International Conference on Multidisciplinary Perspectives in Advanced Computing and Technology (IMPACT 2026)
G. B. Pant University of Agriculture and Technology, Uttarakhand, India. Jan. 10-11, 2026
DMPedia Lecture Notes in Multidisciplinary Research,
Volume 1, pp. 593-544, 2026

A Unified Review of Advanced Paradigms in Spiking Neural Networks From Efficiency and Latency to Novel Architecture and Robustness

Vartika Pandey, Abhinav Malkoti, Parv Saini, Divyansh Upadhyay, and Hardik Saxena
Department of Computer Science, UPES, Dehradun
27.vp.10@gmail.com, abhi.malkoti@gmail.com, parvsaini05@gmail.com,
divyanshupadhyay67000@gmail.com, hardiksaxena267@gmail.com

ABSTRACT

The domain of Spiking Neural Networks (SNNs) has undergone a paradigm shift in the 2024-2025 research cycle, moving beyond theoretical validation to successfully integrate into complex domains like Large Language Models (LLMs) and advanced robotics. This review provides an exhaustive synthesis of state-of-the-art SNN optimization, focusing on Architectural Innovation, Training Algorithm Refinement, and Hardware-Software Co-Design. Key breakthroughs include the "Transformerization" of SNNs via addition-only attention $(A^{2}OS^{2}A)$, the emergence of Saliency-Based Spiking LLMs (SpikeLLM), and the theoretical re-grounding of training dynamics on Riemannian Manifolds (MSG). By detailing these advancements, this paper delineates the current performance-oriented neuromorphic frontier and identifies the critical need for unified, robust, and hardware-aware optimization frameworks.

Keywords: *Spiking Neural Networks, Large Language Models, Neuromorphic Computing, Hardware-Software Co-Design, Transformer Architecture, Riemannian Manifolds, Optimization, Robustness.*

1. Introduction

SNNs represent the third generation of neural networks, leveraging binary spike communication to achieve superior theoretical energy efficiency compared to continuous-valued ANNs. The fundamental computational unit, the Leaky Integrate-and-Fire (LIF) neuron, introduces a temporal dimension to processing, enabling rich dynamics but challenging traditional gradient-based training and demanding high memory in deep architectures. This paper synthesizes the most significant optimization trends (2024-2025) that address the core bottlenecks in SNNs:

1.1. **Architectural Scalability:** Enabling SNNs to handle billions of parameters (LLMs) and complex global dependencies (Transformers).

1.2. **Efficiency and Deployment:** Reducing model size and latency for constrained edge devices via quantization, pruning, and temporal coding.

1.3. **Robustness and Training Stability:** Developing algorithms that overcome gradient instability and defend against increasingly sophisticated adversarial attacks.

1.4. **Novel Application Domains:** Applying SNNs to continuous control, geometric learning, and medical imaging.

2. Training and Temporal Dynamics: Overcoming Non-Differentiability

The non-differentiability of the spike generation function, $S_{t}=\Theta(\tilde{U}_{t})$, is the fundamental optimization challenge. All modern SNNs rely on the Surrogate Gradient (SG) approach during training.

2.1. Gradient Stabilization and Loss Refinement

Recent work has moved beyond simple SG tuning to advanced stabilization techniques:

2.1.1. **Enhancing the Output Feature (EnOF-SNN) [12]:** This framework tackles information loss at the readout layer. It replaces the final LIF layer with a ReLU activation layer during training, enabling the generation of full-precision feature representations. This hybrid approach spiking hidden layers with a continuous readout is guided by a Knowledge Distillation strategy (using a Loss for Aligned Features or LAF loss) to align the SNN's feature space with a high-performing ANN.

2.1.2. **Bit-Reversal and Normalization (ReverB-SNN) [13]:** To combat gradient vanishing in deep SNNs, ReverB-SNN introduces Reversing the Bit of Weight and Activation as a regularizer. Complementing this is Temporal Accumulated Batch Normalization (TAB), which normalizes membrane potentials based on their firing history, stabilizing firing rates and preventing the network from falling into silent or "epileptic" (constant firing) states.

