

5. Concluding Remarks

In this paper, we have proposed a deep learning model with the multi-headed attention for predicting power consumption. After addressing the issues to model the power consumption and reviewing the power prediction models based on deep learning, we have presented the proposed model of CNN-LSTM for extracting the spatiotemporal features and the multi-headed attention for learnable weighting. The model has been evaluated in various temporal conditions and the deep learning parameters are analyzed by class activation map to understand the prediction failures.

Meanwhile, events that are outside the long-term behavior can be considered as possible events happening and failure to correctly predicting them within accuracy, which the rest of the prediction profiles present, can be attributed as outlier points. As a future work, we will take care of this issue through aggregation practices which smooth out the effect of such mis-prediction, and treat it within the tolerable error by the provided flexibilities of the hybrid approach.

Author Contributions: Conceptualization, S.-B.C.; Formal analysis, S.-J.B.; Funding acquisition, S.-B.C.; Investigation, S.-J.B.; Methodology, S.-J.B. and S.-B.C.; Supervision, S.-B.C.; Visualization, S.-J.B.; Writing—review & editing, S.-B.C. All authors have read and agreed to the published version of the manuscript.

Funding: This work was partly supported by the Institute of Information and Communications Technology Planning and Evaluation (IITP) grant funded by the Korean government (MSIT) (No. 2020-0-01361, Artificial Intelligence Graduate School Program (Yonsei University)) and the Korea Electric Power Corporation (Grant number: R18XA05).

Conflicts of Interest: The authors declare no conflict of interest.

References

1. IEA. *World Energy Outlook 2019*; IEA: Paris, France, 2019; Available online: http://www.iea.org/reports/world-energy-outlook-2019 (accessed on 13 November 2019).
2. Nejat, P.; Jomehzadeh, F.; Taheri, M.M.; Gohari, M.; Majid, M.Z.A. A global review of energy consumption, CO_2 emissions and policy in the residential sector (with an overview of the top ten CO_2 emitting countries). *Renew. Sustain. Energy Rev.* **2015**, *43*, 843–862. [CrossRef]
3. Zhao, G.; Liu, Z.; He, Y.; Cao, H.; Guo, Y.B. Energy consumption in machining: Classification, prediction, and reduction strategy. *Energy* **2017**, *133*, 142–157. [CrossRef]
4. Deb, C.; Zhang, F.; Yang, J.; Lee, S.E.; Shah, K.W. A review on time series forecasting techniques for building energy consumption. *Renew. Sustain. Energy Rev.* **2017**, *74*, 902–924. [CrossRef]
5. Arghira, N.; Hawarah, L.; Ploix, S.; Jacomino, M. Prediction of appliances energy use in smart homes. *Energy* **2012**, *48*, 128–134. [CrossRef]
6. Prashar, A. Adopting PDCA (Plan-Do-Check-Act) cycle for energy optimization in energy-intensive SMEs. *J. Clean. Prod.* **2017**, *145*, 277–293. [CrossRef]
