MIA-Former: Efficient and Robust Vision Transformers via Multi-Grained Input Adaptation

Zhongzhi Yu[1], **Yonggan Fu**[1], **Sicheng Li**[2], **Chaojian Li**[1], **Yingyan Lin**[1]

[1] Department of Electrical and Computer Engineering, Rice University [2] Alibaba DAMO Academy

Abstract

Vision transformers (ViTs) have recently demonstrated great success in various computer vision tasks, motivating a tremendously increased interest in their deployment into many real-world IoT applications. However, powerful ViTs are often too computationally expensive to be fitted onto real-world resource-constrained devices, due to (1) their quadratically increased complexity with the number of input tokens and (2) their overparameterized self-attention heads and model depth. In parallel, different images are of varying complexity and their different regions can contain various levels of visual information, e.g., a sky background is not as informative as a foreground object in object classification tasks, indicating that treating all regions/tokens equally in terms of model complexity is unnecessary while such opportunities for trimming down ViTs' complexity have not been fully explored. To this end, we propose a **M**ulti-grained **I**nput-**A**daptive Vision Trans**Former** framework dubbed **MIA-Former** that can input-adaptively adjust the structure of ViTs at three coarse-to-fine-grained granularities (i.e., model depth and the number of model heads/tokens). In particular, our MIA-Former adopts a low-cost network trained with a hybrid supervised and reinforcement training method to skip unnecessary layers, heads, and tokens in an input adaptive manner, reducing the overall computational cost. Furthermore, an interesting side effect of our MIA-Former is that its resulting ViTs are naturally equipped with improved robustness against adversarial attacks over their static counterparts, because MIA-Former's multi-grained dynamic control improves the model diversity similar to the effect of ensemble and thus increases the difficulty of adversarial attacks against all its sub-models. Extensive experiments and ablation studies validate that the proposed MIA-Former framework can (1) effectively allocate computation budgets adaptive to the difficulty of input images, achieving state-of-the-art (SOTA) accuracy-efficiency trade-offs, e.g., 20% computation savings with the same or even a higher accuracy compared with SOTA dynamic transformer models, and (2) boost ViTs' robustness accuracy under various adversarial attacks over their vanilla counterparts by 2.4% and 3.0%, respectively. Our code is available at https://github.com/RICE-EIC/MIA-Former.

1 Introduction

Vision transformers (ViTs) have been proven to be a powerful architecture on various computer vision tasks (Dosovitskiy et al. 2020; Touvron et al. 2021; Chen et al. 2021a; Caron et al. 2021; Strudel et al. 2021), especially when being scaled up with larger model sizes and more training data (Dosovitskiy et al. 2020; Steiner et al. 2021; Ridnik et al. 2021). However, powerful ViTs often come with prohibitive computational overhead. Specifically, (1) the tokens consisting of merely background information and (2) redundant heads within the multi-head self-attention (MSA) module of ViTs which learn similar features can lead to unnecessary yet non-negligible inference costs. Taking the widely used DeiT-Small model (Touvron et al. 2021) as an example, running inference on a single image with a resolution of 224×224 requires over 4.6 Giga floating-point operations (GFLOPs), making it challenging to deploy ViTs onto many real-world resource-constrained devices for supporting intelligent internet of things (IoT) applications. Thus, there is an urgent need to reduce the computational cost of ViTs.

On the other hand, in real-world applications, the complexity of images can vary significantly. As such, processing all the images with the same model complexity of ViTs could be overcooked. For example, for most of the time during video surveillance, the video may be just staring at an empty background, and the corresponding images can be processed with a naively simple model, saving a large portion of computational cost while still achieving satisfying accuracy. Thus, a straightforward solution for trimming down ViTs' complexity is to perform input-adaptive dynamic inference. Although dynamic inference has been extensively explored for convolutional neural networks (CNNs) through various dynamic dimensions (e.g., model depth, channel number, and model bit-width) (Hu et al. 2020; Wang et al. 2018; Shen et al. 2020; Wang et al. 2020), only a few pioneering works have considered this aspect for ViTs (Rao et al. 2021; Wang et al. 2021b), which yet merely focus on reducing the computational budget by adaptively adjusting the number of input tokens. However, as suggested in (Zhou et al. 2021), the similarity between heads and feature maps can increase significantly in deeper ViT layers, implying that the token dimension is not the only source of redundancy, and the unexplored depth and head dimensions could lead to more efficient ViTs.

To this end, we aim to fully explore the redundancy in ViTs and make the following contributions:

- We propose a **M**ulti-grained **I**nput-**A**daptive vision trans**Former** framework, dubbed MIA-Former, in order to trim down the redundancy of ViTs from multiple dimensions at three coarse-to-fine-grained granularities.
- We propose a low-cost MIA-Controller to make input-adaptive decisions, which is jointly trained with the ViT models via a hybrid supervised and reinforcement learning (RL) scheme.
- We empirically find that thanks to the proposed hybrid supervised and reinforcement training method, MIA-Former is equipped with improved robustness to various types of adversarial attacks, achieving a win-win in both robustness and efficiency.
- Extensive experiments and ablation studies based on both DeiT-based (Touvron et al. 2021) and LeViT-based (Graham et al. 2021) models show that the proposed MIA-Former can be used as a plug-in module on top of a wide range of ViTs to achieve better accuracy-efficiency trade-offs and boosted adversarial robustness, compared with state-of-the-art (SOTA) vanilla ViTs, input-adaptive ViTs, as well as CNNs. Specifically, MIA-Former achieves a 20.1% FLOPs reduction and 2.4% higher robustness accuracy under Projected Gradient Descent (PGD) (Kurakin, Goodfellow, and Bengio 2018) attacks together with the same natural accuracy, compared with the original DeiT-Small model.

