AN EFFICIENT DEEP BIDIRECTIONAL TRANSFORMER MODEL FOR ENERGY DISAGGREGATION

Stavros Sykiotis, Maria Kaselimi, Anastasios Doulamis, and Nikolaos Doulamis

National Technical University of Athens, 15773 Athens, Greece
stasykiotis@mail.ntua.gr

ABSTRACT

In this study, we present TransformNILM, a novel Transformer based model for Non-Intrusive Load Monitoring (NILM). To infer the consumption signal of household appliances, TransformNILM employs Transformer layers, which utilize attention mechanisms to successfully draw global dependencies between input and output sequences. TransformNILM does not require data balancing and operates with minimal dataset pre-processing. Compared to other Transformer-based architectures, TransformNILM instigates an efficient training scheme, where model training consists of unsupervised pre-training and supervised model fine-tuning, thus leading to decreased training time and improved predictive performance. Experimental results validate TransformNILM's superiority compared to several state of the art methods.

Index Terms**—** Non-intrusive Load Monitoring, NILM, Transformers, Attention, computational efficiency, class imbalance

1. INTRODUCTION

Non-Intrusive Load Monitoring (NILM), or energy disaggregation, is an efficient and cost effective framework to reduce energy consumption [1], where the aggregate power consumption signal of a household is decomposed into the power signals of the respective domestic appliances. Solving the NILM problem has been studied in various works.Some of the most successful utilize deep learning structures [2]–[4] to extract the individual appliance consumption patterns. Even though these techniques demonstrate good performance, there are some limitations and challenges.

Challenge 1: Recurrent neural networks (RNN), Long Short-term Memory (LSTM), bidirectional LSTM (BiLSTM) and gated recurrent unit (GRU) networks are considered state of the art NILM approaches [1], [5]–[7]. These techniques utilize recurrent mechanisms [1] to extract worthwhile information and discard useless parts of the signal. Therefore, local dependencies are more powerful than global ones, and infrequent, non-regularly appearing event information fades over time. In order to maintain the important details, data balancing is a necessary prerequisite, since the datasets are skewed with sparce appliance activations, in the sense that the appliance run-time is noticeably shorter than the time it is switched off. Most of the aforementioned studies deploy a pre-processing strategy to handle data balancing properly [5].

Challenge 2:

Convolutional Neural Network (CNN)-based architectures have significantly advanced towards accurately capturing long range temporal dependencies in time series [8], [9]. CNN-based solutions that tackle NILM-related challenges have been proposed in various works [3], [10]. These networks combine causal, dilated convolutions with additional modern neural network improvements, such as residual connections and weight normalization, to reduce the required computational power without performance degradation. Even though temporal CNN structures are capable of capturing long range temporal dependencies in time series, considerable model depth is required.

Transformers [11] have swiftly surfaced across a multiple sequence modeling tasks [12]–[14], thanks to their superior scaling properties against recurrent architectures, as well as their ability to instantly and arbitrarily access information across time. The primary advantage of Transformers stems from the fact that,contrary to the aforementioned NILM architectures, a sequence is processed simultaneously, in an order-invariant way. Transformers process a sequence as a whole entity, alleviating the risk of neglecting infrequently occuring information. Even though Transformer architectures seem suitable to NILM challenges, efficiency and computational complexity issues have limited their applicability [15].

1.1. Our contribution

In this paper, we introduce TransformNILM, a Transformer-based framework for energy disaggregation. TransformNILM comprises of two components: ***(i) the pre-training process***, which is an unsupervised training scheme where only the aggregate power signal is required as input, and ***(ii) the training process***, where the pre-trained Transformer model is fine-

