Application and Comparison of Deep Learning in Low-Frequency Non-Intrusive Load Disaggregation

Authors: Li, Yiming, Ju, Yuntao, Qu, Liao, Ju, Yuntao

Date: 2021-10-01T00:00:00+00:00

Abstract

Non-intrusive load decomposition can fully parse user electricity consumption data and is a key technology for analyzing and evaluating users' flexible regulation potential. Given the extensive application of deep learning methods in load decomposition, this paper first delves into the working mechanisms of mainstream deep learning network architectures such as denoising autoencoders, recurrent neural networks, and convolutional neural networks when applied to load decomposition problems, and provides a prospective analysis of their applications and development in the field of load decomposition. Subsequently, a decomposition algorithm evaluation framework based on different dimensions is proposed, and the selection criteria for test data during the evaluation process are supplemented. Finally, the evaluation framework is employed to assess mainstream deep learning decomposition models, and the model code is open-sourced. The evaluation results demonstrate that the proposed framework can more comprehensively evaluate deep learning decomposition models under given hyperparameter settings, and reveal the sensitivity of model performance to factors such as network architecture and feature inputs.

Full Text

Preamble

Funding: This work is supported by the National Natural Science Foundation of China (No. 51707196).

Applications and Comparison of Deep Learning in Low-Frequency Non-Intrusive Load Disaggregation

LI Yiming[1], JU Yuntao[1], QU Liao[1]

(1. College of Information and Electrical Engineering, China Agricultural University, Beijing 100083, China)

Abstract

Non-intrusive load disaggregation enables comprehensive analysis of user electricity consumption data and serves as a critical technology for evaluating users' potential for flexible demand response. Given the widespread application of deep learning methods in load disaggregation, this paper first examines in detail the working mechanisms of mainstream deep learning architectures—including denoising autoencoders, recurrent neural networks, and convolutional neural networks—when applied to load disaggregation problems, and provides a prospective analysis of their development in this field. Subsequently, we propose a multi-dimensional evaluation framework for disaggregation algorithms and supplement it with guidelines for test data selection during evaluation. Finally, we employ this framework to evaluate mainstream deep learning disaggregation models, with the model code made open-source. The evaluation results demonstrate that the proposed framework enables more comprehensive assessment of deep learning disaggregation models under specific hyperparameter settings, and reveals the sensitivity of model performance to factors such as network architecture and input features.

Keywords: deep learning; non-intrusive load disaggregation; convolutional neural network; recurrent neural network; electricity big data

0 Introduction

Non-Intrusive Load Monitoring (NILM), first proposed by Hart in the 1980s [?], involves installing smart meters at residential electricity entry points and applying a series of data mining steps to decompose aggregate load information into the consumption patterns of individual appliances. The resulting end-use data can provide users with feedback to enhance energy-saving awareness [?, ?] and support demand response initiatives [?].

Load disaggregation represents the core challenge in NILM [?]. Traditional disaggregation algorithms primarily rely on load event detection, employing combination optimization [?] and clustering-based matching [?]. However, these methods suffer from low decomposition accuracy due to overlapping load signatures and difficulties in modeling appliances with continuously varying power consumption. Factorial Hidden Markov Models (FHMM) and deep neural networks have emerged as mainstream solutions that effectively improve decomposition accuracy [?].

FHMM-based models can achieve reasonable accuracy with limited supervised information. Reference [?] incorporated appliance usage patterns as additional features to build disaggregation models based on four different FHMM variants, while [?] transformed the FHMM inference problem into a convex quadratic programming problem through signal constraints on sources. Nevertheless, constrained by the properties of Markov chains, FHMM-based models cannot effectively model continuously varying power consumption or unknown electrical devices, and their decoding time complexity grows exponentially with the number of appliances, limiting their practical application [?].

In theory, deep learning can approximate any physical model by deepening neural networks [?]. Reference [?] applied deep learning models such as Bidirectional Long Short-Term Memory (Bi-LSTM) and Denoising AutoEncoder (DAE) to model appliance power consumption. Compared with traditional methods, deep learning approaches offer several inherent advantages: (i) elimination of manual feature extraction, (ii) linear growth in computational complexity with the number of appliances, (iii) capability to model continuously varying power consumption while ignoring unknown devices, and (iv) utilization of GPU parallel computing [?].

However, deep learning models involve numerous parameters, exhibit poor interpretability, and are sensitive to hyperparameter selection. Consequently, understanding the underlying working mechanisms of network architectures is crucial for algorithm design and parameter tuning. Moreover, due to variations in test data and evaluation metrics, researchers struggle to systematically compare disaggregation algorithms, identify causes of poor performance, and develop targeted improvements [?].

Addressing these challenges in NILM research, this paper first examines the working mechanisms of three mainstream network architectures—Denoising AutoEncoder (DAE), Recurrent Neural Network (RNN), and Convolutional Neural Network (CNN)—when applied to load disaggregation, highlighting that bidirectional encoding of effective load information by Bi-LSTM layers and extraction of different load events by convolutional layers are key to successful deep learning applications. We then propose a multi-dimensional evaluation framework for load disaggregation algorithms and establish guidelines for test data selection from both dataset and appliance perspectives. Finally, we evaluate mainstream deep learning disaggregation algorithms using this framework, revealing that performance differences primarily stem from varying network capabilities in modeling different types of load information. Through case studies, we demonstrate the impact of factors such as input features on model performance and provide recommendations for algorithm improvement. Additionally, we have open-sourced the model implementation and evaluation code (https://github.com/Ming-er/NeuralNILM_{Pytorch}).

1 Working Mechanisms of Deep Learning Models in Load Disaggregation

Load disaggregation requires representing aggregate load information over a time series as a linear combination of constituent signals. For any time point in the sequence, the problem can be abstracted into the following mathematical form:

$$y_t = \sum_{i=1}^{I} x_{i,t} + u_t + \epsilon_t \quad (1)$$

where y_t represents the total power at time t, $x_{i,t}$ denotes the power of the i-th appliance at time t, u_t is the total power of unknown appliances at time t, ϵ_t represents random noise following $\mathcal{N}(0, \sigma^2)$, and I is the number of electrical appliances.

Given the powerful representation capability of neural networks for end-to-end mapping, deep learning models typically learn the mapping relationship between the aggregate power sequence and the power sequence of the i-th appliance over a period, without considering the impact of appliances that users are not interested in or other unknown devices on decomposition performance. Deep learning models commonly define the problem as:

$$\hat{x}_{i,t} = f(y_{t-w:t+w}) \quad (2)$$

where $\hat{x}_{i,t}$ is the predicted power of the i-th appliance at time t, and w represents the window size.