7. Makridakis, S.; Spiliotis, E.; Assimakopoulos, V. Statistical and Machine Learning forecasting methods: Concerns and ways forward. *PLoS ONE* **2018**, *13*, e0194889. [CrossRef]
8. Gonzalez-Briones, A.; Hernandez, G.; Corchado, J.M.; Omatu, S.; Mohamad, M.S. Machine Learning Models for Electricity Consumption Forecasting: A Review. In Proceedings of the 2019 2nd International Conference on Computer Applications & Information Security, Riyadh, Saudi Arabia, 19–21 March 2019; pp. 1–6.
9. Burgio, A.; Menniti, D.; Sorrentino, N.; Pinnarelli, A.; Leonowicz, Z. Influence and Impact of Data Averaging and Temporal Resolution on the Assessment of Energetic, Economic and Technical Issues of Hybrid Photovoltaic-Battery Systems. *Energies* **2020**, *13*, 354. [CrossRef]
10. Kim, T.-Y.; Cho, S.-B. Predicting residential energy consumption using CNN-LSTM neural networks. *Energy* **2019**, *182*, 72–81. [CrossRef]
11. Lago, J.; De Ridder, F.; De Schutter, B. Forecasting spot electricity prices: Deep learning approaches and empirical comparison of traditional algorithms. *Appl. Energy* **2018**, *221*, 386–405. [CrossRef]
12. Fan, H.; MacGill, I.; Sproul, A. Statistical analysis of driving factors of residential energy demand in the greater Sydney region, Australia. *Energy Build.* **2015**, *105*, 9–25. [CrossRef]
13. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Polosukhin, I. Attention is all you need. In Proceedings of the Advances in Neural Information Processing Systems, Long Beach, CA, USA, 4–9 December 2017; pp. 5998–6008.
14. Zheng, H.; Fu, J.; Mei, T.; Luo, J. Learning Multi-attention Convolutional Neural Network for Fine-Grained Image Recognition. In Proceedings of the 2017 IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 5219–5227.
15. Ray, A. Symbolic dynamic analysis of complex systems for anomaly detection. *Signal Process.* **2004**, *84*, 1115–1130. [CrossRef]
16. Rajagopalan, V.; Ray, A. Symbolic time series analysis via wavelet-based partitioning. *Signal Process.* **2006**, *86*, 3309–3320. [CrossRef]
17. Lin, J.; Khade, R.; Li, Y. Rotation-invariant similarity in time series using bag-of-patterns representation. *J. Intell. Inf. Syst.* **2012**, *39*, 287–315. [CrossRef]
18. Tso, G.K.; Yau, K.K.; Tso, G.; Yau, K.K.W. Predicting electricity energy consumption: A comparison of regression analysis, decision tree and neural networks. *Energy* **2007**, *32*, 1761–1768. [CrossRef]
19. Ekonomou, L. Greek long-term energy consumption prediction using artificial neural networks. *Energy* **2010**, *35*, 512–517. [CrossRef]
20. Li, W.; Yang, X.; Li, H.; Su, L. Hybrid Forecasting Approach Based on GRNN Neural Network and SVR Machine for Electricity Demand Forecasting. *Energies* **2017**, *10*, 44. [CrossRef]
21. Mocanu, E.; Nguyen, P.H.; Gibescu, M.; Kling, W.L. Deep learning for estimating building energy consumption. *Sustain. Energy Grids Netw.* **2016**, *6*, 91–99. [CrossRef]