This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No 955422

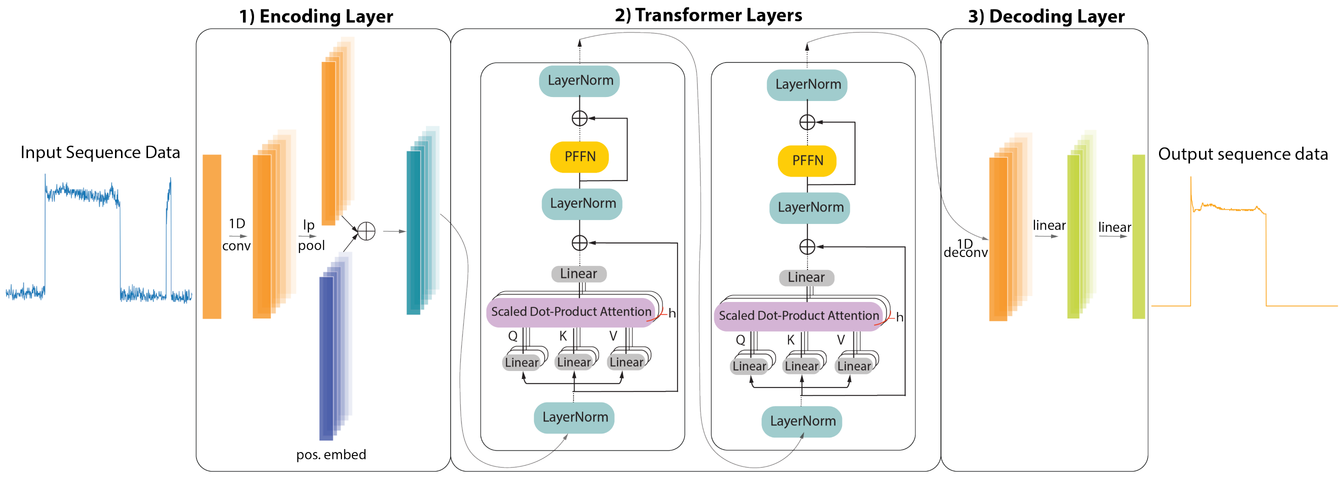

Fig. 1: Proposed model architecture for TransformNILM's Generator and Discriminator models.

tuned in a supervised manner to predict the electrical consumption of the chosen domestic appliance. This process leads to an efficient fast Transformer framework for energy disaggregation. TransformNILM's comparative advantages are summarized below:

Learns long-range temporal dependencies. Learning temporal dependencies is a challenging task, where the model often may forget the first part of the sequence before finishing its processing. TransformNILM utilizes attention mechanisms and is able to identify complex dependencies between input sequence elements regardless of their position.

Handles imbalanced datasets. Our experiments demonstrate that the combination of unsupervised pre-training process with downstream task fine-tuning creates a practical NILM solution that successfully handles dataset imbalance. This is a comparative advantage against the existing NILM state of the art models, where data balancing is usually required to achieve good performance.

Is an efficient and fast Transformer. TransformNILM combines a Generator and a Discriminator model to create a computationally efficient unsupervised pre-training process, which significantly decreases the required training time compared to other Transformer architectures without affecting model performance.

2. TRANSFORMNILM: AN EFFICIENT TRANSFORMER FOR NILM

TransformNILM proposes an efficient model training procedure for energy disaggregation. To properly formulate the task, let a household contain M appliances and i be the index indicating the i-th appliance ($i = 1, \ldots, M$) [16]. In a NILM framework [17], at any given time t, the total power consumption x is expressed as the sum of the power consumption y_i, $\forall \; i = 1, ..., M$ of the individual appliances, $x(t) = \sum_{i=1}^{M} y_i(t) + \epsilon_{noise}(t)$, where ϵ_{noise} describes a noise term. Our goal is to solve the inverse problem and, given the aggregate power signal x, estimate the appliance consumption patterns y_i. NILM is therefore formulated as a highly undetermined blind-source separation problem, since there are infinite combinations of y_i that reconstruct x.

TransformNILM utilizes the Generator/Discriminator model concept that was first introduced in Generative Adversarial Networks (GAN) [18], where two models compete against each other during training. A key difference in our approach is that training is not conducted adversarially. Instead, model training is split into a pre-training and a training routine. Generator and Discriminator cooperate during pre-training in an unsupervised way to maximize performance in the training stage [19], where the Discriminator is fine-tuned to estimate the individual appliance signal.

