Andreas Reinhardt and Mazen Bouchur. "On the Impact of the Sequence Length on Sequence-to-Sequence and Sequence-to-Point Learning for NILM." in: *Proceedings of the 5th Workshop on Non-Intrusive Load Monitoring (NILM)*. 2020.

On the Impact of the Sequence Length on Sequence-to-Sequence and Sequence-to-Point Learning for NILM

Andreas Reinhardt
Department of Informatics
TU Clausthal, Germany
reinhardt@ieee.org

Mazen Bouchur
Department of Informatics
TU Clausthal, Germany
mazen.bouchur@tu-clausthal.de

ABSTRACT

The Sequence-to-Sequence (S2S) and Sequence-to-Point (S2P) optimization methods achieve remarkable accuracy results for load disaggregation tasks. Internally, they rely on neural networks, trained to identify the power consumption of a single appliance under consideration from a sequence of aggregate power data. Their most important configuration parameter – the number of input data samples to consider – is, however, mostly set to a fixed value. As a result thereof, the amount of historical data available at the algorithm's input is governed by the sampling interval of the used input data. For example, UK-DALE [5] provides samples every 6 s, so a sequence length of 599 samples (as proposed in [9]) makes approximately 1 h of historical data available to the disaggregation algorithm. No analyses of the impact of the sequence length on the NILM performance have been documented in literature to date. We hence present a methodological assessment of the sensitivity of S2S and S2P to variations of their input sequence length parameter. Our results show that setting a per-device parameter value leads to improved disaggregation results; however, the required values need to be determined empirically, as they are unrelated to the appliances' operational durations. Even if only a single value may be set, an informed choice (rather than using the default value) can drastically improve NILM performance.

CCS CONCEPTS

• **Hardware** → **Smart grid**; • **Computing methodologies** → **Machine learning**.

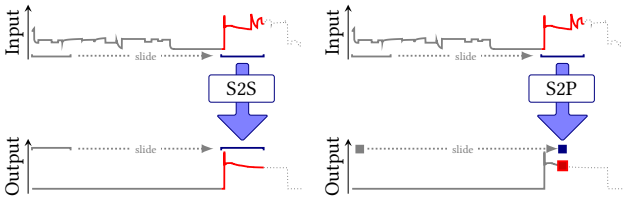
KEYWORDS

non-intrusive load monitoring, optimal input sequence length, sequence-to-sequence learning, sequence-to-point learning

ACM Reference Format:
Andreas Reinhardt and Mazen Bouchur. 2020. On the Impact of the Sequence Length on Sequence-to-Sequence and Sequence-to-Point Learning for NILM. In *The 5th International Workshop on Non-Intrusive Load Monitoring (NILM '20), November 18, 2020, Virtual Event, Japan.* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3427771.3427857

