RNA-ViT: Reduced-Dimension Approximate Normalized Attention Vision Transformers for Latency Efficient Private Inference

Dake Chen*[1], Yuke Zhang*[1], Souvik Kundu*[2], Chenghao Li[1], Peter A. Beerel[1]
[1]University of Southern California, Los Angeles, CA, USA
[2]Intel Labs, San Diego, USA
{dakechen, yukezhan}@usc.edu, souvikk.kundu@intel.com, {cli78217, pabeerel}@usc.edu

2023 IEEE/ACM International Conference on Computer Aided Design (ICCAD) | 979-8-3503-2225-5/23/$31.00 ©2023 IEEE | DOI: 10.1109/ICCAD57390.2023.10323702

***Abstract*—The concern over data and model privacy in machine learning inference as a service (MLaaS) has led to the development of private inference (PI) techniques. However, existing PI frameworks, especially those designed for large models such as vision transformers (ViT), suffer from high computational and communication overheads caused by the expensive multi-party computation (MPC) protocols. The encrypted attention module that involves the `softmax` operation contributes significantly to this overhead. In this work, we present a family of models dubbed RNA-ViT, that leverage a novel attention module called reduced-dimension approximate normalized attention and a latency efficient GeLU-alternative layer. In particular, RNA-ViT uses two novel techniques to improve PI efficiency in ViTs: a reduced-dimension normalized attention (RNA) architecture and a high order polynomial (HOP) softmax approximation for latency efficient normalization. We also propose a novel metric, accuracy-to-latency ratio (`A2L`), to evaluate modules in terms of their accuracy and PI latency. Based on this metric, we perform an analysis to identify a nonlinearity module with improved PI efficiency. Our extensive experiments show that RNA-ViT can achieve average $3.53\times$, $3.54\times$, $1.66\times$ lower PI latency with an average accuracy improvement of 0.93%, 2.04%, and 2.73% compared to the state-of-the-art scheme MPCViT [1], on CIFAR-10, CIFAR-100, and Tiny-ImageNet, respectively.**

***Index Terms*—Deep learning, Computer vision, Vision transformer, Private inference, Multi-party computation**

I. Introduction

In recent years, vision transformers (ViTs) have achieved remarkable success in complex computer vision tasks. Such remarkable performance can be attributed to their ability in capturing long-range dependencies through the self-attention (SA) modules. ViTs have demonstrated superior performance in image classification [2]–[4], object detection [5], [6], and semantic segmentation [6]–[8] tasks, surpassing traditional convolutional neural network (CNN) architectures.

The success of ViTs and other deep neural network models has sparked the emergence of machine learning inference as a service (MLaaS). In this paradigm, a service provider trains models and provides inference service on various clients' data for various tasks including online diagnoses and financial product recommendations [9], [10]. However, the growing concerns regarding privacy have hindered such commercialization. Clients may be reluctant to share their personal data with the service provider, while the providers aim to safeguard the proprietary details of their trained models [9]. Both parties are averse to transmitting sensitive, non-encrypted information to each other. To address these concerns, private inference (PI) [11]–[16] have been proposed. These methods leverage techniques such as homomorphic encryption (HE) and secure multi-party computation (MPC) protocols to protect the privacy of both client data and the intellectual property (IP) of the model.

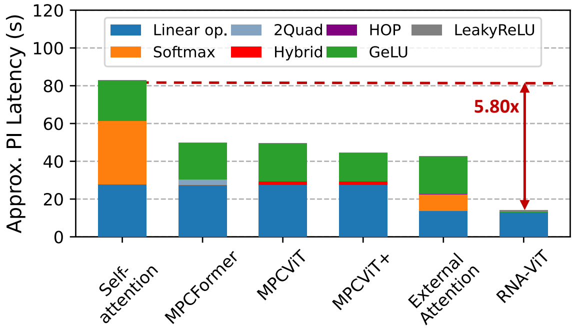

Fig. 1. PI latency comparison between the proposed RNA-ViT and prior works on CIFAR-10. We use a transformer model with 7 layers for the evaluation. RNA-ViT presents PI latency reduction for linear operations, `softmax`, and nonlinear operations (e.g., `GeLU`).

While there have been several works on efficient PI for CNNs [17], [18], the exploration of PI for transformers, specifically ViTs, has been relatively limited. Implementing existing PI methods directly on ViTs results in significantly higher latency and communication overhead compared to standard inference. This poses a significant obstacle to their widespread adoption, especially in resource-constrained client applications [1], [19]. The primary contributor to this high latency is the compute heavy `softmax` and `GeLU` functions [1], [19] in secure MPC paradigms. To reduce this cost for a transformer, a recent work [19] proposed replacing the the `softmax` with a 2^{nd} order polynomial approximation called `2Quad` [20]. With similar goals for ViTs, [1] formulated a neural architecture search (NAS) algorithm to replace the `softmax` with either the

*Authors contributed equally.

This work was supported in part by National Science Foundation (NSF) under Grant No. CCF-1763747.

2ReLU function [11] or the scaling function [21].

While previous works [1], [19] primarily focus on reducing the PI latency overhead associated with softmax, we present a comprehensive solution to reduce PI latency incurred by attention architectures, softmax, and GeLU. For the attention architecture, the size of its attention map determines the number of softmax operations which dominates the PI latency and communication overhead, consequently affecting the overall PI efficiency. Motivated by this observation, we propose a novel attention architecture specifically designed for efficient PI, called reduced-dimension normalized attention (RNA).

In addition to optimizing the attention architecture, we aim to eliminate the PI latency heavy exponential operations in the softmax function for normalization. To achieve this, we propose a novel latency efficient alternative, namely, higher-order polynomial softmax approximation (HOP). HOP introduces a re-weighted normalization scheme for the attention map, effectively replacing the exponential with polynomial computations. We then integrate the HOP normalization layer into the RNA attention architecture, creating a novel attention module to significantly reduce the PI latency.

GeLU nonlinearity in the ViT architecture contributes to significant PI latency overhead due to its reliance on PI computationally expensive exponential functions. Towards easing the non-linear layer latency, we first analyze the performance of different non-linear layers in ViTs in terms of both accuracy and PI latency. In particular, we propose a novel metric called accuracy-to-latency ratio (A2L) that comprehensively assesses the trade-off between accuracy and PI latency. Leveraging the A2L efficiency metric, we analyze various nonlinearity modules and discover that the LeakyReLU nonlinearity module yields the best A2L PI efficiency. Our contributions are as follows.

