Multi-modal Fusion and Transformer Architecture

Alan Devkota, Student ID: 2215320
Department of Electrical and Computer Engineering
University of Houston
Houston, TX, USA
adevkota2@uh.edu

***Abstract*—Hardware accelerators can perform complex computations more efficiently while improving the performance and reducing the power consumption in neural networks. In transformer networks, one of the most time-consuming operations to perform is matrix multiplication operations during attention calculation. When we deal with multimodality, the cross-attention calculations become a bottleneck requiring exchange of input tokens between different modalities. Creating specialized hardware accelerators is necessary to provide fast and energy-efficient computations in computationally demanding multimodal transformer network applications. In this paper, we develop a two-modality object detection transformer model that integrates the information from RGB and thermal IR modalities together to enhance the prediction as well as address the challenges posed by missing modalities. Thereby, we extend the object detection to multi-modality and implement a novel hardware accelerator to handle the computational speed and memory usage issues when applying transformers in multimodality contexts.**

***Index Terms*—Transformer, DETR, Object Detection, FPGA, Attention, Cross-Attention.**

I. Introduction

Transformers gained popularity due to their revolutionary performance in language processing tasks through GPT-3 and BERT models [1]. More recently, they were extended to perform image processing and object detection task and were shown to deliver superior results compared to conventional CNNs. Multi-modal object detection involves identifying and pinpointing objects in a scene using data from various spectral bands or wavelengths. In contrast to standard RGB imaging, multi-modal data captures a broader range of the electromagnetic spectrum, including non-visible wavelengths like infrared, and other modalities data like speech, texts and so on. Information about these extended modalities enhances the precision of machine learning techniques across diverse applications.

Multi-modal object detection utilizes data from multiple modalities, with sensors often incorporating infrared bands such as near-infrared (NIR) and shortwave infrared (SWIR) as well as audio and text information alongside the visible spectrum. This capability allows for the detection of subtle environmental variations, making it particularly valuable in fields like precision agriculture for monitoring crop health, environmental monitoring for land cover changes, biomedical application and military applications for reconnaissance and target detection.

Hardware accelerators are specialized hardware that can perform complex computations more efficiently while improving the performance and reducing the power consumption in neural networks. Numerous studies have explored hardware acceleration for transformers [2]–[20]. In transformer networks, one of the most time-consuming operations is matrix multiplication operations while generating Query, Keys, and Value matrices, as well as the attention calculation. Also, the linear layer operations involve plenty of multiplications and additions. However, when we deal with multimodality in this proposed research, the cross-attention calculations become a bottleneck requiring the exchange of input tokens between different modalities that have large datasets with high numbers of parameters. Creating specialized hardware accelerators that can effectively perform computationally demanding operations intrinsic to multimodal transformer networks becomes an essential step to provide the just-in-time and energy-efficient computing.

Thus, in this work, we focus on extending the DETR transformer model to process different modalities. Our focus is to develop a two modality transformer model at first and extend it into multi-modality scenario that integrates the information from different modalities together to enhance the prediction as well as address the challenges posed by missing modalities. Our modified DETR transformer encoders harness the vast potential of heterogeneous data via early fusion of modalities features (RGB and IR) unlike the original DETR transformer model. Moreover, we also propose to design a novel hardware accelerator that handles the computational speed and memory usage issues when applying transformers in multimodality contexts.

II. Background and Motivation

A. Object Detection Transformer

The article produced by Facebook AI regarding Object Detection using Transformers [1] introduces a paradigm-changing methodology for object detection by leveraging transformer architectures, similar to those used in sequence prediction [21]. Departing from the traditional two-stage approach, this model streamlines the entire process into a single end-to-end framework. The core innovation lies in the direct prediction of object detections without the need for explicit region proposals and extensive feature extraction, differing from methods relying on large sets of proposals [22]–[24], anchors [25], or window centers [21], [26].

The DETR architecture is shown in Figure 1. In DETR, the image is divided into a fixed number of regions, and each