22. Marino, D.L.; Amarasinghe, K.; Manic, M. Building energy load forecasting using Deep Neural Networks. In Proceedings of the IECON 2016—42nd Annual Conference of the IEEE Industrial Electronics Society, Florence, Italy, 23–26 October 2016; pp. 7046–7051.
23. Kong, W.; Dong, Z.Y.; Jia, Y.; Hill, D.; Xu, Y.; Zhang, Y. Short-Term Residential Load Forecasting Based on LSTM Recurrent Neural Network. *IEEE Trans. Smart Grid* **2017**, *10*, 841–851. [CrossRef]
24. Li, C.; Ding, Z.; Zhao, D.; Yi, J.; Zhang, G. Building Energy Consumption Prediction: An Extreme Deep Learning Approach. *Energies* **2017**, *10*, 1525. [CrossRef]
25. Rahman, A.; Srikumar, V.; Smith, A.D. Predicting electricity consumption for commercial and residential buildings using deep recurrent neural networks. *Appl. Energy* **2018**, *212*, 372–385. [CrossRef]
26. Shah, I.; Iftikhar, H.; Ali, S.; Wang, D. Short-Term Electricity Demand Forecasting Using Components Estimation Technique. *Energies* **2019**, *12*, 2532. [CrossRef]
27. Fan, C.; Wang, J.; Gang, W.; Li, S. Assessment of deep recurrent neural network-based strategies for short-term building energy predictions. *Appl. Energy* **2019**, *236*, 700–710. [CrossRef]
28. Wang, Y.; Gan, D.; Sun, M.; Zhang, N.; Lu, Z.-X.; Kang, C. Probabilistic individual load forecasting using pinball loss guided LSTM. *Appl. Energy* **2019**, *235*, 10–20. [CrossRef]
29. Kim, T.-Y.; Cho, S.-B. Particle Swarm Optimization-based CNN-LSTM Networks for Forecasting Energy Consumption. In Proceedings of the 2019 IEEE Congress on Evolutionary Computation, Wellington, New Zealand, 10–13 June 2019; pp. 1510–1516.
30. Shi, H.; Xu, M.; Li, R. Deep Learning for Household Load Forecasting—A Novel Pooling Deep RNN. *IEEE Trans. Smart Grid* **2017**, *9*, 5271–5280. [CrossRef]
31. Bouktif, S.; Fiaz, A.; Ouni, A.; Serhani, M.A. Optimal Deep Learning LSTM Model for Electric Load Forecasting using Feature Selection and Genetic Algorithm: Comparison with Machine Learning Approaches. *Energies* **2018**, *11*, 1636. [CrossRef]
32. Guo, Z.; Zhou, K.; Zhang, X.; Yang, S. A deep learning model for short-term power load and probability density forecasting. *Energy* **2018**, *160*, 1186–1200. [CrossRef]
33. Fan, C.; Sun, Y.; Zhao, Y.; Song, M.; Wang, J. Deep learning-based feature engineering methods for improved building energy prediction. *Appl. Energy* **2019**, *240*, 35–45. [CrossRef]
34. Hinton, G.E. Deep belief networks. *Scholarpedia* **2009**, *4*, 5947. [CrossRef]
35. Taieb, S.B. *Machine Learning Strategies for Multi-Step ahead Time Series Forecasting*; Universit Libre de Bruxelles: Bruxelles, Belgium, 2014; pp. 75–86.
36. Wang, K.; Qi, X.; Liu, H. Photovoltaic power forecasting based LSTM-Convolutional Network. *Energy* **2019**, *189*, 116225. [CrossRef]
37. Bu, S.-J.; Cho, S.-B. A convolutional neural-based learning classifier system for detecting database intrusion via insider attack. *Inf. Sci.* **2020**, *512*, 123–136. [CrossRef]
38. Qing, X.; Niu, Y. Hourly day-ahead solar irradiance prediction using weather forecasts by LSTM. *Energy* **2018**, *148*, 461–468. [CrossRef]
39. Chorowski, J.; Bahdanau, D.; Serdyuk, D.; Cho, K.; Bengio, Y. Attention-Based Models for Speech Recognition. In Proceedings of the Advances in Neural Information Processing Systems, Montreal, QC, Canada, 7–12 December 2015; pp. 577–585.
40. Zhou, B.; Khosla, A.; Lapedriza, A.; Oliva, A.; Torralba, A. Learning Deep Features for Discriminative Localization. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 2921–2929.
41. Kim, J.-Y.; Cho, S.-B. Electric Energy Consumption Prediction by Deep Learning with State Explainable Autoencoder. *Energies* **2019**, *12*, 739. [CrossRef]
42. Sainath, T.N.; Mohamed, A.R.; Kingsbury, B.; Ramabhadran, B. Deep convolutional nueral networks for LVCSR. In Proceedings of the 2013 IEEE International Conference on Acoustics, Speech and Signal Processing, Vancouver, BC, Canada, 26–31 May 2013; pp. 8614–8618.
43. Ronao, C.A.; Cho, S.-B. Human activity recognition with smartphone sensors using deep learning neural networks. *Expert Syst. Appl.* **2016**, *59*, 235–244. [CrossRef]
44. Shen, T.; Zhou, T.; Long, G.; Pan, J.J.S.; Zhang, C. DiSAN: Directional self-attention network for RNN/CNN-free language understanding. In Proceedings of the Thirty-Second AAAI Conference on Artifial Intelligence, New Orleans, LA, USA, 2–7 February 2018; pp. 5446–5455.

45. Miyazaki, K.; Komatsu, T.; Hayashi, T.; Watanabe, S.; Toda, T.; Takeda, K. Weakly-Supervised Sound Event Detection with Self-Attention. In Proceedings of the ICASSP 2020 IEEE International Conference on Acoustics, Speech and Signal Processing, Barcelona, Spain, 4–8 May 2020; pp. 66–70.
46. Bache, K.; Lichman, M. *Individual Household Electric Power Consumption Dataset*; University of California, School of Information and Computer Science: Irvine, CA, USA, 2013; Volume 206.
47. Gal, Y.; Ghahramani, Z. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In Proceedings of the International Conference on Machine Learning, New York City, NY, USA, 19–24 June 2016; pp. 1050–1059.

© 2020 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (http://creativecommons.org/licenses/by/4.0/).