EVGEN Workshop

Simple Transformer with Single Leaky Neuron for Event Vision

Himanshu Kumar, Aniket Konkar

Independent Researcher - India, The George Washington University - USA

WACV 2025 TUCSON, ARIZONA • FEB 28 - MAR 4

Introduction

Event-based vision has seen significant improvements through deep learning and spiking neural networks (SNNs). However, existing methods are complex and computationally expensive. We propose Simple Transformer, a novel approach that leverages:

- A **pretrained ResNet** backbone for feature extraction.
- A **Spiking Temporal Processor** using a Leaky Integrate-and-Fire (LIF) neuron.
- A **lightweight transformer** with multi-head attention to model long-range dependencies.

Our model achieves competitive accuracy with significantly lower computation costs.

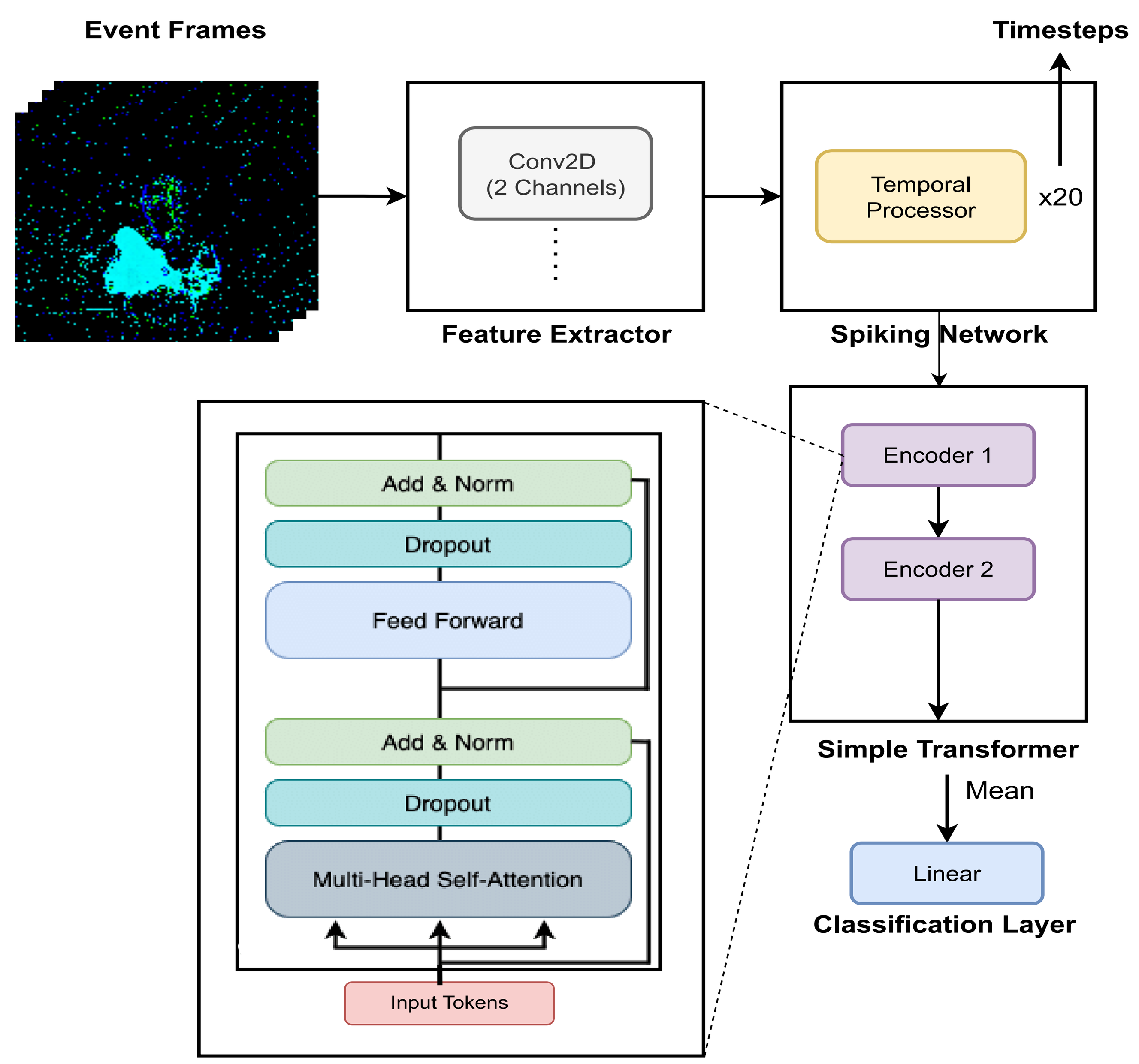
Key Contributions

- **Pretrained Spatial Feature Extractor:** Modified ResNet to process 2-channel event data.
- **Spiking Temporal Processor:** Uses a single PLIF node to process temporal dynamics.
- **Lightweight Transformer:** Multi-head attention captures long-range dependencies efficiently.
- **Superior Performance:** Outperforms many spiking and event-based methods on DVS Gesture, N-MNIST, and CIFAR10-DVS datasets.