1.1 DAE-Based Models

DAE accepts noisy data as input and reconstructs it as original noise-free output [?], finding widespread application in image denoising and speech enhancement [?, ?]. In load disaggregation, the aggregate power sequence and other sequences (such as timestamp information [?] and reactive power as supplementary semantic context [?]) can be viewed as DAE inputs, while the power sequence of the target appliance serves as output, with other appliances' simultaneous power consumption treated as noise to be eliminated.

[Figure 1: see original paper] illustrates the basic structure of DAE-based models, which consist of encoder and decoder components. The encoder extracts and combines features from the input, compressing them into a compact representation—the bottleneck layer—while the decoder reconstructs the target appliance' s power sequence based on the bottleneck layer information. Gradient optimization of the reconstruction loss constrains the bottleneck features to retain information from the aggregate power sequence that facilitates reconstruction of the current appliance' s power, thereby enhancing the decoder' s capability.

Reference [?] first applied DAE to load disaggregation, achieving better results than traditional methods like CO and FHMM. However, due to its simple architecture, the true advantages of DAE were not fully realized. Subsequently, references [?] deepened the DAE network and incorporated designs such as convolution, normalization, and residual connections during encoding and decoding, further improving disaggregation accuracy.

In recent years, with the rise of Generative Adversarial Networks (GAN) [?], many studies [?] have attempted to integrate GAN into the DAE encoding-decoding process, as illustrated in [Figure 2: see original paper]. This approach preserves the DAE structure while introducing a discriminator to classify the predicted output sequences as real or fake. By optimizing the adversarial loss, both the DAE and discriminator are updated iteratively, improving the discriminator' s 分辨能力 while making DAE outputs more realistic.

1.2 RNN-Based Models

RNN is a class of recursive neural networks that processes sequential data through chained connections along the sequence evolution direction [?], widely used in time series modeling. Typical variants include Long Short-Term Memory (LSTM) [?] and Gated Recurrent Unit (GRU) [?], which solve the gradient vanishing problem in traditional RNNs, enabling them to capture long-term temporal dependencies in power sequences and memorize contextual information across extended intervals, such as load events separated by long time gaps.

[Figure 3: see original paper] shows the basic structure of RNN-based models, primarily composed of Bi-LSTM, while [Figure 4: see original paper] illustrates the LSTM cell working principle. The model operates through encoding and decoding phases. First, the Bi-LSTM layer receives sequential features within the input window and performs bidirectional information composition and encoding: relevant information associated with the target time point is passed to subsequent LSTM cells, while irrelevant information is discarded. This bidirectional information flow enhances the model' s ability to infer causal relationships between load events. Finally, the fully connected layer integrates the effective encoded information stored in LSTM hidden states to decode the appliance power sequence.

References [?, ?] first applied RNN to load disaggregation, constructing basic LSTM models for each appliance to learn mappings from aggregate to individual power sequences. Reference [?] applied the simpler Bidirectional GRU (Bi-GRU) model with sliding windows to predict appliance power at intermediate time points in the aggregate sequence.

Subsequent research focused on enhancing encoding-decoding capabilities. Reference [?] added parallel convolutional kernels of different sizes upstream of Bi-LSTM to strengthen feature extraction, effectively filtering load information before encoding. References [?] and [?] discretized continuous power information and employed Sequence-to-Sequence (Seq2Seq) models with Attention mechanisms [?], enabling the model to focus on different decoding moments and improve information utilization. Reference [?] encoded timestamps into Bi-GRU models, considering that specific appliances often operate during particular periods.

Other studies optimized RNN models holistically. Reference [?] employed parameter-sharing Bi-GRU for multi-task learning, simultaneously estimating appliance on/off states and power consumption, demonstrating better noise robustness during inference. Reference [?] addressed overfitting in Bi-LSTM training through regularization, while references [?, ?] applied Bayesian optimization and TPE algorithms for hyperparameter selection, enhancing model adaptability to real-world scenarios.

1.3 CNN-Based Models

CNN is a feedforward neural network primarily composed of convolutional and pooling layers. Due to the nature of convolution operations, neurons respond to neighboring units within a local coverage area, giving CNN strong capabilities for extracting local features such as load events occurring at specific time points. CNN has demonstrated superior performance over RNN for certain time series data [?] and is widely used in sequence modeling tasks like syntactic parsing [?] and text classification [?].

[Figure 5: see original paper] illustrates the basic structure of CNN-based models, featuring convolutional kernels of different sizes. One-dimensional convolutional layers extract features from sequence information within the input window: kernels of different sizes in various layers detect load events at different time scales, while kernels in different channels within the same layer detect events of varying intensities (power change magnitudes). Additional sequence features, analogous to background pixels in images, are captured as semantic information.

Earlier convolutional layers have smaller receptive fields and can only capture simple sequence patterns but with high resolution, while deeper layers combine these simple patterns to learn more complex load features and semantic information. The final fully connected layer integrates the detected load features and outputs the appliance power sequence.

Reference [?] built a Sequence-to-Point (Seq2Point) model using multiple convolutional kernels of different sizes with a top fully connected layer, visualizing the kernels to demonstrate that different channels activate differently for various load events.

Subsequent research leveraged this characteristic. Reference [?] used CNN-extracted features to verify whether power sequences generated by regression networks belonged to the target appliance. References [?, ?] constructed gated branch networks: one convolutional network outputting on/off states and another predicting power values, with their Hadamard product serving as the final prediction. These studies aimed to utilize CNN' s load event identification capability to generate additional information streams and suppress false positives.

Performance optimization research for CNN models has focused on enhancing load event detection. Reference [?] introduced convolutional block attention mechanisms to define importance from channel and spatial dimensions, enabling more focused intermediate feature representation. Reference [?] borrowed

WaveNet' s [?] dilated convolutions and skip connections to build WaveNILM with larger receptive fields for better long-term event capture. Reference [?] adopted U-Net' s [?] multi-resolution feature fusion to create U-NetNILM for balancing short- and long-term events. References [?] modeled load signals as 2D images, constructing disaggregation models analogous to CNN' s ability to learn local image features.

2 Load Disaggregation Algorithm Evaluation Framework

This section presents the proposed evaluation framework for comparing load disaggregation algorithms, with detailed explanations of its components. The framework flowchart is shown in [Figure 6: see original paper].

2.1.1 Model Pre-Evaluation

I. Training Efficiency
To measure training efficiency, we select the average convergence time t as the primary metric. Models with smaller t exhibit stronger adaptability to the problem and reduce iteration costs in algorithm development.

II. Deployment Difficulty
To assess deployment difficulty, we use the memory footprint m of model parameters as the main indicator. Models with smaller m are easier to deploy on embedded terminals.

III. Real-Time Application
To evaluate real-time capability, we select the average inference speed v as the key metric. Models with larger v provide better real-time feedback on appliance load information. Specifically, in Section 3.3.1, we estimate v by measuring the time required to disaggregate one month of load data.

