Efficient Spiking Transformers with Spatio-temporal Neighborhood Attention

M. Hassanshahi[(1)], M. Shahsavari[(1)] and M. van Gerven[(1)]

[(1)] Donders Institute for Brain, Cognition and Behaviour, Radboud University, The Netherlands

Spiking Transformers have shown strong potential for vision tasks by combining the global modeling capability of attention mechanisms with the energy efficiency of spike-based computation. However, most existing approaches rely on full self-attention, which scales quadratically with the number of tokens and often flattens the temporal dimension. This makes them impractical for resource-constrained edge deployment and limits their ability to process temporally rich data such as event streams from dynamic vision sensors (DVS).

We propose Spiking Neighborhood Attention (SNA), a localized spatio-temporal attention mechanism designed specifically for spiking Transformers. Instead of attending to all tokens globally, SNA restricts each query to a fixed spatial neighborhood in 2D, and extends this to spatio-temporal neighborhoods in 3D by incorporating a limited window of adjacent timesteps. This reduces the attention complexity from $O(N^2DT)$ to $O(NK^2sDKt)$, where Ks and Kt denote the spatial and temporal neighborhood sizes, respectively. Crucially, the 3D variant preserves the temporal structure of event-based data rather than collapsing it, which is important for tasks where precise spike timing carries meaningful information.

We integrate SNA into a hierarchical spiking Transformer architecture inspired by QKFormer, where early stages use lightweight QK-attention for global feature extraction and the final stage employs our SNA module for localized, spike-based refinement. The architecture follows a three-stage design with progressive downsampling through Spiking Patch Embedding with Residual Shortcuts (SPEDS). This mixed attention strategy balances representational capacity with computational efficiency: early layers establish long-range token relationships, while the SNA block in the final stage operates within local windows where global attention would offer diminishing returns at high cost.

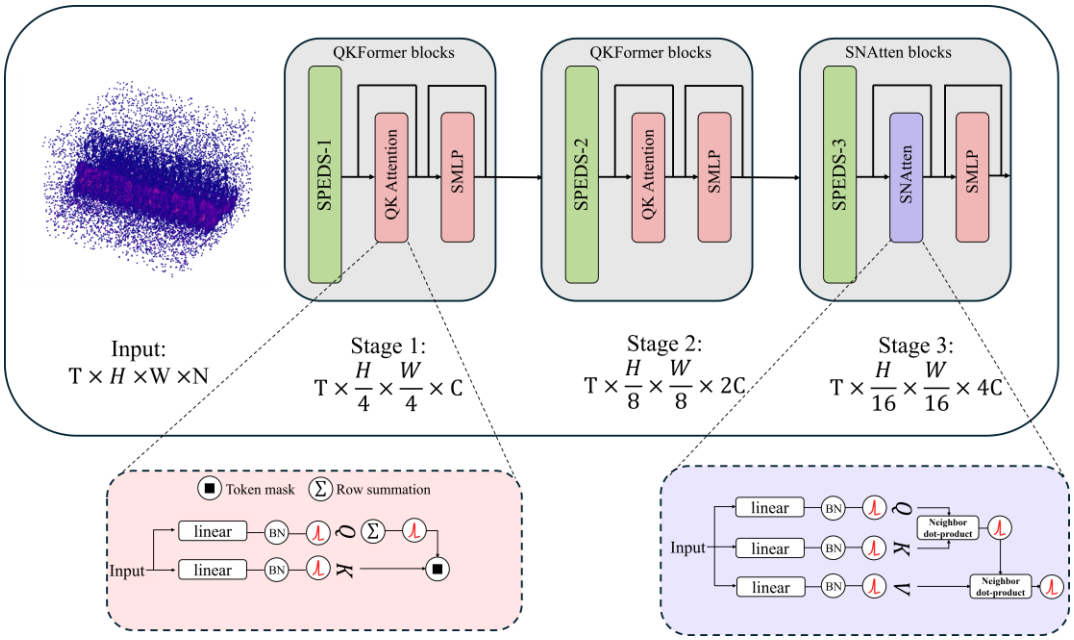

Figure 1: Architecture overview, including the QKAttention block and SNA block.

We evaluate SNA on both neuromorphic (DVS128 Gesture) and static (CIFAR-10, CIFAR-100) benchmarks. On DVS128, our 2D-SNA model achieves 97.38% accuracy while delivering 2.6×

higher throughput (93 vs. 45 img/s) compared to the QKFormer baseline. On CIFAR-10 and CIFAR-100, SNA provides 1.7× throughput improvement with only a marginal accuracy trade-off. The 3D-SNA variant further reduces overall training time on event-based data by restricting attention to local spatio-temporal windows, processing less redundant information per sample. Memory analysis confirms that SNA scales far more favorably than full self-attention as token count increases, making deeper and higher-resolution spiking Transformer architectures feasible without memory explosion.

We also explore replacing the standard GEMM-based dot-product in the attention computation with a Hadamard (element-wise) product, inspired by spike-driven Transformer designs. This variant achieves comparable or slightly improved accuracy and reduces complexity from quadratic to linear in the neighborhood size, though it currently lacks optimized CUDA kernels and thus shows lower throughput in practice. These results point to a promising direction for future hardware-optimized implementations.

Our findings demonstrate that localized spatio-temporal attention offers a practical pathway toward scalable spiking Transformers for resource-constrained neuromorphic systems, bridging the gap between the expressiveness of attention-based models and the efficiency demands of real-time edge deployment.

[1] C. Zhou et al., QKFormer: Hierarchical spiking transformer using Q-K attention, NeurIPS, 2024.

[2] A. Hassani et al., Neighborhood attention transformer, CVPR, 2023.

[3] Z. Zhou et al., Spikformer: When spiking neural network meets transformer, ICLR, 2023.

[4] M. Yao et al., Spike-driven Transformer, NeurIPS, 2023.