- We present a novel attention architecture called RNA with a compressed attention map to achieve significantly lower PI latency and compute overhead.
- We present a novel and more PI-friendly re-weighted normalization called HOP, which replaces the PI computationally expensive softmax function in the attention module. By integrating HOP into RNA, we create an attention module that greatly reduces PI latency.
- We use a metric, namely accuracy per unit latency (A2L), to identify and benchmark various models including ViT in the context of performance per PI latency.
- Through our analysis, we identify the LeakyReLU nonlinear module as a highly efficient component for PI applications, as it significantly reduces PI latency while maintaining competitive accuracy.

These contributions are combined to form a family of models known as **reduced-dimension normalized attention** (RNA)-ViT that outperforms the state-of-the-art scheme MPCViT [1] by achieving an average increase of 0.93%, 2.04% and 2.73% in accuracy with an average of 3.53×, 3.54×, 1.66× lower PI latency on CIFAR-10, CIFAR-100, and Tiny-ImageNet, respectively. Furthermore, in comparison to MPCFormer [19], the proposed RNA-ViT achieves similar accuracies while demonstrating an average PI latency reduction of 3.64×, 3.61×, and 1.68× on CIFAR-10, CIFAR-100, and Tiny-ImageNet, respectively. Specifically, as shown in Figure 1, RNA-ViT outperforms prior works in terms of PI latency reduction for linear, softmax, and nonlinear layers.

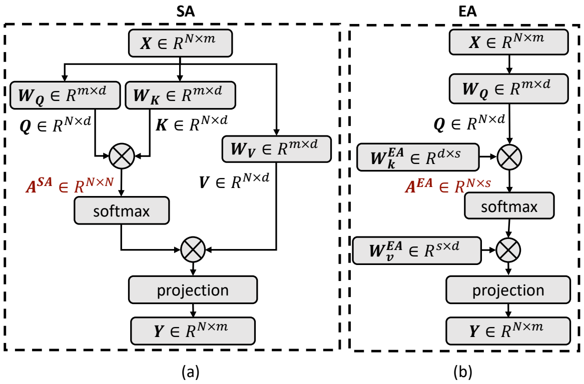

Fig. 2. Architectures of (a) self-attention (SA) and (b) external attention (EA).

II. Background

A. PI frameworks

In the past few years, a series of PI frameworks using homomorphic encryption (HE) and MPC protocols such as secret sharing (SS), Garbled Circuits (GC), and oblivious transfer (OB) have been proposed for convolutional neural networks (CNNs) [11]–[16], [22], [23]. With the rapid adoption of transformers, cryptographic protocols supporting transformer operations have been emerging at a fast pace. THE-X [24] replaces the complex operations such as softmax in transformer inference with crypto-friendly operations, and leverages HE to preserve the transformer inference's privacy. Iron [25] devises efficient protocols for softmax, GeLU, and LayerNorm, and optimizes HE-based protocol to speed up high-dimensional matrix multiplication in transformers.

However, both CNNs and transformers face challenges in privacy-preserving computations and communications, particularly in nonlinear operations. While cryptography experts focus on refining cryptographic protocols for efficiency, an alternative solution involves adapting neural network architectures to be more PI-friendly. For CNNs, common approaches include linearizing the network by pruning ReLUs [17], [18], [26], [27], or substituting ReLUs with a less computationally expensive quadratic function [13]. In the case of transformers, MPCFormer [19] proposes replacing softmax with more crypto-friendly 2Quad functions, while MPCViT [1] introduces a search method to strategically replace each softmax with either a linear scaling function or a 2ReLU function.

B. Attention variants in transformers

While SA, whose architecture is shown in Figure 2(a), yields the benefits of capturing long-distance dependencies, its computational and storage overheads increase quadratically ($O(N^2)$) with the size of the feature map [28], significantly impacting the computation and latency for both regular inference and PI. Therefore, attention variants with linear complexity

($O(N)$), which has been widely studied recently [28]–[32], are candidates for lower PI latency overhead.

SA leverages the scaled dot-product with `softmax` normalization to measure the similarity among the Query, Key, and Value matrices of the input sequence. To reduce complexity, Linformer [28] learns to shrink the length of Key and Value matrices via projections and presents less non-linear and linear operations. However, it degrades the model's accuracy by compressing the attention map. CosFormer [29] replaces SA with a linear projection kernel and a PI-unfriendly *cosine*-based re-weighting mechanism. Hamburger [30] reformulates learning the global context as a low-rank completion problem and solves it via matrix decomposition. SOFT [31] uses Gaussian kernel and exponential function to replace SA and solves it via Newton-Raphson iteration [33]. Both Hamburger and SOFTA utilize iterations to solve the matrix decomposition problem, which introduces additional costs despite having linear complexity. In contrast to these attention variants, external attention (EA) [32], as illustrated in 2, leverages two lightweight memory units to learn the most discriminative features across the entire dataset, and substitutes SA with two linear layers and a normalization layer. We quantified these benefits by measuring the PI latency of SA and its variants using a ViT model [3] having 7 layers each with 4 heads on CIFAR-10 using CrypTen [34]. We found EA to be most latency efficient with a latency advantage of up to 2.73× and 1.8× compared to the SA and the lowest latency of these linear alternatives (CosFormer [29]).

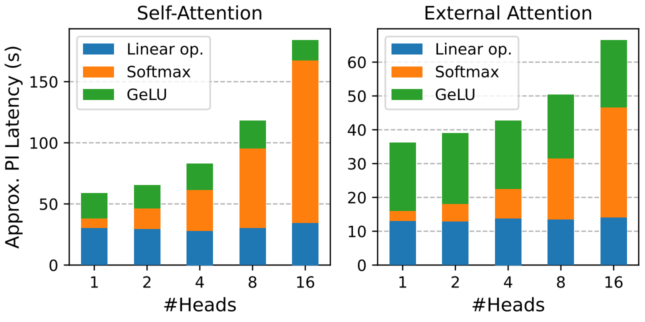

Fig. 3. PI latency breakdown of SA and EA. The results are obtained by using CCT model [3] and CrypTen [34] on CIFAR-100.

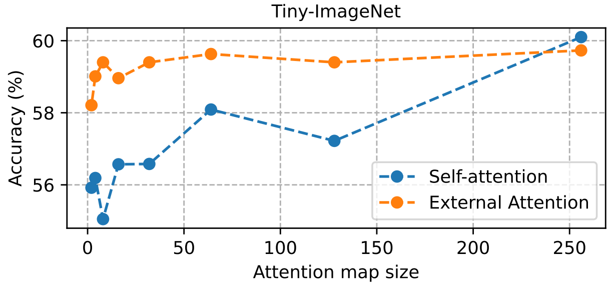

Fig. 4. Accuracy comparison of SA and EA with different attention map sizes s on Tiny-ImageNet. The attention map sizes for both SA and EA are $N \times s$ where N is 256 in this experiment.