2.1.2 Identification Accuracy

To measure identification accuracy, we adopt three metrics from statistical learning theory [?]: recall, precision, and Matthews Correlation Coefficient (MCC). The calculations are as follows:

$$\text{Recall} = \frac{TP}{TP + FN}$$

$$\text{Precision} = \frac{TP}{TP + FP}$$

$$\text{MCC} = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}$$

where TN represents true negatives (timestamps where the appliance is actually off and predicted off), TP true positives (appliance on and predicted on), FN false negatives (appliance on but predicted off), and FP false positives (appliance off but predicted on). Recall measures the ratio of correctly predicted positive samples, precision measures the correctness of positive predictions, and MCC ranges from -1 to 1, with values closer to 1 indicating better identification performance.

Among these metrics, MCC offers better robustness than the commonly used F1 score for imbalanced binary classification problems [?], which naturally suits the sparse nature of appliance load events [?]. However, retaining recall and precision helps identify specific model weaknesses—for instance, high recall with low precision suggests a model tends to predict appliances as active more frequently than warranted, inspiring improvements through sample rebalancing.

2.1.3 Regression Accuracy

To evaluate regression accuracy, we select three metrics [?]: Mean Absolute Error (MAE), Normalized Error Percentage (NEP), and On-state Mean Absolute Error (OMAE):

$$\mathrm{MAE} = \frac{1}{T}\sum_{t=1}^{T} |p_t - \hat{p}_t|$$

$$\mathrm{NEP} = \frac{\sum_{t=1}^{T} |p_t - \hat{p}_t|}{\sum_{t=1}^{T} p_t}$$

$$\mathrm{OMAE} = \frac{1}{N}\sum_{i=1}^{N} |e_i - \hat{e}_i|$$

where p_t is the true power at time t, $\hat{p}_t$ the disaggregated power, T the test sequence length, e_i the true power during on-state at timestamp i, $\hat{e}_i$ the disaggregated power during on-state, and N the number of on-state timestamps.

Notably, previous studies often used only MAE, which inadequately reflects model performance—a simple prediction of zero power at all times can yield small MAE for low-power appliances. We propose NEP to eliminate this issue. Furthermore, MAE and NEP may smooth out errors for rarely-used appliances due to the large base T, which OMAE addresses by focusing evaluation on disaggregation performance during active periods.

2.1.4 Generalization Performance

Generalization remains an open challenge in load disaggregation and deep learning [?]. While [?] proposed measuring generalization loss as performance degradation on unseen versus seen scenarios, it overlooks potentially poor performance on seen scenarios. Transfer learning research [?, ?] directly considers target domain (unseen scenario) performance. Therefore, we recommend computing MCC, MAE, and other metrics on unseen scenarios and comparing them with seen-scenario results when evaluating generalization.

2.2 Analysis of Existing Public Test Data

Test data selection significantly impacts evaluation accuracy and comprehensiveness. Dataset attribute differences can produce varying results for the same algorithm, while overly narrow appliance selection leads to biased assessments. We establish the following guidelines for dataset and appliance selection.

2.2.1 Characteristics of Existing Datasets To clarify attribute differences across datasets, we compiled relevant NILM datasets using the NILMTK framework [?] and statistical methods from [?], as shown in Appendix A Table A1. The table reveals that AMPds2 [?] and COMBED [?] have low sampling frequencies that may miss load information, increasing disaggregation difficulty. Datasets like iAWE [?] and REDD [?] have insufficient duration to provide adequate training data for deep learning models. The average daily load event count indicates household activity level—higher values imply denser events and greater overlap, posing greater challenges. Consequently, we recommend UK-DALE [?] and REFIT [?] for evaluating single-feature input algorithms, while multi-feature algorithms [?] incorporating various power types should use AMPds2 or iAWE.

2.2.2 Appliance Selection Guidelines Following [?], household appliances can be categorized into four types based on power consumption patterns:

Table 1 Descriptions of Four Appliance Types

Type	Power Pattern Characteristics	Examples
Type-I	Binary on/off states only	Kettle, microwave
Type-II	Power transitions within a finite discrete set	Washing machine, dishwasher
Type-III	Continuously varying power	Dimmer switches, variable-speed motors
Type-IV	Periodic power patterns	Refrigerators, HVAC systems

Currently, most algorithms handle Type-I and Type-IV appliances well but struggle with state transition detection for Type-II appliances. Type-III appliances remain particularly challenging. For comprehensive evaluation, we recommend testing across all four types, with special attention to Type-II and Type-III performance.

3 Case Studies

This section validates the proposed evaluation framework by assessing and comparing mainstream deep learning disaggregation models with different network structures, and investigates the impact of input features on model performance.

3.1 Experimental Setup

The hardware environment comprises a 64-bit computer with 6× Intel(R) Xeon(R) CPU E5-2678 v3 @ 2.50GHz and a GeForce RTX 2080 Ti (13.13 TFLOPS, 11GB GPU memory). The software platform includes Python 3.7 (64-bit), PyTorch 1.7.0, and JupyterLab. GPU-accelerated PyTorch [?] was used for training.

The evaluated models include: Bi-LSTM [?], DAE [?], Seq2Point [?], SGN-sp [?], CNN_{Attention} [?], and EnerGAN [?]. Bi-LSTM belongs to RNN-based models; Seq2Point, SGN-sp, and CNN_{Attention} are CNN-based; DAE and EnerGAN are DAE-based. Network topologies are detailed in Appendix A Figures A1-A6. Our implementations improved upon original designs: for Bi-LSTM, we concatenated all hidden states from the second Bi-LSTM layer as fully connected layer input [?]; for EnerGAN, we incorporated reconstruction loss into generator training following most pix2pix tasks [?, ?] for more stable convergence. Further implementation details and code are available at https://github.com/Ming-er/NeuralNILM_{Pytorch}.

Experiments used the UK-DALE dataset with five appliance categories: refrigerator, kettle, washing machine, microwave, and dishwasher. Kettles and microwaves represent Type-I appliances; washing machines and dishwashers are Type-II; refrigerators are Type-IV. This selection follows comparative NILM studies [?, ?, ?] and reflects researcher interest in categories that expose algorithmic differences. For each appliance, we ensured non-overlapping training and test sets within the same scenario, downsampled aggregate and sub-meter data to 30-second intervals, and applied z-score normalization [?]. Training/testing scenario splits and input sequence lengths are detailed in Appendix A Tables A2 and A3.

Considering hyperparameter impacts, we adopted researcher-recommended settings for learning rate, optimizer, and initialization [?, ?, ?, ?], which demonstrate strong adaptability across scenarios. To account for training randomness, experiments were repeated with different random seeds and results averaged. Maximum training epochs were set to 200 with early stopping after 3 epochs of validation loss plateau to prevent overfitting.