2 Related Works

Vision Transformers. Transformers are first introduced to natural language processing tasks in (Vaswani et al. 2017). It has been shown that the self-attention module in transformers can serve as an effective way to model the token-wise relationship of sentences. (Dosovitskiy et al. 2020) then proposes ViTs, which is a pioneering work to extend transformer architectures to large scale compute vision tasks and matches SOTA CNNs' performance. Specifically, ViTs first split an input image into a series of patches which are embedded into tokens before passing into the ViT blocks; Each ViT block consists of a stacked MSA and multi-layer perceptron (MLP) module to extract the global relationship among input tokens; multiple heads further enable ViT blocks to extract different features via different heads, improving the expressiveness of ViTs. The success of ViTs on large scale image recognition tasks (e.g., ImageNet dataset (Deng et al. 2009)) has inspired a series of following works to further exploit the expressive power of ViTs from different perspectives: (Touvron et al. 2021) performs an exhaustive search for the optimal training recipe for training ViTs; (Liu et al. 2021) proposes hierarchical structures for ViTs like ResNet (He et al. 2016), and (Dong et al. 2021) modifies the shape of patches from square to cross to balance the global and local attentions. Among them, LeViT (Graham et al. 2021) achieves impressive performance by exploring the potential of applying convolutional layers before ViTs. Specifically, LeViT is the first ViT network that achieves a better accuracy-efficiency trade-off than EfficientNet (Tan and Le 2019) under certain FLOPs ranges.

Input-adaptive Inference. Adaptively activating different components of a deep neural network (DNN) in an input-dependent manner has been proved to be an effective way to reduce the inference cost, which has been widely explored for CNNs. Existing techniques in this regard can be roughly summarized into two categories: (1) making early predictions by introducing multiple side branch classifiers and dynamically exiting from one branch by analyzing the confidence of intermediate feature maps (Kaya, Hong, and Dumitras 2019; Teerapittayanon, McDanel, and Kung 2016) and (2) adaptively skipping specific components of the model, such as blocks, channels and even bit-width (Wu et al. 2018; Wang et al. 2020; Shen et al. 2020; Fu et al. 2020). Nevertheless, the opportunities of input-adaptive inference have not yet been extensively explored in the scope of ViTs as existing works merely focus on dynamically adjusting the input tokens, either by pruning out certain tokens (Rao et al. 2021) or by changing the input patch size (Wang et al. 2021b). Other design dimensions in ViTs are still neglected, such as the number of attention heads and model depth, which play important roles in the overparameterization of ViTs.

Adversarial Robustness and Model Efficiency. DNNs' robustness and efficiency are two critical features required in real-world applications. Some pioneering works are trying to optimize both for CNNs simultaneously. For example, (Rakin et al. 2019; Ye et al. 2019; Sehwag et al. 2020; Fu et al. 2021b) combine pruning techniques with adversarial training methods. (Fu et al. 2021a) leverages the poor adversarial transferability between different precisions to win both robustness and efficiency, which can be accelerated by customized accelerators (Fu et al. 2021c) for further boosted efficiency. And (Rakin et al. 2018) uses dynamic quantization of activation functions to defend against adversarial examples; (Hu et al. 2020) introduces input-adaptive inference for simultaneously boosting robustness and efficiency. However, existing works have shown that CNNs and ViTs behave differently under adversarial attacks (Shao et al. 2021; Mao et al. 2021), and how to win both adversarial robustness and model efficiency for ViTs is still an open question.

3 The Proposed MIA-Former Framework

3.1 Overview

Fig. 1 illustrates an overview of the proposed MIA-Former framework. On top of vanilla ViTs, MIA-Former integrates a controlling module, dubbed MIA-Controller, to each of its building blocks, i.e., MIA-Blocks. Specifically, the proposed MIA-Controller first decides whether to mask out a given ViT block at a coarse granularity, as shown in Fig. 1(a). In such cases, the outputs of a previous block skip the current block and are directly fed into the next MIA-Controller and MIA-Block. On the other hand, if the current MIA-Block is not skipped as a whole, MIA-Former further uses the corresponding MIA-Controller to dynamically mask out certain tokens and heads to deactivate the corresponding modules during inference in an input-dependent manner, as shown in Fig. 1(b).

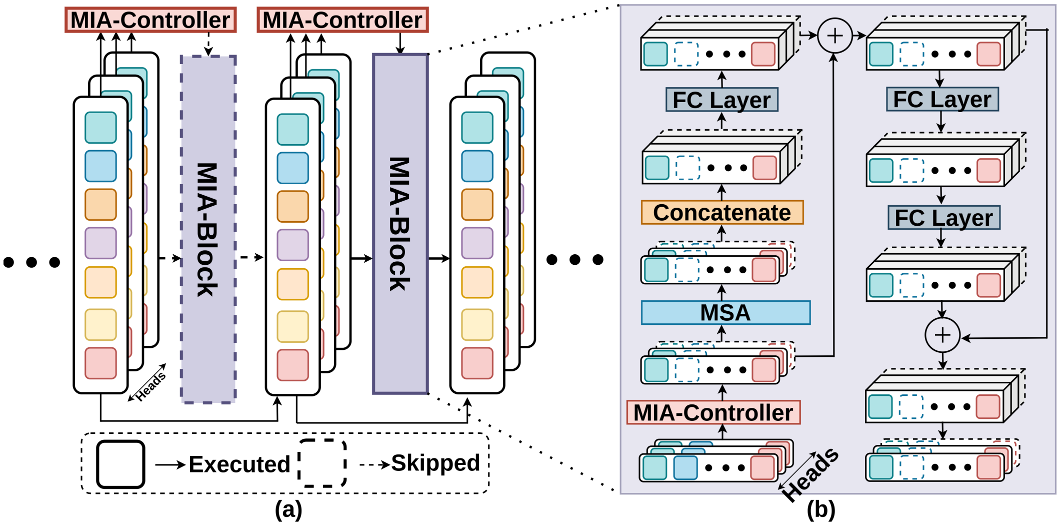

Figure 1: Overview of the proposed MIA-Former framework integrating a MIA-Controller and MIA-Block modules: (a) The MIA-Controller first decide whether to skip the whole upcoming MIA-Block, and (b) if the upcoming MIA-Block is not fully skipped, its MIA-Controller further dynamically masks out certain tokens and heads to deactivate the corresponding modules and thus reduce computational cost.

