HyAtten: Hybrid Photonic-digital Architecture for Accelerating Attention Mechanism

Huize Li, Dan Chen, and Tulika Mitra

School of Computing, National University of Singapore

Transformer and Self-Attention

Transformer-based large models achieve state-of-the-art performance on various natural language processing and computer vision tasks. The core operation of Transformers is the self-attention mechanism, which calculates pairwise correlations between input tokens to enhance inference accuracy. The attention function is computed between these input vectors as follows:

$$Attention(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = softmax(\mathbf{Q}\mathbf{K}^{\mathsf{T}}/\sqrt{d_k})\mathbf{V}, \quad (1)$$

where d_k is **Q** and **K**'s dimension. An intermediate score matrix **S** is obtained with $S = Q \times K^{\mathsf{T}}$.

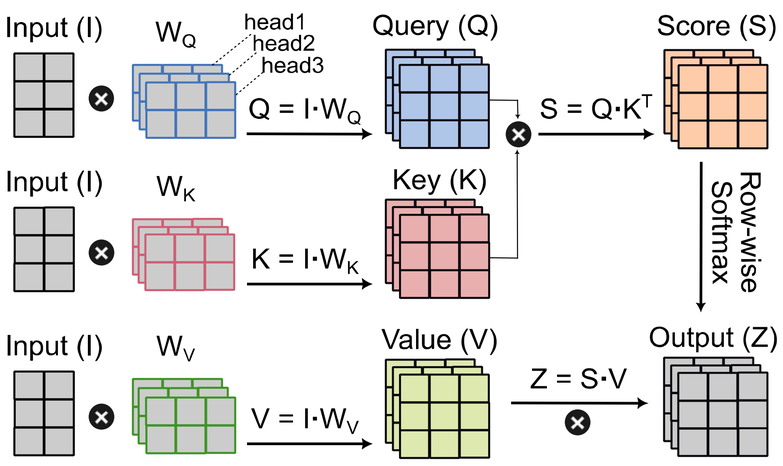

Fig. 1: Multi-head attention

Optical Computing Basics

First, the wavelength-division multiplexing (WDM) technique encodes each input pairs (x_i, y_i) in the same wavelength λ_i. The WDM light signals are then sent through the two arms of $50:50$ directional coupler (DC) with a $-90°$ phase shifter (PS). Consequently, the two output signals become orthogonal in the complex plane. This setup allows each input pair (x_i, y_i) with the same wavelength λ_i to interfere in parallel, while different wavelengths do not interfere.

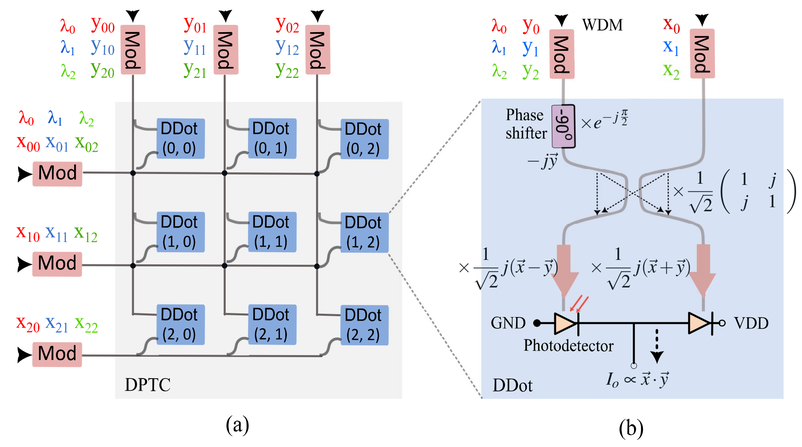

Fig. 2: (a) Dynamically-operated photonic tensor core (DPTC), (b) Dynamically-operated dot-product (DDot) unit