III. Our Methods

A. Notations

In this paper, we use $\boldsymbol{X} \in \mathbb{R}^{N \times m}$ to denote an input sequence of N tokens with each token represented as a m-dimensional feature vector. There are three major components for the input feature, i.e., Query ($\boldsymbol{Q} \in \mathbb{R}^{N \times d}$), Key ($\boldsymbol{K} \in \mathbb{R}^{N \times d}$), and Value ($\boldsymbol{V} \in \mathbb{R}^{N \times d}$), obtained from three learnable linear matrices $\boldsymbol{W_Q} \in \mathbb{R}^{m \times d}$, $\boldsymbol{W_K} \in \mathbb{R}^{m \times d}$, and $\boldsymbol{W_V} \in \mathbb{R}^{m \times d}$ through $\boldsymbol{Q} = \boldsymbol{X W_Q}$, $\boldsymbol{K} = \boldsymbol{X W_K}$, and $\boldsymbol{V} = \boldsymbol{X W_V}$, where d is the embedding dimension of $\boldsymbol{Q}$, $\boldsymbol{K}$, and $\boldsymbol{V}$. We use $\boldsymbol{A}$ to denote the attention map, and the superscript to distinguish the attention architectures.

B. Motivation

1) Breakdown of PI latency: Figure 3 presents the PI latency breakdown of SA and EA, where it can be concluded that EA on average saves around 50% of the PI latency over SA due to having a more compact attention architecture, detailed in the next subsection. Moreover, `GeLU` is a significant contributor to the PI latency in EA compared to that in SA. This motivates us to explore methods reducing the non-linear operation cost.

2) Effects of attention compression: Figure 2(a) and (b) compare the attention architecture of SA and EA, where the attention map size is $N \times N$ and $N \times s$ for SA and EA, respectively. Compared to the number of tokens N (e.g., 256), s is typically a smaller value (e.g., 64) [32]. Therefore, the size of the attention map is reduced before entering the `softmax` normalization, and the PI latency of `softmax` is significantly reduced. Further, EA uses $\boldsymbol{W_k^{EA}}$ and $\boldsymbol{W_v^{EA}}$ in memory unit, it does not involve matrix multiplications $\boldsymbol{K} = \boldsymbol{X W_k^{EA}}$ and $\boldsymbol{V} = \boldsymbol{X W_V^{EA}}$, thus yielding lower PI latency associated with linear operations. Moreover, the size of $\boldsymbol{W_k^{EA}}$ and $\boldsymbol{W_v^{EA}}$ are smaller than $\boldsymbol{W_K}$ and $\boldsymbol{W_V}$ of SA, which further reduces the latency associated with linear operations.

Inspired by EA, we apply a similar strategy to SA, where we add two linear layers before and after the `softmax` normalization to reduce (compress) and regain (expand) the attention map size, respectively. We compare the classification accuracy of SA and EA with different sizes of attention map on Tiny-ImageNet in Figure 4. It can be observed that compressing the attention map size of SA leads to a more significant accuracy drop than compressing the attention map size of EA. The empirical results suggest that the attention map in SA, that represents the token-to-token relationship in the input, is more delicate and potentially more sensitive to compression to a smaller and simpler dimension. Thus compressed attention at low-dimension may impact its ability to learn complex relationships. In contrast, EA has two memory units, one to compress and the other to expand along one dimension of the attention map. These are more stable and provide better robustness against compression. This finding sheds light on the inherent differences between the two attention architectures and the potential trade-offs involved in compressing their attention maps for PI.

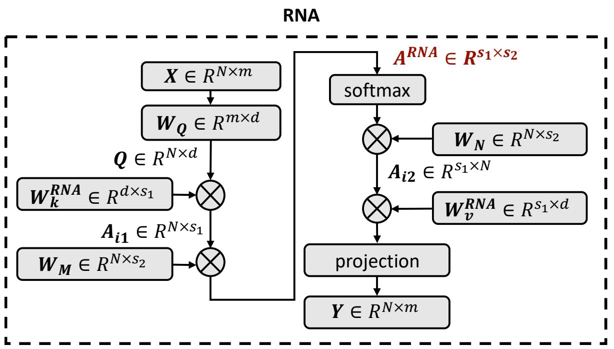
RNA
$X \in R^{N\times m}$
$W_Q \in R^{m\times d}$
$Q \in R^{N\times d}$
$W_k^{RNA} \in R^{d\times s_1}$
$A_{i1} \in R^{N\times s_1}$
$W_M \in R^{N\times s_2}$
$A^{RNA} \in R^{s_1\times s_2}$
softmax
$W_N \in R^{N\times s_2}$
$A_{i2} \in R^{s_1\times N}$
$W_v^{RNA} \in R^{s_1\times d}$
projection
$Y \in R^{N\times m}$

Fig. 5. Architecture of the proposed reduced-dimension normalized attention. Here, the post Q-K computation attention normalization happens through the Softmax layer.

C. Reduced-dimension Normalized Attention (RNA)

Motivated by these observations, we propose a PI-friendly attention architecture, named **reduced-dimension normalized attention (RNA)**. The architecture of RNA is shown in Figure 5. RNA aims at reducing the dimension size of the attention map to tailor a more efficient attention architecture for PI. Unlike EA which compresses along one dimension of the attention map, we introduce two memory units to compress along both dimensions of the attention map. The sequence of operations can be mathematically described as:

$$\begin{aligned} \boldsymbol{Q} &= \boldsymbol{X}\boldsymbol{W_Q}, &(1)\\ \boldsymbol{A_{i1}} &= \boldsymbol{Q}\boldsymbol{W_k^{RNA}}, &(2)\\ \boldsymbol{A^{RNA}} &= \boldsymbol{A_{i1}^T}\boldsymbol{W_M}, &(3)\\ \boldsymbol{A_{i2}} &= softmax(\boldsymbol{A^{RNA}})\boldsymbol{W_N^T}, &(4)\\ \boldsymbol{Y} &= \boldsymbol{A_{i2}^T}\boldsymbol{W_v^{RNA}}, &(5)\end{aligned}$$

where $\boldsymbol{W_M}$ and $\boldsymbol{W_N}$ are the weight matrices of the newly-introduced linear layers, $\boldsymbol{A_{i1}}$ and $\boldsymbol{A_{i2}}$ denotes two intermediate attention maps, and $\boldsymbol{A^{RNA}}$ represents the attention map of RNA. Dimensions of these matrices are shown in Figure 5, where s_1 and s_2 denotes the reduced dimensions hyperparameters. In RNA, the first two linear layers compress the two dimensions of the attention map, respectively, to yield $\boldsymbol{A^{RNA}} \in \mathbb{R}^{s_1\times s_2}$. This low-dimension attention map is then passed through a softmax layer to perform normalization for the attention map. Finally, the last two linear layers recover the original two dimensions of the attention map. The four linear layers associated with four memory units are used to compress and recover the dimension of the attention map. The advantage of RNA is that its attention map is significantly smaller in size than other attention architectures, making the `softmax` operation more efficient that results in significantly lower PI latency.

