Attention-guided Temporal Convolutional Network for Non-intrusive Load Monitoring

1st Huamin Ren
School of Economics, Innovation and Technology
Kristiania University College
Oslo, Norway
huamin.ren@kristiania.no

2nd Xiaomeng Su
Dept. of Computer Science
Norwegian University of Science and Technology
Trondheim, Norway
xiaomeng.su@ntnu.no

3rd Robert Jenssen
Dept. of Physics and Technology
UiT The Arctic University of Norway
Tromsø, Norway
robert.jenssen@uit.no

4th Jingyue Li
Dept. of Computer Science
Norwegian University of Science and Technology
Trondheim, Norway
jingyue.li@ntnu.no

5th Stian Normann Anfinsen
Dept. of Energy Technology
NORCE Norwegian Research Centre
Tromsø, Norway
stia@norceresearch.no

***Abstract*—With the prevalence of smart meter infrastructure, data analysis on consumer side becomes more and more important in smart grid systems. One of the fundamental tasks is to disaggregate users' total consumption into appliance-wise values. It has been well noted that encoding of temporal dependency is a key issue for successful modelling of the relations between the total consumption and its decomposed consumption on an appliance historically, and therefore has been implemented in many state-of-the-art models. However, how to encode the varied long-term and short-term dependency coming from different appliances is yet an open and under-addressed question. In this paper, we propose an attention-guided temporal convolutional network (ATCN), which generates different temporal residual blocks and provides an attention mechanism to indicate the importance of those blocks with respect to the appliance. Ultimately, we aim to address these two questions: i) How to employ both long-term and short-term temporal dependency to better disaggregate future loads while maintaining an affordable memory cost? ii) How to employ attention during the training of an appliance to obtain a better representation of the consumption pattern? We have demonstrated the effectiveness of our approach through comprehensive experiments and show that our proposed ATCN model achieves state-of-the-art performance, particularly on multi-status appliances that are normally hard to cope with regarding disaggregation accuracy and generalization capability.**

***Index Terms*—energy disaggregation, non-intrusive load monitoring, deep learning, temporal convolutional network, attention model**

I. Introduction

Non-intrusive load monitoring (NILM), also referred to as energy disaggregation, aims to disaggregate the power consumption of a customer as a whole into detailed appliance-level consumption [1]. It is considered as a promising technology to gain knowledge about disaggregated consumption patterns and to identify active electrical appliances, and has become one of the key tools to make effective use of the emerging smart meter infrastructure. NILM has great potential in applications such as energy awareness, energy conservation, and identification of controllable loads [2]. Moreover, the analysis on NILM may even affect smart grid management: insight into consumption patterns and the amount of flexible loads can be used to design incentives to motivate shifts in consumer behavior and facilitate peak shaving, and even potentially be used to shape the consumption behavior from households to be environment friendly.

NILM has been framed historically both as classification and regression problems. When regarded as classification problem, the ON/OFF state of each appliance is classified simultaneously at each time stamp and NILM as a whole is considered as a multi-label classification task [3] [4]. However, this solution cannot derive the amount of power consumption of each appliance at each time interval, which makes it inadequate to meet the needs of power producers for load planning. In comparison, treating NILM as a regression problem provides the estimated consumption of individual appliances from the mains signal. In order to capture all distinct consumption patterns from all types of appliances, NILM algorithms tend to adopt a training dataset with a long time span (as long as memory permits) and attempt to learn temporal dependencies for each appliance. The trend is that recent work tends to utilize a range of deep neural network architectures, such as encoder-decoder networks, long short-term memory (LSTM) networks, bi-directional, sequence-to-sequence, and sequence-to-point [5] [6] [7] based prediction algorithms and their variants, including the very recent Bitcn-NILM algorithm [8], which combines sequence-to-point with bidirectional dilated convolution network. The key challenges of the prediction strategy are these: if the time window is too small, essential dependencies cannot be learned, e.g. if an appliance has a cyclic consumption pattern and the time window does not cover a full period. However, if it is too large,

This project was partly supported by Equinor under the Academia Agreement with UiT The Arctic University of Norway, and partly by Kristiania University College under interdisciplinary research fund.

the efficiency of the scheme can significantly degrade, since loading long historical data burdens the memory requirement. Additionally, it also requires a much longer prediction time, and therefore cannot meet the needs of real applications.

Remark that different appliances exhibit vast difference in their temporal dependencies. The relevant dependency ranges are specific to each appliance and should be adapted accordingly. ***Therefore, the ability to use both long-term and short-term dependencies, while varying the attention on them according to the appliance, is crucial in NILM methodologies.*** *How to accomplish this, therefore, is the key question in our paper.*

To the best of our knowledge, such adaptive attention is still missing in the current literature and the question of how to put varied attention on different appliances, i.e. short-term or long-term dependency, has not been fully addressed. We therefore propose an attention-guided temporal convolutional network (ATCN) to encode such dependencies.

