EMERGING TECH CONFERENCE – Edge Intelligence

Volume 03, 2024, pages 24-30

Proceedings of Emerging Tech Conference: Edge Intelligence 2024

Edge-Optimized NILM: Combining Structured Pruning and Quantization for Energy Disaggregation

Sotirios Athanasoulias[1], Dimitris Karagkounis[2], George Samaras[2], IsidorosKokos[2]
[1] *National Technical University of Athens, Athens, Greece sotiriosathanasoulias@mail.ntua.gr*
[2] *Intracom Telecom, Athens, Greece*
{dkaragkounis, gsamaras, isik}@intracom-telecom.com

Abstract

Non-Intrusive Load Monitoring (NILM) enables the disaggregation of total energy consumption, measured in a single household, into the individual energy usage of domestic appliances. This paper presents a novel methodology aimed at optimizing NILM applications by integrating structured pruning and quantization techniques to enhance the efficiency and performance of deep neural networks for edge-enabled deployment. Structured pruning is employed to identify and remove non-essential weights on a layer-by-layer basis while preserving critical feature information and maintaining overall model performance. In parallel, quantization is applied further to compress the model in terms of its memory requirements. The proposed approach is evaluated using the Plegma dataset, a novel resource in NILM research that captures local consumption behaviors and device usage pat terns specific to the Mediterranean region. Experimental results demonstrate significant reduction in the size of the model without jeopardizing inference performance, highlighting the potential of edge-based NILM to contribute to energy efficiency and transition in Mediterranean environments.

1 Introduction

Non-Intrusive Load Monitoring (NILM), also known as energy disaggregation, enables the identification of the operational state and energy consumption of individual appliances from aggregated consumption data [4]. NILM empowers consumers by enhancing energy awareness, improving cost management, supporting demand-side flexibility services, and promoting energy efficiency, ultimately aiding the transition to more sustainable energy systems. Advanced metering infrastructure is crucial for NILM applications, providing high-resolution, ideally real time, data streams. With over half of EU households now equipped with smart meters, 13 countries surpassing 80% penetration, and a strategic focus on smart meter deployment both in the EU and globally [1], NILM has the potential to play a pivotal role in advancing innovative energy services.

NILM algorithms span from traditional statistical analysis methods to more recent deep learning approaches, with the latter demonstrating promising outcomes [4], though they come with higher computational demands. These demands pose challenges for deployment in edge environments and limit scalability due to privacy concerns and data transmission issues. NILM is inherently a context-sensitive problem, as appliance energy consumption is influenced by a range of external and internal factors,

including weather conditions, seasonal variations, appliances' usage patterns and technological variations. Although substantial research has been conducted in regions such as Northern Europe, the UK, and the USA [2], studies focusing on the Mediterranean are relatively limited. This region presents distinct environmental conditions and energy use patterns, particularly in relation to appliances like air conditioners and electric boilers, which contribute significantly to household energy consumption and hold potential as sources of demand-side flexibility.

To address these gaps, our work proposes an edge-based NILM approach for performance aware model minimization, facilitating edge deployment for such applications in the domestic sector. This work also provides and evaluates a NILM algorithm in a new context, utilizing the newly published Plegma dataset [3]. This open dataset provides granular measurements from in-home appliances in the Mediterranean region, enriching the availability of open data for energy disaggregation found in the literature and enabling novel energy services that consider the specific behavioral patterns and devices of the region. In summary, the contributions of this work are as follows:

- Introduce an optimized model minimization methodology for edge-NILM applications based on structured pruning and quantization
- Establish a benchmark for NILM algorithms in the Mediterranean context using the Plegma dataset, while offering insights into the impact of region-specific appliance usage patterns on NILM performance and scalability.

The rest of the article is organized as follows: Section 2 presents the related work, Section 3 presents the mathematically formulation of the problem of NILM and describes the proposed edge-optimized methodology, Section 4 discusses the evaluation results, while concluding re marks are presented in Section 5.

2 Related work

Deep learning has seen remarkable success across various fields, including natural language processing, time-series analysis, and computer vision. Recently, several deep learning methods have been developed for NILM, demonstrating superior performance in energy disaggregation tasks. However, despite their effectiveness, these deep learning-based NILM models are hindered by high computational complexity, leading to increased training costs and presenting significant challenges for real-world deployment, particularly in resource-constrained edge environments [8].

To overcome these challenges, recent advancements have driven the integration of NILM and related energy applications into edge devices. Deploying these models on the edge not only reduces computational overhead but also eliminates the need for data transmission between users and a central server, addressing privacy concerns and challenges tied to centralized data processing. The research landscape for NILM on edge devices is expanding rapidly, with various methodologies being explored, including deep learning models optimized for edge deployment, feature extraction techniques, and federated learning approaches [4,5,8].