NILM '20, November 18, 2020, Virtual Event, Japan

ACM ISBN 978-1-4503-8191-8/20/11…$15.00
https://doi.org/10.1145/3427771.3427857



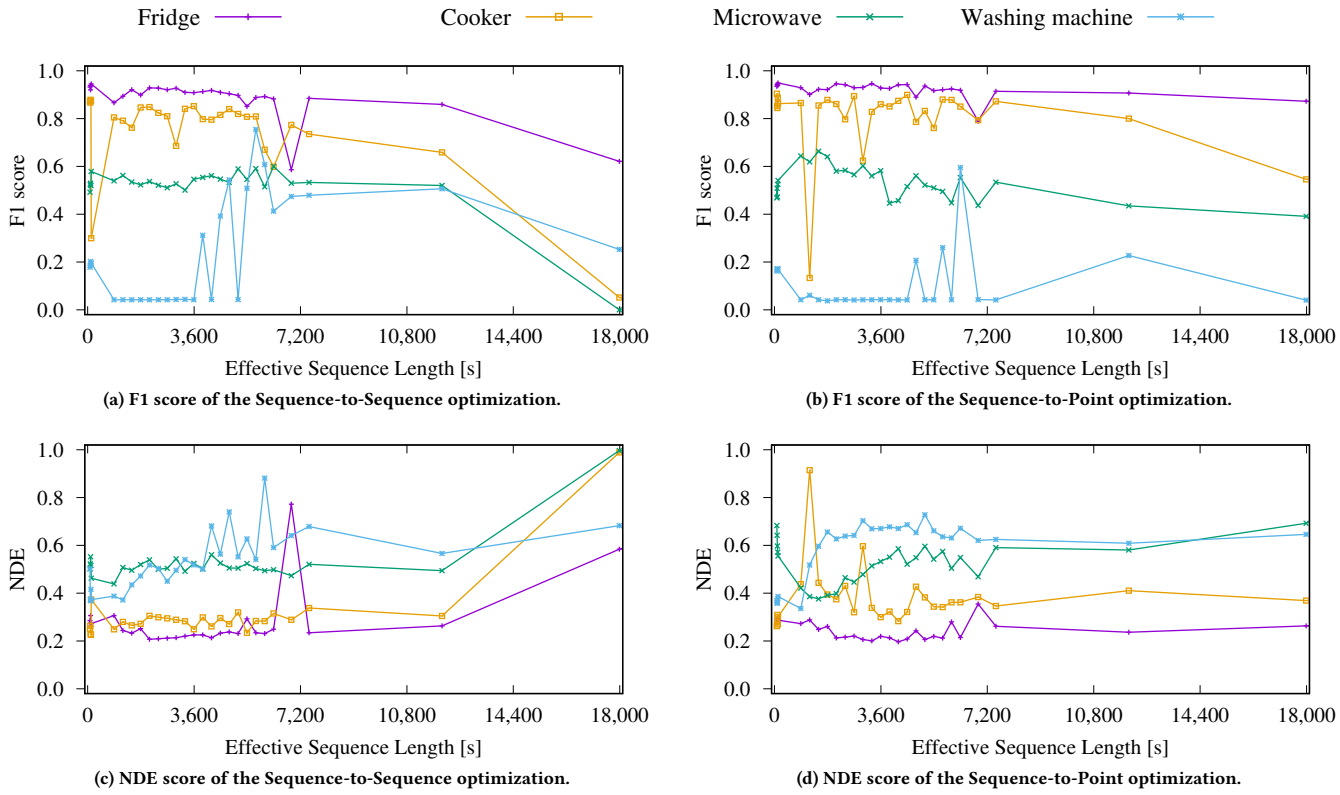
(a) Sequence-to-Sequence. (b) Sequence-to-Point.

Figure 1: Schematic operation of the S2S and S2P methods.

1 INTRODUCTION

In recent years, many Non-Intrusive Load Monitoring (NILM) methods have been proposed that rely on (deep) neural networks [4, 6, 10]. These approaches have proven particularly effective to disaggregate appliances with variable power consumption levels during their activity (i.e., multistate and continuously variable devices [3]). We consider one such family of methods in particular: Sequence-to-Sequence (S2S) and Sequence-to-Point (S2P) optimization [9]. In both methods, a windowed excerpt of (aggregate) input data is provided to a neural network, which has been trained to remove the power consumption contributions of all appliances except for the device under consideration. This is schematically shown in Fig. 1: A sliding window is moved across the aggregate power signal (top) and used to emit the disaggregated device-level power (bottom), either for sequence of the same size as the input (S2S) or only its mid-point (S2P). Note that due to the individual consumption characteristics of most electrical devices, a separate neural network must be trained for each device. As such, it is not strictly necessary to find a sliding window size that fits all appliances equally well.

The informed choice of the sequence length is crucial to achieve optimal disaggregation results. It determines the extent of historical data available to the neural network (unless memory cells are part of the neural network [4, 6]; a case we disregard in this paper). The choice of sequence lengths used in related work does not follow a well-documented methodology, however. Rather than that, the authors of [9] experiment using two datasets (UK-DALE and REDD), but disregard the fact that the former offers readings at a rate of 1/6 Hz, whereas the latter reports data up to once per second. Applying the same sequence length of 599 samples thus effectively spans between 599 s (REDD) and 3,594 s (UK-DALE) of input data (i.e., approximately 10 to 60 min). Multiple sequence lengths were used in [4], set to match the operational durations of the appliances under consideration. The default sequence length



(a) F1 score of the Sequence-to-Sequence optimization.