II. Related Work

The revival of neural networks in the last decade has produced many new deep neural network architectures, which have been utilized to solve NILM problems. One commonly used architecture is encoder-based structures, such as [9] [10] [11]. The basic principle is to use the aggregated power demand as a (noisy) input to the network, which is asked to reconstruct the clean power demand of the target appliance.

Since energy consumption patterns are characterized by inherently recurrent phases, most of the proposed energy disaggregation models try to utilize this recurrence property. Such models are either based on recurrent neural networks (RNNs) or long short-term memory (LSTM) networks, such as [12] and [5], where past predictions are used to predict the current consumption levels. As a solution to solve the hard to train problem caused by the vanishing gradient, GRU nodes are used to replace hidden nodes in traditional RNNs [13], which contain an update node and a reset node. The update mode determines how much the units update the activation and the reset node decides whether previous computed states should be forgotten. As an alternative, bidirectionality was proposed in [6], where the combination of a forward and a backward pass of operations is employed, allowing for the consideration of not only past instances, but also future ones. Nevertheless, those models process the elements of the sequence successively, and are still constrained from maintaining a hidden state of the past within a sequence, which consequently prevents parallel computation.

Compared to recurrent layers, convolutions create representations for fixed size contexts, which can be effective and easily achieved by stacking several layers on top of each other. Convolutional neural networks (CNN) do not depend on the computations of the previous time step and, therefore, allow parallelization over every element in a sequence. A CNN model inspired by Wavenet [14], which was first developed for raw audio generation, is then adopted for energy disaggregation in [15] to capture patterns from long sequences. Among various CNN based models, two recent architectures have achieved state-of-the-art performance in NILM tasks, namely sequence-to-sequence (seq2seq) [16] [6] and sequence-to-midpoint (seq2point) [7]. Seq2seq defines a neural network to map sliding windows of the input mains power to corresponding windows of the output appliance power. Despite the promising performance shown in the seminal paper, it contains many issues in practice: A subset of all possible windows is required during training and may increase computational complexity. Moreover, each element of the output signal is predicted many times, since each sliding window will contribute to one prediction. As a result, a naturally adopted averaging strategy on multiple predictions would smooth edges and hence affect the precision. Seq2point [7] adopts a 6-layer sequence-to-midpoint learning CNN model for the task and achieves state-of-the-art performance. It defines a neural network that maps sliding windows of the input to the midpoint of the corresponding window of the output by assuming that the midpoint element in the output is a nonlinear function of the input data window. The intuition behind this assumption is that the state of the midpoint element of a given appliance is highly related to its temporal neighbors (both before and after that midpoint) in the mains data.

As previously argued, different appliances have specific and characteristic temporal dependencies and self-similarities in their consumption patterns. Unfortunately, there is so far very limited work on exploiting such dependencies. The studies that are most relevant to this topic are SCANet [17] and WaveNILM [15]. SCANet develops a multi-branch architecture with multiple receptive field sizes and branch-wise gates to connect the branches in the sub-networks and then builds a self-attention module to integrate global context. However, such self-attention is built on non-causal dilated convolutions, which means that future elements are used during model learning. This is prohibitive in most practical applications of NILM algorithms. In contrast, WaveNILM presents a causal 1-D convolutional neural network inspired by WaveNet for NILM on low-frequency data, which indicates that such scale variation can be captured using a dilated neural network.

III. Method

Our proposed algorithm sequentially acquires the input and actively attends relevant pieces of temporal information to refine the target consumption estimate at each time step. The key components are the *casual dilation* nature of the model and the *attention mechanisms*, both of which we empirically show the contribution to the appliance-wise consumption prediction.

A. Problem Definition

We follow the same NILM problem definition as in [18]: Given the aggregated power consumption y_t at time step t, where $0 < t < T$, we have a time series $\{y_1, y_2, ..., y_t, ..., y_T\}$. The actual power consumption of M electrical appliances is measured as $[x_t^{(1)}, x_t^{(2)}, ..., x_t^{(M)}]$. Hence, the NILM problem is to estimate the power consumption per appliance $\hat{x}_t^{(i)}$

throughout time as $[\hat{x}_1^{(i)}, \hat{x}_2^{(i)}, ..., \hat{x}_T^{(i)}]$, so that it satisfies:

$$\min \sum_{t=1}^{T} \left[\sum_{i=1}^{M} \|\hat{x}_t^{(i)} - x_t^{(i)}\|_2 + \epsilon_t \right] . \tag{1}$$

Here, $\epsilon_t = y_t - \sum_{i=1}^{M} x_t^{(i)}$ is an error term that represents the potential noise in the system, caused either by corrupted data due to sensor miscalculation or malfunction, sensor noise, or discretization errors from the analog-to-digital conversion.

