Graduate School of Science and Technology Master's Thesis Abstract

Laboratory name (Supervisor)	Computing Architecture (Yasuhiko Nakashima (Professor))		
Student ID	2411433	Submission date	2026 / 7 / 11
Name	ZHONG JIAJUN		
Thesis title	Spikceiver: Compact Audio-Visual Spiking Transformer with Depthwise Splitting and Bottleneck Attention		

Abstract

Nowadays, intelligent sensing systems are becoming an important foundation for robotics, autonomous systems, smart cities, wearable devices, and edge artificial intelligence. In many real-world applications, a single sensing modality is often insufficient because visual signals may be affected by occlusion, illumination changes, or background clutter, while audio signals may provide complementary temporal and semantic information. Therefore, audio--visual learning has attracted increasing attention as a promising approach for robust machine perception. However, deploying audio--visual models on edge and neuromorphic platforms remains challenging because these systems must achieve high recognition accuracy under strict constraints on computation, memory, latency, and power consumption.

Transformer-based architectures have shown strong performance in multimodal learning because attention mechanisms can model long-range dependencies and flexible interactions between audio and visual tokens. Nevertheless, conventional attention-based audio--visual Transformers usually require dense pairwise token interactions. When visual images and audio spectrograms are both divided into long token sequences, the computational cost of multimodal attention increases rapidly with token length. This cost becomes a major obstacle for resource-constrained inference. Spiking neural networks (SNNs) provide a potential solution because they represent information using sparse binary spikes over discrete timesteps and are suitable for event-driven low-power computation. However, recent spiking multimodal Transformers still inherit the expensive fusion pattern of dense attention when long audio and visual token sequences interact directly. As a result, there remain important challenges in designing an audio--visual spiking Transformer that is both accurate and computationally efficient.

This thesis presents Spikceiver, a compact audio--visual spiking Transformer intended for efficient multimodal classification on resource-constrained platforms. The proposed architecture applies two main optimization techniques. First, Spiking Depthwise-Separable Splitting (SDSS) is introduced as a lightweight spike-domain tokenization front-end. It replaces standard convolutional patch splitting with depthwise-separable convolution to reduce the cost of generating visual and audio tokens. Second, Spiking Bottleneck Attention (SBA) is proposed for efficient multimodal fusion. Instead of computing direct all-to-all attention between audio and visual token sequences, SBA uses a small set of learnable bottleneck tokens to compress modality-specific spike representations before aggregation. This design reduces the dominant fusion complexity from O(N_VN_AD) to O((N_V+N_A)MD), where N_V and N_A are the visual and audio token lengths, D is the feature dimension, and M is the number of bottleneck tokens.

For evaluation, two audio--visual benchmarks are used. They are UrbanSound8K-AV and CIFAR10-AV(enhanced). The evaluation examines classification accuracy, computational cost, parameter size, and the overall accuracy--efficiency trade-off. Experimental results show that Spikceiver achieves 95.05¥% and 89.78¥% Top-1 accuracy on the two datasets with only 4.84M MACs and about 1.27M parameters. Compared with the Multimodal Bottleneck Transformer (MBT-4-256), Spikceiver improves accuracy while reducing both MACs and parameters. Compared with the Spiking Multimodal Transformer (SMMT-1-256), Spikceiver reduces computation by more than 800 times while maintaining competitive recognition performance. Ablation studies further demonstrate that SDSS improves tokenization efficiency, SBA provides the main accuracy--efficiency benefit in multimodal fusion, and learnable attention scaling is important for stable Softmax-free spiking attention. These results show that bottleneck-mediated spike-domain fusion is an effective design principle for compact and energy-aware audio--visual recognition.