2.2. Temporal Coding and Dynamic Inference

The shift from Rate Coding (value proportional to spike frequency) to Temporal Coding (value proportional to spike timing) is central to low-latency optimization:

2.2.1. **Time-to-First-Spike (TTFS) Coding:** In TTFS, information is encoded in the precise latency of the first spike, drastically reducing the total spike count and power consumption.

2.2.2. **Dynamic Inference and Top-K Cutoff [17]:** Traditional SNNS operate with a fixed number of timesteps (7). Dynamic inference, such as the Top-K Cutoff strategy, monitors the output confidence and terminates inference early once the top class prediction is stable. A training regularization term is introduced to encourage the network to "front-load" information, making confident predictions earlier in the spike train.

3. Architectural Scalability: Transformers and LLMs

The most profound optimization achievement is the successful integration of SNNs with attention and large language models.

3.1. The Spiking Transformer Era

The goal is to maintain the global receptive field of the self-attention mechanism while eliminating floating-point matrix multiplication and the Softmax function.

3.1.1. **Accurate Addition-Only Spiking Self-Attention (A2OS²A) [6]:** This architecture solves the multiplication problem by using a hybrid neuron model: binary Query (Q), ReLU Key (K), and ternary

Value (V). The attention map Q K^{T} becomes a simple summation (multiplication by 0 or 1), ensuring the computation remains strictly additive and maximizing energy efficiency while achieving 78.66% accuracy on ImageNet-1K.

3.1.2. **Spatial-Temporal Attention (STAtten) [14]:** Addressing the limitation of purely spatial attention, STAtten introduces a block-wise computation strategy that attends across a small window of adjacent timesteps, effectively capturing the motion dynamics and temporal correlations inherent in spike trains.

3.1.3. **Training-Free Conversion:**

- **SpikedAttention [12]:** Uses a Winner-Oriented Spike Shift (WOSS) mechanism to mimic the selectiveness of Softmax without explicit calculation.
- **TTFSFormer [8]:** Achieves near-lossless conversion by introducing a Generalized TTFS Neuron capable of mapping complex non-linearities (like Softmax and GELU) to spike delays.

3.2. Scaling to Large Language Models (LLMs)

SNNs are now scaling to billions of parameters, primarily by coupling the sequence dimension of the LLM with the temporal dimension of the SNN.

3.2.1. **Saliency-Based Spiking (SpikeLLM) [10]:** This technique uses Hessian-based sensitivity analysis to identify "salient" channels in a large model (e.g., 70B parameters). Salient channels are assigned multi-bit, high-fidelity spiking behavior, while non-salient channels are assigned extremely sparse, single-step spiking.

3.2.2. **Autoregressive SNNs (SpikeGPT) [11]:** SpikeGPT adapts the RWKV architecture, aligning the token stream with the SNN temporal dimension. It processes a sequence where the "time" step is the token position, allowing the model to be trained as a Recurrent Neural Network (RNN) that scales like a Transformer.

3.2.3. **Probabilistic Spiking State Space Models (P-SpikeSSM) [16]:** P-SpikeSSM leverages the inherent recurrence of spiking neurons to model long-range dependencies, treating spike timing as a probabilistic event derived from the State Space Model equation.

4. Hardware-Aware Optimization and Co-Design

The optimization frontier is moving toward hardware-software co-design, where algorithms are optimized not just for accuracy but for the specific constraints of neuromorphic hardware (e.g., Loihi, FPGA, NVM).

4.1. Quantization, Pruning, and Model Compression

4.1.1. **QP-SNN (Quantized and Pruned SNNs) [1]:** This is the SOTA in model compression for edge deployment. It uses Weight Rescaling (ReScaW) to optimize low-bit quantization and a structured pruning criterion based on the Singular Value of Spatiotemporal Spike Activity (SVS) to identify and remove kernels based on temporal variance.

4.1.2. **Custom Accelerator Designs:**

- **NVM Accelerators [20]:** Research shows digital Non-Volatile Memory (NVM) accelerators achieve significantly higher performance per watt compared to conventional SRAM designs.
- **Robustness Optimizations [21]:** Includes Spike Regenerator circuits and Adaptive STDP generators implemented in CMOS to reshape irregular spikes from unstable devices.