D. Metric for PI: accuracy-to-latency ratio (A2L)

To facilitate a convenient comparison of modules for PI and MPC, we propose a novel metric called accuracy-to-latency ratio (`A2L`), which is defined by the equation shown below:

$$\texttt{A2L} = \frac{Accuracy}{PI\ Latency}. \tag{6}$$

A higher value of `A2L` indicates that the nonlinear operation or module achieves higher accuracy while also having lower PI latency, which is desirable for PI applications.

E. Higher order polynomial softmax approximation (HOP)

In PI, the `softmax` involves multiple exponential operations that can result in significant latency. This is particularly true for vision transformer models, which contain multiple softmax layers, and can be observed in the breakdown of private inference latency illustrated in Figure 3. Therefore, in addition to optimizing the attention architecture, it is crucial to develop a softmax approximation that is better suitable for PI and MPC scenarios.

In essence, the `softmax` performs a re-weighted normalization of the attention map and the general form is as follows [29]:

$$\frac{S(a_{ij})}{\sum_j S(a_{ij}) + \epsilon}, \tag{7}$$

where a_{ij} denotes the element of attention map, S is a re-weighting function, and ϵ is a small positive value to avoid a zero denominator. Due to the nature of normalization, the summation of each row in $\boldsymbol{A}$ equals one. The widely-adopted `softmax` normalization function applies an exponential function for the re-weighting. As evident from Figure 3, the number of exponential operations increases with an increase in the number of attention heads, resulting in a significant rise in the PI latency of `softmax`. This motivates the use of a PI-friendly softmax approximation that can lower private inference, regardless of the number of attention heads employed. In order to improve convergence during training, it is recommended that the normalized attention map is positive. This necessitates the use of a re-weighting function that produces non-negative outputs and is differentiable [29], [35]. Additionally, a nonlinear re-weighting function with a suitable shape has been shown to enhance the contrast among attention map elements more effectively than a linear one [29].

Inspired by these observations, we propose a PI-friendly approximation of `softmax`, named **higher order polynomial softmax approximation** (HOP). The general form of HOP is mathematically described as:

$$\texttt{HOP}(a_{ij}) = \begin{cases} \dfrac{(a_{ij}+c)^p}{\sum_j (a_{ij}+c)^p + \epsilon}, & if\ p\ is\ even \\ \dfrac{|a_{ij}+c|^p}{\sum_j |a_{ij}+c|^p + \epsilon}, & if\ p\ is\ odd \end{cases} \tag{8}$$

where c is a constant and p represents the power of the re-weighting function. The value of c is a hyper-parameter and is typically set to 5. For the `HOP` with odd power p, the introduced absolute function is to ensure a non-negative attention map. Empirical results demonstrate that even powers have slightly lower PI latencies compared to odd powers. It is worth noting that `2Quad` [20] is a specific instance of the `HOP` approach with a power of 2.

To determine the optimal power of `HOP` for SA and RNA, we conducted experiments by exhaustively training models with different values for p and measuring their final test accuracies,

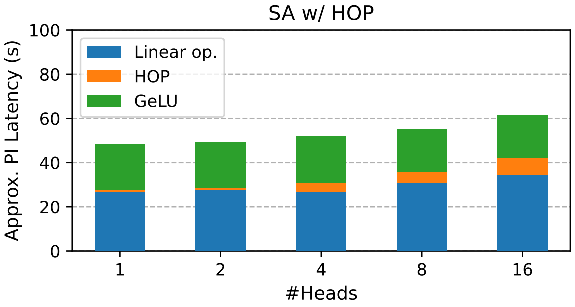

Fig. 6. PI latency breakdown of SA with higher order polynomial softmax approximation for different numbers of heads

TABLE I
PI LATENCY AND PERFORMANCE ANALYSIS OF SA WITH VARIOUS NONLINEARITIES

Nolinearity	CIFAR-10			CIFAR-100			Tiny-ImageNet		
	Acc. (%)	Approx. PI Lat. (s)	A2L	Acc. (%)	Approx. PI Lat. (s)	A2L	Acc. (%)	Approx. PI Lat. (s)	A2L
GeLU	**95.56**	81.48	1.17	77.36	81.20	0.95	**61.6**	142.87	0.43
ReLU	95.43	61.20	1.56	77.3	61.34	1.26	60.43	127.26	0.47
LeakyReLU	**95.56**	**59.33**	**1.61**	**78.46**	**59.66**	**1.32**	61.15	**122.48**	**0.50**
ReLU6	95.4	60.28	1.58	77.63	61.07	1.27	60.58	123.99	0.49
RReLU	95.54	60.79	1.57	78.45	60.64	1.29	60.45	123.73	0.49

PI latency, and A2L. For SA with HOP, we observed that the accuracy initially improved as we increased p, however, the improvements saturated around $p = 4$. Similarly, for RNA, we performed similar experiments and found that the accuracy also reached a saturation point around $p = 4$. Furthermore, we examined the A2L metric for both SA and RNA and the peak values are obtained when using HOP with a power of 4, implying that HOP with a power of 4 demonstrates the best PI efficiency for both SA and RNA modules.

In Figure 6, we present the breakdown of PI latency for SA using HOP with a power of 4 and varying numbers of attention heads. We observe that with an increase in the number of attention heads, the PI latency associated with HOP is $8.34\times$ lower than that associated with softmax, as shown in Figure 3.

F. Analysis for a PI-friendly nonlinearity

After optimizing the PI latency of the attention module with softmax layers using HOP, we see from Figures 6 that the GeLU activation function accounts for up to 40.48% of the total PI latency. Because this high PI latency stems from GeLU's use of exponential operations, we conduct an analysis of various alternative nonlinearities that avoid exponential functions, including ReLU, LeakyReLU, ReLU6, and RReLU.

The results of this analysis, presented in Table I, show that all evaluated nonlinearity alternatives have similar PI latency, however, LeakyReLU achieves the best A2L across all datasets. Notably, LeakyReLU with a small gradient in the negative domain provides a significant improvement in PI latency while maintaining comparable performance to models that use GeLU.