In vanilla ViT blocks, reducing the number of heads can only lead to linear FLOPs reduction in MSA modules; moreover, the computational overhead in the upcoming MLP module cannot be reduced, which is often the computational bottleneck. Thus, we propose to further proportionally reduce both the input and output size of the upcoming fully connected (FC) layers in the MIA-Block module. In this way, MIA-Former can lead to linear FLOPs reduction in the MSA module and nearly quadratic FLOPs reduction in the following MLP module when skipping certain heads during inference, which can lead to large savings of FLOPs in MIA-Former. Note that a skip connection is also added to pass the information of masked heads and tokens to the outputs of the next block, avoiding the permanent loss of information.

3.2 MIA-Controller

The key characteristic of the proposed MIA-Former is to dynamically adjust its structure and thus the complexity at three coarse-to-fine-grained granularities (i.e., model depth, number of heads, and number of tokens). To push forward the trade-offs between efficiency and accuracy, MIA-Former adopts a lightweight controller to adaptively generate masks for skipping the corresponding uninformative blocks/heads/tokens on top of its vanilla ViT backbone model. However, it is non-trivial to derive such finer-grained learnable masks due to the large skipping policy space. To tackle this problem, we propose a lightweight controller, i.e., MIA-Controller, to generate the masks and thus corresponding skipping policy. Specifically, MIA-Controller first examines whether a ViT block shall be completely skipped; if not, it will skip certain heads and tokens of this ViT block based on the input feature complexity. The merit of introducing the above block-wise skipping is that it eliminates the necessity of the additional efforts for computing head- and token-wise skipping policy.

For each block l in ViTs, we maintain three binary masks, $D_b^l \in \{0,1\}$, $D_h^l \in \{0,1\}^H$, and $D_n^l \in \{0,1\}^N$ for skipping blocks, heads, and tokens, respectively, where H and N denote the number of heads and tokens, respectively. In particular, a MIA-Controller is inserted ahead of each block l with the outputs of the previous MIA-Block $I^l \in \mathcal{R}^{N_h \times N_w \times (HE)}$ as the inputs, where $N = N_h N_w$ is the spatial dimension of the token array and E is the hidden dimension of each head. The MIA-Controller first passes I^l through a two-layer CNN (CNN$_b$) with pooling to extract the features F_b^l which are then passed through an FC layer (FC$_b$) with Gumbel softmax (Jang, Gu, and Poole 2016) to generate D_b^l, deciding whether to skip the whole block. Such a skipping pipeline can be formulated as:

$$\begin{aligned} F_b^l &= \text{CNN}_b(I^l) \in \mathcal{R}^{1\times 1\times HE'}, \\ G_b^l &= \text{FC}_b(F_b^l) \in \mathcal{R}, \\ D_b^l &= \text{Round}(G_b^l) \in \{0,1\}, \end{aligned} \tag{1}$$

where $E' < E$ is the hidden dimension of each head after the CNN with pooling in the MIA-Controller, and we set $E' = E/4$ in this work. $Round(\cdot)$ is a rounding operation that rounds the value to the nearest integer.

If a ViT block is not skipped, the MIA-Controller further generates a token mask D_n^l and a head mask D_h^l for determining the skipping policy for the corresponding block's heads and tokens, respectively. For D_h^l, an FC layer (FC$_{h1}$) accepts F_b^l as its inputs to extract the head features F_h^l, and then another FC layer (FC$_{h2}$) with Gumbel softmax is adopted to generate the mask D_h^l:

$$\begin{aligned} F_h^l &= \text{FC}_{h1}(F_b^l) \in \mathcal{R}^{H \times E''} \\ G_h^l &= \text{FC}_{h2}(F_h^l) \in \mathcal{R}^{H} \\ D_h^l &= \text{Round}(G_h^l) \in \{0,1\}^H. \end{aligned} \tag{2}$$

For D_n^l, MIA-Former follows a similar procedure as (Rao et al. 2021). Specifically, I^l is reshaped to $I^{l'} \in \mathcal{R}^{N \times (HE)}$ and then applied to a two-layer MLP (MLP$_n$) to extract the features F_n^l. After that, an FC layer (FC$_n$) with Gumbel softmax is used to process F_n^l and generate the masks D_n^l.

$$\begin{aligned} F_n^L &= \text{MLP}_n(I^{l'}) \in \mathcal{R}^{N \times HE'} \\ G_n^l &= \text{FC}_n(F_n^l) \in \mathcal{R}^{N} \\ D_n^l &= \text{Round}(G_n^l) \in \{0,1\}^N. \end{aligned} \tag{3}$$

The featuremap I_F^l after the MIA-Controller to the MIA-Block l are computed as:

$$\begin{aligned} I_F^l = & I^l.\text{reshape}(N_h, N_w, H, E) \odot \\ & D_b^l.\text{reshape}(1,1,1,1) \odot D_h^l.\text{reshape}(1,1,H,1) \odot \\ & D_n^l.\text{reshape}(N_h, N_w, 1, 1), \end{aligned} \tag{4}$$

where $\odot$ represents the element-wise matrix multiplication (broadcast will be performed if the matrix shapes do not match).

3.3 Hybrid Supervised and Reinforcement Training

Given the complexity of MIA-Former, directly training all its components leads to unstable training and thus inferior performance. Thus, we propose a hybrid supervised and reinforcement training pipeline, which consists of three steps: (1) MIA-Controller pretraining, where we fix the parameters in the pretrained MIA-Block and pretrain the MIA-Controller until they keep activating all components of the MIA-Former without any skipping, (2) MIA-Former co-training, where we co-train the MIA-Block and the MIA-Controller with a hybrid loss function in a differentiable way, and (3) skipping policy finetuning with hybrid RL, where we finetune the MIA-Block and the MIA-Controller with a hybrid supervised and RL training method. We will illustrate the detail of the above stages in the remaining part of this section.