A common technique for compressing NILM models is pruning, which is a deep learning method used to create more computationally efficient neural networks by eliminating less significant parameters [8]. However, in most studies, pruning is implemented through the use of binary masks, where model weights are set to zero but not physically removed from the network. While this method offers a reasonable approximation of the model's behavior, it does not represent true pruning, as the network's structure

remains unchanged.

To address this issue and enable real-world deployment, we implement a comprehensive and fully automated structured pruning method. This approach explicitly models the dependencies between layers and comprehensively group coupled parameters for pruning [6]. Additionally, we combine this method with dynamic model quantization to further reduce the model size, enhancing the overall efficiency of deploying NILM applications on edge devices.

3 Optimising NILM for edge deployments

3.1. Problem Formulation

Non-intrusive load monitoring (NILM) problem [7], involves estimating the electricity consumption of individual appliance y_i *(i ∈ [1, M])*, solely based on the aggregated consumption at household level *x*, at timepoint *t* within a fixed time window *t ∈ [1, T].* A noise term ϵ is also utilized to capture error of measurement instruments and appliances not individually metered. Hence, NILM formula is defined as:

$$x(t) = \sum_{i=1}^{M} y_{i(t)} + \varepsilon(t)$$

Deep learning networks are frequently used to address this problem, as they can analyze time series data and identify appliance signatures. However, the high-dimensional nature and computational demands of these models require the application of compression techniques to enhance efficiency.

3.2. Structured Pruning

Pruning in deep neural networks (DNNs) is a widely used technique aimed at reducing model size and computational complexity by eliminating unnecessary parameters, thereby enhancing the model's efficiency while maintaining acceptable performance. Pruning can target either individual weights or entire structural units within the network [4]. Unstructured pruning involves the removal of individual weights, while structured pruning eliminates entire units, such as filters or channels, based on a predefined criterion, such as the L1 norm. In the case of structured pruning, the sparsity ratio s is defined as the proportion of pruned units (filters, neurons, or channels) relative to the total units. For structured pruning, a channel C_i is pruned if its L1 norm $\left|\left|C_i\right|\right|_1$ falls below a threshold determined by the sparsity ratio s. This can be expressed as $\left|\left|C_i\right|\right|_1 < s$, for all $Ci \in C$. After pruning, the resulting set of active channels is denoted as $C' \in \mathbb{N}^{k'}, where\ k' < k$, is the reduced number of channels.

The structured pruning process reduces both the model size and the computational com plexity. The number of parameters in the network is reduced proportionally to the number of pruned channels, and the computational cost is reduced by eliminating the associated floating-point operations (FLOPs) of the pruned channels.

3.3. Dynamic Quantization

Quantization is a technique that reduces storage and memory requirements by converting model weights from floating-point to lower-precision integers. This process can be described as an irreversible mapping where floating-point weights are discretized into integer bins [9]. Quantization may be performed either

post-training, by compressing a pre-trained model, or during training through quantization-aware training. This study focuses on post-training dynamic quantization, which offers flexibility and adaptability to varying input data. It also simplifies the compression process by determining scale and zero-point dynamically during inference, removing the need for a separate calibration phase.

In this study, dynamic quantization was applied specifically to the linear layers of the model because these layers, compared to convolutional layers, are less sensitive to precision loss and have a more predictable activation range. This makes linear layers better suited for efficient quantization without significant degradation in model accuracy.

4 Experimental Results

4.1. Experimental Setup

The selected model for this study is a sequence-to-sequence 1D Convolutional Neural Network, noted for its effectiveness in edge-NILM tasks [4]. Experiments were conducted using the Plegma dataset [3], featuring electricity consumption data sampled every 10 seconds. The model was trained on data from houses 1-13, with house 2 excluded and used as unseen test data.

To implement the model compression strategy, we utilized DepGraph [6] and the associated Torch-Pruning library. These tools prune substructures based on their interdependencies, resulting in smaller models, unlike other libraries that only apply binary masking without fully removing components from the model. Regarding quantization, we employed the Pytorch quantization library.

4.2. Model Compression & Performance Evaluation Metrics

The model's performance in predicting the energy consumption profiles of the tested appliances was assessed using Mean Absolute Error and Mean Relational Error. F1 Score was used to measure its accuracy in identifying the operational states of appliances (on/off). Additionally, the model's computational complexity was evaluated through Floating Point Operations (FLOPS), the number of trainable parameters, and their storage size in Megabytes (MB). To determine the optimal pruning threshold, we used the metric from [4], selecting the point where the Euclidean distance between the sparsity-F1 pruning ratio curve and the ideal point is minimized.