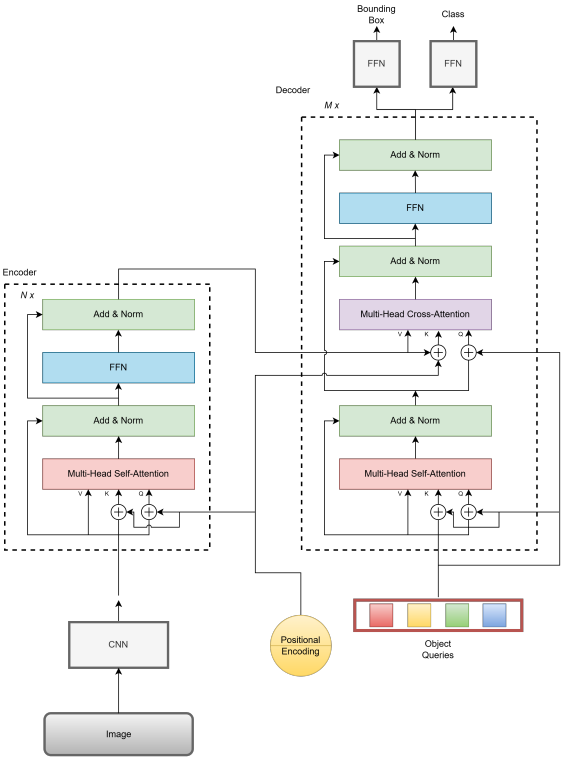

Fig. 1. DETR Transformer Architecture.

region is treated as a separate "object query." The transformer processes these queries in parallel, attending to the entire image simultaneously, a method contrasting with learnable NMS methods and relation networks that explicitly model relations between different predictions with attention [27]. This approach eliminates the need for separate region proposals and feature extraction steps. The self-attention mechanism of the transformer is instrumental in processing images divided into a set of learnable object queries [28]. The capacity of the transformer to capture long-range dependencies is harnessed for understanding global context, enabling a more comprehensive perception of the entire scene.

A significant advancement is the incorporation of a bipartite matching algorithm during training [29]. This mechanism facilitates the direct association of predicted queries with ground truth objects, simplifying the training process and eliminating the need for post-processing steps. The end result is a model capable of simultaneously predicting object classes and bounding boxes in a single pass [30].

Evaluation on benchmark datasets like COCO [31] underscores the model's competitive performance compared to conventional two-stage detection approaches, such as Faster R-CNN [24]. The end-to-end design, coupled with the inherent strengths of transformers, suggests a promising direction for the evolution of object detection methods. Beyond achieving state-of-the-art results, the paper contributes to the broader exploration of transformer architectures in visual tasks, pushing the boundaries of what is achievable in object detection and paving the way for more efficient and effective computer vision systems.

The DETR architecture is characterized by its simplicity and is composed of three main components: a CNN backbone for feature extraction [32], an encoder-decoder transformer, and a straightforward feed-forward network (FFN) for the final detection prediction [33]. Unlike many contemporary detectors, DETR can be implemented in various deep learning frameworks with minimal lines of code [34]. The backbone starts with an initial image, extracting features through a conventional CNN to produce a lower-resolution activation map.

Positional encodings are added to account for the transformer's permutation invariance [28]. Then, after reducing the channel dimension of the activation map using a 1x1 convolution, and generating a new feature map [35], the transformer encoder is applied. This map is then treated as a sequence, and each encoder layer employs a multi-head self-attention module and a feed-forward network. The transformer decoder follows a standard architecture but decodes N objects in parallel at each layer, in contrast to the autoregressive model used by Vaswani et al [21]. The N object queries, serving as positional encodings, are transformed into output embeddings and independently decoded into box coordinates and class labels.

The final prediction is computed by a 3-layer perceptron with ReLU activation and a linear projection layer. This predicts the normalized box coordinates and class labels using a softmax function. To handle a fixed-size set of N bounding boxes, an additional special class label Φ is introduced to represent cases where no object is detected within a slot [33]. Auxiliary decoding losses, including prediction FFNs and Hungarian loss, are employed during training to assist the model in outputting the correct number of objects for each class [36]. These losses are applied after each decoder layer, and shared layer normalization is used to normalize inputs to the prediction FFNs from different decoder layers.

The architecture of the transformers has an encoder-decoder structure. The encoder part maps the input sequence to a continuous form and sends it to the decoder, which generates an output sequence. The input image is divided into a sequence of patches, which are equivalent to the tokens in the language models. For each input, three matrices are derived: Query(Q), Key(K), and value(V), where Query and Key have dimensions d_k and the values have dimension d_v. Using These matrices are used to calculate the attention based on the formula:

$$Attention(Q, K, V) = Softmax(\frac{QK^T}{\sqrt{d_k}})V \quad (1)$$

Detection Transformers are introduced as a novel object detection system built on transformer architecture and bipartite matching loss for direct set prediction. In evaluations on the challenging COCO dataset, DETR achieves results comparable to an optimized Faster R-CNN baseline[13]. Its implementation is straightforward, and its flexible architecture easily

extends to panoptic segmentation, delivering competitive performance [31]. Notably, DETR outperforms Faster R-CNN on large objects, attributed to its ability to process global information through self-attention [31].

