# Explore Activation Sparsity in Recurrent LLMs for Energy-Efficient Neuromorphic Computing

Ivan Knunyants[1,2], Maryam Tavakol[2], Manolis Sifalakis[1], Yingfu Xu[1], Amirreza Yousefzadeh[3], and Guangzhi Tang[4]

[1] imec the Netherlands, Eindhoven, Netherlands
[2] TU Eindhoven, Eindhoven, Netherlands
[3] University of Twente, Enschede, Netherlands
[4] Maastricht University, Maastricht, Netherlands
guangzhi.tang@maastrichtuniversity.nl

**Abstract.** The recent rise of Large Language Models (LLMs) has revolutionized the deep learning field. However, the desire to deploy LLMs on edge devices introduces energy efficiency and latency challenges. Recurrent LLM (R-LLM) architectures have proven effective in mitigating the quadratic complexity of self-attention, making them a potential paradigm for computing on-edge neuromorphic processors. In this work, we propose a low-cost, training-free algorithm to sparsify R-LLMs' activations to enhance energy efficiency on neuromorphic hardware. Our approach capitalizes on the inherent structure of these models, rendering them well-suited for energy-constrained environments. Although primarily designed for R-LLMs, this method can be generalized to other LLM architectures, such as transformers, as demonstrated on the OPT model, achieving comparable sparsity and efficiency improvements. Empirical studies illustrate that our method significantly reduces computational demands while maintaining competitive accuracy across multiple zero-shot learning benchmarks. Additionally, hardware simulations with the SENECA neuromorphic processor underscore notable energy savings and latency improvements. These results pave the way for low-power, real-time neuromorphic deployment of LLMs and demonstrate the feasibility of training-free on-chip adaptation using activation sparsity.

**Keywords:** Large Language Models · Recurrent Neural Networks · Neuromorphic Computing · Activation Sparsity

Recurrent LLMs (R-LLMs) [6, 1, 2] have emerged as lighter alternatives to self-attention LLMs. These methods combine the ability of recurrent inference and operate with linear complexity, addressing the quadratic self-attention issue while still benefiting from the rapid parallel training. Additionally, the recurrent computational paradigm is well-suited for neuromorphic processors, which are brain-inspired low-power AI-dedicated hardware [8, 5]. These processors manage event-based data-flow processing, exploiting the activation sparsity in neural networks for energy-efficient and low-latency computation. To maximize the benefits of neuromorphic computing, a high level of activation sparsity in the neural

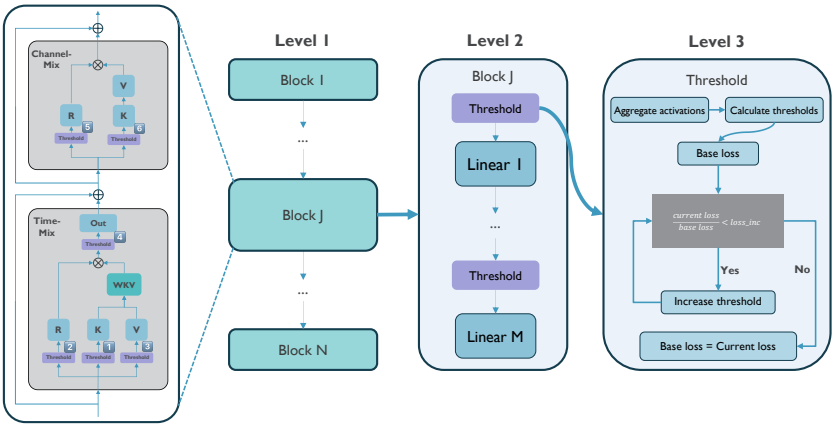


**Fig. 1.** Recurrent LLM (RWKV [6]) block with proposed thresholding function for activation sparsity and the training-free threshold initialization algorithm.

networks is crucial [9]. However, R-LLM models consist of dense linear layers, including high-dimensional upward and downward projections, resulting in costly computation on neuromorphic processors. Therefore, there is a need to explore activation sparsity in R-LLMs for energy-efficient neuromorphic computing.

State-of-the-art methods for improving activation sparsity in LLMs apply the ReLU activation function before linear layers in pre-trained models and perform fine-tuning on large datasets to restore performance [4, 7]. Moreover,[7] applied activation regularization in loss functions to enable sparse-aware fine-tuning and further increased sparsity. However, these approaches involve a training-based fine-tuning stage, which requires high computational costs due to the model size and training tokens. Privacy concerns over local fine-tuning data make on-device LLM adaptations preferable [3], but training computational costs render this impractical. Furthermore, current solutions are designed for transformer-based models, and to our knowledge, no low-cost activation sparsification method exists for R-LLMs that are also well-suited for on-device adaptation.

