

# FPGA-Based Hardware AI Accelerator for Spike-Driven Transformers

## Background

Traditional Transformers are computationally expensive due to dense matrix multiplications and Softmax operations. Recent research in neuromorphic computing has introduced Spike-driven Self-Attention (SDSA) [1], which replaces these multiplications with sparse additions, achieving x87.2 lower energy costs than the vanilla self-attention, making them ideal for hardware acceleration.

This project aims to design and implement an FPGA accelerator that exploits this sparsity to achieve energy-efficient inference.

[1] [https://papers.neurips.cc/paper\\_files/paper/2023/file/ca0f5358dbadda74b3049711887e9ead-Paper-Conference.pdf](https://papers.neurips.cc/paper_files/paper/2023/file/ca0f5358dbadda74b3049711887e9ead-Paper-Conference.pdf)

## Project Objectives

In this project, students will build a digital SNN architecture on an FPGA that specifically implements the building blocks of a Spiking Transformer.

Particularly, students will:

1. Implement spiking neuron models on FPGA: use fixed-point arithmetic to implement the *snntorch*\* neuron models, ensuring they match the *python* training environment.
2. Implement spike-driven attention: design a digital module that uses the spiking neurons and performs "mask-and-add" operations instead of the standard dot-product attention.
3. Quantization & sparsity analysis: study the trade-offs between bit-width and accuracy.
4. Power & timing analysis (FPGA): estimate dynamic power using FPGA-specific tools (e.g., Vivado power analyzer).
  - a. Optimize pipeline latency to ensure the transformer can process temporal sequences in real-time.

\**snntorch* is a python library which makes training spiking neural network easier.

**Prerequisites Courses:** Electronic Circuits and Machine Learning (recommended, can be taken in parallel).

**Contact:** Jeries Saleh-Naser (jeries.saleh@campus.technion.ac.il)