B. Attention-guided Temporal Convolutional Network (ATCN)

Sequence modelling has long been addressed via recurrent and recursive networks. However, [19] suggested reconsidering the habitual association between sequence processing and recurrent networks. By conducting a systematic empirical evaluation of convolutional and recurrent architectures on a broad range of sequence modeling tasks, they concluded that convolutional networks should be regarded as a natural starting point for sequence modeling tasks. They also proposed a generic temporal convolutional network (TCN) to represent convolutional architectures. Inspired by their work, we use the TCN as a backbone architecture to model a long history of appliance profiles. On top of that, we propose employing an attention mechanism on each dilation in order to capture distinct information from dilated layers.

1) Temporal Convolutional Network (TCN): The essential principle of the TCN is to adopt dilated convolution layers. A dilated convolution is a convolution where the filter is applied over a time window larger than its length by skipping input values with a certain step, which effectively allows the network to operate on a coarser scale than normal convolutions [14]. By stacking dilated causal convolutions with increasing dilation factors, a large receptive field with a limited number of parameters can be achieved, while still maintaining causality. This secures that there is no information "leakage" from future to past. See a dilated network architecture in Fig. 1 for illustration.

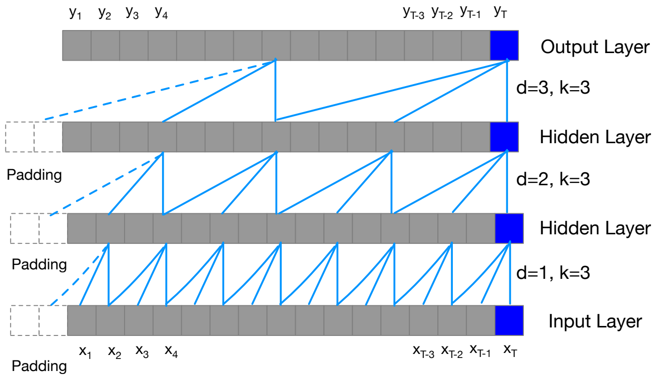

Fig. 1. Dilated casual convolutions, as used in the TCN.

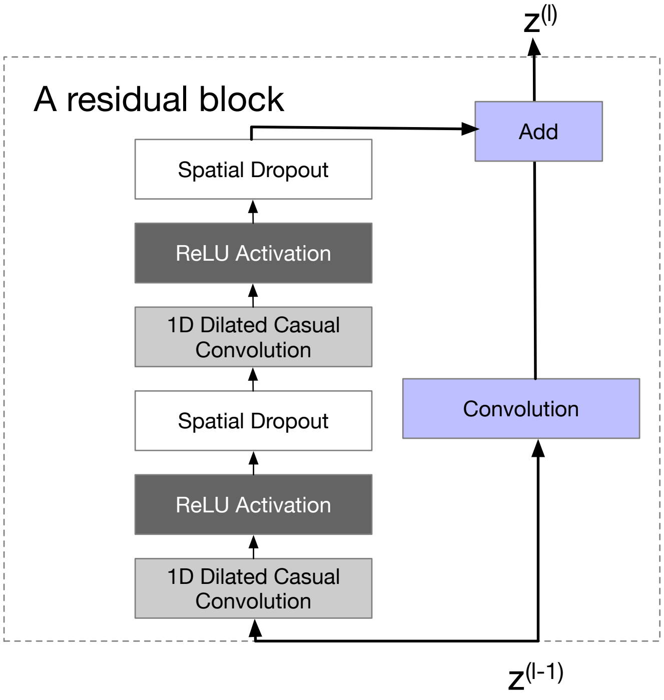

Fig. 2. A representative residual block: $z^{(l)}$.

Dilated convolution layers are able to extract local features and represent local dependencies, capturing subtle information that may occur for appliances such as a washing machine. Therefore, enduring small fluctuations followed by a large step change after a short period, which typically exists in multi-status appliances, can be richly represented. In the meanwhile, the dilation strategy looks back at historical data over a much longer period to find a potential repeated consumption pattern to better estimate the current consumption. Therefore, we are inspired by this architecture and exploit the TCN to solve the NILM problem.

Formally, for a univariate time series $X(t) \in \mathbb{R}$ of length n, where $0 < t < n-1$ is a discrete time variable, and a temporal filter $f(t) \in \mathbb{R}$ of length k, such that $f(t) = 0$ for $t < 0$ and $t \geq k$, the causal dilated convolution operation is defined as:

$$F(t) = (x *_d f)(t) = \sum_{i=0}^{k-1} f(i)X(t - id) \, . \tag{2}$$

Here, d is the dilation factor and the values used in the TCN layers are $d \in \{2^0, \ldots, 2^l, \ldots, 2^{n-1}\}$, where the index l of a general layer is called the layer indicator. If TCN blocks are stacked together to take account for long term historical data (as we did in this paper), l is called the block indicator. k is the filter size, therefore, $i = \{0, 1, ..., k-1\}$.