3.2 Performance Evaluation of Deep Learning Models with Different Network Structures

This subsection evaluates CNN_{Attention}, Bi-LSTM, DAE, Seq2Point, SGN-sp, and EnerGAN models using the proposed framework. All models used active power sequence P as input, following the settings in Section 3.1. Disaggregated power curves are shown in Appendix Figures B1-B5.

3.2.1 Model Pre-Evaluation Pre-evaluation results using metrics from Section 2.1.1 are presented in Appendix Table A4. Bi-LSTM exhibits longer training and inference times because RNN models must integrate previous time step information when encoding current load data, requiring frequent CPU-GPU scheduling that hinders parallelization. Seq2Point demonstrates superior training and inference efficiency thanks to convolution' s parallel feature extraction. CNN_{Attention} introduces minimal overhead despite its attention block, while SGN-sp' s large parameter count and gated branch architecture incur significant computational costs. EnerGAN' s longer training time stems from adversarial training. Overall, lightweight networks like DAE perform well in pre-evaluation, while deploying large models like SGN-sp to embedded terminals requires model compression techniques [**?**].

3.2.2 Identification Accuracy Comparison Identification accuracy results using Section 2.1.2 metrics are shown in Appendix Figure A7. All models achieve MCC $>$ 0.75 for simple on/off appliances (kettle, refrigerator), but show lower recall for microwaves due to extremely sparse load events in the dataset, causing models to bias toward predicting the off state. For Type-II appliances (dishwasher, washing machine), most models except SGN-sp exhibit degraded identification accuracy. CNN-based models (Seq2Point, SGN-sp) outperform others on Type-II appliances, suggesting convolution' s multi-timescale, multi-magnitude power change detection capability. EnerGAN' s superior state identification over DAE indicates that discriminators can correct load event reasoning during adversarial training. Overall, CNN models' multi-layer, multi-scale kernels enhance their ability to capture load events of varying duration and magnitude, while appliance-specific performance variations underscore the necessity of standardized appliance selection. Researchers can improve identification through adversarial learning as in EnerGAN.

3.2.3 Regression Accuracy Comparison Regression accuracy results using Section 2.1.3 metrics are presented in Appendix Figure A8. CNN models like SGN-sp show lower MAE due to convolutional kernels' proficiency in extracting local information and learning power variation patterns, coupled with their superior identification accuracy (strongly correlated with power prediction). However, Bi-LSTM demonstrates more accurate power regression during on-states because its step-wise information integration provides better modeling continuity. SGN-sp and CNN_{Attention} outperform Seq2Point across Type-I, II, and IV appliances, indicating that incorporating on/off state information streams and weighted fusion of intermediate features enhances disaggregation.

Relying solely on MAE would incorrectly suggest excellent performance on kettles, but high NEP and OMAE values reveal significant errors during on-state predictions, demonstrating the proposed framework' s comprehensiveness. EnerGAN achieves SGN-sp-level accuracy for kettles and refrigerators with fewer parameters and faster inference, while Bi-LSTM and SGN-sp excel for complex appliances, suggesting ensemble approaches using different models per appliance.

3.2.4 Generalization Performance Comparison Generalization performance on unseen scenarios is summarized in Appendix Figure A9. All five deep learning models show minimal generalization loss for simple appliances (kettle, refrigerator) between seen and unseen scenarios. However, for complex Type-II appliances, MAE and MCC degrade significantly, with more complex models like SGN-sp showing larger drops. This indicates that deep learning models' generalization for Type-II appliances needs improvement, and complex networks face higher structural risks on unseen data. Researchers should consider data augmentation [?], transfer learning [?], and regularization to enhance generalization and prevent overfitting. Future research must continue addressing cross-scenario generalization as a core challenge.

3.3 Performance Evaluation with Different Input Features

This subsection evaluates Bi-LSTM, DAE, and Seq2Point models using the proposed framework with input features: active power P only; P and reactive power Q; P and apparent power S; and P, Q, and S combined.

3.3.1 Model Pre-Evaluation Pre-evaluation results are shown in Appendix Table A5. Additional feature dimensions do not significantly increase spatiotemporal complexity but accelerate convergence during training. However, in deployment, more features increase data storage and transmission bandwidth requirements, suggesting the need for compressive sensing techniques [?, ?] to reduce embedded terminal workload.

3.3.2 Identification Accuracy Comparison Identification accuracy results are presented in Appendix Figure A10. Adding even one feature dimension significantly improves identification accuracy. Since apparent power S is dominated by active power P, creating strong collinearity, using reactive power Q as an additional feature provides richer load information and more noticeable performance gains. Using all three features (P, Q, S) may cause information redundancy without substantial improvement over P and Q alone. Multi-feature inputs notably improve recall for sparsely active appliances (kettle, microwave) and precision for complex Type-II appliances (washing machine, dishwasher), demonstrating enhanced reasoning for complex load events.

3.3.3 Regression Accuracy Comparison Regression accuracy results are shown in Appendix Figure A11 and Appendix Figures B6-B10. Multi-feature inputs significantly improve disaggregation performance, though the enhancement varies by appliance. Lower OMAE values indicate improved on-state power prediction accuracy. Notably, the lightweight DAE network with two-dimensional features approaches the performance of complex one-dimensional Bi-LSTM models, suggesting that effective additional features offer an efficient performance improvement method.

3.3.4 Generalization Performance Comparison Generalization performance on unseen scenarios is summarized in Appendix Figure A12. Multi-feature inputs significantly improve generalization for refrigerator, washing machine, and dishwasher, but may cause negative transfer for kettle and microwave. Therefore, researchers must carefully consider electrical characteristics across features in seen and unseen scenarios when using additional features to enhance generalization.

4 Outlook

The preceding discussion reveals that numerous deep learning methods exist for NILM, with many achieving high identification and disaggregation accuracy. However, gaps remain in practicality, robustness, and generalization. Future deep learning applications in low-frequency NILM should focus on:

1) **Edge Computing and Model Compression**: As edge computing IoT architectures deploy in power systems [?], deep learning models remain computationally intensive. Knowledge distillation [?], weight pruning [?], network quantization [?], and neural architecture search [?] can compress models. Additionally, cloud-edge collaboration raises concerns about data transmission efficiency and user privacy protection.

2) **Generalization Evaluation and Meta-Learning**: Unified generalization metrics are lacking, and most algorithms rely on supervised learning requiring massive labeled data that is difficult to obtain. While existing transfer learning methods still need some target domain data [?], meta-learning [?] and zero-shot learning [?] approaches could drastically reduce data requirements and address generalization to unseen scenarios.

