Evaluation of hardware-induced errors in optically computing transformer models

Chung-Chih Lin, Nathaniel Y. Na, Erik Chen, and Neil Na*
Artilux Inc., Zhubei City, Hsinchu County, Taiwan 30288
* neil@artiluxtech.com

ABSTRACT

Light-based analog computing schemes for deep-learning applications have recently received wide attention. However, in the literature, there are only a few works studying whether transformer models, which are the engines behind all large-language models (LLMs), can be computed optically. In this paper, to assess the impact of the hardware-induced errors in optically computing transformer models, we employ a vision transformer (ViT) model with different model sizes as benchmarks. The ViT model, known for its self-attention mechanism and high computational complexity, provides an ideal testbed for evaluating the performance of optically computing transformer models under realistic AI workloads. By integrating the optical computing framework into the computational pipeline of the ViT model, we quantify how errors propagate through the model by measuring their effects on accuracy, providing critical insights into the viability of optically computing transformer models. Our findings highlight the validity of optically computing transformer models and the guidelines for training such models with light-based analog computing machines.

Keywords: analog computing, optical/photonic computing, transformer models, vision transformer, quantization

1. INTRODUCTION

The rise of transformer models in deep learning has transformed research fields such as natural language processing (NLP) and computer vision (CV), to name a few, and pushed the boundaries of what artificial intelligence (AI) can achieve. The advancements, fueled by increasingly complex models with vast parameter spaces, rely heavily on the capabilities of digital computing architectures and systems such as central processing units (CPUs), graphics processing units (GPUs), and application-specific integrated circuits (ASICs). While these hardware have delivered remarkable computational performance and enabled rapid development and deployment of sophisticated AI technologies, as the model sizes and so the computational demands continue to grow, issues of digital electronics, such as energy consumption, memory-bandwidth bottleneck, and scalability challenge, have become more and more stringent, prompting new explorations into alternative or complementary technologies.

One of such technologies is optical (or photonic) computing, a paradigm shift that leverages the physical properties of light-based signals to perform analog calculations, which is distinct from the discrete, binary nature of digital computers [1]. Instead of replacing digital architectures and systems, it is generally believed that optical computing offers the potential to co-work alongside CPUs, GPUs, and ASICs to enhance the overall performance and efficiency of certain operations, e.g., multiply-and-accumulate (MAC), which is used in vector-vector, vector-matrix, and matrix-matrix calculations, the main operations in deep-learning models including transformer models. Under such a circumstance, the properties of digital-to-analog converter (DAC) and analog-to-digital converter (ADC) are expected to play major roles, in addition to the properties of various optoelectronic components used in optical computing [2].

In this paper, we study the hardware-induced errors in optical computing, with a special focus on their impacts over transformer models in deep learning. By introducing the quantization done by DAC and ADC (modeled as INT8 and INT4 formats), as well as the nonuniformity arising from various optical components (modeled as Gaussian noises), the resultant hardware-induced errors are evaluated, demonstrating that transformer models computed optically are capable of robust inference when the training is done appropriately.

AI and Optical Data Sciences VII, edited by Masaya Notomi, Tingyi Zhou,
Proc. of SPIE Vol. 13906, 139060P · 2026 Published by SPIE
0277-786X · doi: 10.1117/12.3077269

2. METHODOLOGY

2.1 Optical Computing Architecture

In the literature, various types of optical computing have been demonstrated. Generally, they can be categorized into incoherent scheme [3-5], in which only the intensity degree of freedom of light is used, and coherent scheme [6-9], in which the amplitude and phase degrees of freedom are used. For incoherent scheme, the setup typically consists of optical emitters coupled to optical modulators coupled to optical detectors, representing activation (emitted light intensity), weight (modulated light intensity), and output (detected light intensity) of an artificial neuron. For coherent schemes, the setup typically consists of optical emitters coupled to optical interferometers coupled to optical detectors, representing activation (emitted light amplitude or phase), weight (modulated amplitude or phase), and output (the interfered and then detected light amplitude or phase) of an artificial neuron. See Figure 1 for the illustration of the two schemes. To simplify the modeling, we include one (i.e., to emulate output) or two (i.e., to emulate activation and weight) independent noises to represent the nonuniformity of various optical components, and two digitization steps (one before MAC and the other after MAC) to represent the quantization of DAC and ADC. Furthermore, the schemes of post-training quantization (PTQ) and quantization-aware training (QAT) are both considered in the following.