(b) F1 score of the Sequence-to-Point optimization.

(c) NDE score of the Sequence-to-Sequence optimization.

(d) NDE score of the Sequence-to-Point optimization.

Figure 2: Results of the parameter range study for the four appliances under consideration.

used in the NILMTK framework [1] is 99[1]. Lastly, the authors of [6] have experimented with sequence lengths of 50, 100, or 200 samples, at a dataset sampling interval of 6 s (i.e., an actual duration of 5–20 min). In none of the cited publications, a clear reasoning for these sequence length choices is provided.

It is important to understand that the actual duration of historical data available to a NILM algorithm not only depends on size of the sliding window, but also on the sampling rate of the input data. Moreover, if further data downsampling is being applied (e.g., by setting a value greater than 1 for the *sample_ period* in NILMTK), the input sequence effectively covers an even longer time interval. We hence define the notion of the Effective Sequence Length (ESL), i.e., effective amount of historical data available to the NILM algorithm, for the analyses we perform in the remainder of this work. It is computed according to Eq. (1).

$$ESL = SL \cdot SI \cdot RF \tag{1}$$

where:

SL = length of the NILM method's input sequence
SI = sampling interval of the data (=$1/f_s$ for sampling rate f_s)
RF = data reduction factor (1 if no downsampling is applied)

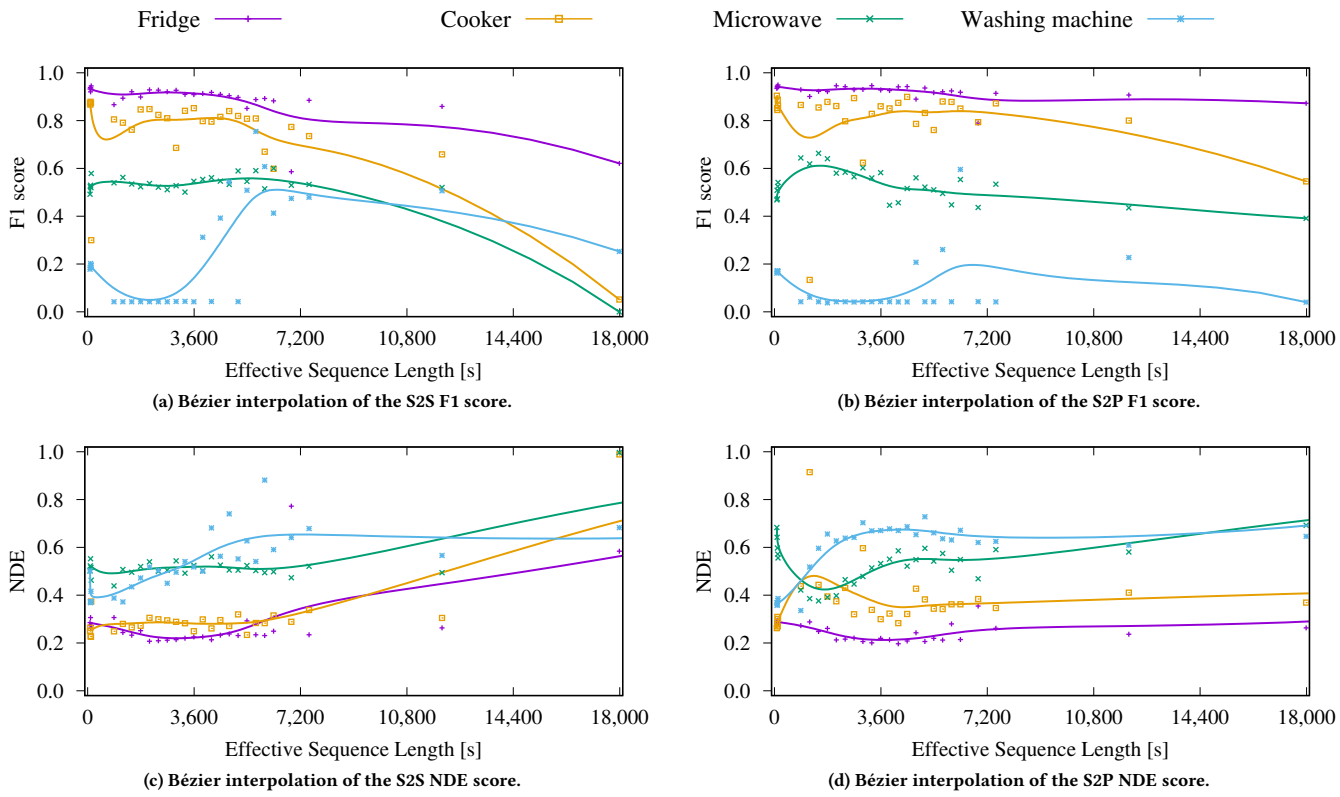
We explore the impact of the ESL on the NILM accuracy in this work, and derive guidelines for its choice the context of NILM, specifically for S2S and S2P.