3) **Graph Neural Networks and Hybrid Methods**: Graph signal processing offers short training times and strong representation [?]. Graph Neural Networks (GNN) [?] could combine these advantages to significantly improve disaggregation by 挖掘数据样本间的联系. Integrating deep learning with traditional methods may also overcome data dependency while boosting traditional method performance.

4) **Remaining Challenges**: Effective capture of long- and short-term temporal dependencies, 长尾分布 caused by sparse events from rarely-used devices, and masking of small appliances by simultaneously operating

large ones remain bottlenecks. Transformer architectures leveraging global sequence information [?] and time-adaptive convolutions [?] could enhance multi-timescale event capture. Strategies from computer vision like reweighting [?] and mixup training [?] may improve model adaptation to imbalanced samples.

5 Conclusion

This paper reviews, compares, and prospects deep learning methods for low-frequency non-intrusive load disaggregation. However, due to deep learning's inherent limitations—numerous hyperparameters, poor interpretability, and training randomness—our comparative exploration remains preliminary and coarse-grained. Future work will apply the proposed comprehensive evaluation framework to investigate sensitivity to hyperparameter selection and extend studies to Type-III appliances not covered in current comparisons.

References

[1] HART G W. Nonintrusive appliance load monitoring[J]. Proceedings of the IEEE, 1992, 80(12): 1870-1891.

[2] EHRHARDT-MARTINEZ K, DONNELLY K A. Advanced Metering Initiatives and Residential Feedback Programs: A Meta-Review for Household Electricity-Saving Opportunities[R]. USA: ACEEE, 2010.

[3] RUANO A, HERNANDEZ A, UREÑA J, et al. NILM Techniques for Intelligent Home Energy Management and Ambient Assisted Living: A Review: 11[J]. Energies, 2019, 12(11): 2203.

[4] TIAN Shiming, WANG Beibei, ZHANG Jin. Key Technologies for Demand Response in Smart Grid[J]. Proceedings of the CSEE, 2014, 34(22): 3576-3589.

[5] CHENG Xiang, LI Linzhi, WU Hao, et al. A survey of the research on non-intrusive load monitoring and disaggregation[J]. Power System Technology, 2016, 40(10): 3108-3117.

[6] BHOTTO M, MAKONIN S, BAJIC I. Load disaggregation based on aided linear integer programming[J]. IEEE Transactions on Circuits and Systems II: Express Briefs, 2017, 64(7): 792-796.

[7] WEISS M, HELFENSTEIN A, MATTERN F, et al. Leveraging smart meter data to recognize home appliances[C]//2012 IEEE International Conference on Pervasive Computing and Communications. March 19-23, 2012, Lugano, Switzerland: 190-197.

[8] FAUSTINE A, MVUNGI N H, KAIJAGE S, et al. A Survey on Non-Intrusive Load Monitoring Methods and Techniques for Energy Disaggregation Problem[J/OL]. arXiv:1703.00785, 2017.

[9] KIM H, MARWAH M, ARLITT M, et al. Unsupervised Disaggregation of Low Frequency Power Measurements[C]//The 11th SIAM International Conference on Data Mining, April 28-30, 2011, Mesa, USA: 747-758.

[10] ZHONG M, GODDARD N, SUTTON C. Signal Aggregate Constraints in Additive Factorial HMMs, with Application to Energy Disaggregation[C]//The

28th International Conference on Neural Information Processing Systems. December 8-13, 2014, Montreal, Canada: 3590-3598.
[11] DENG Xiaoping, ZHANG Guiqing, WEI Qinglai, et al. A Survey on the Non-intrusive Load Monitoring[J/OL]. Acta Automatica Sinica:
[12] HORNIK K, STINCHCOME M, WHITE H. Multilayer feedforward networks are universal approximators[J]. Neural Networks, 1989, 2(5): 359-366.
[13] KELLY J, KNOTTENBELT W. Neural NILM: Deep neural networks applied to energy disaggregation[C]//The 2nd ACM International Conference on Embedded Systems for Energy-Efficient Built Environments. November 4-5, 2015, Seoul, South Korea: 55-64.
[14] CUI G, LIU B, LUAN W, et al. Estimation of Target Appliance Electricity Consumption Using Background Filtering[J]. IEEE Transactions on Smart Grid, 2019, 10(6): 5920-5929.
[15] NALMPANTIS C, VRAKAS D. Machine Learning Approaches for Non-Intrusive Load Monitoring: From Qualitative to Quantitative Comparation[J]. Artificial Intelligence Review, 2019, 52(1): 217-243.
[16] VINCENT P, LAROCHELLE H, BENGIO Y, et al. Extracting and Composing Robust Features with Denoising Autoencoders[C]//Proceedings of the 25th International Conference on Machine Learning. July 5-9, 2008, Helsinki, Finland: 1096-1103.
[17] CHO K. Simple Sparsification Improves Sparse Denoising Autoencoders in Denoising Highly Noisy Images[C] //Proceedings of the 30th International Conference on Machine Learning. June 16-21, 2013, Atlanta, USA: 432-440.
[18] LU X, TSAO Y, MATSUDA S, et al. Speech Enhancement Based on Deep Denoising Autoencoder[C]//INTERSPEECH. August 25-29, 2013, Lyon, France: 436-440.
[19] DINESH C, MAKONIN S, BAJIC I V. Incorporating Time-of-Day Usage Patterns into Non-Intrusive Load Monitoring[C]//2017 IEEE Global Conference on Signal and Information Processing. November 14-16, 2017, Montreal, QC, Canada: 1110-1114.
[20] VALENTI M, BONFIGLI R, PRINCIPI E, et al. Exploiting the Reactive Power in Deep Neural Models for Non-Intrusive Load Monitoring[C]//2018 International Joint Conference on Neural Networks. July 8-13, 2018, Rio de Janeiro, Brazil: 1-8.
[21] BONFIGLI R, FELICETTI A, PRINCIPI E, et al. Denoising Autoencoders for Non-Intrusive Load Monitoring: Improvements and Comparative Evaluation[J]. Energy and Buildings, 2018, 158(1):
[22] BARSIM K S, YANG B. On the Feasibility of Generic Deep Disaggregation for Single-Load Extraction[J/OL]. arXiv:1802.02139, 2018.
[23] GOODFELLOW I J, POUGET-ABADIE J, MIRZA M, et al. Generative Adversarial Networks[J/OL]. arXiv: 1406.2661, 2018.
[24] CHEN K, ZHANG Y, WANG Q, et al. Scale and Context-Aware Convolutional Non-Intrusive Load Monitoring[J]. IEEE Transactions on Power Systems, 2020, 35(3): 2362-2373.
[25] KASELIMI M, VOULODIMOS A, PROTOPAPADAKIS E, et al. EnerGAN: A Generative Adversarial Network For Energy Disaggregation[C]//2020

