Ensemble Transfer Learning Strategy in Forecasting Power Consumption for Residential Buildings

Bowen Yang¹, Yuxiang Chen¹, Mustafa Gül¹, Haitao Yu², Charlie Shields³
¹Department of Civil and Environmental Engineering, University of Alberta, Edmonton, Canada
²Landmark Group of Companies Inc
³Department of Mechanical Engineering, University of Alberta, Edmonton, Canada

Abstract
This paper presents an ensemble transfer learning (TL) technique for predicting one-hour-ahead building power consumption with limited historical data, thereby addressing the data scarcity issues in the first year of building energy prediction. In the first step, four long short-term memory (LSTM) pre-trained models are constructed from the source buildings to capture diverse data patterns and develop transfer learning-based long short-term memory (TL-LSTM) models. The prediction performance is then enhanced by employing a weighted average ensemble technique to combine the TL-LSTM models. The result demonstrates that the ensemble TL achieves better prediction performance than the conventional LSTM- and TL-LSTM models for comparison.

Highlights
• Transfer learning
• Ensemble learning
• Multi-source
• Building power consumption prediction

Introduction
Buildings account for 40% of worldwide energy consumption and 30% of greenhouse gas emissions (Fang et al. 2021). Increasing urbanization and population expansion are two major factors for the rise in energy demand. Building energy management (BEM) is crucial to enhance energy efficiency to reduce energy consumption and carbon emissions. One of the essential aspects is developing a robust and generalized building energy prediction model. The short-term (from one-hour to one week) power consumption prediction is beneficial for the decision-making of BEM tasks. An accurate building energy prediction model is essential for many applications, including peak demand shaving (Hwang et al. 2020), energy consumption in response to grid signals adjustment (Yongli Wang et al. 2018), building maintenance activities scheduling (Cui et al. 2016) and system control (Huang et al. 2021).

Physical (or “white box”) and machine learning (ML) (also known as “black box”) approaches are the two main types of building energy estimation. The physical method relies on detailed building information (such as building envelope properties, operation characteristics of building systems, and occupants’ schedules) and domain experience to specify the relationship between input and output. The simulation process can be very time-consuming. Currently, the ML method is prevalent because it can uncover the underlying correlation between inputs and outputs based on historical data. Many advanced ML methods, such as artificial neural work (ANN) (Agatonovic-Kustrin and Beresford 2000), random forest (RF) (Breiman 2001), and long short-term memory (LSTM) (Greff et al. 2017), are successfully employed to predict short-term building energy consumption.

Although the techniques mentioned above can produce respectable results, they heavily rely on a huge amount of historical data, especially for deep learning models, which indicates that conventional ML loses generalization power when working with a small sample size. Yet, acquiring sufficient and accurate data, in reality, is difficult due to mechanical and human faults, and collecting sufficient data for new buildings is also challenging due to the time-consuming data collection process (Fang et al. 2021). Even if the data are sufficient for developing ML models to predict power consumption for building A, they may not generalize well for estimating power consumption for building B due to the fact that traditional ML models learn features and patterns (e.g., size of the building, number of occupants, local weather conditions) that are relevant to power consumption in building A. However, since the models have been optimized for the specific characteristics of building A, the features and patterns learned may not directly apply to building B, resulting in poor generalization. In this regard, additional studies are required to learn how to get a reliable prediction model using insufficient data. This study therefore proposes a model parameter transfer technique for training transferable building power consumption prediction models using limited operational data. The contributions of this study are twofold. First, it provides a reliable model parameter transfer method to build multi-source transfer learning (TL) models with insufficient data. Second, ensemble learning is used to construct an ensemble TL model that improves the accuracy of predictions.

Literature review
TL is a new technique in ML that can apply the features/patterns learned from one task to a related task. It has been widely and successfully applied to the applications of image classification (Shaha and Pawar...
(February 1, 2022, to February 28, 2022) are used for the target domain to test the TL models. The dataset contains five weather parameters: outdoor temperature (°C), dewpoint temperature (°C), relative humidity (%), wind speed (km/h), and atmospheric pressure (kPa), as well as three time-related characteristics: month type, day, hour, and one previous power consumption parameter: total power usage (Watts). Table 1 shows the input dataset variables.