Motivations

- Problem: Signal conversion costs remain the primary bottleneck for emerging photonic systems.
- Observation#1: Utilizing low-resolution signal converters can significantly reduce latency and area overhead, but it may also lead to substantial model accuracy loss.
- Observation#2: Only a small fraction of analog signals require high-resolution ADCs.
- Our goal: Building on Observation#2, analog signals in photonic-based Transformer accelerators can be categorized into two groups: low-resolution signals ($\leq$4-bit) and high-resolution signals ($>$4-bit). Low-resolution signals can be efficiently processed using 4-bit ADCs with lower latency and area overhead. As noted in Observation#1, however, high-resolution signals require careful handling to avoid accuracy loss. Rather than introducing high-resolution ADCs, we employ digital circuits to process these signals. Since high-resolution signals constitute less than 15% of the total, the computational overhead imposed on the digital circuits remains minimal.

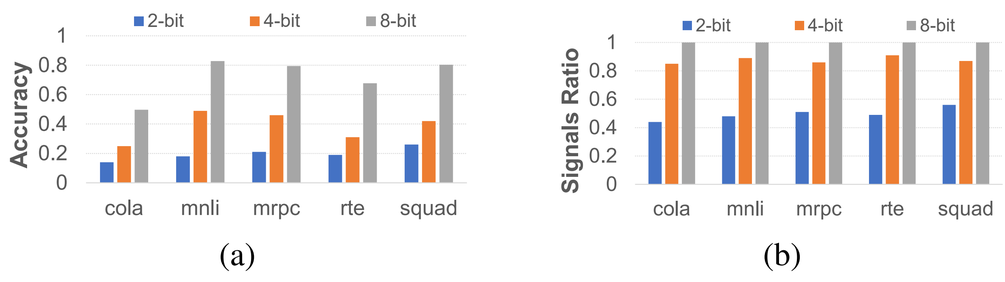

Fig. 3: (a) The model accuracy when employing different ADC resolutions, and (b) the proportion of signals that remain

HyAtten Architecture

HyAtten (ⓐ) consists of multiple Tiles, each of which includes a photonic die (ⓑ) and a digital die (ⓒ). The photonic die includes a shared SRAM, a shared photonic digital-to-analog converter (PDAC), and a photonic processing element (PE) (ⓓ). The photonic PE handles the core computations in the attention mechanism, specifically the GEMM operations for $Q \times K^{\mathsf{T}}$ and $S \times V$. The digital die consists of a shared SRAM, a softmax unit (ⓖ), and a digital PE (ⓕ). The digital PE is responsible for completing GEMM operations that cannot be processed by the photonic die.

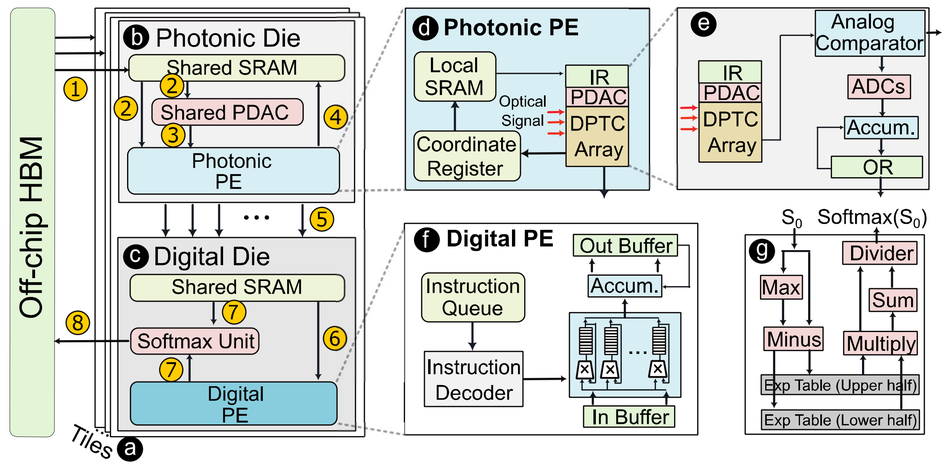

Fig. 4: HyAtten Architecture

HyAtten Dataflow

During cycle_0, The first sub-matrix of matrix A is sent to the shared PDAC for signals conversion and broadcast to all Tiles. The resulting photonic currents produced by the DPTC arrays are processed by the ADCs to generate the outputs of the sub-matrix multiplication. In the following two cycles, cycle_1 and cycle_2, the second and third sub-matrices of matrix A are sequentially sent to the shared PDAC, where they are converted into photonic signals and broadcast to all Tiles.