As illustrated in Figure 1, both Generator and Discriminator share the same architecture. The input sequence data enters the model. As a first step, a 1D-convolutional layer along with a squared-average pooling layer are used for feature extraction. Then, the data sequence is added to a positional embedding and passed to a series of Transformer layers. The Transformer output is deconvoluted and passed through two linear layers to produce the model output sequence data.

2.1. TransformNILM's unsupervised model pre-training process

Utilizing a model pre-training procedure is common strategy in various Transformer architectures [14], [15]. In such approaches, some values from the input signal are replaced during the unsupervised pre-training phase, while the model is then further subsequently fine-tuned to adapt to any downstream task. However, only a small fraction of the data is properly utilized for model training, as the loss function [14], [15] is calculated only considering the replaced positions. Even though the pre-training technique is interesting, we ar-

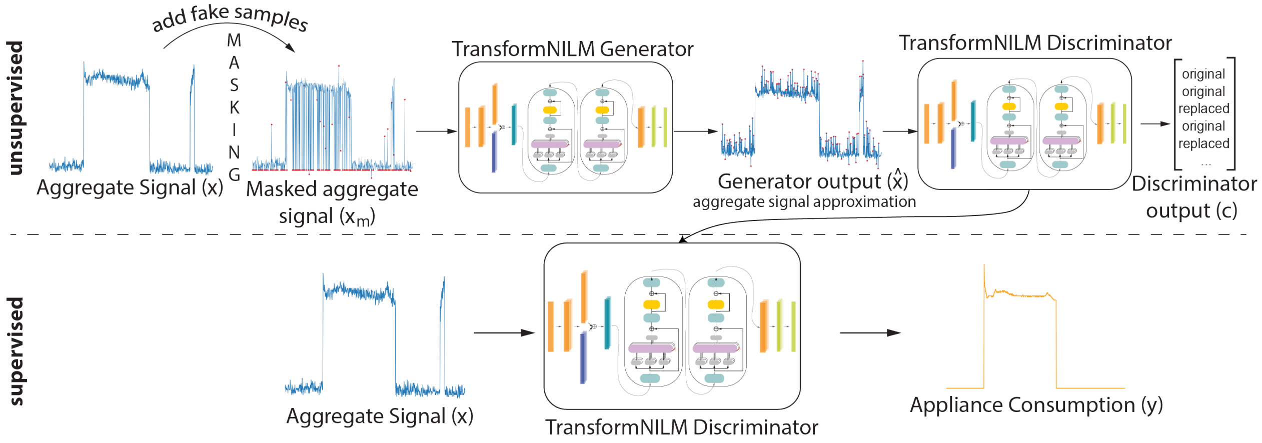

Fig. 2: Overview of TransformNILM's model training routine.

gue that a more data efficient strategy, that does not ignore most input values, could yield a higher performance.

TransformNILM's efficient pre-training approach is illustrated in Figure 2. Contrary to the aforementioned traditional Transformers approaches, which make use of a single Transformer model, TransformNILM consists of a Transformer-based Generator and Discriminator. In our approach, we replace a certain percentage of the aggregate sequence values $x \in \mathbb{R}^N$ and create a masked aggregate signal x_m. The Generator takes the masked aggregate signal as input and tries to reconstruct the original aggregate sequence by predicting the original signal values at the masked positions. Through this procedure, the model is forced to understand the interdependencies of the aggregate signal without relying on labeled data. The Discriminator then receives the Generator estimation and tries to understand which samples were replaced and which correspond to the original aggregate sequence.

To address the data inefficiency of traditional masked pre-training mechanisms [15], the Generator loss function considers only the masked positions, whereas in the Discriminator loss funcion the whole signal is utilized. The pre-training loss functions $\mathcal{L}_{gen}$ and $\mathcal{L}_{disc}$ are defined as:

$$\mathcal{L}_{gen} = \frac{1}{T}\sum_{i=1}^{M}(\hat{x_i} - x)^2 + D_{KL}(\sigma(\frac{\hat{x}}{\tau})\|\sigma(\frac{x}{\tau}))$$
$$\mathcal{L}_{disc} = -\frac{1}{N}\sum_{i=1}^{N} m_i log(p(c_i)) + (1 - m_i) log(1 - p(c_i)) \tag{1}$$

$x \in \mathbb{R}^N$ denotes the aggregate signal, $\hat{x} \in \mathbb{R}^N$ the Generator output, $m \in \mathbb{R}^N$ is a binary mask with M masking positions and x_m describes the masked input sequence, while c is the Discriminator output. Finally, τ is a hyperparameter to control softmax function (σ) temperature. The Generator loss function is formulated as the combination of Mean Squared Error (MSE) and Kullback-Leibler Divergence (D_{KL}), while Binary Cross-Entropy (BCE) loss is implemented in the Discriminator loss function.