Methodology

Architecture Overview

- **Input:** Event-based frames converted into a 20-timestep representation.
- **Feature Extraction:** Modified ResNet-18 extracts spatial embeddings.
- **Temporal Processing:** A single PLIF node models spiking behavior. Residual layers included for faster convergence.
- **Transformer Encoder:** Two encoder layers refine temporal embeddings.
- **Classification:** Pooled outputs passed through a fully connected layer.



Results

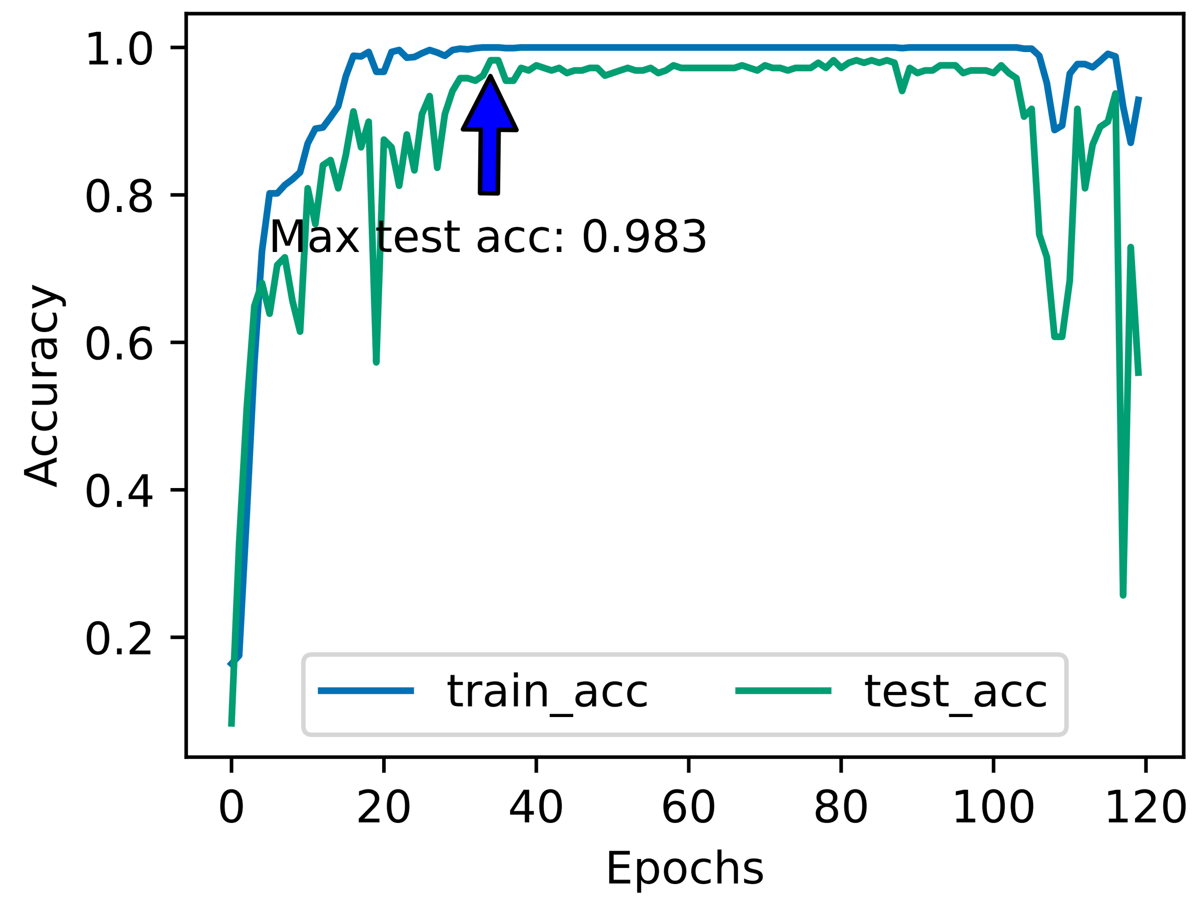
Experiment results on Event based datasets

Dataset	Results
DVS Gesture	98.3%
N-MNIST	99.3%
CIFAR10-DVS	75.9%

Computational Complexity

Methods	SOPs(G)	Energy(mJ)
Spiking ResNet-34	65.28	59.30
Spikformer	22.09	21.48
Spikingformer		13.68
Ours	1.82	8.372

Faster convergence with Residual spiking layers



Conclusion

- Simple Transformer effectively bridges the gap between deep learning and event vision by integrating pretrained vision models with spiking neural networks. It demonstrates that a minimalist approach to transformers can yield competitive results while maintaining computational efficiency.

Future work

- Scaling to Larger Datasets: Extend the model to more complex and diverse event-based datasets to further validate its robustness.
- Hybrid Models: Integrate self-supervised learning and multimodal transformers for improved generalization across different domains such as video processing, speech recognition, and biosignals.
- Advanced Regularization Techniques: Introduce sparsity constraints and adaptive spike-rate control mechanisms to enhance energy efficiency without compromising accuracy.
- Exploring Vision-Language Models: Investigate the potential of hybrid vision-language models, such as CLIP, in event-based tasks by integrating spiking mechanisms.