**Table 1: The input variables**

<table>
<thead>
<tr>
<th>Variable category</th>
<th>Input</th>
<th>Type</th>
<th>Unit</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Weather condition</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Outdoor temperature</td>
<td>Numeric</td>
<td>°C</td>
<td></td>
</tr>
<tr>
<td>Dewpoint temperature</td>
<td>Numeric</td>
<td>°C</td>
<td></td>
</tr>
<tr>
<td>Relative humidity</td>
<td>Numeric</td>
<td>%</td>
<td></td>
</tr>
<tr>
<td>Wind speed</td>
<td>Numeric</td>
<td>km/h</td>
<td></td>
</tr>
<tr>
<td>Atmospheric pressure</td>
<td>Numeric</td>
<td>kPa</td>
<td></td>
</tr>
<tr>
<td><strong>Timestamp based features</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Month</td>
<td>Categorical</td>
<td>January ~</td>
<td>December</td>
</tr>
<tr>
<td>Day</td>
<td>Categorical</td>
<td>Monday ~</td>
<td>Sunday</td>
</tr>
<tr>
<td>Hour</td>
<td>Categorical</td>
<td>00:00 ~</td>
<td>23:00</td>
</tr>
<tr>
<td><strong>Power consumption</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total power usage</td>
<td>Numeric</td>
<td>Watts</td>
<td></td>
</tr>
<tr>
<td>Handling unit</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Heat recovery ventilator/Heat pump</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Process of transfer and ensemble learning**

Figure 1 depicts the process flowchart of transfer and ensemble learning, which includes three steps:

Step 1: The data are collected from five townhouse units in a residential community. To ensure the quality of the data, the occasionally missing raw data are filled in, and then the data are processed to increase their accuracy and consistency.

Step 2: The well-pre-trained models are then used to develop TL-LSTM models by fine-tuning some of the parameters of the pre-trained models. Normally, the knowledge can be transferred to the target model by freezing the top layers of a pre-trained LSTM model, adding new layers on top of the frozen levels, and then fine-tuning and retraining the new layers on the target domain training dataset.

Step 3: For the TL-LSTM models’ ensemble, each TL-LSTM is assigned a weight based on the prediction error on the target domain training dataset from the target building, the lower the prediction error, the greater the weight. Then, all the TL-LSTM models are assembled to produce a final ensemble model using weighted average ensemble, a technique used in ML and statistics.
that combines the prediction outcomes of multiple models and can minimize the weaknesses of an individual model, thereby improving predictive performance. (Pawlikowski and Chorowska 2020).

**Figure 1: Process of transfer and ensemble learning.**

**LSTM model**

LSTM is an architecture of neural networks proposed by Schmidhuber and Hochreiter (Hochreiter and Schmidhuber 1997) that is designed to remember information over long periods of time. LSTM networks contain a memory cell that maintains its state over time and three “gates”, which control the flow of information into and out of this cell. These gates use logistic sigmoid activations to control the proportion of information to be retained. Due to its unique structure, LSTM is capable of learning long-term dependencies in sequence prediction problems. Figure 2 (a), (b), (c), and (d) depict the calculation process of the forget gate, input gate, cell state, and output gate, respectively. The initial values of the weight matrices (i.e., $W_f, W_i, W_o, W_c$) and bias (i.e., $b_f, b_i, b_o, b_c$) are updated using the gradient descent method. The information from the current input $x_t$ and previous hidden state $h_{t-1}$ is transmitted through the sigmoid function ($\sigma$), and the forget gate determines whether the gate should keep the information or not based on Eq. (1). The input gate is used to update the status of the cell, which includes two tasks. First, the precious cell output and current state are provided to a classifier sigmoid function to decide what relevant information needs to be updated, as shown in Eq. (2). Second, the previous cell output and current input are once more sent to the activation tangent function ($\tanh$) in order to generate a new candidate value ($\tilde{C}_t$), Eq. (3). The cell state ($C_t$) is the path of information transmission, thus the information can be transmitted in a serial connection, Eq. (4). Lastly, the current output vector $h_t$ is determined by linear transformation through the sigmoid function of the previous cell state and the output is passed through tangent function, and LSTM is used to predict the next-hour building power consumption at $t+1$ time, as shown in Eq. (5) and (6).