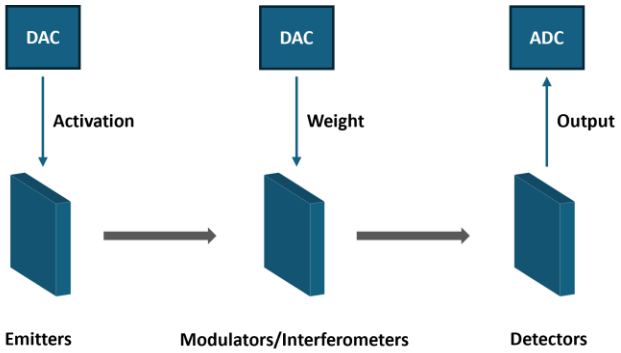

Figure 1. A typical incoherent/coherent optical computing scheme, consisting of optical emitters coupled to optical modulators/interferometers coupled to optical detectors, representing activation, weight, and output, respectively. In such a scheme, the DAC and ADC introduce errors due to quantization, while the emitters, modulators, and detectors introduce errors due to the nonuniformity of the optical components.

2.2 Vision Transformer Model

To assess the impact of the hardware-induced errors in optically computing transformer models, we employ a vision transformer (ViT) model [10, 11] with different model sizes as benchmarks. ViTs, known for their self-attention mechanism and high computational complexity, provide an ideal testbed for evaluating the performance of optically computing transformer models under realistic AI workloads. In this evaluation, the mini-ImageNet1K dataset is utilized, comprising 100 distinct classes with 120 images per class allocated for training and 50 images per class reserved for testing. Such a dataset, while being compact, retains sufficient diversity to probe the robustness of the ViT model across all classification categories. By integrating the optical computing framework into the computational pipeline of the ViT model, we quantify how errors propagate through the model by measuring their effects on accuracy, providing critical insights into the practical viability of optically computing transformer models in deep learning.

In the optical computing architecture, errors are introduced primarily from two sources: the quantization due to DAC and ADC, and the nonuniformity due to various optical components. For the former case, we shall assume INT8 and INT4 formats and evaluate their impacts, reflecting the digitization process of DAC and ADC. Note that the effect of DAC and ADC digitization process is emulated by discretizing the activations, weights, and outputs of the ViT model into finite precision levels, mirroring the resolution constraints of the converter hardware. For the latter case, we shall assume Gaussian noises and evaluate their impacts, reflecting the spatial variations of various optical components. Note that the Gaussian noises are independently applied to outputs, or activations and weights in a multiplicative fashion. The standard deviation σ may physically represent the degree of variability in the intensity, amplitude, or phase distribution of the light emitted, modulated, interfered, or detected.

To better emulate the losses due to quantization and nonuniformity at the hardware side, the original LLM.INT8() algorithm for PTQ [12] is modified and applied to all the matrix multiplications in the ViT model, including

Query/Key/Value matrices in the self-attention layers and the multilayer perceptron in the feed-forward layers. The algorithm treats activation and weight values exceeding the threshold (set to 6) as outliers, which typically constitute only a small percentage of the total number of all possible values. These outliers are processed separately with higher precision to prevent information loss, as illustrated in Figure 2. In contrast, activation and weight values below the threshold are quantized to INT8 format, emulating the 8-bit hardware requirement of the DAC. This way, the matrix multiplications of inlier/outlier values (in large/small quantity) are processed using analog/digital computing with lower/higher precision by default. Following a matrix multiplication, an additional INT8 quantization is performed on the INT32 matrix multiplication output to emulate the 8-bit hardware requirements of the ADC. A tunable range hyperparameter is introduced during the quantization in order to best utilize the full precision of the INT8 format and avoid issues such as under-utilization or saturation of the INT8 format. Note that since in the case of PTQ, a "twin" digital computer must be used in conjunction with an optical computer to further process the outliers, QAT may be a more efficient approach compared to PTQ.