MIA-Controller Pretraining. As a randomly initialized MIA-Controller randomly skips different modules within the model, directly co-training the MIA-Block and MIA-Controller leads to inferior performance. We conjecture this is due to the fact that at the beginning of training, a large portion of the model is randomly skipped, leading to a significant deviation from the original learned distribution of pretrained ViTs. To tackle this, we propose to first pretrain the MIA-Controller with the MIA-Block weights fixed until it does not skip any components, and the whole MIA-Former should behave exactly the same as its backbone ViT at the end of this stage. To achieve this, we train the MIA-Controller with the pretraining loss defined as:

$$\mathcal{L}_{pretrain} = \sum_{l=0}^{L} [(1 - D_b^l) + (1 - D_h^l) + (1 - D_n^l)], \tag{5}$$

where L is the total number of blocks in a MIA-Former. When $\mathcal{L}_{pretrain}$ is minimized, the MIA-Controller pretraining stage is finished.

MIA-Former Co-training. The learning objective of MIA-Former is to reduce the computational cost while preserving the model accuracy. To this end, we define the training loss as follow,

$$\mathcal{L}_{diff} = \mathcal{L}_{task} + \alpha \mathcal{L}_{cost}, \tag{6}$$

where L_{task} is the task loss, L_{cost} is the computational cost loss, and α is a weighted factor that trades off importance between accuracy and computational budget. In this work, we define L_{cost} as

$$L_{cost} = \frac{\text{FLOPs}_{exec}}{\text{FLOPs}_{total}}, \tag{7}$$

where FLOPs$_{exec}$ and FLOPs$_{total}$ are the total FLOPs of the executed parts in a MIA-Former and the total FLOPs when executing the whole MIA-Former model.

To ensure that the MIA-Former can adaptively allocate computational budget among all the input samples, instead of setting α to a fixed value, we dynamically change the sign of α during training. Specifically, when FLOPs$_{exec}$ is larger than the given target FLOPs (FLOPs$_{target}$), we set $\alpha > 0$ to penalize MIA-Former for adapting to a smaller computation budget. On the other hand, when FLOPs$_{exec}$ is smaller than FLOPs$_{target}$, we set $\alpha < 0$ to encourage the model to execute more components in the MIA-Former.

Skipping Policy Finetuning with Hybrid RL. We further finetune the MIA-Former with a hybrid RL method to achieve better performance. Specifically, we use an A2C-based (Mnih et al. 2016) RL learning method to train the MIA-Controller and MIA-Block modules in a differentiable manner. Furthermore, to make the MIA-Controller compatible with the above A2C algorithm, we replace the last FC layer in each of the skipping dimensions (i.e., depth, heads, and tokens) in a MIA-Controller with two parallel FC layers as the actor and critic network in the A2C algorithm, respectively, to construct a simple RL agent. To guarantee stable training, we further inherit all trained weights from the MIA-Block and 75% of weights in the remaining parts of the MIA-Controller. In this way, the newly introduced RL agent can start from the learned features from the trained MIA-Controller, while the MIA-Block is also adapted to the dynamically skipping mechanism. The reward function of RL is defined as:

$$\mathcal{R} = Y + \beta(\text{FLOPs}_{target} - \text{FLOPs}_{exec}) \tag{8}$$

where Y is a binary value indicating the task accuracy, and β is a weighted factor that balances ViTs' model accuracy and efficiency.

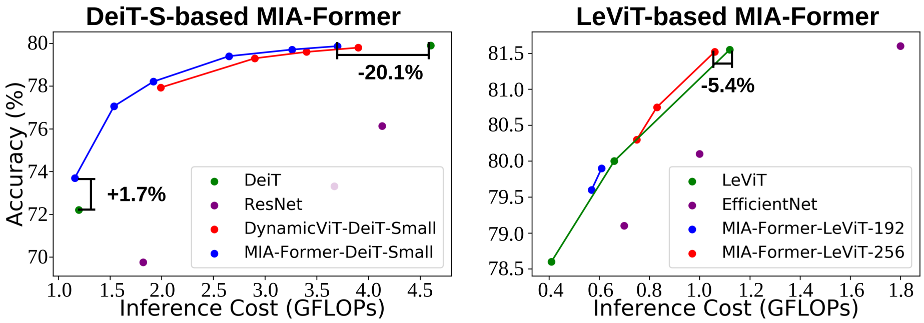

Figure 2: Benchmarking model performance (top-1 accuracy) and inference efficiency (FLOPs) trade-off, where we compare our proposed MIA-Former with SOTA dynamic ViTs (DynamicViT) and other SOTA image classification ViTs/CNNs on the ImageNet dataset.

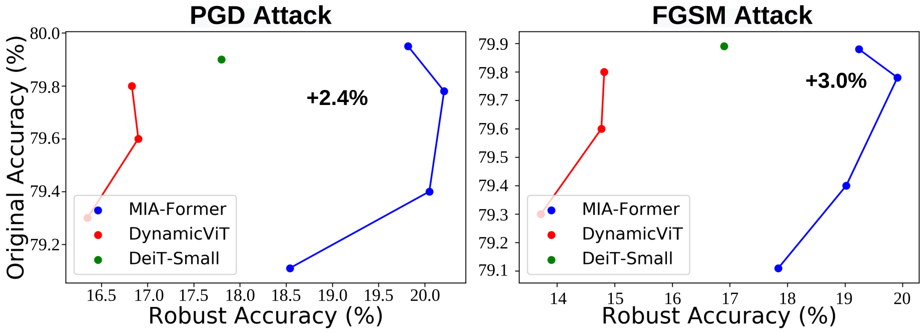

Figure 3: Comparison of MIA-Former with DynamicViT and its backbone transformer models on the achieved ImageNet top-1 accuracy and robustness accuracy under both PGD and FGSM attacks.