G. Knowledge distillation

To further enhance the accuracy of the proposed RNA-ViT with attention compression and approximation, we utilize logits-based knowledge distillation (KD) [36] with the following loss function:

$$\mathcal{L}_{train} = \mathcal{L}(\hat{\boldsymbol{Y}}_{\boldsymbol{s}}, \boldsymbol{Y}) + \alpha\mathcal{L}(\hat{\boldsymbol{Y}}_{\boldsymbol{s}}, \hat{\boldsymbol{Y}}_{\boldsymbol{t}}), \tag{9}$$

where $\mathcal{L}$ represents the cross-entropy loss, α is a hyperparameter used to control the KD strength, and $\hat{\boldsymbol{Y}}_{\boldsymbol{s}}$ and $\hat{\boldsymbol{Y}}_{\boldsymbol{t}}$ are the logits from the student model and teacher model, respectively.

IV. EXPERIMENTS

A. Experimental setup

Our RNA-ViT models are developed on two types of CCT [3] ViT architecture on three datasets, i.e., CIFAR-10, CIFAR-100, and Tiny-ImageNet. The baseline ViT depth, the number of heads, and hidden dimension are set to 7, 4, and 256, respectively, for the CIFAR-10 and CIFAR-100 datasets, and 9, 12, and 192, respectively, for the Tiny-ImageNet dataset. For RNA-ViT, we use one attention head in RNA attention, the same depth, the same hidden dimension, and 64 for s_1 and s_2 to achieve the best A2L, other hyperparameters are the same as the baseline ViT. The batch size for all datasets we use is 256. We use the same image augmentations as [3]. We train all ViTs and variants for 600 epochs in experiments on CIFAR-10 and CIFAR-100 and for 300 epochs in experiments on Tiny-ImageNet. For KD, the KD strength α in Equation 9 is set to 2 and we use the teacher model in [2]. The training procedures are conducted on an Nvidia A40 GPU. PI latency is measured using CrypTen [34] under the semi-honest threat model [13] on a 8-Core Intel CPU with 16 GB RAM. Please note that latency results can vary based on the machine used. While our reported latency is approximate, the comparison remains fair as all candidate models are tested using the same machine and setup.

B. Comparison of RNA-ViT with Prior-Art

In this section, we combine the proposed RNA and normalization HOP with a power of 4 to form a novel attention module and apply the PI-friendly nonlinearity LeakyReLU. Collectively, these modules form an innovative family of models for PI: RNA-ViT. We measure the performance of RNA-ViT with and without KD. For comparison, we also present the results of baseline ViT consisting of original attention, softmax, and GeLU and prior PI-efficient ViT frameworks, namely, MPCFormer [19] (with and without KD), MPCViT (with and without KD), MPCViT+ [1] and EA ViT [32] in Table II.

The results demonstrate that RNA-ViT consistently exhibits lower latency while maintaining comparable accuracy. More precisely, our RNA-ViT models yield up to $5.80\times$, $5.75\times$, and $3.23\times$ lower latency, and up to $5.78\times$, $5.88\times$, $3.47\times$ higher A2L than even the baseline ViT on CIFAR-10, CIFAR-100, and Tiny-ImageNet, respectively. Moreover, RNA-ViT

TABLE II
PERFORMANCE COMPARISON ON CIFAR-10, CIFAR-100, AND TINY-IMAGENET

Work	Attention	Softmax Approx.	Nonlinearity	CIFAR-10			CIFAR-100			Tiny-ImageNet		
				Acc. (%)	Approx. PI Lat. (s)	A2L	Acc. (%)	Approx. PI Lat. (s)	A2L	Acc. (%)	Approx. PI Lat. (s)	A2L
Baseline	SA	`softmax`	`GeLU`	**95.56**	81.48	1.17	77.36	81.20	0.95	61.60	142.87	0.43
EA [32]	EA	`softmax`	`GeLU`	92.86	41.72	2.23	74.39	41.54	1.79	59.27	84.08	0.70
MPCFormer [19] (w/o KD)	SA	`2Quad`	`GeLU`	95.12	51.12	1.86	76.70	51.02	1.50	59.37	74.31	0.80
MPCFormer [19] (w/ KD)	SA	`2Quad`	`GeLU`	95.13	51.12	1.86	77.07	51.02	1.51	60.84	74.31	0.82
MPCViT [1] (w/o KD)	SA	Hybrid*	`GeLU`	*93.38*	50.94	1.83	*75.38*	51.33	1.47	*59.02*	76.14	0.78
				93.21	50.04	1.86	*74.45*	50.46	1.48	*58.39*	74.34	0.79
				93.01	49.13	1.89	*74.51*	49.59	1.50	*58.05*	72.54	0.80
				92.86	48.23	1.93	*73.17*	48.72	1.50	*56.75*	70.74	0.80
MPCViT [1] (w/ KD)	SA	Hybrid*	`GeLU`	*94.27*	50.94	1.85	*77.76*	51.33	1.51	*63.03*	76.14	0.83
				94.22	50.04	1.88	*76.92*	50.46	1.52	*63.45*	74.34	0.85
				94.08	49.13	1.91	*76.93*	49.59	1.55	*63.38*	72.54	0.87
				93.59	48.23	1.94	*76.40*	48.72	1.57	*62.65*	70.74	0.89
MPCViT+ [1]	SA	Hybrid*	75% `GeLU`	*94.27*	45.08	2.09	–	–	–	–	–	–
				93.94	44.51	2.11	–	–	–	–	–	–
				93.92	43.94	2.14	–	–	–	–	–	–
RNA-ViT (Ours) (w/o KD)	RNA	`HOP`	`LeakyReLU`	93.78	**14.04**	6.68	76.79	**14.13**	5.43	61.96	**44.27**	1.40
RNA-ViT (Ours) (w/ KD)	RNA	`HOP`	`LeakyReLU`	94.97	**14.04**	**6.76**	**79.04**	**14.13**	**5.59**	**65.86**	**44.27**	**1.49**

∗ A mix of `2ReLU` and `scaling`.
∗ The italic accuracy values are taken from the paper [1]

presents significantly lower PI latency than EA ViT [32] while increasing accuracy by more than 2% on all three datasets. Compared to MPCFormer [19], RNA-ViT achieves up to 3.64×, 3.61×, and 1.68× lower latency on CIFAR-10, CIFAR-100, and Tiny-ImageNet, respectively. As MPCViT [1] can configure the trade-off between PI latency and accuracy, we compare RNA-ViT to the MPCViT variants with the lowest latency and highest accuracy, respectively. For the lowest latency, RNA-ViT further lowers latency by 3.44×, 3.45×, and 1.60× while increasing accuracy by 0.7%, 1.28%, 2.41%, on CIFAR-10, CIFAR-100, and Tiny-ImageNet, respectively. Compared to the MPCViT with knowledge distillation, RNA-ViT achieving an average increase of 0.93%, 2.04% and 2.73% higher accuracy and an average of 3.53×, 3.54×, 1.66× lower PI latency on CIFAR-10, CIFAR-100, and Tiny-ImageNet, respectively.