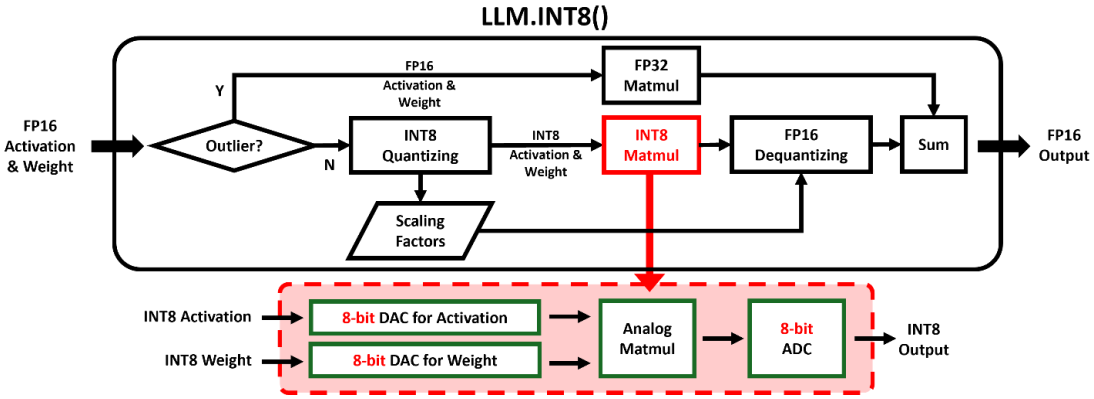

Figure 2. LLM.INT8() algorithm schematic. During the quantization process, values below the threshold undergo an affine transformation to be represented in the INT8 format. The scaling factors are extracted from the INT8 quantization step, and then stored for later utilization in the FP16 dequantization step. The red region represents the part assigned to an optical computer.

Through the procedures above to introduce the hardware-induced errors, a comprehensive evaluation of optically computing transformer models is enabled, capturing the core issues of light-based analog computing and offering a framework to optimize the quantization used in optical computing for transformer-based deep-learning applications.

3. RESULTS

When PTQ is applied, for 8-bit quantization, the evaluation results are illustrated in Figure 3a, where σ of the output noise is varied from 0% to 10%. The ViT-Base model demonstrates superior resilience to noise, with its classification accuracy on the mini-ImageNet1K dataset degrading by only 1.5%, from the baseline of 81.7% to 80.2%, when σ of the output noise reaches 10%. In contrast, the ViT-Tiny model exhibits inferior resilience to noise, with classification accuracy degrading by ~8% under the same conditions. This disparity suggests that larger models with more parameters, such as ViT-Base, possess enhanced robustness to noise-induced errors, implying that the increased number of neurons contributes to greater stability. To further dissect this behavior, the effects of activation noise and weight noise are assessed and illustrated in Figure 3b. It can be observed that similar trends of classification accuracies manifest according to each noise, suggesting that the two noises influence the ViT model in a seemingly independent way. These findings showcase the role of the larger model size in mitigating errors and the limit of noise susceptibility in optically computing transformer models.