This abstract proposes an R-LLM with activation sparsity for energy-efficient neuromorphic computing and a training-free threshold adaptation algorithm for improving activation sparsity (see Figure 1). Our contributions are as follows:

- We introduce an event-based R-LLM with thresholding functions. The resulting R-LLM (RWKV 430M-3B [6]) obtains an average 63% activation sparsity, increasing 2.2× compared to the model with natural sparsity.
- We propose a training-free algorithm to find thresholds using local data adaptively. The algorithm can be deployed on neuromorphic processors and is 30× faster on GPU than the training-based method [4].
- We demonstrate a 1.9× energy and latency improvement of the sparse model via hardware simulation with the SENECA neuromorphic processor [8].
- We extend our approach to a self-attention LLM (OPT 2.7B [10]), demonstrating comparable results to the training-based method [4].

## References


1. Beck, M., Poppel, K., Spanring, M., Auer, A., Prudnikova, O., Kopp, M.K., Klambauer, G., Brandstetter, J., Hochreiter, S.: xlstm: Extended long short-term memory. ArXiv **abs/2405.04517** (2024), https://api.semanticscholar.org/CorpusID:269614336
2. Dao, T., Gu, A.: Transformers are ssms: Generalized models and efficient algorithms through structured state space duality. ArXiv **abs/2405.21060** (2024), https://api.semanticscholar.org/CorpusID:270199762
3. Li, Q., Hong, J., Xie, C., Tan, J., Xin, R., Hou, J., Yin, X., Wang, Z., Hendrycks, D., Wang, Z., et al.: Llm-pbe: Assessing data privacy in large language models. arXiv preprint arXiv:2408.12787 (2024)
4. Mirzadeh, I., Alizadeh-Vahid, K., Mehta, S., Mundo, C.C.D., Tuzel, O., Samei, G., Rastegari, M., Farajtabar, M.: Relu strikes back: Exploiting activation sparsity in large language models. ArXiv **abs/2310.04564** (2023), https://api.semanticscholar.org/CorpusID:263830421
5. Modha, D.S., Akopyan, F., Andreopoulos, A., Appuswamy, R., Arthur, J.V., Cassidy, A.S., Datta, P., DeBole, M.V., Esser, S.K., Otero, C.O., Sawada, J., Taba, B., Amir, A., Bablani, D., Carlson, P.J., Flickner, M., Gandhasri, R., Garreau, G.J., Ito, M., Klamo, J.L., Kusnitz, J.A., McClatchey, N.J., McKinstry, J.L., Nakamura, Y.Y., Nayak, T.K., Risk, W.P., Schleupen, K., Shaw, B., Sivagnaname, J., Smith, D.F., Terrizzano, I., Ueda, T.: Ibm northpole neural inference machine. 2023 IEEE Hot Chips 35 Symposium (HCS) pp. 1–58 (2023), https://api.semanticscholar.org/CorpusID:263183607
6. Peng, B., Alcaide, E., Anthony, Q.G., Albalak, A., Arcadinho, S., Biderman, S., Cao, H., Cheng, X., Chung, M., Grella, M., Kranthikiran, G., He, X., Hou, H., Kazienko, P., Kocoń, J., Kong, J., Koptyra, B., Lau, H., Mantri, K.S.I., Mom, F., Saito, A., Tang, X., Wang, B., Wind, J.S., Wozniak, S., Zhang, R., Zhang, Z., Zhao, Q., Zhou, P., Zhu, J., Zhu, R.: Rwkv: Reinventing rnns for the transformer era. In: Conference on Empirical Methods in Natural Language Processing (2023), https://api.semanticscholar.org/CorpusID:258832459
7. Song, C., Han, X., Zhang, Z., Hu, S., Shi, X., Li, K., Chen, C., Liu, Z., Li, G., Yang, T., Sun, M.: Prosparse: Introducing and enhancing intrinsic activation sparsity within large language models. ArXiv **abs/2402.13516** (2024), https://api.semanticscholar.org/CorpusID:267770175
8. Tang, G., Vadivel, K., Xu, Y., Bilgic, R., Shidqi, K., Detterer, P., Traferro, S., Konijnenburg, M., Sifalakis, M., van Schaik, G.J., et al.: Seneca: building a fully digital neuromorphic processor, design trade-offs and challenges. Frontiers in Neuroscience **17**, 1187252 (2023)
9. Xu, Y., Shidqi, K., van Schaik, G.J., Bilgic, R., Dobrita, A., Wang, S., Meijer, R., Nembhani, P., Arjmand, C., Martinello, P., Gebregiorgis, A., Hamdioui, S., Detterer, P., Traferro, S., Konijnenburg, M., Vadivel, K., Sifalakis, M., Tang, G., Yousefzadeh, A.: Optimizing event-based neural networks on digital neuromorphic architecture: a comprehensive design space exploration. Frontiers in Neuroscience **18** (2024), https://api.semanticscholar.org/CorpusID:268836639
10. Zhang, S., Roller, S., Goyal, N., Artetxe, M., Chen, M., Chen, S., Dewan, C., Diab, M., Li, X., Lin, X.V., et al.: Opt: Open pre-trained transformer language models. arXiv preprint arXiv:2205.01068 (2022)