Figures 7 show that RNA-ViT provides a better accuracy-latency trade-off and significantly higher `A2L` than the existing alternatives.

C. Latency evaluation and analysis

The overall latency of private inference, assuming sequential execution, can be modeled as follows:

$$T = (N_{lin} * t_{lin} + N_{sm_{eq}} * t_{sm_{eq}} + N_{nl} * t_{nl}), \qquad (10)$$

where N_{lin}, $N_{sm_{eq}}$ and N_{nl} represent the total number of linear, softmax equivalent, and nonlinear operations required to perform a single forward pass. t_{lin}, $t_{sm_{eq}}$ and t_{nl} denote the execution time per linear operation, per softmax equivalent operation, and per nonlinear operation, respectively. For the baseline SA shown in Figure 1, the latency for `softmax` operation is 33.52s, and number of `softmax` operations $N_{sm_{eq}}$ is 1.8M. Thus the $t_{sm_{eq}}$ for `softmax` is 18.27μs. Similarly, we calculated t_{lin} and t_{nl} to be 0.015μs and 23.54μs respectively.

We also conduct an analysis to demonstrate the breakdown of PI latency across different numbers of attention heads. The results are presented in Figure 8. It demonstrates that with the incorporation of our proposed attention module RNA with `HOP` softmax approximation, RNA-ViT achieves a significant reduction in the PI latency of the `softmax` operation, which is a dominant factor in regular ViTs. Moreover, the PI latency stemming from the `GeLU` nonlinearity is also substantially reduced. Importantly, as the number of attention heads increases, the overall PI latency remains consistently lower than that of the baseline and existing alternatives.

D. Ablation studies

1) Importance of RNA: To assess the importance of the RNA attention module in RNA-ViT, we conducted an ablation study by replacing the RNA module with an SA module. The corresponding accuracy and latency results are presented in the first row of Table III. Comparing the complete RNA-ViT with RNA-ViT using SA, we observe that the proposed RNA attention module achieves comparable accuracies while achieving 2.26×, 2.23×, and 1.43× lower latencies and 2.23×, 2.20×, and 1.44× higher `A2L` on CIFAR-10, CIFAR-100, and Tiny-ImageNet, respectively. This demonstrates the value and effectiveness of the RNA attention module in improving the PI efficiency of RNA-ViT.

2) Importance of `HOP`: To evaluate the significance of the `HOP` softmax approximation in RNA-ViT, we replace `HOP` with the original `softmax` and present the results in the second row of Table III. It is evident that `HOP` assists in achieving comparable accuracies while reducing the PI latency of RNA-ViT with `softmax` by 13.65%, and 14.67% on CIFAR-10

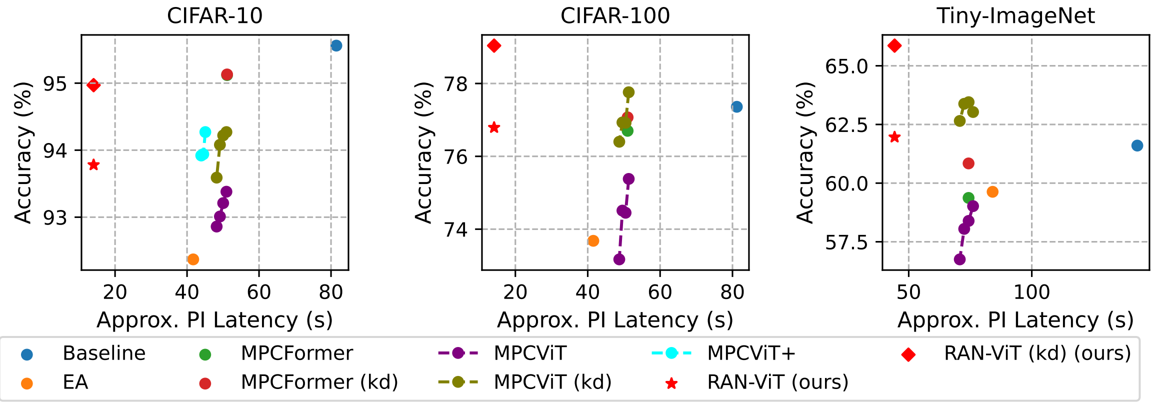

Fig. 7. Comparison between prior arts (Baseline, EA ViT [32], MPCFormer [19], MPCViT [1]) and RNA-ViT on CIFAR-10, CIFAR-100 and Tiny-ImageNet.

TABLE III
RESULTS OF ABLATION STUDIES
("✓" INDICATES THAT THE MODULE IS APPLIED, WHILE "✗" INDICATES THAT THE MODULE IS NOT APPLIED. FOR THE APPLIED REPLACEMENTS, THEY ARE MENTIONED IN PARENTHESES)

Model	Ablation			CIFAR-10			CIFAR-100			Tiny-ImageNet		
	RNA	`HOP`	`LeakyReLU`	Acc. (%)	Approx. PI Lat. (s)	A2L	Acc. (%)	Approx. PI Lat. (s)	A2L	Acc. (%)	Approx. PI Lat. (s)	A2L
Ablated RNA-ViT	✗(SA)	✓	✓	95.34	31.76	3.00	78.05	31.58	2.47	61.25	63.20	0.97
Ablated RNA-ViT	✓	✗(`softmax`)	✓	94.26	16.26	5.80	75.83	16.56	4.58	61.14	47.54	1.29
Ablated RNA-ViT	✓	✓	✗(`GeLU`)	94.26	34.33	2.75	76.23	34.18	2.23	60.71	59.17	1.03

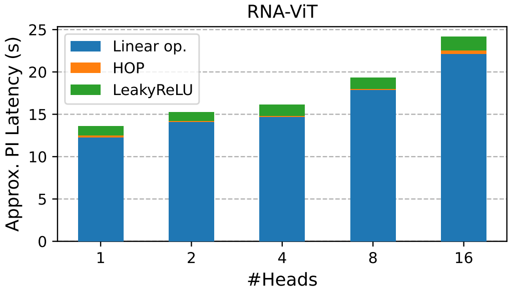

Fig. 8. PI latency breakdown of RNA-ViT for different numbers of heads

and CIFAR-100 respectively. Despite the compression and reduction in the size of the attention map due to the proposed RNA attention, we still observe a significant contribution of `HOP` in reducing PI latency.