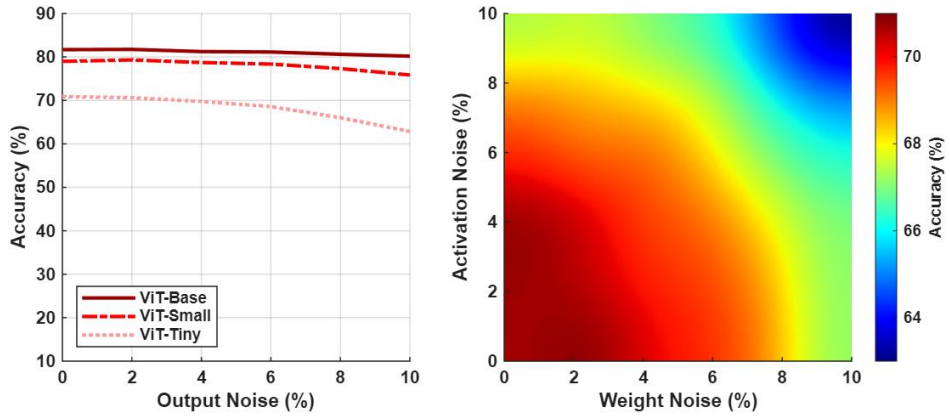

Figure 3. (a) The 8-bit ViT model classification accuracy plotted as a function of the output noise (as σ), given ViT-Base, ViT-Small, and ViT-Tiny models. (b) The 8-bit ViT-Tiny model classification accuracy plotted as a function of the activation noise (as the 1st σ) and weight noise (as the 2nd σ).

The effect of 4-bit quantization is simulated by directly reducing the precision from INT8 to INT4 using affine quantization, and is illustrated in Figure 4 for the case of ViT-Tiny model. The results of classification accuracy at zero output noise shows a noticeable degradation from 70.8% (INT8) to 43.6% (INT4), attributed to the cumulative loss induced by lower-precision quantization. To mitigate this issue, the quantization threshold is adjusted from 6 to 1.5, yielding improved accuracy and greater resilience to noise. This adjustment demonstrates that a lower threshold enhances the performance of the ViT model, likely by preserving more granular information during the DAC and ADC digitization process. These findings showcase the trade-off between the classification accuracy and the utilization ratio (i.e., the ratio between inliers and the sum of inliers and outliers), offering insights into how to optimize the quantization used in optical computing for transformer-based deep-learning applications.

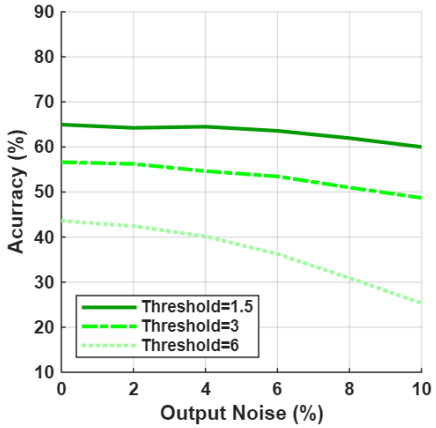

Figure 4. The 4-bit ViT-Tiny model classification accuracy plotted as a function of the output noise (as σ), given different outlier thresholds.

Finally, QAT is also applied through fine-tuning to evaluate the impact of hardware-induced errors, and so enhancing the performance of the ViT model by further mitigating the information loss from the quantization. The original ViT-Tiny model is fine-tuned on the mini-ImageNet1K dataset, where the reduced number of classes (100, compared to 1000 in the full mini-ImageNet1K) simplifies the classification task, facilitating more effective learning. As illustrated in Figure 5, both 8-bit and 4-bit quantization schemes achieve improved classification accuracy, reaching up to ~87% at zero output noise and ~85% with σ of the output noise equal to 10%, showing less than 2% drop even in the 4-bit ViT-Tiny model. In summary, QAT effectively reduces the quantization loss by aligning the software training process with the hardware quantization characteristics, compensating for the lower precision used in optical computing. Additionally, the fine-tuned ViT model with QAT exhibits excellent noise tolerance, suggesting that the cumulative loss induced by lower-precision quantization is significantly reduced. These findings highlight the important role of QAT in optimizing the optical computing architecture.