While DETR presents promising results, it also introduces new challenges, particularly in training, optimization, and performance on small objects. The paper acknowledges that addressing these challenges might require further work, drawing parallels with the evolution of current detectors that needed several years of refinement to overcome similar issues. The authors anticipate future research efforts to successfully tackle these challenges and further enhance the capabilities of DETR [31].

The DETR model often operate in a unimodal context and are not inherently equipped to handle the complexities of multimodal data. The attention mechanism at the core of transformer models is particularly beneficial for integrating data from different modalities. It computes relevance scores across all parts of the input sequences, enabling the model to focus on the most pertinent elements of data at each step of processing. In the context of multimodal data fusion, this means the model can attend to and emphasize the important features from one modality (like an EEG reading) while correlating them with pertinent aspects of another (such as motion sensor data), regardless of their differing natures or sampling rates. We propose a novel cross-modality attention transformer network (CMAT) that utilizes a multimodal fusion transformer approach for multi-modality context. By introducing mechanisms such as Pairwise Key-Values exchange and Cross-modality Attention, CMAT is engineered to dynamically allocate attention across different data sources, thus ensuring that relevant information is synergized into a comprehensive representation that enhances predictive performance.

While integrating multi-modal data presents challenges such as increased data complexity and computational demands, the potential benefits, including improved accuracy and richer information, make multi-modal object detection a valuable and continually advancing field. This is especially true in remote sensing applications where satellites and aircraft capture multi-modal data for Earth observation and analysis.

Building upon our previously mentioned accelerator architecture, our aim is to delve deeper into multiple acceleration techniques for multimodal transformers. This exploration seeks to optimize processing speed for real-time personalized monitoring, all the while reducing mobile device energy consumption, thereby significantly extending battery life.

III. Multi-modality Transformer

The focus of this research is to develop a transformer model that integrates the information from different modalities together to enhance the prediction as well as address the challenges posed by missing modalities. We use similar concept like retrieval from database by using query, keys, and value utilizing attention calculation in transformers. We use query (a query we wish to run on a database) from one modality and keys (the keys to search on in the database) and values (values corresponding to each key in the database) from other modalities. Cross-attention in transformer encoder is used to gain context from another modality/ input type as a method of TokenFusion in channels. This is accomplished by pairwise exchange of keys and values from different modalities. For example, to gain context from text for object detection, we simply extract the queries matrix from text modality, and keys and values matrix from the RGB and IR modalities. Moreover, self-attention blocks at the end of our model architecture would allow the model to further process the combined representations as well as enable the model to understand the dependencies between different parts of the input from different modalities.

A. Two Modality Case

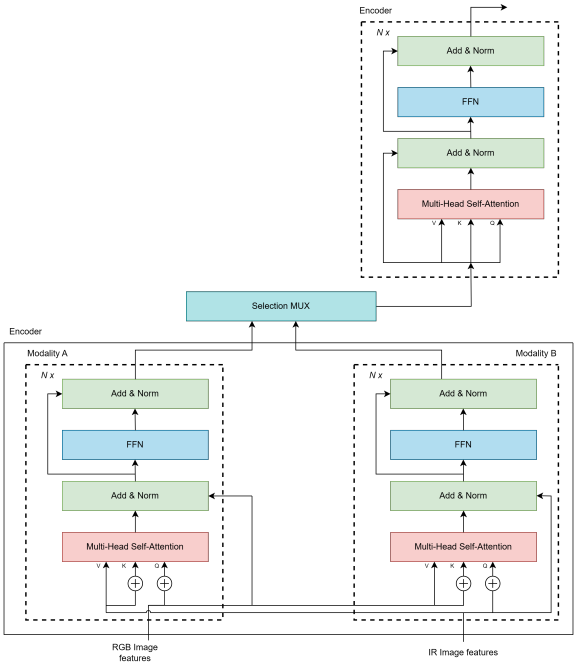

Fig. 2. Two-modality Transformer Architecture.

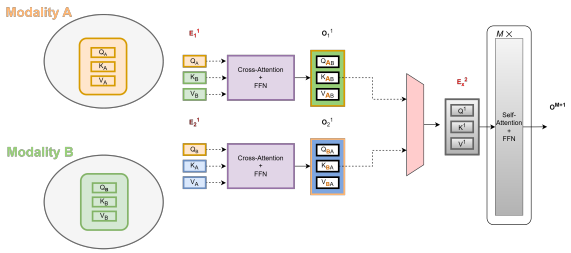

Fig. 3. Two Modality Cross-Modality Attention Transformer Network.