3) Importance of `LeakyReLU`*:* To quantify the importance of `LeakyReLU` in RNA-ViT, we replace it with `GeLU` and report the results in the third row of Table III. The results reveal that `LeakyReLU` achieves comparable accuracies while reducing the PI latency by 59.10%, 58.66%, and 25.18%, and exhibiting $2.43\times$, $2.43\times$, and $1.36\times$ higher `A2L` on CIFAR-10, CIFAR-100, and Tiny-ImageNet, respectively. These outcomes align with our analysis in Section III-F, which indicates `LeakyReLU` is the most PI efficient nonlinearity and well-suited for PI applications.

E. Visualization of attention maps

In Figure 9, we provide visualizations of the attention maps at the last layer for different attention modules, including prior approaches and RNA variants. In particular, we leverage grad-CAM [37] to generate the heat-map plot for the attention maps. These attention maps correspond to randomly selected images. It is clear that both MPCFormer and SA, which employ the same attention architecture, exhibit similar heat maps generated in the attention maps across the majority of images. Notably, our complete RNA-ViT, which incorporates RNA, HOP, and LeakyReLU, shows attention maps of high quality despite the reduced dimensionality of the compressed attention maps. These attention maps in RNA-ViT closely resemble those generated by SA. The preservation of attention quality in complete RNA-ViT provides valuable insights into its comparable performance with SA, further highlighting its effectiveness as a viable alternative in the context of PI.

V. CONCLUSIONS AND FUTURE WORK

In this paper, we propose RNA-ViT, a novel family of models that addresses the challenges of PI in ViTs. RNA-ViT utilizes an attention module composed of a novel attention architecture and a softmax approximation, both of which are specifically designed to reduce PI latency. Additionally, we propose a new evaluation metric, `A2L`, that facilitates the assessment of ViT and other neural network modules in PI applications. Through our analysis, we identify `LeakyReLU` as the most PI-efficient nonlinearity. Our extensive experiments demonstrate

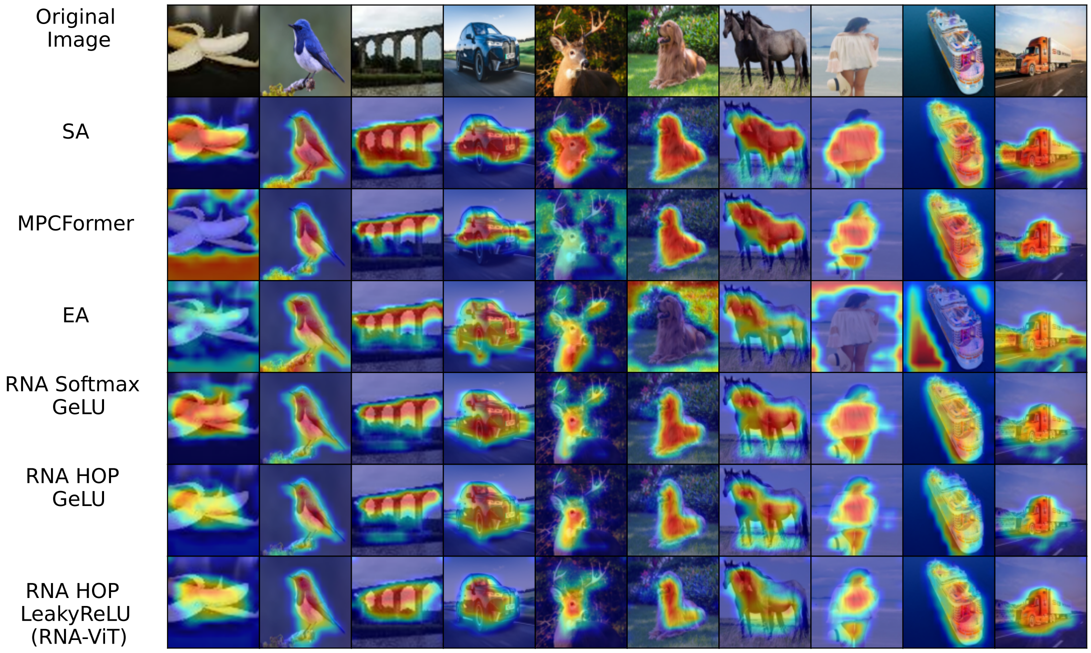

Fig. 9. Attention maps Grad-CAM visualization and comparison of prior works (Baseline SA, MPCFormer [19], EA [32]) with RNA-ViT variants.

the effectiveness of RNA-ViT in reducing PI latency and improving `A2L`. To the best of our knowledge, RNA-ViT sets a new state of the art in PI on ViT. Note that research on improving MPC protocols for ViT, e.g., Iron [25], is orthogonal to our work, and can be applied on top of RNA-ViT.

Based on the insights gained from the results presented in Figure 8, we observe that linear components currently dominate the PI latency. Therefore, our future work includes architectural optimization for reduced linear as well as non-linear operations. We also plan to extend the application of RNA-ViT to other tasks including object detection and semantic segmentation.

References

[1] W. Zeng, M. Li, W. Xiong, W. Lu, J. Tan, R. Wang, and R. Huang, "MPCViT: Searching for MPC-friendly vision transformer with heterogeneous attention," *arXiv preprint arXiv:2211.13955*, 2022.

[2] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.

[3] A. Hassani, S. Walton, N. Shah, A. Abuduweili, J. Li, and H. Shi, "Escaping the big data paradigm with compact transformers," 2021. [Online]. Available: https://arxiv.org/abs/2104.05704

[4] S. Kundu, M. Nazemi, P. A. Beerel, and M. Pedram, "DNR: a tunable robust pruning framework through dynamic network rewiring of dnns," in *Proceedings of the 26th Asia and South Pacific Design Automation Conference*, 2021, pp. 344–350.

[5] 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.

[6] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," *arXiv preprint arXiv:2103.14030*, 2021.

[7] S. Zheng, J. Lu, H. Zhao, X. Zhu, Z. Luo, Y. Wang, Y. Fu, J. Feng, T. Xiang, P. H. Torr *et al.*, "Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 6881–6890.

[8] E. Xie, W. Wang, Z. Yu, A. Anandkumar, J. M. Alvarez, and P. Luo, "SegFormer: Simple and efficient design for semantic segmentation with transformers," *arXiv preprint arXiv:2105.15203*, 2021.