To make the MIA-Block compatible with RL-based MIA-Controller, we also co-train the MIA-Block along with MIA-Controller in this stage with the optimization objective defined as:

$$\min_{\omega} \mathcal{L}_{hybrid} = \mathcal{L}_{task} - \mathcal{R}, \tag{9}$$

where ω is the weight of the MIA-Block.

4 Experiment Results

In this section, we first introduce the experiment setup, including models, datasets, training hyperparameters, baselines, and evaluation metrics. Then, we validate the superiority of MIA-Former in boosting both efficiency and robustness. In particular, we find that (1) MIA-Former can boost the efficiency, i.e., achieving a better accuracy vs. efficiency trade-off than both vanilla models and models with the SOTA dynamic ViT technique, and (2) MIA-Former can boost the robustness, i.e., achieving higher adversarial robustness while preserving the accuracy compared with both vanilla ViTs and dynamic ViT variants. Furthermore, we perform ablation study on (1) the comparison of the redundancy between different dimensions of MIA-Former based on the relative accuracy gap when running MIA-Former with different subsets of the input-adaptive granularities, and (2) which training strategy for learning the skipping policy can win better adversarial robustness in MIA-Former. Finally, we visualize the skipping policies on a subset of input samples as well as the skipping ratio distributions of different modules in the proposed MIA-Former.

4.1 Settings

Models, Datasets, and Baselines. We evaluate our proposed MIA-Former over **three ViT models** (i.e., DeiT-Small (Touvron et al. 2021), LeViT-192 and LeViT-256 (Graham et al. 2021)) on ImageNet-1K dataset (Deng et al. 2009). We benchmark our method with vanilla ViTs, SOTA dynamic ViT method DynamicViT (Rao et al. 2021) and other SOTA ViTs/CNNs designs. For adversarial robustness, we evaluate the proposed MIA-Former using Projected Gradient Descent (PGD) (Madry et al. 2017) attack under L_{Inf} constraint with a perturbation strength of 0.002 and Fast Signed Gradient Matching (FGSM) attack under L_2 constraint with a perturbation strength of 0.03 (Athalye, Carlini, and Wagner 2018), following the adversarial attack setting in (Chen et al. 2020).

Training Recipe. We adopt a three-stage training strategy: Stage 1: MIA-Controller pretraining: we use an Adam (Kingma and Ba 2014) optimizer with a learning rate of 1e-4 to train the MIA-Controller with fixed MIA-Block until $\mathcal{L}_{pretrain}$ is decreased to 0. Stage 2: MIA-Former co-training: we use an

Model	GFLOPs	Acc. (%)
MobileNet (Howard et al. 2017)	0.58	70.6
PVT-v2 (Wang et al. 2021a)	0.61	76.9
PiT-Ti (Heo et al. 2021)	0.70	74.6
EfficientNet-B1 (Tan and Le 2019)	0.70	79.1
LeViT-192 (Graham et al. 2021)	0.66	80.0
IPE (Chen et al. 2021b)	0.88	78.6
MIA-Former-LeViT-192	**0.61**	**79.9**
DeiT-Tiny (Touvron et al. 2021)	1.30	72.2
PVT-v2 (Wang et al. 2021a)	0.98	74.5
PiT-XS (Heo et al. 2021)	1.40	79.1
LocalViT (Li et al. 2021)	1.05	74.9
CoaT (Xu et al. 2021)	0.95	73.8
EfficientNet-B2 (Tan and Le 2019)	1.00	80.1
LeViT-256 (Graham et al. 2021)	1.12	81.5
MIA-Former-LeViT-256	**1.06**	**81.5**

Table 1: Comparing MIA-Former with SOTA ViTs/CNNs on ImageNet. We refer to MIA-Former with different backbones as MIA-Former-BACKBONE.

AdamW (Loshchilov and Hutter 2017) optimizer with a batch size of 1024 and a learning rate of 1e-5/1e-3 to train the MIA-Block/MIA-Controller, respectively, for 200 epochs. We set the α to $0.1 \times \frac{L_{cls}}{L_{cost}}$.
Stage 3: Skipping policy finetuning with hybrid RL: after inserting the RL agents, we first train the RL agent for 20 epochs with all other parameter fixed and then unfreeze other parameters and co-train the MIA-Former for a total of 50 epochs.

4.2 Benchmark with SOTA Designs

Enhanced accuracy-efficiency trade-off. We apply the MIA-Former framework on top of three ViT models, including DeiT-Small (one of the most widely used ViTs), LeViT-192, and LeViT-256, which are SOTA ViT models with the optimal accuracy-efficiency trade-off. As shown in Fig. 2, we observe that the MIA-Former can achieve a 20.1% reduction in FLOPs on top of vanilla DeiT-Small with comparable accuracy. We further benchmark the proposed MIA-Former with other SOTA CNNs and ViTs in Tab. 1 and show that the proposed MIA-Former pushes forward the frontier of the achievable accuracy-efficiency trade-off with a 5.4% reduction in FLOPs on top of LeViT-256 and a comparable accuracy.

Robustness improvement. We compare the adversarial robustness of MIA-Former with vanilla DeiT-Small and DynamicViT (Rao et al. 2021) models. As shown in Fig. 3, DynamicViT suffers from the reduction in robust accuracy in both cases of attacks, while our proposed method can achieve even higher model robustness compared with the original model, leading to a win-win in robustness and efficiency. Specifically, MIA-Former-DeiT-Small achieves up to a 2.4%/3.0% higher robust accuracy and a 26.1% less FLOPs with comparable natural accuracy compared with the original DeiT-Small model under PGD/FGSM attacks, respectively.