2) Deep TCN: To facilitate training a deep TCN, a common practice is to organize temporal convolutional layers into blocks and add residual connections between blocks. Residual blocks effectively allow layers to learn modifications to the identity mapping rather than the entire transformation, which has greatly improved stability in training of deep neural networks [20]. The residual block $z^{(l)}$ from our model is shown in Fig. 2. First, a dilated causal convolutional layer is acting on the input of the block (noted as z^{l-1}), which uses operation F as in Eq. 2, where $d = 2^l$. Then, ReLU activation and spatial dropout have been applied. After another repetition of causal dilated convolution, ReLU activation and spatial dropout, a residual action is performed, where the output of such a series of transformations ($\mathcal{F}$) is added to the input of this block, and finally, the output of this block $z^{(i)}$ is formed by Eq. 3. Be noted that we use an additional convolutional layer on the input of the block to account for discrepant input-output width.

$$z^{(i)} = \text{Add}(\text{Conv}(z^{(i-1)}) + \mathcal{F}(z^{(i-1)})) \tag{3}$$

Furthermore, we define a series of blocks, each of which contains a sequence of layers, and then a series of the block is illustrated in Fig. 2. The input of $(l)^{th}$ block is the output of the previous $(l-1)^{th}$ block, except for the first block in which the input is the input data.

After stacking layers of various dilated blocks, we can obtain a sequence of TCN encoders $(z^{(1)}, z^{(2)}, ...z^{(L)})$, which encode the input sequence into representations that capture different dilated historical data.

3) Attention Mechanism: Through experiments we have discovered that an aggressive increase of dilation factors fails to aggregate local features of appliances with a short usage time. This is a side-effect of the increased interval of the kernel weights. Whereas increasing dilation factors is important in terms of representing context in the long-term historical data, it can be detrimental to local changes in a short period, which is common in appliances such as microwaves. To address this issue, we perform an attention mechanism to discover the intrinsic relationships among consumption patterns and make the prediction based on the most potential related behavior that happened before.

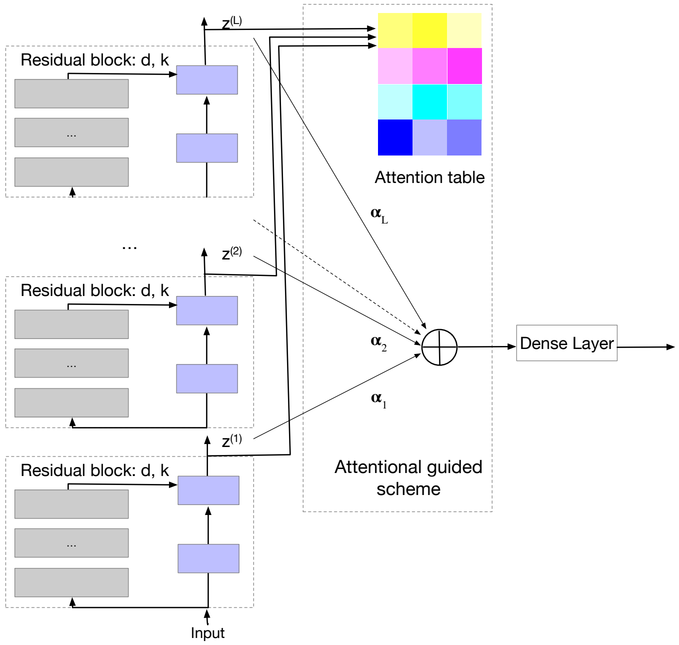

Fig. 3. Attention guided temporal convolutional networks.

Our attention mechanism does not attempt to encode a whole input sentence into a single fixed-length vector. Instead, it encodes the input sentence into a sequence of vectors based on deep residual and temporal convolutional networks, and then chooses a subset of these vectors adaptively (equivalent to selecting a subset of residual blocks) while performing the prediction. Inspired by the attention mechanism in WaveNet model [14] that is designed for speech synthesis, we apply a similar attention mechanism. Notably, our attention is put on residual blocks, aiming to select the most representative temporal convolutional neurons. Hence, an attention table is learned through previous encoders (residuals) - different from [14] where the attention was learned from both encoders and the previous decoder result.

We define c as a vector generated from the sequence of the blocks, such that

$$c = \sum_{l=1}^{L} \alpha_l z^{(l)}, l = 1...L, \tag{4}$$

We call c a context vector, where it depends on a sequence of $(z_t^{(1)}, ..., z_t^{(L)})$ - each component is an output of a residual block with varied dilation and a context vector is a learned representation of the weights of each residual block. With deep TCN, we aim to learn each appliance i to minimize the objective function in Eq. (1), i.e., for a given input of total consumption $\{y_1, ..., y_t, ..., y_T\}$, the prediction on the $i^{(th)}$ appliance can be represented as a probability in:

$$p(\hat{x}^{(i)}|y_1, y_2, ..., y_T) = g(z^{(1)}, ..., z^{(L)}, c), i = 1....M \tag{5}$$

The output (also the prediction of consumption per each appliance) can be computed either as a linear or non-linear combination (g) of these residual outputs.