In terms of dataflow, the aggregate signal x is masked to create x_m which is given as input to the Generator. The Discriminator receives the Generator output $\hat{x}$ and predicts which values were replaced and which correspond to the original signal, thus creating the vector c. This process can be summarized as:
$x \rightarrow x_m \rightarrow \textit{Generator} \rightarrow \hat{x} \rightarrow \textit{Discriminator} \rightarrow c$

2.2. TransformNILM supervised model training process

Conceptually, the pre-training process can be seen as a technique to boost the performance of the model through a a task-specific weight initialization. In the training phase, the Generator is discarded and the Discriminator is re-trained to produce the appliance signature. Since the model's objective changes, a different loss function, fitting to the energy disaggregation problem, is required. The Discriminator loss function is formulated in Equation 2.

$$\mathcal{L}(y, s) = \mathcal{L}_{gen}(y) + \sum_{i=1}^{N}\frac{log(1 + exp(\hat{s}_i s_i))}{N} + \frac{\lambda}{N}\sum_{i \in \mathcal{O}}|\hat{y}_i - y_i| \tag{2}$$

λ is a hyperparameter used to control the impact of the absolute error stemming from the set $\mathcal{O}$ of incorrectly predicted samples when the appliance was turned on. The loss function also considers the appliance ground truth status, as well as the on-off status s of the predicted consumption signal. The dataflow during training is simpler. The aggregate signal x is given as input to the Discriminator, which predicts the individual appliance consumption pattern y, as in: $x \rightarrow \textit{Discriminator} \rightarrow y$

Device	Model	MRE	MAE	Acc.	F1
Fridge (F)	GRU+ [7]	0.901	39.54	0.636	0.401
	LSTM+ [7]	0.956	43.74	0.573	0.174
	CNN [9]	0.758	29.20	0.772	0.718
	BERT4NILM [15]	0.732	25.49	0.813	0.766
	TransformNILM	**0.710**	**23.11**	**0.836**	**0.802**
Washer (W)	GRU+ [7]	0.662	68.65	0.342	0.018
	LSTM+ [7]	0.067	15.66	0.938	0.150
	CNN [9]	0.094	11.90	0.913	0.173
	BERT4NILM [15]	0.040	6.98	0.966	0.325
	TransformNILM	**0.012**	**5.22**	**0.997**	**0.915**
Microwave (M)	GRU+ [7]	0.014	6.41	0.996	0.266
	LSTM+ [7]	0.014	6.55	0.995	0.060
	CNN [9]	0.014	6.36	0.995	**0.341**
	BERT4NILM [15]	0.014	6.57	0.995	0.014
	TransformNILM	**0.013**	**6.28**	**0.996**	0.277
Dishwasher (D)	GRU+ [7]	0.035	38.42	0.977	0.639
	LSTM+ [7]	0.033	36.36	0.976	0.605
	CNN [9]	0.069	25.43	0.947	0.560
	BERT4NILM [15]	0.049	**16.18**	0.966	0.667
	TransformNILM	**0.027**	18.96	**0.983** .	**0.817**