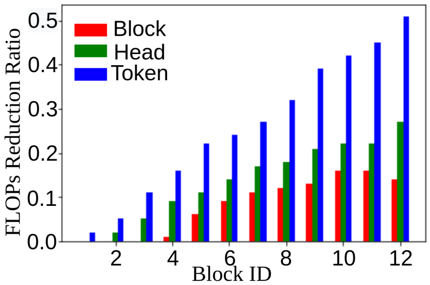

Figure 4: Visualizing the blockwise skipping ratios along different dynamic dimensions on MIA-Former-DeiT-Small.

Dynamic Dimension Head	 Depth	 Token	GFLOPs	Acc. (%)
			4.6	79.9
✓			3.9	78.7
	✓		4.1	76.2
		✓	3.8	79.7
✓	✓		4.0	78.6
✓		✓	3.7	79.8
	✓	✓	3.9	79.3
✓	✓	✓	**3.9**	**79.9**

Table 2: Ablation study on dynamic dimension combinations.

4.3 Ablation Study

Analysis of the redundancy along each dimension. To better understand the contribution of each dimension (i.e., head-wise, depth-wise, and token-wise) to the finally achieved accuracy-efficiency trade-off by MIA-Former, we conduct an ablation study by only enabling input-adaptive skipping at certain dimensions of MIA-Former on top of DeiT-Small. In Tab. 2, the first row is the vanilla DeiT-Small model without any input-adaptive mechanism and the last row is the proposed MIA-Former with input-adaptive skipping enabled along all dimensions. As shown in Tab. 2, by activating different dimensions, the performance of the MIA-Former varies significantly under similar inference FLOPs. Specifically, only activating dynamic depth suffers from 3.7% lower accuracy with 0.2 GFLOPs higher than the fully activated MIA-Former, indicating that the redundancy of the model in the depth-wise dimension is the lowest. On the other hand, even when only activating token-wise

	Target	Inherit Weight (%)			
	GFLOPs	50	75	100	Pretrain
Clean Acc. Robust Acc.	3.4	79.2 21.3	79.9 19.8	79.9 15.7	79.9 14.2
Clean Acc. Robust Acc.	3.9	78.3 22.8	79.6 20.2	79.6 17.5	79.5 16

Table 3: Ablation study on the effect of hybrid training on model's robustness.

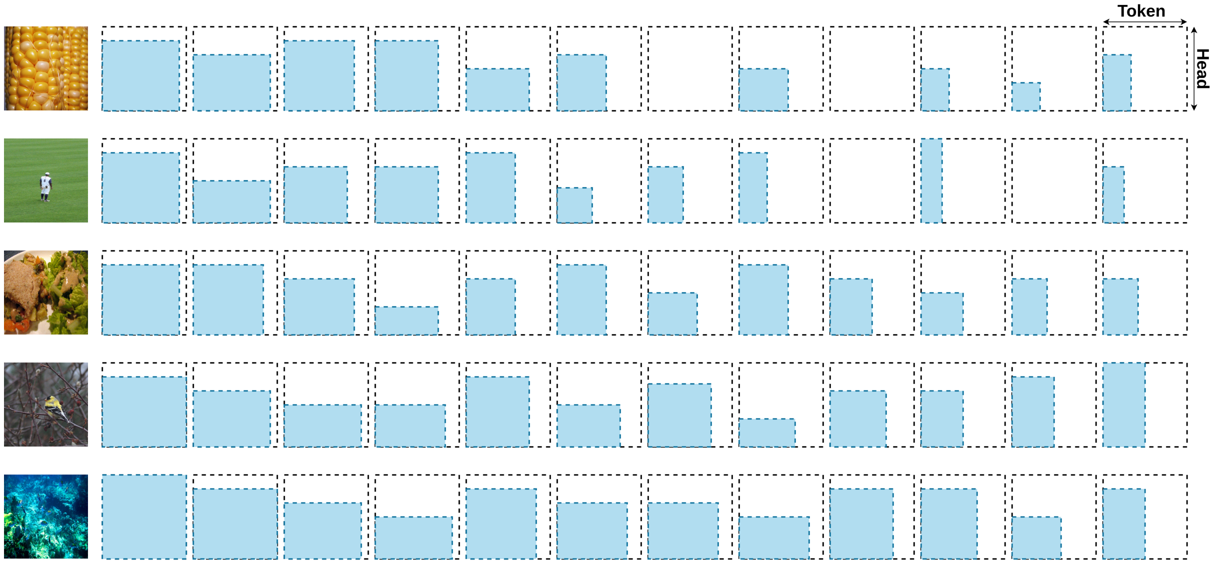

Figure 5: Visualizing MIA-Former's skipping policy on different input samples. Each square in the figure represents a block in MIA-Former, the height and width of the blue square indicate the executed number of heads and number of tokens, respectively. Dashed square without any color represents fully skipped block.

skipping, the accuracy only drops for 0.2% with comparable FLOPs compared to MIA-Former, indicating the high redundancy in this dimension. This observation aligns with our intuition that depth-wise skipping is of the most coarse granularity while token-wise skipping has the finest granularity in ViTs.

Effectiveness of finetuning the skipping policy with hybrid RL. How to inherit weights from the pretrained MIA-Controller when introducing the RL agent is critical to the finally achieved natural and robust accuracy. We study the impact of the portion of the inherited MIA-Controller weight at the beginning of the skipping policy finetuning stage on top of DeiT-Small under different FLOPs constraints. As shown in Tab. 3, there exists a trade-off between the natural and robust accuracy when using different weight inheriting strategies. In particular, inheriting a sufficient portion of pretrained MIA-Controller weights leads to significantly better natural accuracy. On the other hand, starting with a fully pretrained MIA-Controller without any reinitialization will degrade the robust accuracy. We suspect that this is because the inherited MIA-Controller makes the RL agent observe a similar feature as the differentiablly co-trained MIA-Controller in the previous training stage. The RL agent may try to make the decision in the same way as the differentiablly co-trained MIA-Controller. Thus, we pick a sweet spot from the trade-off and inherit 75% weights throughout the paper.