α_l is calculated in Eq. (6). We parametrize the alignment model a as a feedforward neural network, similarly to alignment model in [14], which is also jointly trained with all the other components of the proposed system.

$$\alpha_l = \frac{exp(e_{lj})}{\sum_{j=1}^{L} exp(e_{lj})}, \text{where } e_{lj} = a(z^{(l)}, z^{(j)}). \tag{6}$$

The associated energy e_{lj} to α_l reflects the importance of the residual block $z^{(l)}$ with respect to other existing residual blocks. Our attention is computed in such a way that a residual block which agrees with most of the residual blocks should be given a high weight. Therefore, its block-wise output has also been highlighted when it comes to a prediction. Such a mechanism is based on the assumption that through those blocks where varied dilations were applied, important and representative characteristics could be finely expressed.

IV. Experimental setup

We first introduce the datasets and the evaluation metrics we use throughout experiments, and later demonstrate the performance and compared results with the state-of-the-art under the unified settings for an impartial comparison.

A. Datasets

Three datasets have been used to evaluate the performance of the proposed ATCN model and compared with the state-of-the-art models: the REDD [21], UK-DALE [22], and DRED [23] datasets. Four appliances (microwave, fridge, dish washer and washer dryer) are trained on the REDD and UK-DALE datasets, while two appliances (fridge and dish washer) are trained on the DRED dataset due to availability.

The REDD dataset has been collected from six houses in the state of Massachusetts, USA. It includes mains with 1 s sampling period and several appliances with 3 s sampling period. High-frequency current and voltage measurements are also available at 15 KHz sample frequency. The lengths of observations were between 3 and 19 days.

The UK-DALE dataset contains 5 buildings in the U.K. during the period from 2013 to 2015. The sampling periods for mains and appliances were 1s and 6s from November 2012 to January 2015, respectively.

The DRED dataset includes both appliance-wise and mains-wise consumption data from a household in the Netherlands. The objective of using this data was to measure the generalization capability of the models trained from the REDD and UK-DALE datasets.

B. Evaluation Metrics

Let $x_t^{(i)}$ represents the ground truth for the i^{th} appliance, and $\hat{x}_t^{(i)}$ be the prediction at time stamp t. Three evaluation metrics have been chosen to evaluate the results, as described in Tab. I.

Metric	Definition
MAE	$\frac{1}{T}\sum_{t=1}^{T} \lvert \hat{x}_t^{(i)} - x_t^{(i)} \rvert$
RMSE	$\sqrt{\frac{1}{T}\sum_{t=1}^{T} (\hat{x}^{(i)} - x_t^{(i)})^2}$
EA^i	$1 - \frac{\sum_{t=1}^{T} \lvert \hat{x}_t^{(i)} - x_t^{(i)} \rvert}{2\sum_{t=1}^{T} x_t^{(i)}}$

TABLE I
EVALUATION METRICS AND THEIR DEFINITIONS.

Mean absolute error (MAE) and root mean square error (RMSE) are the most common performance measures among NILM researchers. They evaluate the deviation between the prediction $\hat{x}_t$ and the observed signal x_t at each timestamp. In contract, EA^i allows for reporting the EA of each appliance. Aggregation over all target and non-target appliances would unavoidably result in inflated values since non-target appliances generally account for a large proportion of the electricity consumption, the appliance-specific EA^i evaluation has a significant value in avoiding that.

V. Quantitative experimental results

The most relevant work that also considers such a casual dilated neural network to incorporate lengthy temporal relationships in data is the WaveNILM algorithm [15]. WaveNILM concatenates blocks that consist of gated dilated layers, which impose heavy computations and become infeasible when capturing long time dependencies. Therefore, we improve the architecture by building residual blocks that employ a skipping mechanism, which is closer to the deep TCN architecture illustrated in Sec. 3.2.2. Therefore, we refer to this improved model as TCN in the following experiments and use it as a reference to compare with our proposed ATCN algorithm. Specifically, a detailed specification of hyperparameter settings, attention contribution and appliance-level performance of our proposed ATCN algorithm is provided in Sec. V-A; A comparative study with state-of-the-art methodologies through performance measures and visualized results in presented Sec. V-B; and finally, an in-depth analysis of the generalization capability of all models is shown in Sec. V-C.

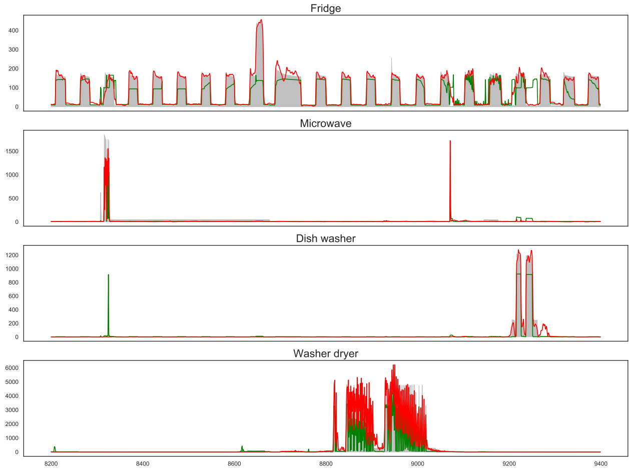

Fig. 4. Comparison of ATCN (red) with TCN (green) algorithm per each appliance on the REDD dataset (replace the last appliance.