[9] S. Kundu, Q. Sun, Y. Fu, M. Pedram, and P. A. Beerel, "Analyzing the confidentiality of undistillable teachers in knowledge distillation," *Advances in Neural Information Processing Systems*, vol. 34, pp. 9181–9192, 2021.

[10] H. Ma, T. Chen, T.-K. Hu, C. You, X. Xie, and Z. Wang, "Undistillable: Making a nasty teacher that cannot teach students," *arXiv preprint arXiv:2105.07381*, 2021.

[11] P. Mohassel and Y. Zhang, "SecureML: A system for scalable privacy-preserving machine learning," in *2017 IEEE symposium on security and privacy (SP)*. IEEE, 2017, pp. 19–38.

[12] C. Juvekar, V. Vaikuntanathan, and A. Chandrakasan, "GAZELLE: A low latency framework for secure neural network inference," in *27th USENIX Security Symposium (USENIX Security 18)*, 2018, pp. 1651–1669.

[13] P. Mishra, R. Lehmkuhl, A. Srinivasan, W. Zheng, and R. A. Popa, "DELPHI: A cryptographic inference service for neural networks," in *29th USENIX Security Symposium (USENIX Security 20)*, Aug. 2020.

[14] S. Tan, B. Knott, Y. Tian, and D. J. Wu, "CryptGPU: Fast privacy-preserving machine learning on the GPU," in *2021 IEEE Symposium on Security and Privacy (SP)*. IEEE, 2021, pp. 1021–1038.

[15] Z. Huang, W. jie Lu, C. Hong, and J. Ding, "Cheetah: Lean and fast secure two-party deep neural network inference," Cryptology ePrint Archive, Paper 2022/207, 2022.

[16] L. Shen, Y. Dong, B. Fang, J. Shi, X. Wang, S. Pan, and R. Shi, "ABNN2: secure two-party arbitrary-bitwidth quantized neural network predictions,"

in *Proceedings of the 59th ACM/IEEE Design Automation Conference*, 2022, pp. 361–366.

[17] S. Kundu, S. Lu, Y. Zhang, J. Liu, and P. A. Beerel, "Learning to linearize deep neural networks for secure and efficient private inference," *International Conference on Learning Representation*, 2023.

[18] M. Cho, A. Joshi, B. Reagen, S. Garg, and C. Hegde, "Selective network linearization for efficient private inference," in *International Conference on Machine Learning*. PMLR, 2022, pp. 3947–3961.

[19] D. Li, R. Shao, H. Wang, H. Guo, E. P. Xing, and H. Zhang, "MPCFormer: fast, performant and private Transformer inference with MPC," 2022.

[20] E. Chou, J. Beal, D. Levy, S. Yeung, A. Haque, and L. Fei-Fei, "Faster CryptoNets: Leveraging sparsity for real-world encrypted inference," *arXiv preprint arXiv:1811.09953*, 2018.

[21] X. Wang, R. Girshick, A. Gupta, and K. He, "Non-local neural networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 7794–7803.

[22] M. S. Riazi, M. Samragh, H. Chen, K. Laine, K. Lauter, and F. Koushanfar, "XONN: XNOR-based oblivious deep neural network inference," in *28th USENIX Security Symposium (USENIX Security 19)*, 2019, pp. 1501–1518.

[23] Y. Zhang, D. Chen, S. Kundu, H. Liu, R. Peng, and P. A. Beerel, "C2PI: An efficient crypto-clear two-party neural network private inference," in *Proceedings of the 60th ACM/IEEE Design Automation Conference*, 2023.

[24] T. Chen, H. Bao, S. Huang, L. Dong, B. Jiao, D. Jiang, H. Zhou, and J. Li, "The-x: Privacy-preserving transformer inference with homomorphic encryption," *Findings of ACL*, pp. 3510–3520, 2022.

[25] M. Hao, H. Li, H. Chen, P. Xing, G. Xu, and T. Zhang, "Iron: Private inference on transformers," in *Advances in Neural Information Processing Systems*, 2022.

[26] N. K. Jha, Z. Ghodsi, S. Garg, and B. Reagen, "DeepReDuce: ReLU reduction for fast private inference," in *International Conference on Machine Learning*. PMLR, 2021, pp. 4839–4849.

[27] S. Kundu, Y. Zhang, D. Chen, and P. A. Beerel, "Making models shallow again: Jointly learning to reduce non-linearity and depth for latency-efficient private inference," in *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*. Efficient Deep Learning for Computer Vision (ECV), 2023.

[28] S. Wang, B. Z. Li, M. Khabsa, H. Fang, and H. Ma, "Linformer: Self-attention with linear complexity," *arXiv preprint arXiv:2006.04768*, 2020.

[29] Z. Qin, W. Sun, H. Deng, D. Li, Y. Wei, B. Lv, J. Yan, L. Kong, and Y. Zhong, "cosFormer: Rethinking softmax in attention," in *International Conference on Learning Representations*, 2022.

[30] Z. Geng, M.-H. Guo, H. Chen, X. Li, K. Wei, and Z. Lin, "Is attention better than matrix decomposition?" *International Conference on Learning Representations*, 2021.

[31] J. Lu, J. Yao, J. Zhang, X. Zhu, H. Xu, W. Gao, C. Xu, T. Xiang, and L. Zhang, "Soft: Softmax-free transformer with linear complexity," *Advances in Neural Information Processing Systems*, vol. 34, pp. 21 297–21 309, 2021.

[32] M.-H. Guo, Z.-N. Liu, T.-J. Mu, and S.-M. Hu, "Beyond self-attention: External attention using two linear layers for visual tasks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022.

[33] M. Crisfield, "A faster modified Newton-Raphson iteration," *Computer methods in applied mechanics and engineering*, vol. 20, no. 3, pp. 267–278, 1979.

[34] B. Knott, S. Venkataraman, A. Hannun, S. Sengupta, M. Ibrahim, and L. van der Maaten, "CRYPTEN: Secure multi-party computation meets machine learning," *Advances in Neural Information Processing Systems*, vol. 34, pp. 4961–4973, 2021.

[35] A. Katharopoulos, A. Vyas, N. Pappas, and F. Fleuret, "Transformers are RNNs: Fast autoregressive transformers with linear attention," in *International Conference on Machine Learning*. PMLR, 2020, pp. 5156–5165.

[36] G. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," *arXiv preprint arXiv:1503.02531*, 2015.

[37] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-cam: Visual explanations from deep networks via gradient-based localization," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 618–626.