$$f_t = \sigma(W_f \cdot [h_{t-1}, x_t] + b_f)$$  \hspace{1cm} (1)

$$i_t = \sigma(W_i \cdot [h_{t-1}, x_t] + b_i)$$  \hspace{1cm} (2)

$$\tilde{C}_t = \tanh(W_c \cdot [h_{t-1}, x_t] + b_c)$$  \hspace{1cm} (3)

$$C_t = f_t \cdot C_{t-1} + i_t \cdot \tilde{C}_t$$  \hspace{1cm} (4)

$$o_t = \sigma(W_o \cdot [h_{t-1}, x_t] + b_o)$$  \hspace{1cm} (5)

$$h_t = o_t \cdot \tanh(C_t)$$  \hspace{1cm} (6)

**Figure 2: Procedure of forward propagation of LSTM.**

**Performance evaluation metrics**

This study evaluates the performance of the proposed method using two performance metrics: the coefficient of variation of root mean squared error (CV-RMSE) and Jensen-Shannon divergence (JSD). For CV-RMSE, the American Society of Heating, Refrigerating, and Air-Conditioning Engineers (ASHRAE) recommends that the CV-RMSE should be within 30% on an hourly basis. However, it is based on mean values and can be heavily influenced by outliers. JSD is a method of measuring the similarity between two probability distributions (Mahdavi, Tahmasebi, and Kayalar 2016), which can quantify the similarity of the predicted and real values rather than calculating the average errors as in CV-RMSE. For two probability distributions, $P$ and $Q$, the JSD is calculated based on Kullback–Leibler divergence (KLD). Two evaluation metrics are calculated by Eqs. (7)-(8), respectively.

$$CV - RMSE = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n}}$$  \hspace{1cm} (7)

where $n$ is the number of the test data observations, $\hat{y}_i$ is the predicted values of building power consumption, and $y_i$ is the actual values.

$$JSQ(P, Q) = \frac{1}{2} KLD(P, M) + \frac{1}{2} KLD(Q, M)$$  \hspace{1cm} (8)

$$M = \frac{1}{2} (P + Q)$$  \hspace{1cm} (9)

$$KLD(P, Q) = \sum_{i} P(i) \ln \frac{P(i)}{Q(i)}$$  \hspace{1cm} (10)

**Average weight ensemble calculation**

The average weighted ensemble approach assigns a weight to each base TL model, and the final ensemble prediction is obtained by taking a weighted average of the base learners’ predictions. The weight calculation and final ensemble prediction can be found in Eq. (9) and (10).
where \( w_i \) is the weight assigned to the \( i \)-th based TL model, \( p_i \) is the performance metric (CV-RMSE is used in this study) of the \( i \)-th based TL model, and \( p \) is the sum of the performance metrics for all base TL models in the ensemble.

\[
y_{\text{ensemble}} = \sum (w_i \cdot y_i)
\]

(10)

where \( y_{\text{ensemble}} \) is the ensemble’s final prediction. \( y_i \) is the prediction of the \( i \)-th base TL model.

### Result

#### Data pre-processing

In order to maintain the integrity of time-series data, interpolation is used to fill in missing data initially. After filling in the missing data, outliers are replaced with Nah values and treated as missing data. Lastly, min-max normalization (scale the inputs), cyclical encoding (convert the data of month, day, and hour into sin/cos format and inform the model that they occur in the cycle), and logarithmic transformation (handle the left-skewed data and making them more closely to a normal distribution) are implemented. The sliding window approach (Li, Li, et al. 2022) is then used to process the original time series data and create the new input-output consecutive pairs after the missing data has been filled in.

Figure 3 depicts the frequency curve of the selected buildings’ power consumption. Due to the differences in occupants’ behaviours and schedules, each building has a unique power consumption distribution. However, they have a similar distribution shape, and the high frequency is concentrated between 200 and 300 watts, indicating a great potential for transfer learning to extract and apply useful information from these source buildings to the target building. In this study, the time horizontal is set as 1, which means the time interval is 1 hour, and the window size is set to 24. In other words, the previous 24 historical measurement data are used to predict power consumption 1 hour ahead.