Table 1: Performance comparison, UK-DALE

Device	Model	MRE	MAE	Acc.	F1
Fridge (F)	GRU+ [7]	0.829	44.28	0.794	0.705
	LSTM+ [7]	0.841	44.82	0.789	0.709
	CNN [9]	0.822	35.69	0.796	0.689
	BERT4NILM [15]	**0.806**	**32.35**	0.841	0.756
	TransformNILM	0.823	32.47	**0.875**	**0.799**
Washer (W)	GRU+ [7]	0.090	27.63	0.922	0.216
	LSTM+ [7]	0.020	35.73	0.989	0.125
	CNN [9]	0.042	36.12	0.970	0.274
	BERT4NILM [15]	0.022	34.96	0.991	0.559
	TransformNILM	**0.016**	**23.07**	**0.997**	**0.902**
Microwave (M)	GRU+ [7]	0.059	17.72	0.988	0.574
	LSTM+ [7]	0.058	17.39	0.989	0.604
	CNN [9]	0.060	18.59	0.986	0.378
	BERT4NILM [15]	0.057	17.58	0.989	0.476
	TransformNILM	**0.056**	**16.41**	**0.990**	**0.611**
Dishwasher (D)	GRU+ [7]	0.042	25.29	0.955	0.034
	LSTM+ [7]	0.056	25.25	0.956	0.421
	CNN [9]	0.053	25.29	0.953	0.298
	BERT4NILM [15]	**0.039**	**20.49**	0.969	0.523
	TransformNILM	0.050	24.05	**0.969**	**0.655**

Table 2: Performance comparison, REDD

3. EXPERIMENTAL RESULTS

To compare our results, we chose two open-source datasets, UK-DALE [20] and REDD [21], and train the model on the fridge, washer/dryer, microwave and dishwasher. For validation, we utilized state of the art models that are based on different architectures. We adopted two recurrent approaches, GRU+ and LSTM+ [7], a convolutional neural network [9] and a transformer-based model [15]. To evaluate model performance, we used four widely used metrics , namely Mean Relative Error (MRE),Mean Absolute Error (MAE), Accuracy and F1-score. The latter is recorded to assess the model's ability to address class imbalance. To assess their generalization capabilities, all models were tested on unseen data, from a house of the dataset not used during training.

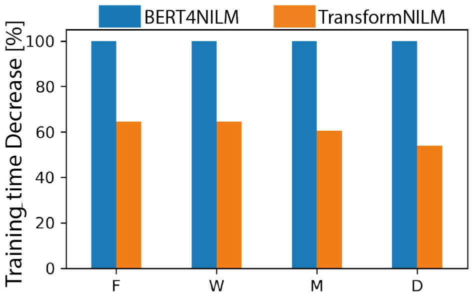

Fig. 3: Transformer models training time comparison

Tables 1 and 2 present the experimental results for UK-DALE and REDD respectively.TransformNILM surpasses the comparative models in most of the appliances. The most notable performance improvement was noted in the washing machine. While all other models fail to accurately predict the on-off status of the appliance, indicated by the low F1 score, TransformNILM can capture the status changes of the washing machine and surpasses all models by a wide margin.

In UK-DALE, TransformNILM achieves performance increase for all appliances and adequately handles imbalanced data. The low F1-score of the microwave can be explained by the extended sparsity of the data, since activations span over only some minutes. The performance of TransformNILM is higher than the other models on REDD as well. Comparing the performance of the model on an appliance level between both datasets, we see that the performance declines slightly on REDD. UK-DALE contains more appliance activations at a similar ratio of positive-negative samples. The higher number or training samples in UK-DALE helps the model to better capture the interdependencies between the data.

Overall, the introduction of an efficient pre-training technique leads to both performance and training time improvements, thus making TransformNILM a fast and efficient energy disaggregation Transformer architecture. On average, TransformNILM required 42% reduced training time than the other Transformer-based model (BERT4NILM).

4. CONCLUSION

TransformNILM is an efficient fast Transformer for NILM that outperforms state-of-the-art models without employing a data balancing approach. Averaging across all appliances, TransformNILM attains a performance boost across all metrics in both datasets,while requiring significantly less training time than BERT4NILM, making our approach superior both in computational efficiency and performance.

References

[1] M. Kaselimi, N. Doulamis, A. Voulodimos, E. Protopapadakis, and A. Doulamis, "Context aware energy disaggregation using adaptive bidirectional lstm models," *IEEE Transactions on Smart Grid*, vol. 11, no. 4, pp. 3054–3067, 2020.

[2] M. Kaselimi, N. Doulamis, A. Doulamis, A. Voulodimos, and E. Protopapadakis, "Bayesian-optimized bidirectional lstm regression model for non-intrusive load monitoring," in *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019, pp. 2747–2751.