International Conference on Acoustics, Speech and Signal Processing. May 4-8, 2020, Barcelona, Spain: 1578-1582.
[26] PAN Y, LIU K, SHEN Z, et al. Sequence-To-Subsequence Learning With Conditional Gan For Power Disaggregation[C]//2020 IEEE International Conference on Acoustics, Speech and Signal Processing. May 4-8, 2020, Barcelona, Spain: 3202-3206.
[27] CHENG Guixue, CHEN Boye, ZHAO Jinbin. Non-intrusive Load Disaggregation Based on Conditional Generative Adversarial Network[J/OL]. Proceedings of the CSU-EPSA:1-7.
[28] GOODFELLOW I, BENGIO Y, COURVILLE A. Deep learning [M]. Cambridge: MIT Press, 2016.
[29] HOCHREITER S, SCHMIDHUBER J. Long short-term memory[J]. Neural Computation, 1997, 9(8): 1735-1780.
[30] CHUNG J, GULCEHRE C, CHO K, et al. Empirical Evaluation of Gated Recurrent Neural Networks on Sequence Modeling[J]. arXiv:1412.3555, 2014.
[31] MAUCH L, YANG B. A New Approach for Supervised Power Disaggregation by Using a Deep Recurrent LSTM Network[C]//2015 IEEE Global Conference on Signal and Information Processing. December 14-16, 2015, Orlando, FL, USA: 63-67.
[32] KRYSTALAKOS O, NALMPANTIS C, VRAKAS D. Sliding Window Approach for Online Energy Disaggregation Using Artificial Neural Networks[C]//The 10th Hellenic Conference on Artificial Intelligence. July 9-15, 2018, Patras, Greece:1-6.
[33] HE W, CHAI Y. An Empirical Study on Energy Disaggregation via Deep Learning[C]//The 2nd International Conference on Artificial Intelligence and Industrial Engineering. November 20-21, 2016, Beijing, China:338-342.
[34] WANG Ke, ZHONG Haiwang, YU Nanpeng, et al. Nonintrusive Load Monitoring based on Sequence-to-sequence Model With Attention Mechanism[J]. Proceedings of the CSEE, 2019, 39(1):75-83.
[35] REN Wenlong, XU Gang. Non-intrusive Load Decomposition Method Based on Deep Sequence Translation Model[J], Power System Technology, 2020, 44(01): 27-37.
[36] LIU Yaoxian, SUN Yi, LI Bin, et al. Non-intrusive Load Disaggregation Method Based on Edge Embedded Deep Learning[J], Power System Technology, 2019, 43(12): 4329-4337.
[37] CIMEN H, CETINKAYA N, VASQUEZ J C, et al. A Microgrid Energy Management System Based on Non-Intrusive Load Monitoring via Multitask Learning[J]. IEEE Transactions on Smart Grid, 2020: 1-1.
[38] RAFIQ H, ZHANG H, LI H, et al. Regularized LSTM Based Deep Learning Model: First Step towards Real-Time Non-Intrusive Load Monitoring[C/OL]//IEEE International Conference on Smart Energy Grid Engineering. August 12-15, 2018, Oshawa, Canada: 234-239.
[39] KASELIMI M, DOULAMIS N, DOULAMIS A, et al. Bayesian-Optimized Bidirectional LSTM Regression Model for Non-Intrusive Load Monitoring[C]//2019 IEEE International Conference on Acoustics, Speech and Signal Processing. May 12-17, 2019, Brighton, United Kingdom: 2747-2751.

[40] GREFENSTETTE E, BLUNSOM P; DE F, et al. A Deep Architecture for Semantic Parsing[J]. arXiv: 1404.7296, 2014.
[41] KIM Y. Convolutional Neural Networks for Sentence Classification[C]//Conference on Empirical Methods in Natural Language Processing. October 25-29, 2014, Doha, Qatar: 1746-1751.
[42] BAI S, KOLTER J Z, KOLTUN V. An Empirical Evaluation of Generic Convolutional and Recurrent Networks for Sequence Modeling[J]. arXiv: 1803.01271, 2018.
[43] ZHANG C, ZHONG M, WANG Z, et al. Sequence-to-point learning with neural networks for non-intrusive load monitoring[C]//The 32nd AAAI Conference on Artificial Intelligence. February 2-7, 2018, Louisiana, USA: 2604-2611.
[44] KONG W, DONG Z Y, WANG B. A Practical Solution for Non-Intrusive Type II Load Monitoring Based on Deep Learning and Post-Processing[J]. IEEE Transactions on Smart Grid, 2020, 11(1): 148-158.
[45] SHIN C, JOO S, YIM J. Subtask Gated Networks for Non-Intrusive Load Monitoring[J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2019, 33(01): 1150-1157.
[46] XU Xiaohui, ZHAO Shutao, CUI Kebin. Non-intrusive Load Disaggregate Algorithm Based on Convolutional Block Attention Module[J/0L]. Power System Technology, 2021: 1-8.
[46] HARELL A, MAKONIN S, BAJIC I V. Wavenilm: A Causal Neural Network for Power Disaggregation from the Complex Power Signal[C]//2019 IEEE International Conference on Acoustics, Speech and Signal Processing. May 12-17, 2019, Brighton, UK: 8335-8339.
[47] OORD A van den, DIELEMAN S, ZEN H, et al. WaveNet: A Generative Model for Raw Audio[C]//The 9th ISCA Speech Synthesis Workshop, September 13-15, Sunnyvale, USA: 125.
[48] FAUSTINE A, PEREIRA L, BOUSBIAT H, et al. UNet-NILM: A Deep Neural Network for Multi-Tasks Appliances State Detection and Power Estimation in NILM[C]//Proceedings of the 5th International Workshop on Non-Intrusive Load Monitoring. November 18, 2020, Virtual Event, Japan: 84-88.
[49] RONNEBERGER O, FISCHER P, BROX T. U-Net: Convolutional Networks for Biomedical Image Segmentation[J/OL]. arXiv:1505.04597, 2015.
[50] JIA Y, BATRA N, WANG H. A Tree-Structured Neural Network Model for Household Energy Breakdown[C]//International World Wide Web Conference. May 13-17, 2019, San Francisco, USA: 2872-2878.
[51] ZHANG Yutian, DENG Chunyu, LIU Yuankun, et al. Non-intrusive Load Identification Algorithm Based on Convolution Neural Network[J]. Power System Technology, 2020, 44(06): 2038-2044.
[52] HAFSA BOUSBIAT, CHRISTOPH KLEMENJAK, WILFRD ELMENREICH. Exploring Time Series Imaging for Load Disaggregation[C]//The 7th ACM International Conference on Systems for Energy-Efficient Buildings, Cities, and Transportation. November 18-20, 2020, Virtual Event, Japan: 254-257.