4.4 Skipping Policy Visualization

We first summarize the statistical characteristic of the generated skipping policy on the validation set of ImageNet-1k (Deng et al. 2009). As in Fig. 4, we have the following observations: (1) deeper blocks have significantly higher redundancy than the shallower layers among all skipping dimensions in MIA-Former, and (2) the token-wise dimension has the highest skipping probability compared with the depth-wise and head-wise dimension, which is consistent with the ablation study on different combinations of skipping dimensions.

To better understand the behavior of MIA-Former, we visualize the generated skipping policy of MIA-Former on different input samples. As shown in Fig. 5, MIA-Former generates different skipping policies based on different input samples. We can observe that MIA-Former can adaptively generate different policy based on the difficulty of input samples. For example, according to the comparison between Fig. 5(a) and (b), MIA-Former skips more from the token-wise dimension when processing (b) while skipping more from the head-wise dimension when processing (a) since (a) is full of corn in the image while (b) has a clean background with a small object in the front.

5 Conclusion

In this paper, we propose, develop, and validate MIA-Former, a multi-grained input-adaptive ViT framework, that is compatible with most of SOTA ViTs to achieve a higher accuracy-efficiency trade-off by dynamically skipping ViTs' blocks, heads, and tokens at coarse-to-fine granularities in an input-adaptive manner. The proposed hybrid supervised and reinforcement training method for effectively training the MIA-Former not only improves the achievable accuracy-efficiency trade-off of MIA-Former, but also boosts the robust accuracy, leading to a triple-win benefits. Extensive experiments and ablation studies show that the proposed MIA-Former achieves both efficiency and robustness improvement when being applied on top of various SOTA ViTs.

References

Athalye, A.; Carlini, N.; and Wagner, D. 2018. Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples. In *International conference on machine learning*, 274–283. PMLR.

Caron, M.; Touvron, H.; Misra, I.; Jégou, H.; Mairal, J.; Bojanowski, P.; and Joulin, A. 2021. Emerging properties in self-supervised vision transformers. *arXiv preprint arXiv:2104.14294*.

Chen, H.; Wang, Y.; Guo, T.; Xu, C.; Deng, Y.; Liu, Z.; Ma, S.; Xu, C.; Xu, C.; and Gao, W. 2021a. Pre-trained image processing transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 12299–12310.

Chen, P.; Chen, Y.; Liu, S.; Yang, M.; and Jia, J. 2021b. Exploring and Improving Mobile Level Vision Transformers. *arXiv preprint arXiv:2108.13015*.

Chen, T.; Zhang, Z.; Liu, S.; Chang, S.; and Wang, Z. 2020. Robust overfitting may be mitigated by properly learned smoothening. In *International Conference on Learning Representations*.

Deng, J.; Dong, W.; Socher, R.; Li, L.-J.; Li, K.; and Fei-Fei, L. 2009. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, 248–255. Ieee.

Dong, X.; Bao, J.; Chen, D.; Zhang, W.; Yu, N.; Yuan, L.; Chen, D.; and Guo, B. 2021. CSWin Transformer: A General Vision Transformer Backbone with Cross-Shaped Windows. *arXiv preprint arXiv:2107.00652*.

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

Fu, Y.; You, H.; Zhao, Y.; Wang, Y.; Li, C.; Gopalakrishnan, K.; Wang, Z.; and Lin, Y. 2020. Fractrain: Fractionally squeezing bit savings both temporally and spatially for efficient dnn training. *arXiv preprint arXiv:2012.13113*.

Fu, Y.; Yu, Q.; Li, M.; Chandra, V.; and Lin, Y. 2021a. Double-Win Quant: Aggressively Winning Robustness of Quantized Deep Neural Networks via Random Precision Training and Inference. In *International Conference on Machine Learning*, 3492–3504. PMLR.

Fu, Y.; Yu, Q.; Zhang, Y.; Wu, S.; Ouyang, X.; Cox, D. D.; and Lin, Y. 2021b. Drawing Robust Scratch Tickets: Subnetworks with Inborn Robustness Are Found within Randomly Initialized Networks. In *Thirty-Fifth Conference on Neural Information Processing Systems*.

Fu, Y.; Zhao, Y.; Yu, Q.; Li, C.; and Lin, Y. 2021c. 2-in-1 Accelerator: Enabling Random Precision Switch for Winning Both Adversarial Robustness and Efficiency. In *MICRO-54: 54th Annual IEEE/ACM International Symposium on Microarchitecture*, 225–237.

Graham, B.; El-Nouby, A.; Touvron, H.; Stock, P.; Joulin, A.; Jégou, H.; and Douze, M. 2021. LeViT: a Vision Transformer in ConvNet's Clothing for Faster Inference. *arXiv preprint arXiv:2104.01136*.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 770–778.

Heo, B.; Yun, S.; Han, D.; Chun, S.; Choe, J.; and Oh, S. J. 2021. Rethinking spatial dimensions of vision transformers. *arXiv preprint arXiv:2103.16302*.

Howard, A. G.; Zhu, M.; Chen, B.; Kalenichenko, D.; Wang, W.; Weyand, T.; Andreetto, M.; and Adam, H. 2017. Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv preprint arXiv:1704.04861*.

Hu, T.-K.; Chen, T.; Wang, H.; and Wang, Z. 2020. Triple wins: Boosting accuracy, robustness and efficiency together by enabling input-adaptive inference. *arXiv preprint arXiv:2002.10025*.

Jang, E.; Gu, S.; and Poole, B. 2016. Categorical reparameterization with gumbel-softmax. *arXiv preprint arXiv:1611.01144*.

Kaya, Y.; Hong, S.; and Dumitras, T. 2019. Shallow-deep networks: Understanding and mitigating network overthinking. In *International Conference on Machine Learning*, 3301–3310. PMLR.