A. ATCN model

There are several parameters affecting our ATCN model. A long sequence as inputs with a large dilation rate and residual blocks can provide rich context information, but is memory demanding. Empirically, we have set the sequence length to 299, the dilation rate to 5, the number of residual blocks to 6, and the number of attention states to 64.

To investigate the contribution of attention in the model, we further compare the appliance-level prediction of the ATCN with the standard TCN, where there is no attention mechanism, and show the results in Fig. 4. The ATCN outperforms the TCN on all appliances. Why does our attention mechanism make a significant contribution? Through the stacks of residual blocks, common information and local properties can be conveyed through the learning process. A block in the topper levels may indicate higher semantic meaning/representations of appliance usage (in a similar way as an object in an image); while lower levels of blocks may refer to lower signal patterns (similar to geometry information in an image). An attention table would guide a TCN model on which temporal residual blocks an appliance should put emphasis on: the lower signal pattern, or the higher semantic representation. We postulate that a high attention block indicates that the future consumption behavior has a high potential of repeating such a historical behavior represented by this block, among all residual blocks, and therefore, could achieve better performance.

B. Comparison with state-of-the-art models

To get a comprehensive understanding of how our proposed model works on NILM tasks, we conduct a comparative study and show resulting predictions in Fig. 5. For each appliance, a comparison of predictions from six algorithms is shown together with the real observations (ground truth). All models perform reasonably well on the fridge, see upper-left in Fig. 5, mainly because the consumption pattern of fridge shows apparent periodical repetitions. In contrast, the models do not perform equally well for other appliances, and particularly for microwave (lower-left), as most models fail to detect the ON-status and falsely report the consumption as values close to zero. A suspected reason is that the short time the microwave is active provides insufficient information for the learning process. Hence, appliances that behave similarly or exhibit stages of similar consumption are comparatively hard to detect and predict. Dish washer contains multi-modes, but the ON status per each mode lasts longer than for the microwave. Therefore, all models display better performance on average for this appliance. In particular, Seq2Point and our ATCN model outperform the others.

We show detailed performance measures in Table II. In overall, the RNN model provides the best performance with respect to all five performance measures on the fridge; The ATCN model is competitive with the other models on microwave and dish washer, and outperforms them on washer dryer.

C. Generalization capability

Generalization is an important perspective in the NILM task. Assuming the electricity consumption of an appliance can be well represented from an available large dataset, generalization capability refers to whether the learned model can predict the consumption of the same appliance on another dataset, either from a different country or population. Such generalization capability is an important factor when considering the applicability of different models. To investigate this property, we train various models on one dataset (the UK-DALE dataset) and test and show their performance on another (the DRED dataset).

We have chosen the three best-performing models in the previous experiments, i.e., Seq2Seq, Seq2Point and ATCN, and then show their performance on fridge and washing machine (due to the availability) in Fig. 6. Results show that all of the compared models have demonstrated their generalization capability to some extent. The ATCN and Seq2Point models are on par, while Seq2Seq tends to underestimate the consumption. We also notice that the Seq2Point model more often predicts an active status of an appliance, when it is off (see e.g. the intervals between the working status of the fridge).

VI. Conclusion

As an important problem in smart home management, NILM still remains a challenge with great potential for further exploration and improvement. We propose a residual block concatenation strategy and apply an attention mechanism based on such residuals instead of dilated layers to improve NILM performance. The essential dilation and temporal convolution structure helps capture the long-term as well as short-term dependencies in the consumption signatures, while attention residuals ensure that the model's emphasis on relevant time scales is adapted to the appliance. Our proposed ATCN algorithm outperforms state-of-the-art methodologies on multi-status appliances, especially those with short usage time, and has demonstrated excellent generalization capability.

References

[1] J. Z. Kolter, T. Jaakkola, Approximate inference in additive factorial hmms with application to energy disaggregation, Vol. 22 of Proceedings of Machine Learning Research, PMLR, 2012, pp. 1472–1482.

[2] W. Kong, Z. Y. Dong, D. J. Hill, J. Ma, J. Zhao, F. Luo, A hierarchical hidden markov model framework for home appliance modeling, IEEE Trans. Smart Grid 9 (4) (2018) 3079–3090.

[3] S. M. Tabatabaei, S. Dick, W. Xu, Toward non-intrusive load monitoring via multi-label classification, IEEE Trans. Smart Grid 8 (1) (2017) 26–40.

[4] S. Singh, A. Majumdar, Non-intrusive load monitoring via multi-label sparse representation-based classification, IEEE Trans. Smart Grid 11 (2) (2020) 1799–1801.