#### Performance evaluations of pre-trained models

The hyperparameter parameters must be optimized to build appropriate pre-trained models. Table 2 displays the hyperparameter settings of pre-trained models.

<table>
<thead>
<tr>
<th>Unit</th>
<th>LSTM layer, units</th>
<th>Activation function</th>
<th>Learning rate</th>
<th>L2</th>
</tr>
</thead>
<tbody>
<tr>
<td>29</td>
<td>2, (64,64)</td>
<td>ReLu</td>
<td>0.0001</td>
<td>0.001</td>
</tr>
<tr>
<td>30</td>
<td>2, (64,64)</td>
<td>ReLu</td>
<td>0.0001</td>
<td>0.001</td>
</tr>
<tr>
<td>31</td>
<td>2, (32,64)</td>
<td>ReLu</td>
<td>0.0001</td>
<td>0.001</td>
</tr>
<tr>
<td>32</td>
<td>2, (64,64)</td>
<td>ReLu</td>
<td>0.0001</td>
<td>0.001</td>
</tr>
</tbody>
</table>

The performance of LSTM pre-trained models for predicting power consumption is presented in Figure 4. The CV-RMSE findings of all pre-trained models are less than 30%, which satisfies the ASHRAE 14 requirement that the CV-RMSE of hourly energy consumption should be within 30% (Haberl, Culp, and Claridge 2005). Unit 29 (LSTM 1) performs marginally better than other units regarding CV-RMSE and JSD, whereas units 31 (LSTM 3) and 32 (LSTM 4) have larger inaccuracies compared to the pre-trained model of LSTM 1. It is challenging for LSTM models to learn and predict future power consumption for units 31 and 32 since their power consumption profiles fluctuate substantially over time, and the periodicity needs to be more easily discernible.

#### Performance evaluations of transfer learning models

Due to the discrepancies in data distribution between domain and target building, the hyperparameters of the pre-trained models have to be fine-tuned to obtain appropriate TL models. To fine-tune the hyperparameters of the TL models, the number of layers, the number of hidden units per layer, the activation function, the learning rate, and the L2 are considered. This study follows some of the network structure and hyperparameter settings used in Ref. (Fang et al. 2021), and the parameter configuration can be found in Table 3.

The average performance evaluation metrics for transfer learning models are presented in Figure 5. It can be seen that the prediction performance of each TL-LSTM model is mediocre. The values of CV-RMSE of four TL-LSTM models are beyond 30%, indicating that these models are unacceptable. After ensembling four TL-LSTM models, the ensemble TL-LSTM outperforms...
other models for forecasting power consumption. The CV-RMSE drops to an acceptable range (0.301) as the weighted average ensemble attempts to capitalize on the strengths of each model while minimizing their weaknesses.

Figure 6 shows the actual and predicted values of the target residential building’s power usage from February 1\textsuperscript{st} to February 14\textsuperscript{th} based on five TL models. The red line indicates the real values of power consumption of the target building, while the grey line represents the predicted values. As can be observed from Figure 6, the traditional LSTM fails to forecast the power consumption as deep learning requires enormous data for training, and only one month of training data are not enough to develop an LSTM model. The predicted lines of the TL models are generally close to the actual lines, and the predicted values closely match the actual power consumption using limited data, except for the prediction of the true extreme values, including peak and valley power consumption, due to the randomness of data distribution of peak and valley values over time. Consequently, obtaining precise prediction results for peak and valley values is difficult, and more efficient algorithms need to be discovered to address this challenge.

As shown in Figure 6 (b), the scatter distribution of the real and predicted values of the TL and ensemble TL models is concentrated near the center line (actual value = predicted value), the closer the points are to the center line, the better the models perform. The scatter distribution of the LSTM model shows wide dispersion, indicating there is no linear correlation between the predicted and true values.

Table 3: Hyperparameters settings of the transfer learning models.