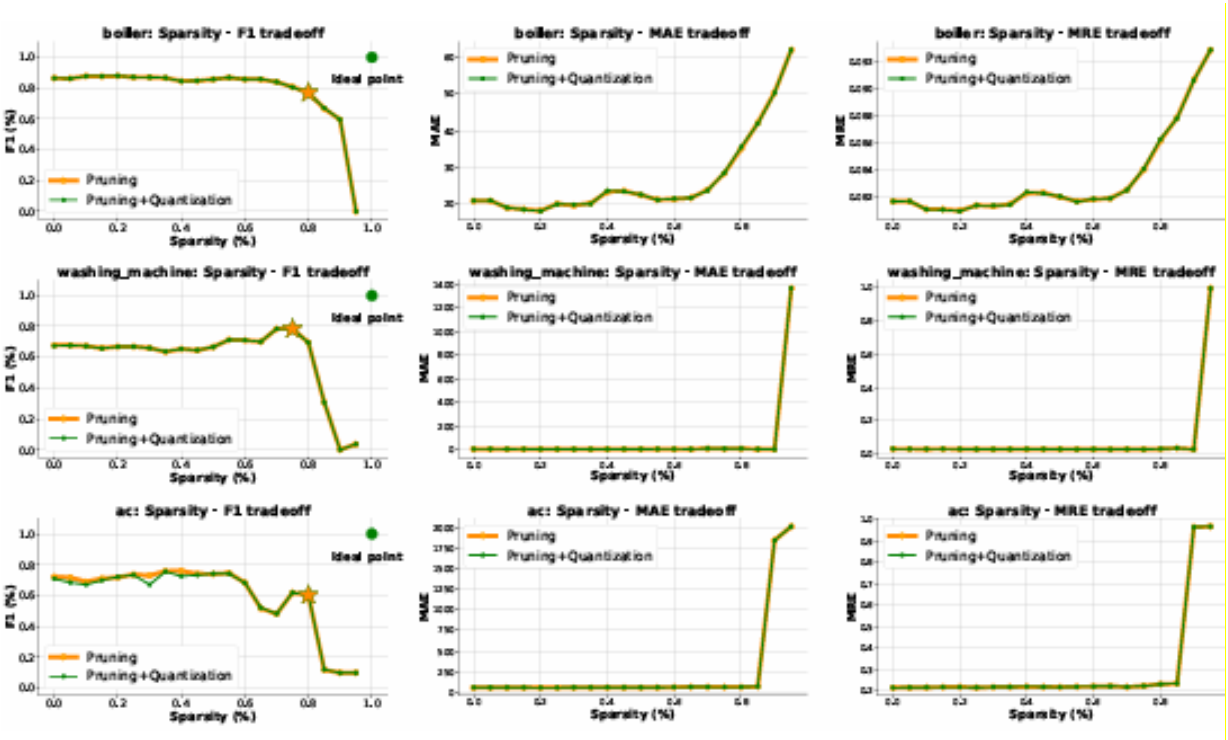

Figure1: Comparison of different pruning thresholds (0-95%) to performance degradation (F1, MAE, MRE) for different devices for structured pruning and structured pruning with dynamic quantization approaches. The optimal pruning thresholds for both approaches are denoted with the star symbol.

4.3. Results Evaluation

Figure 1 presents a comparative analysis between structured pruning and structured pruning combined with dynamic quantization. The evaluation spans various compression levels, with sparsity ratios ranging from 0% to 95%, and uses the selected performance metrics. The resulting 'sparsity vs. performance' curves indicate that both methods demonstrate comparable performance across all pruning thresholds. In most cases, pruning with dynamic quantization achieves nearly identical results to pruning alone, though it occasionally exhibits a slight decrease in performance.

Appliance	Approach	Pruning (%) P_thr = P_opt	Performance metrics		
			F1- Score	MAE	MRE
Boiler	Baseline Model	0	0.864	21.002	0.081
	Structured Pruning	80	0.772	35.067	0.086
	Structured Pruning + Quantization	80	0.768	35.480	0.086
Washing Machine	Baseline Model	0	0.757	7.834	0.029
	Structured Pruning	75	0.806	9.678	0.026
	Structured Pruning + Quantization	75	0.784	9.680	0.026
A/C	Baseline Model	0	0.72	56.687	0.215
	Structured Pruning	80	0.603	63.842	0.231
	Structured Pruning + Quantization	80	0.602	63.855	0.231

Table 1: Comparative performance evaluation results- baseline model vs optimal pruned model vs optimal pruned quantized model.

A closer examination of the performance metrics at the optimal pruning thresholds, as shown in Table 1, reveals that the optimal pruning thresholds are identical for both compression methods across all tested appliances. The performance metrics confirm the insights drawn from the sparsity diagram, showing that structured pruning with quantization has a minimal impact on the performance of structured pruning alone for all appliances.