[5] J. Kelly, W. Knottenbelt, Neural nilm, BuildSys '15 (2015).

[6] M. Kaselimi, N. Doulamis, A. Voulodimos, E. Protopapadakis, A. D. Doulamis, Context aware energy disaggregation using adaptive bidirectional LSTM models, IEEE Trans. Smart Grid 11 (4) (2020) 3054–3067.

[7] C. Zhang, M. Zhong, Z. Wang, N. H. Goddard, C. A. Sutton, Sequence-to-point learning with neural networks for non-intrusive load monitoring, in: S. A. McIlraith, K. Q. Weinberger (Eds.), AAAI-18, AAAI Press, 2018, pp. 2604–2611.

[8] Z. Jia, L. Yang, Z. Zhang, H. Liu, F. Kong, Sequence to point learning based on bidirectional dilated residual network for non-intrusive load monitoring, International Journal of Electrical Power and Energy Systems 129 (2021) 106837.

[9] R. Bonfigli, A. Felicetti, E. Principi, M. Fagiani, S. Squartini, F. Piazza, Denoising autoencoders for non-intrusive load monitoring: Improvements and comparative evaluation, Energy and Buildings 158 (11 2017).

[10] J. Kelly, W. J. Knottenbelt, Neural NILM: deep neural networks applied to energy disaggregation, in: D. E. Culler, Y. Agarwal, R. Mangharam (Eds.), BuildSys 2015, Seoul, South Korea, November 4-5, 2015, ACM, 2015, pp. 55–64.

[11] M. Valenti, R. Bonfigli, E. Principi, S. Squartini, Exploiting the reactive power in deep neural models for non-intrusive load monitoring, in: IJCNN, Rio de Janeiro, Brazil, July 8-13, 2018, IEEE, 2018, pp. 1–8.

[12] L. Mauch, B. Yang, A new approach for supervised power disaggregation by using a deep recurrent LSTM network, in: GlobalSIP 2015, Orlando, FL, USA, December 14-16, 2015, IEEE, 2015, pp. 63–67.

[13] T. Le, J. Kim, H. Kim, Classification performance using gated recurrent unit recurrent neural network on energy disaggregation, in: ICMLC, IEEE, 2016, pp. 105–110.

[14] A. van den Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. W. Senior, K. Kavukcuoglu, Wavenet: A generative model for raw audio, in: The 9th ISCA Speech Synthesis Workshop, Sunnyvale, ISCA, 2016, p. 125.

[15] A. Harell, S. Makonin, I. V. Bajić, Wavenilm: A causal neural network for power disaggregation from the complex power signal, in: ICASSP, 2019, pp. 8335–8339.

[16] J. Du, Y. Tu, L. Dai, C. Lee, A regression approach to single-channel speech separation via high-resolution deep neural networks, IEEE/ACM Transactions on Audio, Speech, and Language Processing 24 (8) (2016) 1424–1437.