4.2. Automated Hardware-Aware Architecture Search

4.2.1. **AutoSNN / SNASNet:** These Neural Architecture Search (NAS) frameworks optimize SNN topology by explicitly incorporating hardware metrics (like spike count or energy model) into the fitness function. They search for optimal temporal feedback connections often absent in standard feedforward ANNs.

5. Geometric Learning and Application-Specific Optimization

5.1. Geometric Deep Learning on Manifolds (MSG)

For complex data with non-Euclidean structures, Manifold-valued Spiking Graph Neural Networks (MSG) offer a theoretical breakthrough [5].

5.1.1. **Differentiation via Manifold (DvM):** MSG re-frames training optimization on a Riemannian manifold. The MSNeuron interprets the spike train as a velocity vector in the Tangent Space, which drives the feature representation across the curved manifold via the Exponential Map.

5.2. Ultra-Low Latency and Continuous Control

5.2.1. **Ultra-Low-Latency Object Detection (SUHD) [2]:** SUHD achieves competitive mAP on MS COCO with only four timesteps by using Timesteps Compression and Spike-Time-Dependent Integrated (STDI) Coding.

5.2.2. **Continuous Robotic Control (Pred-Control SNN) [4]:** Uses a model-based learning approach with Adaptive Leaky Integrate-and-Fire (ALIF) neurons and Learnable Time Constants to tune internal temporal scales for stability in high-dimensional motor control.

5.3. Medical and Temporal Signal Processing

5.3.1. **Neuroimaging (NeuCube) [18]:** Maps spatiotemporal data into a 3D reservoir of spiking neurons, optimizing the capture of complex brain activity patterns.

5.3.2. **Spike-Based Attention for Segmentation [19]:** Integrated into backbones to enhance focus on salient features (like tumor boundaries) maintaining low inference latency (<50 ms).

6. Robustness, Security, and Future Trajectories

6.1. Adversarial Robustness and Defense Mechanisms

SNNs are highly vulnerable to adversarial attacks [3]. Defensive optimization centers on utilizing the temporal dynamics of the network:

6.1.1. **Neural Dynamic Signatures (NDS) [15]:** Defense mechanisms that minimize the deviation from a neuron's "signature" (its membrane potential averaged over clean data) during an attack.

6.1.2. **Certified Robustness [23]:** Introducing mathematical guarantees of robustness using methods like Interval Bound Propagation (IBP) and Randomized Smoothing.

7. Critical Research Gaps and Future Outlook

7.1. **The Robustness-Efficiency Trade-off:** The unified effect of model compression on adversarial vulnerability remains unknown.

7.2. **The Online Learning Gap:** Complex architectures still rely heavily on replay buffers. The promise of true online learning remains elusive for high-dimensional tasks.

7.3. **Universal Optimization Frameworks:** The field requires a framework that simultaneously optimizes for Latency + Size + Robustness for a specific hardware target.

Conflict of Interest

The authors declare no conflict of interest.

References

[1]. Wei, W., Zhang, M., Zhou, Z., Belatreche, A., Shan, Y., Liang, Y., ... & Yang, Y. (2025). Qp-snn: Quantized and pruned spiking neural networks. *arXiv preprint arXiv:2502.05905*.

[2]. Zhang, A., Cao, H., Shan, N., Wang, J., Pu, M., & Song, Y. (2025). Spiking neural networks for object detection and semantic segmentation across event-driven and frame-based modalities: A review. *Intelligent Opto-Electronics*, *1*(2), 250007-1.

[3]. Lun, L., Feng, K., Ni, Q., Liang, L., Wang, Y., Li, Y., ... & Cui, X. (2025). Towards effective and sparse adversarial attack on spiking neural networks via breaking invisible surrogate gradients. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 3540-3551).

[4]. Huebotter, J., Lanillos, P., van Gerven, M., & Thill, S. (2025). Spiking Neural Networks for Continuous Control via End-to-End Model-Based Learning. *arXiv preprint arXiv:2509.05356*.

[5]. Sun, L., Huang, Z., Wan, Q., Peng, H., & Yu, P. S. (2024). Spiking graph neural network on riemannian manifolds. *Advances in Neural Information Processing Systems*, *37*, 34025-34055.