We will start with two modality case for simplicity as shown in Figure 2 and Figure 3. Our modified DETR transformer encoders for two modality case extract the features from RGB and IR modality via two parallel ResNet50 backbone and harness the vast potential of heterogeneous data via channel

fusion of modalities features (RGB and IR) using two parallel cross-attention encoders unlike the original DETR transformer model. Afterward, the model performs object detection by using a DETR decoder with object queries.

When implementing dual-modality object detection approach, both modalities assume a distinct role which provides a more holistic understanding of the object(s) of focus. The process is subdivided into three main stages, a backbone of CNN to extract the features from both RGB and Thermal IR images, a transformer architecture that consists of an encoder and decoder to learn the contextual information between the embedded features, and two classifier units to predict class and bounding box for each object detected inside an image. Here we are using ResNet50 to extract features of both RGB and thermal images and computing cross-attention between the tokens of RGB and IR modalities to get learned feature representations.

The initial step involves tokenizing the RGB and IR images by breaking down the visual information into manageable units that can be analyzed by the model. The feature representations are then extracted using a pre-trained Resnet50 model. The features extracted are passed to the two parallel DETR encoder and multi-headed cross-attention mechanism is applied, allowing the model to focus on relevant parts of the image as well as on the similarity or closeness of RGB and Thermal pairs. Following this attention phase in the encoder, a fixed set of learnable encodings called object queries would attend with the keys and values output from the encoder via cross-attention. Therefore, object queries would detect corresponding objects in the output features from the encoder. A multi-layer perceptron (MLP) block is used to process these attention-guided features. The MLP block is a sophisticated component designed to process the features that have been refined through the cross-attention mechanism. The MLP block acts as a computational mechanism that applies non-linear transformations to the extracted features, further accentuating the similar and distinct patterns from the image dataset to provide precise object detection.

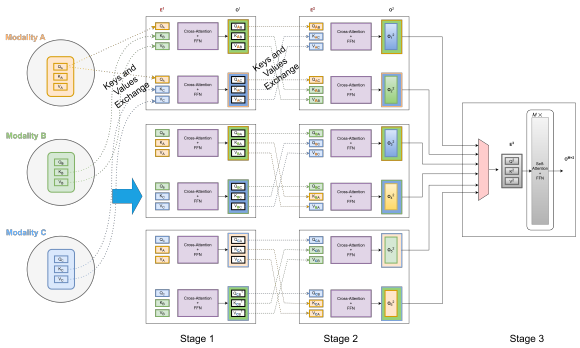

Fig. 4. Three Modality Cross-Modality Attention Transformer Network.

B. Three Modality Case

The versatility of the transformers model we discussed in Two Modality Case allows us to extend behavior and comprehension of the larger-scale environment using multiple CMATs. The goal of CMAT is to effectively integrate and synthesize information from multiple data modalities. CMAT achieves this by adopting two important mechanisms: Pairwise Key-Values Exchange and Cross-modality Attention, as illustrated in Figure 4. It dynamically allocates attention to different data sources based on their relevance to the downstream task. CMAT ensures that the most pertinent information from each modality is highlighted and integrated into a unified representation. The attention-based approach allows for a context-aware fusion of data, thereby enhancing the model's predictive accuracy and interpretability.

IV. Overall Accelerator Architecture for Three Modality

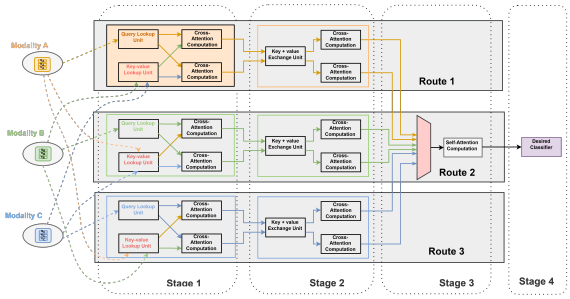

Fig. 5. Three modality Accelerator Architecture.

Figure 5 shows our multimodality transformer architecture which contains three stages. The first stage consists of multiple parallel module groups, one for each modality. Each of the Cross-Attention module groups has three components inside it. They are a Query-lookup unit, a Key-value lookup unit, and a pair of parallel Attention-computation units. The Query-lookup unit chooses queries from the same modality, for instance, modality A. The Key-value lookup unit selects key-value pairs from different modalities, for instance, modality B and modality C. Among the Attention-computation unit pair, the first Attention-computation unit performs the cross-attention computation between the first set of queries, keys, and values, for instance, queries from modality A and keys and values from modality B as X. The second Attention-computation unit performs the attention computation between the second set of queries, keys, and values, for instance, queries from modality B and keys and values from modality C as Y.