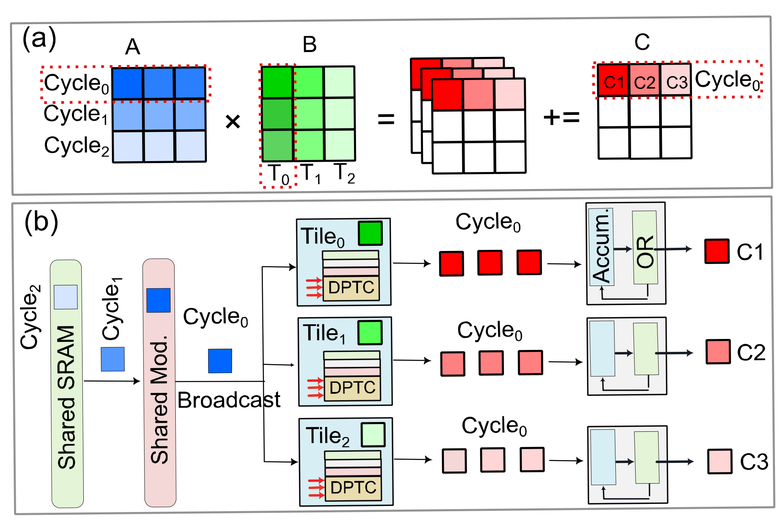

Fig. 5: GEMM operations on multiple photonic Tiles

Evaluation Results

Comparison with Lightening-Transformer baseline. HyAtten delivers a 9.8× speedups and 2.2× energy reduction per unit area. This improvement can be attributed to two key factors. First, HyAtten replaces each high-resolution ADC with multiple low-resolution ADCs, significantly reducing signal conversion latency without increasing chip area. Second, HyAtten employs a digital die to handle the 15% of signals that require high resolution ADCs, thereby avoiding excessive signal conversion overhead while incurring only a small area penalty.

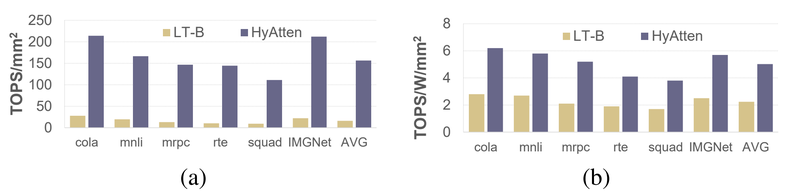

Fig. 6: (a) Performance per unit area, and (b) energy efficiency per unit area

Comparison with CPU and GPU platforms. we compare HyAtten against, a single Nvidia A100 GPU and an Intel Core i7-9750H CPU. It achieves over 100× speedup and over 50× energy efficiency per unit area relative to the A100 GPU, largely due to the high processing speed enabled by photonic computing.

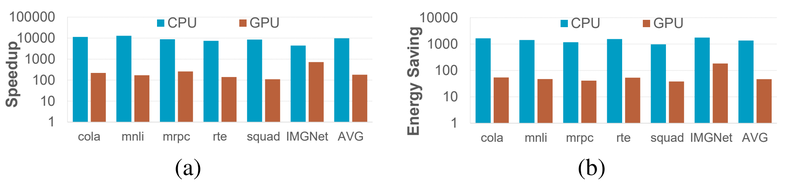

Fig. 7: (a) HyAtten speedups compared to CPU and GPU, and (b) HyAtten energy saving compared to CPU and GPU