[6]. Guo, Y., Liu, X., Chen, Y., Peng, W., Zhang, Y., & Ma, Z. (2025). Spiking transformer: Introducing accurate addition-only spiking self-attention for transformer. In *Proceedings of the Computer Vision and Pattern Recognition Conference* (pp. 24398-24408).

[7]. Zhou, Z., Che, K., Fang, W., Tian, K., Zhu, Y., Yan, S., ... & Yuan, L. (2024). Spikformer v2: Join the high accuracy club on imagenet with an snn ticket. *arXiv preprint arXiv:2401.02020*.

[8]. Zhao, L., Huang, Z., Ding, J., & Yu, Z. (2025, October). Ttfsformer: a ttfs-based lossless conversion of spiking transformer. In *Forty-second International Conference on Machine Learning*.

[9]. Svoboda, K., & Adegbija, T. (2025). Spiking Neural Network Architecture Search: A Survey. *arXiv preprint arXiv:2510.14235*.

[10]. Xing, X., Gao, B., Zhang, Z., Clifton, D. A., Xiao, S., Du, L., ... & Zhang, J. (2024). Spikellm: Scaling up spiking neural network to large language models via saliency-based spiking. *arXiv preprint arXiv:2407.04752*.

[11]. Zhu, R. J., Zhao, Q., Li, G., & Eshraghian, J. K. (2023). Spikegpt: Generative pre-trained language model with spiking neural networks. *arXiv preprint arXiv:2302.13939*.

[12]. Guo, Y., Peng, W., Liu, X., Chen, Y., Zhang, Y., Tong, X., ... & Ma, Z. (2024). Enof-snn: Training accurate

spiking neural networks via enhancing the output feature. *Advances in Neural Information Processing Systems*, *37*, 51708-51726.

[13]. Guo, Y., Zhang, Y., Jie, Z., Liu, X., Tong, X., Chen, Y., ... & Ma, Z. (2025). Reverb-snn: Reversing bit of the weight and activation for spiking neural networks. *arXiv preprint arXiv:2506.07720*.

[14]. Lee, D., Li, Y., Kim, Y., Xiao, S., & Panda, P. (2025). Spiking transformer with spatial-temporal attention. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 13948-13958).

[15]. Geng, H., & Li, P. (2023). HoSNN: Adversarially-robust homeostatic spiking neural networks with adaptive firing thresholds. *arXiv preprint arXiv:2308.10373*.

[16]. Bal, M., & Sengupta, A. (2024). P-spikessm: Harnessing probabilistic spiking state space models for long-range dependency tasks. *arXiv preprint arXiv:2406.02923*.

[17]. Wu, D., Jin, G., Yu, H., Yi, X., & Huang, X. (2025). Optimizing event-driven spiking neural network with regularization and cutoff. *Frontiers in neuroscience*, *19*, 1522788.

[18]. Garcia-Palencia, O., Fernandez, J., Shim, V., Kasabov, N. K., Wang, A., & Alzheimer's Disease Neuroimaging Initiative. (2025). Spiking neural networks for multimodal neuroimaging: A comprehensive review of current trends and the NeuCube brain-inspired architecture. *Bioengineering*, *12*(6), 628.

[19]. Al-Ebrahim, M. A. (2025). Spike-based attention mechanisms for enhanced medical image segmentation. *Engineering, Technology & Applied Science Research*, *15*(5), 28273-28285.

[20]. Kulkarni, S. R., Kadetotad, D. V., Yin, S., Seo, J. S., & Rajendran, B. (2019, November). Neuromorphic hardware accelerator for SNN inference based on STT-RAM crossbar arrays. In *2019 26th IEEE International Conference on Electronics, Circuits and Systems (ICECS)* (pp. 438-441). IEEE.

[21]. Kim, M. J., Lee, H. M., Jeong, Y., & Kwak, J. Y. (2025). Emerging Neuromorphic Devices-Compatible SNN Hardware with Adaptive STDP and Validation Using Novel CMOS Neuron-Synapse Circuits. *IEEE Access*.