The second stage also consists of three parallel module groups, one for each modality. Each module group has a key-value exchange unit and two Attention-computation units in parallel. The key-value exchange unit would first group the outputs from the first stage by modality and exchange the key-value pairs, for instance, queries from output X and keys and values from output Y. The Attention-computation units in this group perform cross-attention between the set of queries, keys, and values. For instance, the first Attention-computation

unit performs cross-attention between queries from output X and key and values from output Y.

The third stage has an output selection unit that selects the output from the cross-attention module depending on the machine learning task requirement and a self-attention unit to refine the representation of the input sequence after selecting the output from the cross-attention module. Finally, The fourth stage consists of a classifier depending on our modality of interest and the machine learning task requirement.

A. FPGA Implementation

Our proposed architecture will utilize an FPGA device to implement the hardware accelerator. At first, we load input data from different modalities and weights into an off-chip DRAM memory on the FPGA board. The DDR controller and DRAM work together to handle input, intermediate, and output data and parameters via the memory interface, thus enabling efficient and parallel data processing for cross-attention accelerators. We utilize an on-chip SRAM buffer with three units: an input storage unit, an intermediate storage unit, and a parameter and weight storage unit to avoid frequent off-chip memory access.

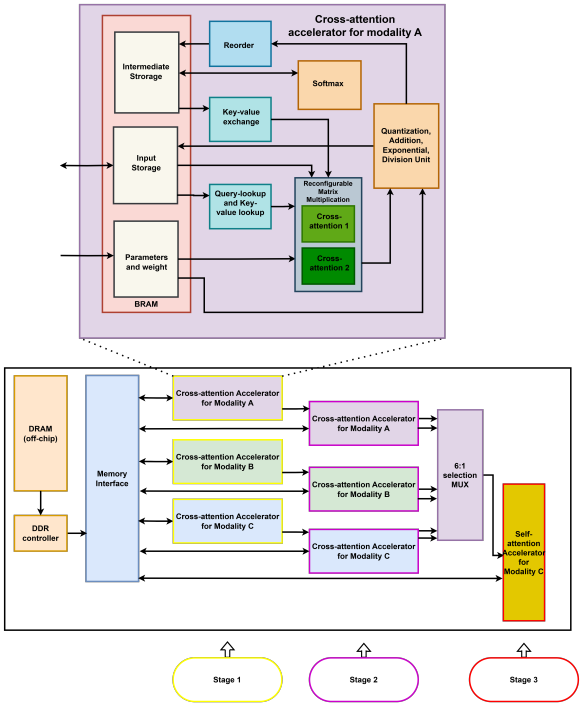

Fig. 6. Implementation of three modality in FPGA board.

As illustrated in the figure 6, we have three parallel Cross-attention Accelerator modules in the first stage (represented by blocks with yellow border), each performing cross-attention calculation focusing on one modality. The first stage Cross-attention accelerator for modality A consists of a Query-lookup and Key-value lookup unit, a Key-value exchange unit, a diverse reconfigurable matrix multiplication unit featuring two parallel cross-attention computing, a data reordering unit, a softmax operation unit, and a vector unit for low computation requiring operations. The reconfigurable matrix multiplication unit is configured to match the shape and size and performs the matrix multiplication for cross-attention using an array of processing elements for different precision and sparsity. Moreover, we will split an array of processing elements into two parallel sections to support two cross-attention calculation units in parallel. The multi-headed cross-attention layers use the softmax operation unit to compute softmax. Our vector unit performs operations with low computational density, including quantization, addition, residual addition, ReLU activation, and division. The reorder unit reorganizes the temporary results before writing them back to the intermediate storage unit after each matrix multiplication and before the concatenation. The Parameter and weight storage unit saves the weights and parameters, the input storage unit saves the input data, whereas the intermediate storage unit saves the temporary results during residual connection. Additionally, we will write the results back to the input storage and transfer them to DRAM.

Our overall transformer architecture consists of three cross-attention accelerator pairs, all operating in parallel for the first and second stages. We use a 6-to-1 selection mux with a self-attention accelerator for the third stage. This approach enables us to leverage parallelism for faster computation of multimodality datasets, eliminating the need for hardware resource re-utilization. However, we can reuse the previous accelerators if we require additional cross-attention and self-attention stages. It is important to note that the Key-value exchange unit is not used in the first stage cross-attention accelerator, Query lookup and Key-value lookup units are not used in the second stage cross-attention accelerator, and both Key-value exchange and Query lookup and Key-value lookup units are not used in the final stage self-attention accelerator.