[1]Specified in https://github.com/nilmtk/nilmtk-contrib/blob/master/nilmtk_contrib/disaggregate/seq2seq.py#L38 for commit `01d20b5` from 8 November 2019.

2 SWEEPING THE PARAMETER RANGE

In order to empirically confirm the assumption that different ESLs have an impact on the disaggregation results, we begin our analysis with a short experiment. Using the NILMTK implementations [2] of S2S and S2P, we have computed the F1 score and the Normalized Disaggregation Error (NDE) to determine the NILM disaggregation performance for four appliances (fridge, cooker, microwave, washing machine). Note that both metrics range from $[0, 1]$, yet an F1 score value of 1.0 indicates a perfect disaggregation of the appliance under consideration, while the minimal NDE value indicates the best disaggregation performance. Using data from the Dutch Residential Energy Dataset (DRED) [7], all neural networks were trained for 30 epochs. The sampling interval of the dataset is $SI = 1$ s. We have used a training period of 14 days (27 July to 9 August 2015), and the following 7 days (10 August to 16 August 2015) for testing. Beyond an ESL of 120 s, NILMTK's downsampling feature was used ($RF = 10$) to accelerate the neural network's training process. This way, the ESL was varied from 90 s to 6 h.

The resulting F1 and NDE values are shown in Fig. 2 for both S2S and S2P. The results of this preliminary analysis confirm that the appliances under consideration exhibit different reactions to the variations in the ESL. Despite the slightly erratic behavior of the F1 score, which can be attributed to the quirks of training the underlying neural network, general trends can be observed for all appliances. This motivates our further study of how to select appropriate ESL values, which we present as follows.



(a) Bézier interpolation of the S2S F1 score.

(b) Bézier interpolation of the S2P F1 score.

(c) Bézier interpolation of the S2S NDE score.

(d) Bézier interpolation of the S2P NDE score.

Figure 3: Approximation of the samples using a Bézier interpolation. Extremal values are tabulated in Table 1.

3 THE QUEST FOR AN OPTIMAL ESL VALUE

In an attempt to improve the visual clarity of Figs. 2a and 2b, we plot the data points again in Fig. 3 and approximate their overall trend by means of a Bézier curve. Visualized this way, it becomes clear that most of the Bézier approximations feature one globally optimal ESL value. We extract these values (for maximum F1 and minimal NDE, respectively) from the curves and list their values in Table 1 for both S2S and S2P, given that they represent the ESL candidate values where the best disaggregation results are expected. Note that we only consider globally extremal points on the Bézier curve, even when two local optima exist (e.g., for the F1 score of the cooker and washing machine); we consider their analysis as future work. As follows, we explicitly study the disaggregation results when using these ESL values. To cater to a fair comparison to related work, we also consider ESLs of 99 s and 599 s (cf. [1, 9]), as well as the average operational duration of the appliances [4].

Table 1: Characteristic features of the considered appliances.