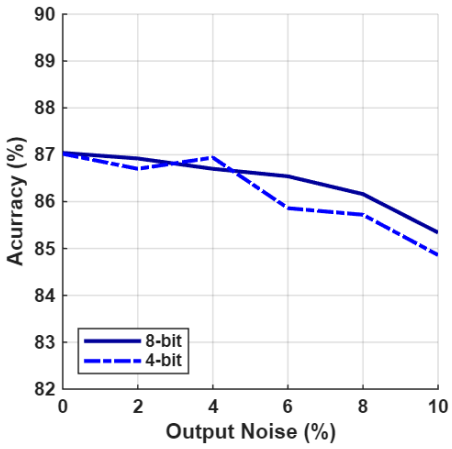

Figure 5. The ViT-Tiny model classification accuracy plotted as a function of output noise (as σ), given INT4 and INT8 formats.

4. CONCLUSION

In this work, we evaluate the impacts of hardware-induced errors in optically computing transformer models, assuming the light-based analog computer is used to complement digital computers such as CPUs, GPUs, and ASICs. The potential of leveraging light-based signals to enhance the power efficiency while maintaining the model performance is examined through the issues of DAC and ADC quantization and optical component nonuniformity. With the framework of the ViT model and the mini-ImageNet1K dataset, and the introduction of output noise stemming from the nonuniformity of various optical components and modeled by Gaussian distribution, our analysis demonstrates that, in the case of PTQ, larger models exhibit greater robustness, with accuracy dropping only by 1.5% for 8-bit ViT-Base model compared to 8% for 8-bit ViT-Tiny model. Further assessments by weight noise and activation noise are also performed. In particular, the lower-precision 4-bit quantization showcases the trade-off between the classification accuracy and the utilization ratio. Finally, the application of QAT through fine-tuning elevates the model performance significantly, achieving classification accuracies up to 87% for both 8-bit and 4-bit quantization at zero noise as well as their tolerance to higher noise levels.

As a whole, all findings above demonstrate the viability of optically computing transformer models with an analog machine, paving the path to a more efficient AI technology [13]. It should be pointed out that more work should be done, such as refining the noise models for the nonuniformity of optical components, calibrating the quantization conditions for specific workloads, and extending the evaluations to larger datasets and diverse transformer variants.

REFERENCES

[1] H. J. Caulfield and S. Dolev, “Why future supercomputing requires optics,” Nat. Photon. **4**, 261 (2010).
[2] A. Tsakyridis et al., “Photonic neural networks and optics-informed deep learning fundamentals,” APL Photon. **9**, 011102 (2024).
[3] M. G. Anderson et al., “Optical Transfomers,” Transactions on Machine Learning Research (TMLR) **3**, 1 (2024).
[4] J. Feldmann et al., “Parallel convolution processing using an integrated photonic tensor core,” Nature **589**, 52 (2021).
[5] C. Mesaritakis, V. Papataxiarhis, and D. Syvridis, “Micro ring resonators as building blocks for an all-optical high-speed reservoir-computing bit-pattern-recognition system,” J. Opt. Soc. Am. B **30**, 3048 (2013).
[6] H. Zhu et al., “Lightening-Transformer: A Dynamically-operated Optically-interconnected Photonic Transformer Accelerator,” IEEE International Symposium on High-Performance Computer Architecture (HPCA), (2024).
[7] C. Feng et al., “A compact butterfly-style silicon photonic–electronic neural chip for hardware-efficient deep learning,” ACS Photon. **9**, 3906 (2022).
[8] W. Bogaerts et al., “Programmable photonic circuits.” Nature **586**, 207 (2020).
[9] Y. Shen et al., “Deep learning with coherent nanophotonic circuits,” Nature Photon. **11**, 441 (2017).
[10] A. Dosovitskiy et al., “An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale,” arXiv preprint, arXiv:2010.11929v2, (2020).

[11] H. Touvron et al., “Training data-efficient image transformers & distillation through attention,” arXiv preprint, arXiv:2012.12877v2, (2020).
[12] T. Dettmers et al., “Llm.int8(): 8-bit matrix multiplication for transformers at scale”, Proceedings of the 36th International Conference on Neural Information Processing Systems (NeurIPS), (2022).
[13] N. Na et al., “Implementation of transformer-based LLMs with large-scale optoelectronic neurons on a CMOS image sensor platform,” arXiv preprint, arXiv:2511.04136v1, (2025).