To achieve high performance and energy efficiency, we will utilize concepts of data loading units that would load query, key, and value vectors into the accelerator quickly from different modalities in parallel. In the first stage, the Query-lookup unit requires queries from one particular modality, whereas the Key-value lookup unit selects pairs from other modalities to learn the relationship between the different modalities. We will use a hashing scheme to reduce the number of pairwise comparisons by bucketing the query and key vector together based on their similarities. This technique would help during the cross-attention score calculation as there will only be a comparison between queries and key vectors that belong to the same bucket. Moreover, we would use a large number of buckets to reduce the number of collisions when two or more vectors are hashed to the same bucket and also distribute the query and key vectors evenly on each bucket to reduce the load on each bucket. Specifically, to create buckets that contain similar vectors and improve the performance of cross-attention computation, we would use locality-sensitive hashing while creating buckets.

B. Accelerating Cross-Attention Operation

The cross-attention operation in multimodality transformers is the most time-consuming operation, especially for long sequences. This is because the cross-attention operation involves calculating attention scores between every pair of elements in two sequences from different modalities. This can lead to a quadratic increase in computational complexity with sequence length. Our architectural design and use of hardware accelerator components encourage parallelism in multiple ways. To understand the components and functionality of accelerator design for our multimodal transformer architecture, we use the metaphor of route, road, and lanes. The route represents the different modalities we aim to process, and each route is a distinct highway that would handle the information in parallel from a specific source. Within each route, we have two parallel roads for the simultaneous computation of data that further enhance parallelism. These roads facilitate parallel computations by distributing the workload and optimizing the processing speed. The most granular level of our design is the lanes. Depending on the number of heads in our multimodality transformer architecture, we choose the same number of lanes because each multiheaded attention operates independently and allows simultaneous calculations. This arrangement would multiply the processing capacity, especially in multimodality transformers with long sequence datasets. Again, many input sequences in a large dataset are independent of each other and would produce weak attention during cross-attention computations, so we can further increase the number of lanes to reduce the computational time. Moreover, our method allows us flexibility to configure the accelerator for different multimodal transformer networks by adjusting the number of routes, roads, and lanes depending on the size of the input sequence and several modalities. Besides exploring the architectural-level parallelism, we will also explore approximate attention techniques to reduce the computational complexity of the cross-attention operation by approximating the attention scores. This can be done using a variety of different methods, such as using a hashing scheme to reduce the number of pairwise comparisons or using a low-rank approximation of the attention matrix.

C. Sparsity Exploration in Multimodal Transformers

Our multimodality transformer network often produces many low attention scores for large multimodality input sequences during the cross-attention and self-attention calculation because many data/sequences are slightly related or sometimes independent of other data/ sequences. Thus, it is critical to reduce the cost of attention blocks as the weak connections contribute little to the final output of the feature aggregation. Previous research has shown that exploiting sparsity patterns on weight parameters is enough to achieve similar accuracy for different transformer neural network models. Reconfigurable matrix multiplication units are the key components in accelerator design that perform dot product and multiplication operations. We will deploy reconfigurable multiplication units in our design because different modalities have different data computation precision requirements. Our reconfigurable matrix multiplication unit would support row-wise precision reconfigurability that will have some rows of the matrix multiplication unit working in high precision mode while using the rest of the rows in the PE for low precision mode.

Furthermore, we will develop a sorting and priority assignment algorithm that sorts the columns of query vector based on sparsity and thus assigns priority, such that columns with the highest number of non-zero elements will receive the highest priority. Likewise, the neighboring token/ embedding is part of the same or related group most of the time. For instance, tokens corresponding to two nearby points in an image are likely to produce higher attention. We can utilize this concept in ordering and setting priorities for the keys. Our implementation also features grouping columns of keys and rows of queries together in an on-chip storage to improve data locality and computational efficiency. We will use a group of query rows to compute the attention score with a column of keys at once. Lastly, we will utilize Block Sparse Matrix Multiplication by grouping Multiple Keys Column, especially for the sparse column groups. For this, we will identify the sparse blocks within a matrix by dividing the matrix into smaller sub-matrices or blocks. Then, we will group columns based on sparsity. Next, we will compute attention by multiplying the queries block with the keys block to speed up the computation. We can also skip the weakest (most sparse) blocks during attention calculations.