[3] A. Harell, S. Makonin, and I. Bajić, "Wavenilm: A causal neural network for power disaggregation from the complex power signal," *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 8335–8339, 2019.

[4] D. Murray, L. Stankovic, V. Stankovic, S. Lulic, and S. Sladojevic, "Transferability of neural network approaches for low-rate energy disaggregation," in *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019, pp. 8330–8334.

[5] P. Huber, A. Calatroni, A. Rumsch, and A. Paice, "Review on deep neural networks applied to low-frequency nilm," *Energies*, vol. 14, no. 9, 2021.

[6] M. Kaselimi, N. Doulamis, A. Voulodimos, A. Doulamis, and E. Protopapadakis, "Energan++: A generative adversarial gated recurrent network for robust energy disaggregation," *IEEE Open Journal of Signal Processing*, vol. 2, pp. 1–16, 2021.

[7] H. Rafiq, H. Zhang, H. Li, and M. K. Ochani, "Regularized lstm based deep learning model: First step towards real-time non-intrusive load monitoring," *2018 IEEE International Conference on Smart Energy Grid Engineering (SEGE)*, pp. 234–239, 2018.

[8] A. van den Oord, S. Dieleman, H. Zen, *et al.*, "Wavenet: A generative model for raw audio," in *9th ISCA Speech Synthesis Workshop*, 2016, pp. 125–125.

[9] C. Zhang, M. Zhong, Z. Wang, N. Goddard, and C. Sutton, "Sequence-to-point learning with neural networks for nonintrusive load monitoring," in *AAAI*, 2018.

[10] M. Kaselimi, E. Protopapadakis, A. Voulodimos, N. Doulamis, and A. Doulamis, "Multi-channel recurrent convolutional neural networks for energy disaggregation," *IEEE Access*, vol. 7, pp. 81 047–81 056, 2019.

[11] A. Vaswani, N. Shazeer, N. Parmar, *et al.*, "Attention is all you need," in *Proceedings of the 31st International Conference on Neural Information Processing Systems*, ser. NIPS'17, Long Beach, California, USA: Curran Associates Inc., 2017, pp. 6000–6010.

[12] T. B. Brown, B. Mann, N. Ryder, *et al.*, "Language models are few-shot learners," in *Advances in Neural Information Processing Systems*, vol. 33, Curran Associates, Inc., 2020, pp. 1877–1901.

[13] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever, "Language models are unsupervised multitask learners," 2019.

[14] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding.," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, Association for Computational Linguistics, 2019, pp. 4171–4186.

[15] Z. Yue, C. R. Witzig, D. Jorde, and H.-A. Jacobsen, "Bert4nilm: A bidirectional transformer model for non-intrusive load monitoring," in *Proceedings of the 5th International Workshop on Non-Intrusive Load Monitoring*, ser. NILM'20, Virtual Event, Japan: Association for Computing Machinery, 2020, pp. 89–93.

[16] H. Liu, C. Liu, L. Tian, H. Zhao, and J. Liu, "Non-intrusive load disaggregation based on deep learning and multi-feature fusion," in *2021 3rd International Conference on Smart Power Internet Energy Systems (SPIES)*, 2021, pp. 210–215.

[17] G. W. Hart, "Nonintrusive appliance load monitoring," *Proceedings of the IEEE*, vol. 80, no. 12, pp. 1870–1891, 1992.

[18] I. Goodfellow, J. Pouget-Abadie, M. Mirza, *et al.*, "Generative adversarial nets," in *Advances in Neural Information Processing Systems*, vol. 27, Curran Associates, Inc., 2014.

[19] K. Clark, M. Luong, Q. V. Le, and C. D. Manning, "ELECTRA: pre-training text encoders as discriminators rather than generators," in *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*, 2020.

[20] J. Kelly and W. Knottenbelt, "The UK-DALE dataset, domestic appliance-level electricity demand and whole-house demand from five UK homes," *Scientific Data*, vol. 2, no. 150007, 2015.

[21] J. Z. Kolter and M. J. Johnson, "Redd: A public data set for energy disaggregation research," 2011.