Appliance	Approach	Pruning (%) P_thr = P_opt	Compression metrics		
			Model params	FLOPS	MB
Boiler	Baseline Model	0	22147640	2501.32	84.49
	Structured Pruning	80	868835	89.96	3.31
	Structured Pruning + Quantization	80	868835	89.96	0.84
Washing Machine	Baseline Model	0	22147640	2501.32	84.49
	Structured Pruning	75	1417050	154.84	5.41
	Structured Pruning + Quantization	75	1417050	154.84	1.36
A/C	Baseline Model	0	22147640	2501.32	84.49
	Structured Pruning	80	868835	89.96	3.31
	Structured Pruning + Quantization	80	868835	89.96	0.84

Table 2: Comparative compression evaluation results- baseline model vs optimal pruned model vs optimal pruned quantized model.

Although dynamic quantization has minimal impact on the performance of the pruned models, the same cannot be said for compression metrics. As shown in Table 2, quantization significantly reduces the storage requirements of the pruned models by 74.62% for both the boiler and the A/C, and by 74.86% for the washing machine. The FLOPS and the number of trainable parameters, however, remain unchanged, as quantization does not alter these aspects of the model. In conclusion, the proposed structured pruning combined with dynamic quantization proves to be an effective approach, as it maintains model performance while significantly reducing storage requirements. When comparing the baseline model with the optimized approach, we observe a slight performance decline, yet the compression metrics exhibit significant improvements. Specifically, the number of trainable parameters and FLOPS were reduced by up to 96%, and storage requirements saw a reduction of 99%. Despite an average performance drop of approximately 10% across all appliances, there were instances, such as with the washing machine, where the compressed model outperformed the baseline. This improvement is likely due to the overparameterization of the baseline model, allowing the compressed model to generalize more effectively.

5 Conclusions

This study tackles the challenges of resource-intensive deep learning models in NILM, specifically addressing their context-aware nature. Unlike previous research, this work benchmarks appliances such as air conditioners and boilers from the Plegma dataset, which are common in the Mediterranean region. It evaluates the performance of a baseline CNN model, structured pruning, and structured pruning combined with quantization. Results show that all methods effectively disaggregate appliance usage, providing insights into NILM's applicability in this region. Notably, the structured pruning approach reduces the model's MFLOPs by up to 96% and model size by 99%, lowering computational demands, improving feasibility for edge deployment, promoting NILM technologies for wider adoption.

Acknowledgements

This project has received funding from the European Union's Horizon 2020 and Horizon Eu rope research and innovation program under the Marie Sklodowska-Curie and ODEON grant agreements, No 955422 and No 101136128, respectively.

References

[1] Energy Retail and Consumer Protection 2023 Market Monitoring Report, ACER, September 2023.

[2] Georgios-Fotios Angelis, Timplalexis, Stelios Krinidis, Dimosthenis Ioannidis, and Dimitrios Tzo varas. Nilm applications: Literature review of learning approaches, recent developments and chal lenges. Energy and Buildings Volume, 261, 2022.

[3] S. Athanasoulias, F. Guasselli, N. Doulamis, A. Doulamis, N. Ipiotis, A. Katsari, L. Stankovic, and V. Stankovic. The plegma dataset: Domestic appliance-level and aggregate electricity demand with metadata from greece. Scientific Data, vol. 11, no. 1, p. 376, April 2024.

[4] S. Athanasoulias, S. Sykiotis, M. Kaselimi, A. Doulamis, N. Doulamis, and N. Ipiotis. Opt-nilm: An iterative prior-to-full-training pruning approach for cost-effective user side energy disaggregation,. Transactions on Consumer Electronics, vol. 70, no. 1, page 4435–4446, 2024.

[5] Sotirios Athanasoulias, Stavros Sykiotis, Nikos Temenos, Anastasios Doulamis, and Nikolaos Doulamis. A pre-training pruning strategy for enabling lightweight non-intrusive load monitor ing on edge devices. In 2024 ICASSPW, pages 249–253, 2024.

[6] Gongfan Fang, Xinyin Ma, Mingli Song, Michael Bi Mi, and Xinchao Wang. Depgraph: Towards any structural pruning. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 16091–16101, 2023.

[7] G. W. Hart. Nonintrusive appliance load monitoring. Proceedings of the IEEE, 80, no. 12:pp. 1870–1891, 1992.

[8] Stavros Sykiotis, Sotirios Athanasoulias, Maria Kaselimi, Anastasios Doulamis, Nikolaos Doulamis, Lina Stankovic, and Vladimir Stankovic. Performance-aware nilm model optimization for edge deployment. IEEE Transactions on Green Communications and Networking, 7(3):1434–1446, 2023.

[9] Yuhui Xu, Shuai Zhang, Yingyong Qi, Jiaxian Guo, Weiyao Lin, and Hongkai Xiong. Dnq: Dynamic network quantization. arXiv preprint arXiv:1812.02375, 2018.