V. Conclusion

In our future work, we aim to expand the current framework from three modalities to multiple modalities, enhancing our capability to capture and integrate information from multiple sources comprehensively. We will also implement proper sorting and priority assignment algorithm along with scheduler, hashing scheme for better computation of attention matrices for different precision. Additionally, we propose a new architecture that incorporates multiple encoders and decoders and applies to other machine learning applications other than DETR based object detection. This development will enable us to process the data features of different modalities with greater precision and such advancements will lay a solid foundation for addressing more complex issues such as cross-modal understanding and generation.

Acknowledgment

References

[1] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-end object detection with transformers," in *European conference on computer vision*. Springer, 2020, pp. 213–229.

[2] J. Park, H. Yoon, D. Ahn, J. Choi, and J.-J. Kim, "Optimus: Optimized matrix multiplication structure for transformer neural network accelerator," *Proceedings of Machine Learning and Systems*, vol. 2, pp. 363–378, 2020.

[3] F. Tu, Z. Wu, Y. Wang, W. Wu, L. Liu, Y. Hu, S. Wei, and S. Yin, "Multcim: Digital computing-in-memory-based multimodal transformer accelerator with attention-token-bit hybrid sparsity," *IEEE Journal of Solid-State Circuits*, 2023.

[4] ——, "16.1 muitcim: A 28nm 2.24μJ/token attention-token-bit hybrid sparse digital cim-based accelerator for multimodal transformers," in *2023 IEEE International Solid-State Circuits Conference (ISSCC)*, 2023, pp. 248–250.

[5] T. J. Ham, S. J. Jung, S. Kim, Y. H. Oh, Y. Park, Y. Song, J.-H. Park, S. Lee, K. Park, J. W. Lee *et al.*, "Aˆ 3: Accelerating attention mechanisms in neural networks with approximation," in *2020 IEEE International Symposium on High Performance Computer Architecture (HPCA)*. IEEE, 2020, pp. 328–341.

[6] T. J. Ham, Y. Lee, S. H. Seo, S. Kim, H. Choi, S. J. Jung, and J. W. Lee, "Elsa: Hardware-software co-design for efficient, lightweight self-attention mechanism in neural networks," in *2021 ACM/IEEE 48th Annual International Symposium on Computer Architecture (ISCA)*. IEEE, 2021, pp. 692–705.

[7] L. Lu, Y. Jin, H. Bi, Z. Luo, P. Li, T. Wang, and Y. Liang, "Sanger: A co-design framework for enabling sparse attention using reconfigurable architecture," in *MICRO-54: 54th Annual IEEE/ACM International Symposium on Microarchitecture*. New York, NY, USA: ACM, 2021.

[8] A. Nagrani, S. Yang, A. Arnab, A. Jansen, C. Schmid, and C. Sun, "Attention bottlenecks for multimodal fusion," *Advances in neural information processing systems*, vol. 34, pp. 14 200–14 213, 2021.

[9] Z. Qu, L. Liu, F. Tu, Z. Chen, Y. Ding, and Y. Xie, "DOTA: detect and omit weak attentions for scalable transformer acceleration," in *Proceedings of the 27th ACM International Conference on Architectural Support for Programming Languages and Operating Systems*, vol. 1703. New York, NY, USA: ACM, 2022, pp. 14–26.

[10] F. Tu, Z. Wu, Y. Wang, L. Liang, L. Liu, Y. Ding, L. Liu, S. Wei, Y. Xie, and S. Yin, "A 28nm 15.59µj/token full-digital bitline-transpose CIM-based sparse transformer accelerator with pipeline/parallel reconfigurable modes," in *2022 IEEE International Solid- State Circuits Conference (ISSCC)*. IEEE, 2022.

[11] Y. Wang, Y. Qin, D. Deng, J. Wei, Y. Zhou, Y. Fan, T. Chen, H. Sun, L. Liu, S. Wei, and S. Yin, "A 28nm 27.5TOPS/W approximate-computing-based transformer processor with asymptotic sparsity speculating and out-of-order computing," in *2022 IEEE International Solid-State Circuits Conference (ISSCC)*. IEEE, 2022.

[12] Y. Xu, Z. Zhang, M. Zhang, K. Sheng, K. Li, W. Dong, L. Zhang, C. Xu, and X. Sun, "Evo-vit: Slow-fast token evolution for dynamic vision transformer," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 36, no. 3, 2022, pp. 2964–2972.