Kingma, D. P.; and Ba, J. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Kurakin, A.; Goodfellow, I. J.; and Bengio, S. 2018. Adversarial Examples in the Physical World. In *Artificial Intelligence Safety and Security*, 99–112. Chapman and Hall/CRC.

Li, Y.; Zhang, K.; Cao, J.; Timofte, R.; and Van Gool, L. 2021. Localvit: Bringing locality to vision transformers. *arXiv preprint arXiv:2104.05707*.

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

Loshchilov, I.; and Hutter, F. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.

Madry, A.; Makelov, A.; Schmidt, L.; Tsipras, D.; and Vladu, A. 2017. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083*.

Mao, X.; Qi, G.; Chen, Y.; Li, X.; Ye, S.; He, Y.; and Xue, H. 2021. Rethinking the Design Principles of Robust Vision Transformer. *arXiv preprint arXiv:2105.07926*.

Mnih, V.; Badia, A. P.; Mirza, M.; Graves, A.; Lillicrap, T.; Harley, T.; Silver, D.; and Kavukcuoglu, K. 2016. Asynchronous methods for deep reinforcement learning. In *International conference on machine learning*, 1928–1937. PMLR.

Rakin, A. S.; He, Z.; Yang, L.; Wang, Y.; Wang, L.; and Fan, D. 2019. Robust sparse regularization: Simultaneously optimizing neural network robustness and compactness. *arXiv preprint arXiv:1905.13074*.

Rakin, A. S.; Yi, J.; Gong, B.; and Fan, D. 2018. Defend deep neural networks against adversarial examples via fixed and dynamic quantized activation functions. *arXiv preprint arXiv:1807.06714*.

Rao, Y.; Zhao, W.; Liu, B.; Lu, J.; Zhou, J.; and Hsieh, C.-J. 2021. DynamicViT: Efficient Vision Transformers with Dynamic Token Sparsification. *arXiv preprint arXiv:2106.02034*.

Ridnik, T.; Ben-Baruch, E.; Noy, A.; and Zelnik-Manor, L. 2021. Imagenet-21k pretraining for the masses. *arXiv preprint arXiv:2104.10972*.

Sehwag, V.; Wang, S.; Mittal, P.; and Jana, S. 2020. Hydra: Pruning adversarially robust neural networks. *arXiv preprint arXiv:2002.10509*.

Shao, R.; Shi, Z.; Yi, J.; Chen, P.-Y.; and Hsieh, C.-J. 2021. On the adversarial robustness of visual transformers. *arXiv preprint arXiv:2103.15670*.

Shen, J.; Wang, Y.; Xu, P.; Fu, Y.; Wang, Z.; and Lin, Y. 2020. Fractional skipping: Towards finer-grained dynamic cnn inference. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, 5700–5708.

Steiner, A.; Kolesnikov, A.; Zhai, X.; Wightman, R.; Uszkoreit, J.; and Beyer, L. 2021. How to train your ViT? Data, Augmentation, and Regularization in Vision Transformers. *arXiv preprint arXiv:2106.10270*.

Strudel, R.; Garcia, R.; Laptev, I.; and Schmid, C. 2021. Segmenter: Transformer for Semantic Segmentation. *arXiv preprint arXiv:2105.05633*.

Tan, M.; and Le, Q. 2019. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International Conference on Machine Learning*, 6105–6114. PMLR.

Teerapittayanon, S.; McDanel, B.; and Kung, H.-T. 2016. Branchynet: Fast inference via early exiting from deep neural networks. In *2016 23rd International Conference on Pattern Recognition (ICPR)*, 2464–2469. IEEE.

Touvron, H.; Cord, M.; Douze, M.; Massa, F.; Sablayrolles, A.; and Jégou, H. 2021. Training data-efficient image transformers & distillation through attention. In *International Conference on Machine Learning*, 10347–10357. PMLR.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. In *Advances in neural information processing systems*, 5998–6008.

Wang, W.; Xie, E.; Li, X.; Fan, D.-P.; Song, K.; Liang, D.; Lu, T.; Luo, P.; and Shao, L. 2021a. Pvtv2: Improved baselines with pyramid vision transformer. *arXiv preprint arXiv:2106.13797*.

Wang, X.; Yu, F.; Dou, Z.-Y.; Darrell, T.; and Gonzalez, J. E. 2018. Skipnet: Learning dynamic routing in convolutional networks. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 409–424.

Wang, Y.; Huang, R.; Song, S.; Huang, Z.; and Huang, G. 2021b. Not All Images are Worth 16x16 Words: Dynamic Vision Transformers with Adaptive Sequence Length. *arXiv preprint arXiv:2105.15075*.

Wang, Y.; Shen, J.; Hu, T.-K.; Xu, P.; Nguyen, T.; Baraniuk, R.; Wang, Z.; and Lin, Y. 2020. Dual dynamic inference: Enabling more efficient, adaptive, and controllable deep inference. *IEEE Journal of Selected Topics in Signal Processing*, 14(4): 623–633.

Wu, Z.; Nagarajan, T.; Kumar, A.; Rennie, S.; Davis, L. S.; Grauman, K.; and Feris, R. 2018. Blockdrop: Dynamic inference paths in residual networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 8817–8826.

Xu, W.; Xu, Y.; Chang, T.; and Tu, Z. 2021. Co-scale conv-attentional image transformers. *arXiv preprint arXiv:2104.06399*.

Ye, S.; Xu, K.; Liu, S.; Cheng, H.; Lambrechts, J.-H.; Zhang, H.; Zhou, A.; Ma, K.; Wang, Y.; and Lin, X. 2019. Adversarial robustness vs. model compression, or both? In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 111–120.

Zhou, D.; Kang, B.; Jin, X.; Yang, L.; Lian, X.; Jiang, Z.; Hou, Q.; and Feng, J. 2021. Deepvit: Towards deeper vision transformer. *arXiv preprint arXiv:2103.11886*.