	Avg. active	ESL @ max. F1		ESL @ min. NDE	
Appliance	duration	S2S	S2P	S2S	S2P
Fridge	1,003 s	84 s	134 s	2,981 s	3,695 s
Cooker	1,294 s	79 s	79 s	97 s	79 s
Microwave	217 s	5,472 s	1,536 s	815 s	1,691 s
Washing mach.	7,176 s	6,380 s	6,689 s	315 s	90 s

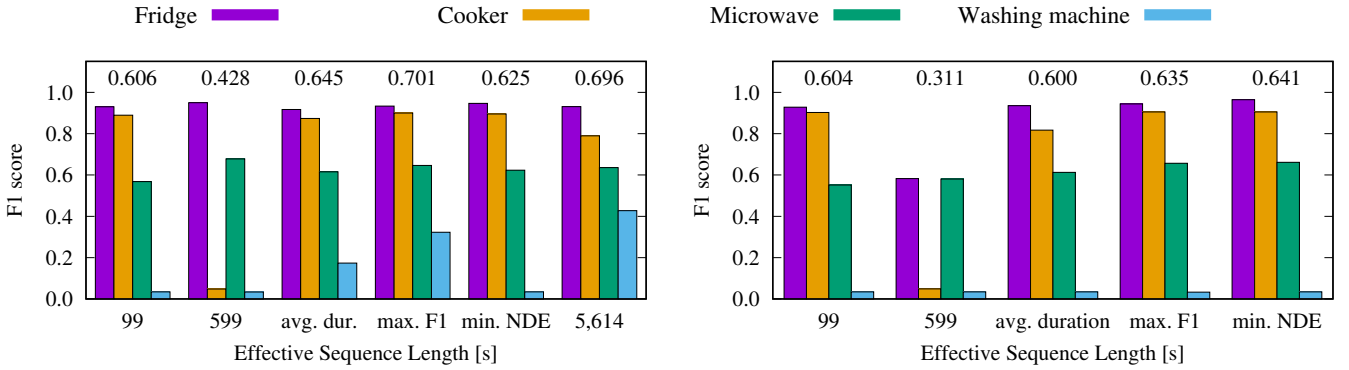
3.1 Analyzing the Candidate Values

For our in-depth study of the ESL candidates identified in Sec. 2, we use data from the extended timeframe between 6 July to 17 August 2015, of which 28 days were used for training and 14 days for testing. We have trained the neural networks for 50 epochs, using $RF = 1$ for all ESL values up to 250 s and $RF = 10$ for greater values. The corresponding F1 scores are plotted in Fig. 4 for the ESL values under consideration. The arithmetic mean values of the F1 scores across all four appliance types are provided above the bars.

It becomes apparent that the ESL value of 99 s shows a moderate average F1 score of 0.6 for both methods, even though this choice allows the sequence-based learning algorithms to only consider input data from slightly more than the last one and a half minutes. However, as expected from the interpolation of the samples collected during the previous study (cf. Fig. 3), using individual values (at ESLs where the F1 score is maximal or the NDE is minimal) leads to a much better disaggregation performance. Interestingly, the value proposed in the publication introducing S2S and S2P for NILM [9] consistently shows the poorest performance among the considered candidate ESL values.

3.2 The Best Overall ESL Compromise

If it is desired to configure a single ESL value for all appliances, a compromise needs to be found. In order to find such an appliance-agnostic optimum ESL value, we have summed up the data points for each of the four appliance types and applied the same Bézier



(a) F1 scores for the S2S method (number on top shows average value).

(b) F1 scores for the S2P method (number on top shows average value).

Figure 4: NILM performance for the appliances under consideration using the ESLs candidates identified in Table 1.

interpolation as before, resulting in the graph shown in Fig. 5. The F1 score of the S2S is the only feature that exhibits a peak value, at 5,614 s. We thus show the corresponding F1 scores in the rightmost group of bars in Fig. 4a. While this value does not show any clear correlation to the activity durations of the appliances under consideration, the resulting overall F1 score is still very close to the observed maximum when selecting ESLs that yield the maximum F1 score individually for each appliance. This confirms the general viability of using a single ESL value across a whole dataset, yet at the same time motivates further research into methodologies and/or heuristics to cater for its optimum selection.