[13] T. Tambe, C. Hooper, L. Pentecost, T. Jia, E.-Y. Yang, M. Donato, V. Sanh, P. Whatmough, A. M. Rush, D. Brooks, and G.-Y. Wei, "Edge-BERT: Sentence-level energy optimizations for latency-aware multi-task NLP inference," in *MICRO-54: 54th Annual IEEE/ACM International Symposium on Microarchitecture*. New York, NY, USA: ACM, 2021.

[14] B. Li, S. Pandey, H. Fang, Y. Lyv, J. Li, J. Chen, M. Xie, L. Wan, H. Liu, and C. Ding, "FTRANS: Energy-efficient acceleration of transformers using FPGA," in *Proceedings of the ACM/IEEE International Symposium on Low Power Electronics and Design*. New York, NY, USA: ACM, 2020.

[15] H. Wang, Z. Zhang, and S. Han, "SpAtten: Efficient sparse attention architecture with cascade token and head pruning," in *2021 IEEE International Symposium on High-Performance Computer Architecture (HPCA)*. IEEE, 2021.

[16] L. Liu, Z. Qu, Z. Chen, F. Tu, Y. Ding, and Y. Xie, "Dynamic sparse attention for scalable transformer acceleration," *IEEE Trans. Comput.*, pp. 1–14, 2022.

[17] Z. Zhou, J. Liu, Z. Gu, and G. Sun, "Energon: Toward efficient acceleration of transformers using dynamic sparse attention," *IEEE Trans. Comput.-aided Des. Integr. Circuits Syst.*, vol. 42, no. 1, pp. 136–149, 2023.

[18] G. Shen, J. Zhao, Q. Chen, J. Leng, C. Li, and M. Guo, "SALO: An efficient spatial accelerator enabling hybrid sparse attention mechanisms for long sequences," in *Proceedings of the 59th ACM/IEEE Design Automation Conference*. New York, NY, USA: ACM, 2022.

[19] C. Fang, A. Zhou, and Z. Wang, "An algorithm–hardware co-optimized framework for accelerating N:M sparse transformers," *IEEE Trans. Very Large Scale Integr. VLSI Syst.*, vol. 30, no. 11, pp. 1573–1586, 2022.

[20] C. Fang, S. Guo, W. Wu, J. Lin, Z. Wang, M. K. Hsu, and L. Liu, "An efficient hardware accelerator for sparse transformer neural networks," in *2022 IEEE International Symposium on Circuits and Systems (ISCAS)*. IEEE, 2022.

[21] A. Vaswani, "Attention is all you need," *Advances in Neural Information Processing Systems*, 2017.

[22] Z. Tian, C. Shen, H. Chen, and T. He, "FCOS: Fully convolutional one-stage object detection," in *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*. IEEE, 2019.

[23] Z. Cai and N. Vasconcelos, "Cascade r-cnn: High quality object detection and instance segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 5, pp. 1483–1498, 2021.

[24] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 6, pp. 1137–1149, 2017.

[25] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollar, "Focal loss for dense object detection," in *2017 IEEE International Conference on Computer Vision (ICCV)*. IEEE, 2017.

[26] X. Zhou, D. Wang, and P. Krähenbühl, "Objects as points," *arXiv preprint arXiv:1904.07850*, 2019.

[27] H. Hu, J. Gu, Z. Zhang, J. Dai, and Y. Wei, "Relation networks for object detection," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 3588–3597.

[28] K. He, R. Girshick, and P. Dollár, "Rethinking imagenet pre-training," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 4918–4927.

[29] H. W. Kuhn, "The hungarian method for the assignment problem," *Nav. Res. Logist. Q.*, vol. 2, no. 1-2, pp. 83–97, 1955.

[30] A. Salvador, M. Bellver, V. Campos, M. Baradad, F. Marques, J. Torres, and X. Giro-i Nieto, "Recurrent neural networks for semantic instance segmentation," *arXiv preprint arXiv:1712.00617*, 2017.

[31] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft COCO: Common objects in context," in *Computer Vision – ECCV 2014*. Cham: Springer International Publishing, 2014, pp. 740–755.

[32] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[33] J. Hosang, R. Benenson, and B. Schiele, "Learning non-maximum suppression," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2017.

[34] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga *et al.*, "Pytorch: An imperative style, high-performance deep learning library," *Advances in neural information processing systems*, vol. 32, 2019.

[35] X. Glorot and Y. Bengio, "Understanding the difficulty of training deep feedforward neural networks," in *Proceedings of the thirteenth international conference on artificial intelligence and statistics*. JMLR Workshop and Conference Proceedings, 2010, pp. 249–256.

[36] A. Kirillov, R. Girshick, K. He, and P. Dollar, "Panoptic feature pyramid networks," in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2019.