<table>
<thead>
<tr>
<th>TL model</th>
<th>LSTM layer, units</th>
<th>Activation function</th>
<th>Learning rate</th>
<th>L2</th>
</tr>
</thead>
<tbody>
<tr>
<td>TL-LSTM 1</td>
<td>2, (64,64)</td>
<td>ReLU</td>
<td>0.000000</td>
<td>0.001</td>
</tr>
<tr>
<td>TL-LSTM 2</td>
<td>2, (64,64)</td>
<td>ReLU</td>
<td>0.000000</td>
<td>0.000001</td>
</tr>
<tr>
<td>TL-LSTM 3</td>
<td>2, (64,64)</td>
<td>ReLU</td>
<td>0.000000</td>
<td>0.000001</td>
</tr>
<tr>
<td>TL-LSTM 4</td>
<td>2, (64,64)</td>
<td>ReLU</td>
<td>0.000000</td>
<td>0.000001</td>
</tr>
</tbody>
</table>

Figure 6: The prediction result of different prediction models.

Improvement of the ensemble transfer learning model

To further analyze the proposed model’s improvement compared to other prediction models, two criteria are used to evaluate it in this study (Ying Wang, Wang, and Li 2020), and improvement percentage criteria between the ensemble TL and other models are listed in Table 4. In Table 5, it can be drawn that the error reduction in CV-RMSE and JSD of the ensemble TL and traditional LSTM reached more than 58% and 85%, respectively. Compared with TL-LSTM1, TL-LSTM2, TL-LSTM3, and TL-LSTM4, the proposed model performs the best and can reduce the prediction error effectively (between 7%–17% using $P_{CV-RMSE}$ and 11%–27% using $P_{JSD}$).

Table 4: The improvement percentages among different prediction models.
Table 5: Improvement percentages among the ensemble transfer learning and the compared models.

<table>
<thead>
<tr>
<th>Prediction models</th>
<th>Evaluation Metrics</th>
<th>$P_{CV-RMSE}$ (%)</th>
<th>$P_{JSD}$ (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ensemble TL VS. Traditional LSTM</td>
<td></td>
<td>58.78</td>
<td>85.45</td>
</tr>
<tr>
<td>Ensemble TL VS. TL-LSTM 1</td>
<td></td>
<td>12.35</td>
<td>11.11</td>
</tr>
<tr>
<td>Ensemble TL VS. TL-LSTM 2</td>
<td></td>
<td>17.45</td>
<td>27.27</td>
</tr>
<tr>
<td>Ensemble TL VS. TL-LSTM 3</td>
<td></td>
<td>14.56</td>
<td>20</td>
</tr>
<tr>
<td>Ensemble TL VS. TL-LSTM 4</td>
<td></td>
<td>7.17</td>
<td>11.11</td>
</tr>
</tbody>
</table>

Conclusion

This study aims to predict the short-term power consumption of residential buildings. A multi-source ensemble transfer learning framework is proposed to improve prediction accuracy using limited historical data. LSTM pre-trained models are initially built to learn useful power consumption patterns that can be transferred to the target task. Then, the LSTM parameters are transferred to the TL-LSTM models. Finally, the ensemble TL is built using the weighted average ensemble based on TL-LSTM performances. The outcomes are as follows:

- Overall, the LSTM pre-trained models achieve outstanding performances. The values of CV-RMSE of all pre-trained models are within 30%.
- For the TL-LSTM models, prediction for one hour ahead is able to obtain mediocre performance using previous 24 hours data points, but TL-LSTM models estimate peak and valley power consumption are not well due to the randomness of the data distribution of peak and valley values over time.
- The ensemble TL outperforms other models, and the weighted average ensemble method can minimize the error in individual TL-LSTM performance, potentially leading to improved generalization and predictive performance. The CV-RMSE of ensemble TL can reach 0.30, and the error could decrease from 6% to 58%.

Additional valuable features (such as solar radiation and the number of occupants) will be considered for future work to improve the overall prediction performance, particularly peak load values.

Acknowledgement

The authors gratefully acknowledge financial support for this study from the NSERC (Natural Sciences and Engineering Research Council of Canada) Alliance Grants, Landmark Inc, and SSRIA (Smart Sustainable Resilient Infrastructure Association).

References


Huang, Sen, Yashen Lin, Venkatesh Chinde, Xu Ma, and Jianming Lian. 2021. “Simulation-Based