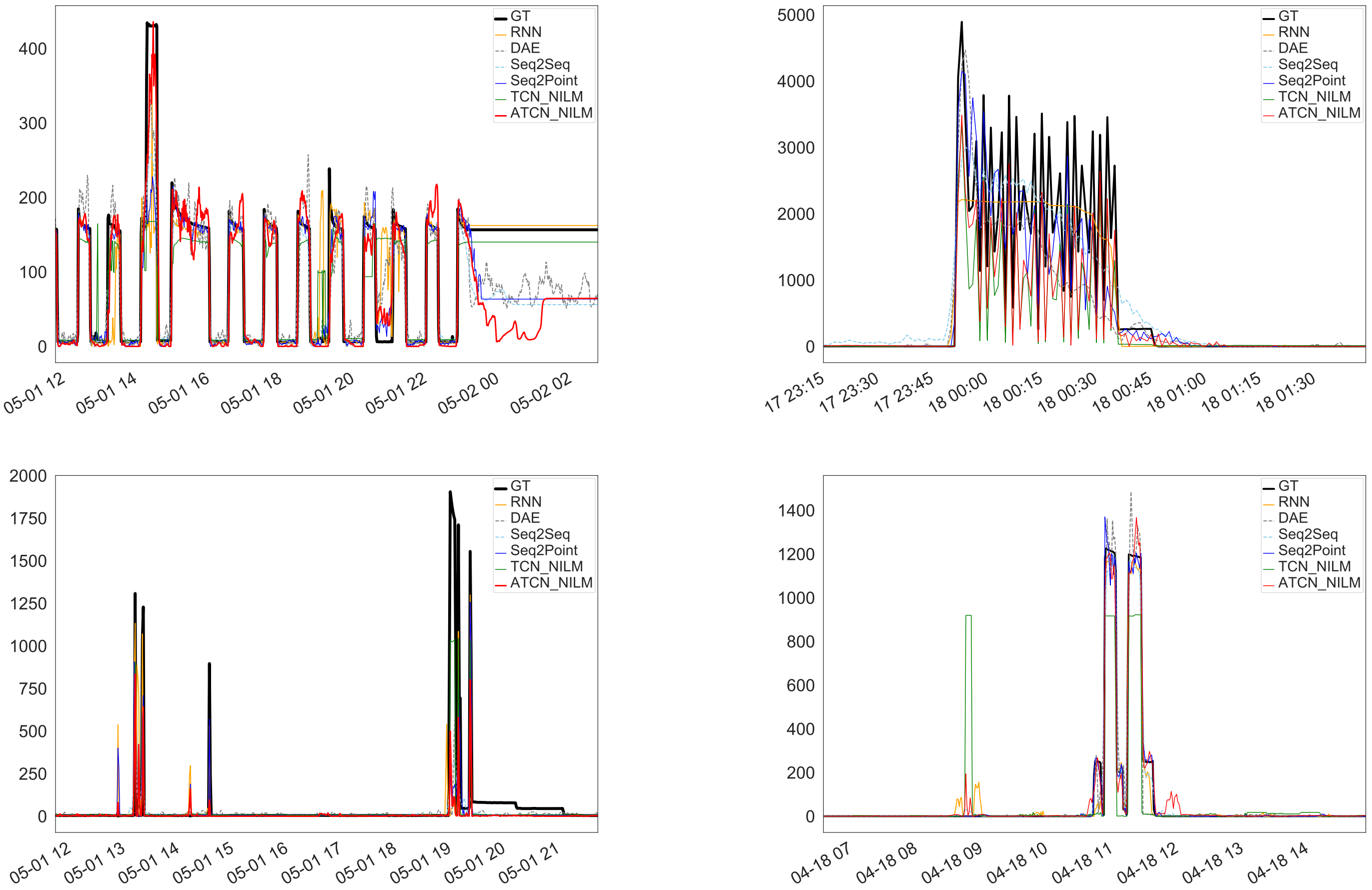

Fig. 5. Comparison with state-of-the-art models. Appliances from left to right, top to bottom: fridge, washer dryer, microwave, dish washer.

Models	MAE				RMSE				EA			
	F	M	DW	WD	F	M	DW	WD	F	M	DW	WM
RNN [5]	6.95	4.76	0.97	106.0	16.18	30.58	9.01	316.85	0.97	0.73	0.86	0.50
DAE [11]	54.96	10.95	1.66	102.83	66.77	52.86	14.83	293.20	0.78	0.39	0.77	0.55
Seq2Seq [16]	63.35	5.46	0.97	99.48	78.72	33.91	10.65	267.87	0.74	0.69	0.86	0.83
Seq2Point [7]	58.08	4.67	0.38	83.17	72.52	28.56	4.70	196.10	0.76	0.74	0.94	0.82
TCN_NILM [15]	21.04	7.44	5.17	91.53	34.71	37.75	40.16	221.82	0.91	0.59	0.29	0.80
ATCN_NILM (Ours)	60.83	8.11	0.71	82.44	73.24	44.00	6.59	182.30	0.75	0.55	0.90	0.85

TABLE II
COMPARISON OF MAE (LEFT), RMSE (MIDDLE) AND EA (RIGHT) FOR VARIOUS MODELS.

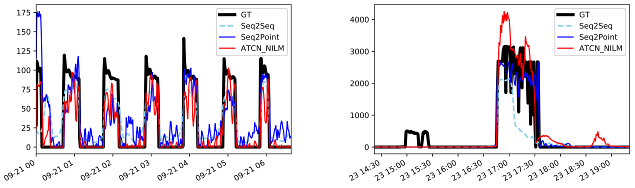

Fig. 6. Performance of fridge (left) and washing machine (right) prediction when trained on UK-DALE dataset and tested on DRED dataset.

[17] K. Chen, Y. Zhang, Q. Wang, J. Hu, H. Fan, J. He, Scale- and context-aware convolutional non-intrusive load monitoring, IEEE Trans. Power Systems 35 (3) (2020) 2362–2373.

[18] H. Ren, F. Bianchi, J. Li, R. Olsen, R. Jense, S. Anfinsen, Towards applicability: A comparative study on non-intrusive load monitoring algorithms, in: ICCE, IEEE, United States, 2021.

[19] S. Bai, J. Z. Kolter, V. Koltun, An empirical evaluation of generic convolutional and recurrent networks for sequence modeling, arXiv:1803.01271 (2018).

[20] R. Chakraborty, X. Zhen, N. Vogt, B. B. Bendlin, V. Singh, Dilated convolutional neural networks for sequential manifold-valued data, in: ICCV, IEEE, 2019, pp. 10620–10630.

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

[22] J. Kelly, W. Knottenbelt, The UK-DALE dataset, domestic appliance-level electricity demand and whole-house demand from five UK homes 2 (150007) (2015).

[23] A. S. Uttama Nambi, A. Reyes Lua, V. R. Prasad, Loced: Location-aware energy disaggregation framework, BuildSys '15, Association for Computing Machinery, 2015.