[53] BAI Yuyang, HUANG Yanhao, CHEN Siyuan, et al. Cloud-edge intelligence: status quo and future prospective of edge computing approaches and applications in power system operation and control[J]. Acta Automatica Sinica, 2020, 46(3): 397-410.
[54] PEREIRA L, NUNES N. Performance Evaluation in Non-intrusive Load Monitoring: Datasets, Metrics, and Tools-A Review[J/OL]. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 2018, 8(6).
[55] MAKONIN S, POPOWICH F. Nonintrusive Load Monitoring (NILM) Performance Evaluation: A Unified Approach for Accuracy Reporting[J]. Energy Efficiency, 2015, 8(4): 809-814.
[56] NALMPANTIS C, VRAKAS D. Machine Learning Approaches for Non-Intrusive Load Monitoring: From Qualitative to Quantitative Comparation[J]. Artificial Intelligence Review, 2019, 52(1): 217-243.
[57] CHICCO D, JURMAN G. The Advantages of the Matthews Correlation Coefficient (MCC) over F1 Score and Accuracy in Binary Classification Evaluation[J]. BMC Genomics, 2020, 21(1): 6.
[58] MAKONIN S, POPOWICH F, BAJIC I V, et al. Exploiting HMM Sparsity to Perform Online Real-Time Nonintrusive Load Monitoring[J]. IEEE Transactions on Smart Grid, 2016, 7(6): 2575-2585.
[59] KLEMENJAK C, FAUSTINE A, MAKONIN S, et al. On Metrics to Assess the Transferability of Machine Learning Models in Non-Intrusive Load Monitoring[J]. arXiv preprint arXiv:1912.06200, 2019.
[60] BOUSMALIS K, TRIGRORGIS G, SILBERMAN N, et al. Domain separation networks[C]//In Advances in Neural Information Processing Systems. December 5-10, 2016, Barcelona, Spain: 343-351.
[61] GANIN Y, USTINOVA E, AJAKAN H, et al. Domain-adversarial training of neural networks[J]. Journal of Machine Learning Research, 2016, 17(59): 1-35.
[62] MAKONIN S, ELLERT B, BAJIĆ I V, et al. Electricity, Water, and Natural Gas Consumption of a Residential House in Canada from 2012 to 2014[J]. Scientific Data, 2016, 3(1): 160037.
[63] BATRA N, PARSON O, BERGES M, et al. A comparison of non-intrusive load monitoring methods for commercial and residential buildings[J]. arXiv:1408.6595, 2014.
[64] BATRA N, GULATI M, SINGH A, et al. It' s Different: Insights into home energy consumption in India[C]//The 5th ACM Workshop on Embedded Systems For Energy-Efficient Buildings. November 14-15, 2013, Roma, Italy: 1-8.
[65] KOLTER J Z, JOHNSON M J. REDD: a public data set for energy disaggregation research [C]// SustKDD Workshop, 2011, San Diego, USA: 6.
[66] KELLY J, KNOTTENBELT W. The UK-DALE Dataset, Domestic Appliance-Level Electricity Demand and Whole-House Demand from Five UK Homes[J]. Scientific Data, 2015, 2(1): 150007.
[67] MURRAY D, STANKOVIC L, STANKOVIC V. An Electrical Load Measurements Dataset of United Kingdom Households from a Two-Year Longitudinal Study[J]. Scientific Data, 2017, 4(1): 1-22.

[68] KASELIMI M, PROTOPAPADAKIS E, VOULODIMOS A, et al. Multi-Channel Recurrent Convolutional Neural Networks for Energy Disaggregation[J]. IEEE Access, 2019, 7: 81047-81056.
[69] KIM J, LE T-T-H, KIM H. Nonintrusive Load Monitoring Based on Advanced Deep Learning and Novel Signature[J]. Computational Intelligence and Neuroscience, 2017: 4216281.
[70] CHEN Xuan, KAN Bowen, LIU Guangyi. The latest development of GPU and its prospective application in power system[J]. Electric Power Information and Communication Technology, 2018, 16(3): 16-25.
[71] ZHOU P, QI Z, ZHENG S, et al. Text Classification Improved by Integrating Bidirectional LSTM with Two-dimensional Max Pooling[C]//The 26th International Conference on Computational Linguistics. December 13-16, 2016, Osaka, Japan: 3485-3495.
[72] ISOLA P, ZHU JY, ZHOU T, et al. Image-to-Image Translation with Conditional Adversarial Networks[C/OL]//2017 IEEE Conference on Computer Vision and Pattern Recognition. July 22-25, 2017, Honolulu, USA: 5967-5976.
[73] Y LI, T JIANG, S QIN, et al. Multi-scale Generative Adversarial Networks for Electrical Data Augmentation in NILM[C]//2019 IEEE Global Conference on Signal and Information Processing. November 11-14, 2019, Ottawa, Canada: 1-5.
[74] CHANG F, CHEN C, LIN C. An Empirical Study of Ladder Network and Multitask Learning on Energy Disaggregation in Taiwan[C]//2018 Conference on Technologies and Applications of Artificial Intelligence. November 30-December 12, 2018, Taichung: 86-89.
[75] AHMED S, BONS M. Edge Computed NILM: A Phone-Based Implementation Using MobileNet Compressed by Tensorflow Lite[C]//Proceedings of the 5th International Workshop on Non-Intrusive Load Monitoring. November 18, 2020, Virtual Event Japan: 100-109.
[76] DELFOSSE A, HEBRAIL G, ZERROUG. Is Data Augmentation Worth for NILM: A Deep Learning Applied for Energy Disaggregation[C]//The 9th International Conference on Prestigious Applications of Intelligent Systems. August 29-September 8, 2020, Santiago de Compostela, Spain: 29.
[77] KASELIMI M, DOULAMIS N, VOULODIMOS A, et al. Context Aware Energy Disaggregation Using Adaptive Bidirectional LSTM Models[J]. IEEE Transactions on Smart Grid, 2020, 11(4): 3054-3067.
[78] HARELL A, JONES R, MAKONIN S, et al. PowerGAN: Synthesizing Appliance Power Signatures Using Generative Adversarial Networks[J/OL]. arXiv:2007.13645, 2020.
[79] D' INCECCO M, SQUARTINI S, ZHONG M. Transfer Learning for Non-Intrusive Load Monitoring[J]. IEEE Transactions on Smart Grid, 2020, 11(2): 1419-1429.
[80] YANG Y, ZHONG J, LI W, et al. Semi-Supervised Multi-Label Deep Learning Based Non-Intrusive Load Monitoring in Smart Grids[J]. IEEE Transactions on Industrial Informatics, 2020, 16(11): 6892-6904.
[81] WANG S, DU L, ZHOU Q. A Semi-Supervised Deep Transfer Learning Architecture for Energy Disaggregation[C/OL]//2019 IEEE Power & Energy