4 CONCLUSION

Algorithms that leverage (deep) neural networks have been demonstrated to be the top-performing NILM methods to date [8]. The most important parameter for two such methods (S2S and S2P) – their Effective Sequence Length (ESL), i.e., the extent of historical data available to the algorithms – has, however, seen little explicit consideration in literature so far. We have thus conducted a comparative assessment of the F1 scores when different ESL values are used on real-world data (from the DRED dataset), using a data-driven approach to identify the promising candidate values for best disaggregation results. Our analysis proves that the optimum ESL choice strongly depends on the appliance type to identify in aggregate data. Only the appliance-wise ESL choice leads to the best overall disaggregation performance. When only a single ESL is supposed to be used (e.g., for the sake of configuration simplicity), an informed choice should be made, such as through conducting empirical studies on ESL values that maximize the overall F1 score.

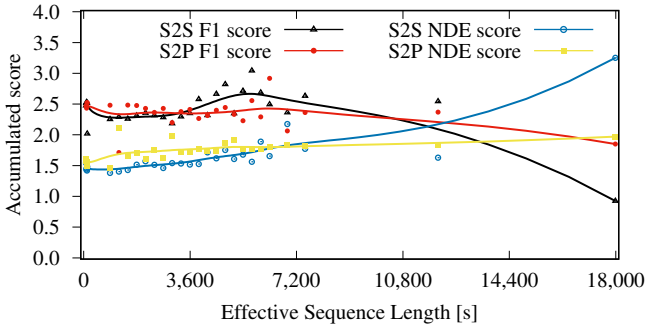


Figure 5: Aggregated scores for all appliances. The only distinctive peak is at an ESL of 5,614 s for the F1 score of S2S.

ACKNOWLEDGMENTS

This work was supported by Deutsche Forschungsgemeinschaft grant no. RE 3857/2-1.

REFERENCES

[1] N. Batra, J. Kelly, O. Parson, H. Dutta, W. Knottenbelt, A. Rogers, A. Singh, and M. Srivastava. 2014. NILMTK: An Open Source Toolkit for Non-Intrusive Load Monitoring. In *Proceedings of the 5th ACM International Conference on Future Energy Systems (e-Energy)*.

[2] N. Batra, R. Kukunuri, A. Pandey, R. Malakar, R. Kumar, O. Krystalakos, M. Zhong, P. Meira, and O. Parson. 2019. Towards Reproducible State-of-the-Art Energy Disaggregation. In *Proceedings of the 6th ACM International Conference on Systems for Energy-Efficient Buildings, Cities, and Transportation (BuildSys)*.

[3] G. W. Hart. 1992. Nonintrusive Appliance Load Monitoring. *Proc. IEEE* 80, 12 (1992).

[4] J. Kelly and W. Knottenbelt. 2015. Neural NILM: Deep Neural Networks Applied to Energy Disaggregation. In *Proceedings of the 2nd ACM International Conference on Embedded Systems for Energy-Efficient Built Environments (BuildSys)*.

[5] J. Kelly and W. Knottenbelt. 2015. The UK-DALE Dataset, Domestic Appliance-level Electricity Demand and Whole-House Demand from Five UK Homes. *Scientific Data* 2, 150007 (2015).

[6] O. Krystalakos, C. Nalmpantis, and D. Vrakas. 2018. Sliding Window Approach for Online Energy Disaggregation Using Artificial Neural Networks. In *Proceedings of the 10th Hellenic Conference on Artificial Intelligence (SETN)*.

[7] S. N. A. U. Nambi, A. Reyes Lua, and V. R. Prasad. 2015. LocED: Location-Aware Energy Disaggregation Framework. In *Proceedings of the 2nd ACM International Conference on Embedded Systems for Energy-Efficient Built Environments (BuildSys)*.

[8] A. Reinhardt and C. Klemenjak. 2020. How does Load Disaggregation Performance Depend on Data Characteristics? Insights from a Benchmarking Study. In *Proceedings of the 11th ACM International Conference on Future Energy Systems (e-Energy)*.

[9] C. Zhang, M. Zhong, Z. Wang, N. Goddard, and C. Sutton. 2018. Sequence-to-Point Learning with Neural Networks for Non-Intrusive Load Monitoring. In *Proceedings of the 32nd AAAI Conference on Artificial Intelligence (AAAI)*.

[10] Y. Zhang, G. Yang, and S. Ma. 2019. Non-Intrusive Load Monitoring based on Convolutional Neural Network with Differential Input. In *Proceedings of the 11th Conference on Industrial Product-Service Systems (CIRP)*, Vol. 83.