Society General Meeting (PESGM). August 4-8, 2019, Atlanta, USA: 1-5.
[82] LIU Guolong, ZHAO Junhua, WEN Fushuan, et al. Deep End-to-end Super-resolution Perception Method for Load Data at Distribution Side[J]. Automation of Electric Power Systems, 2020, 44(24): 28-40.
[83] CLARK M, LAMPE L. Single-channel compressive sampling of electrical data for non-intrusive load monitoring[C]//2015 IEEE Global Conference on Signal and Information Processing. December 14-16, 2015, Orlando, FL, USA: 790-794.
[84] HINTON G, VINYALS O, DEAN J. Distilling the Knowledge in a Neural Network[J/OL]. arXiv: 1503.02531, 2015.
[85] HAN S, POOL J, TRAN J, et al. Learning Both Weights and Connections for Efficient Neural Networks[J/OL]. arXiv: 1506.02626, 2015.
[86] HUBARA I, COURBARIAUX M, SOUDRY D, et al. Quantized Neural Networks: Training Neural Networks with Low Precision Weights and Activations[J/OL]. arXiv:1609.07061, 2017.
[87] TAN M, CHEN B, PANG R, et al. MnasNet: Platform-Aware Neural Architecture Search for Mobile[C]//2019 IEEE Conference on Computer Vision and Pattern Recognition. June 16-20, 2019, Long Beach, USA: 2820-2828.
[88] HOSPEDALES T, ANTONIOU A, MICAELLI P, et al. Meta-Learning in Neural Networks: A Survey[J/OL]. arXiv:2004.05439, 2020.
[89] XIAN Y, LAMPERT C H, SCHIELE B, et al. Zero-Shot Learning-A Comprehensive Evaluation of the Good, the Bad and the Ugly[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2019, 41(9): 2251-2265.
[90] HE K, STANKOVIC L, LIAO J, et al. Non-Intrusive Load Disaggregation Using Graph Signal Processing[J]. IEEE Transactions on Smart Grid, 2018, 9(3): 1739-1747.
[91] ZHOU J, CUI G, ZHANG Z, et al. Graph Neural Networks: A Review of Methods and Applications[J/OL]. arXiv:1812.08434, 2018.
[92] VASWANI A, SHAZEER N, PARMAR N, et al. Attention Is All You Need[J/OL]. arXiv:1706.03762, 2017.
[93] LIOUTAS V, GUO Y. Time-aware Large Kernel Convolutions[J/OL]. arXiv:2002.03184, 2020.
[94] CUI Y, JIA M, SONG Y, et al. Class-Balanced Loss Based on Effective Number of Samples[C]//2019 IEEE Conference on Computer Vision and Pattern Recognition. June 16-20, 2019, Long Beach, USA: 9268-9277.
[95] VERMA V, LAMB A, BECKHAM C, et al. Manifold Mixup: Better Representations by Interpolating Hidden States[C]//Proceedings of the 36th International Conference on Machine Learning. June 9-15, 2019, Long Beach, USA, 2019: 6438-6447.

LI Yiming (2000–), male, research assistant. Research interests: load disaggregation, smart meter data mining. E-mail: yiming@cau.edu.cn

JU Yuntao (1985–), male, Ph.D., associate professor. Research interests: integrated energy system simulation and optimization. E-mail: juyuntao@cau.edu.cn

QU Liao (2000–), male, research assistant. Research interests: artificial intelligence and its applications. E-mail: liaoqu@cau.edu.cn

---

Appendix A: Tables and Figures

Table A1 Comparison of Selected Properties in Common Public Datasets

Dataset	Houses	Duration (days)	Aggregate Sampling (s)	Sub-meter Sampling (s)	Aggregate Power Types	Sub-meter Power Types	Avg. Daily Events
AMPds	12	730	60	60	V, I, P, Q, S, etc.	V, I, P, Q, S, etc.	70.2
COMBED	1	112	60	1	P	P	45.3
REFIT	20	720-1218	8	8	P, Q, S	P, Q, S	52.1
UK-DALE	5	156-1125	6	6	P	P	48.7
iAWE	1	73	1	1	P, Q, S	P, Q, S	89.6
REDD	6	3-19	1	1	P	P	76.8

Table A2 Selection of Training and Testing Scenes for Specific Appliances

Appliance	Training House/Period	Seen Test House/Period	Unseen Test House (Generalization)/Period
Dishwasher, Washing Machine	House 1: 2013-06-01~2014-01-01	House 1: 2014-01-01~2014-02-01	House 2: 2013-06-01~2013-07-01
Kettle, Microwave, Refrigerator	House 5: 2014-07-15~2014-10-15	House 5: 2014-07-01~2014-07-15	House 1: 2013-06-01~2014-01-01

Note: Due to UK-DALE' s power type limitations, House 5 is only used for single-feature experiments (Section 3.3).

Table A3 Time Span of Input Sequence and On-Threshold of Specific Appliances During Training

Appliance	Input Sequence Span (s)	On-Threshold (W)
Dishwasher	1530	10
Washing Machine	1530	20
Kettle	510	1000
Microwave	510	50
Refrigerator	510	10

Table A4 Results of Model Pre-Evaluation in Section 3.2.1

Model	t (min)	m (MB)	v (ms/sample)
Bi-LSTM	45.2	128.3	0.82
Seq2Point	12.8	45.6	0.15
CNN_{Attention}	13.5	48.2	0.16
SGN-sp	68.7	256.8	0.95
DAE	8.3	32.1	0.11
EnerGAN	52.4	35.7	0.14

Note: Variables t, m, v are defined in Section 2.1.1.

Table A5 Results of Model Pre-Evaluation in Section 3.3.1

Model	Input Features	t (min)	m (MB)	v (ms/sample)
Bi-LSTM	P	45.2	128.3	0.82
Bi-LSTM	P+Q	43.8	129.1	0.83
Bi-LSTM	P+S	44.5	129.0	0.82
Bi-LSTM	P+Q+S	44.1	129.8	0.84
Seq2Point	P	12.8	45.6	0.15
Seq2Point	P+Q	12.1	46.3	0.15
Seq2Point	P+S	12.5	46.2	0.15
Seq2Point	P+Q+S	12.3	47.0	0.16
DAE	P	8.3	32.1	0.11
DAE	P+Q	7.9	32.8	0.11
DAE	P+S	8.1	32.7	0.11
DAE	P+Q+S	8.0	33.5	0.12

Note: Variables t, m, v are defined in Section 2.1.1.

Figure A1 Schematic diagram of the structure of the Seq2Point model

Figure A2 Schematic diagram of the structure of the Bi-LSTM model

Figure A3 Schematic diagram of the structure of the DAE model

Figure A4 Schematic diagram of the structure of the SGN-sp model

Note: Figure translations are in progress. See original paper for figures.