Adaptive Masked Autoencoder Transformer for Image Classification

Xiangru Chen[a,b,1], Chenjing Liu[a,b,1], Peng Hu[a], Jie Lin[b], Yunhong Gong[a], Yingke Chen[d], Dezhong Peng[a,c,*] and Xue Geng[b,*]

[a]College of Computer Science, Sichuan University, 610065, Chengdu, China
[b]Institute for Infocomm Research, Agency for Science, Technology and Research, 138632, Singapore
[c]Sichuan Newstrong UHD Video Technology Co., Ltd., 610095, Chengdu, China
[d]Department of Computer and Information Sciences, Northumbria University, Newcastle upon Tyne NE1 8ST, UK

ARTICLE INFO

Keywords:
vision transformer
masked image modeling
image classification

ABSTRACT

Vision Transformers (ViTs) have exhibited exceptional performance across a broad spectrum of visual tasks. Nonetheless, their computational requirements often surpass those of prevailing CNN-based models. Token sparsity techniques have been employed as a means to alleviate this issue. Regrettably, these techniques often result in the loss of semantic information and subsequent deterioration in performance. In order to address these challenges, we propose the Adaptive Masked Autoencoder Transformer (AMAT), a masked image modeling-based method. AMAT integrates a novel adaptive masking mechanism and a training objective function for both pre-training and fine-tuning stages. Our primary objective is to reduce the complexity of Vision Transformer models while concurrently enhancing their final accuracy. Through experiments conducted on the ILSVRC-2012 dataset, our proposed method surpasses the original ViT by achieving up to 40% FLOPs savings. Moreover, AMAT outperforms the efficient DynamicViT model by 0.1% while saving 4% FLOPs. Furthermore, on the Places365 dataset, AMAT achieves a 0.3% accuracy loss while saving 21% FLOPs compared to MAE. These findings effectively demonstrate the efficacy of AMAT in mitigating computational complexity while maintaining a high level of accuracy.

1. Introduction

Image classification [25] is a fundamental challenge in computer vision that holds significant importance across various applications and tasks [1]. It forms the basis for advanced tasks such as object detection [21] and semantic segmentation [11]. By accurately assigning predefined labels or categories to input images, image classification algorithms empower machines to comprehend and interpret visual content [5]. This, in turn, drives advancements in areas such as autonomous vehicles [12], medical diagnose [23, 24].

Recently, Vision Transformers (ViT) [7] have gained considerable attention and showcased remarkable achievements in tackling a diverse array of visual tasks [9]. They have consistently outperformed many Convolutional Neural Networks (CNNs) [25] when pre-trained on extensive datasets using effective data augmentation and regularization techniques [15].

As a foundational work, Dosovitskiy et al. [7] initially introduced the use of a pure Transformer-based model for training on large-scale datasets. Their method involves partitioning the input image into smaller patches, which are subsequently transformed and processed using transformer blocks. However, the training efficiency of the transformer-based model, which heavily relies on large-scale training, poses a challenge and hampers its broader applications. To address this training issue, DeiT [26] was proposed, which was a data-efficient image transformer aimed at enhancing training efficiency. They employed knowledge distillation techniques, leveraging a well-performing CNN teacher network to assist the ViT in learning robust semantic features. Surprisingly, the resulting student ViT model demonstrates superior accuracy compared to the teacher network.

*Corresponding authors
cxrasdfg@gmail.com (X. Chen); cjliu1996@gmail.com (C. Liu); penghu.ml@gmail.com (P. Hu); jie.dellinger@gmail.com (J. Lin); gongyunhong@stu.scu.edu.cn (Y. Gong); yingke.chen@northumbria.ac.uk (Y. Chen); pengdz@scu.edu.cn (D. Peng); geng_xue@i2r.a-star.edu.sg (X. Geng)
ORCID(s):
[1]These authors contributed equally to this work

The Transformer-based models first ascended to prominence within the realm of Natural Language Processing (NLP) and have since established themselves as the preeminent deep learning models for an extensive array of NLP undertakings [27, 14, 13]. Significantly, the Masked Language Modeling (MIM)-based Transformers have emerged as the predominant models, characterized by their prelusive immersion in vast textual corpora for pre-training, followed by a refined calibration during fine-tuning phases on more modest datasets tailored to particular downstream tasks [6]. The rise in popularity of chat-based Large Language Models (LLMs), such as GPT [17], has further fueled the exploration of Transformers within the research community. These developments have prompted a surge of interest in applying self-supervised learning methods, as seen in linguistic tasks, to computer vision. Although Masked Image Modeling (MIM) approaches have garnered attention in computer vision, they have yet to become mainstream. However, the introduction of the MAE model by He et al. [10] has brought significant improvements by leveraging a vanilla ViT architecture.

However, despite the demonstrated success of the aforementioned methods, the dense modeling of long-range dependencies among image tokens often leads to computational inefficiency. In other words, a significant limitation of the ViT model lies in its substantial computational complexity, demanding a greater allocation of GPU resources for performing inference, compared to CNN-based models. For instance, when considering Floating-point operations per second (FLOPs) as a metric, a base ViT model requires approximately 18 GFLOPs, whereas a typical Resnet-34 [12] model only demands 7 GFLOPs when operating on a 224x224 input resolution. As a result, the utilization of ViT-based models on hardware-limited devices presents additional resource challenges.

In theory, the attention mechanism shoulders the computational load of the model, with its complexity increasing quadratically in relation to the number of input patches [15]. This implies that as the number of input patches grows, the computational demands placed on the attention mechanism rise significantly, resulting in a quadratic increase in complexity. To resolve this challenge, researchers have proposed efficient transformers [28, 32]. These efficient transformers aim to modify the original attention mechanism and achieve linear complexity with respect to the input sequence. Another approach focuses on reducing computational complexity by eliminating some of the input tokens per layer [20]. Their underlying concept is that many input image patches correspond to less informative regions of the original images, such as backgrounds. The objective of this method is to alleviate the computational burden by masking irrelevant tokens, thereby enabling faster inference through the utilization of token masks.

To address this challenge, we propose the Adaptive Masking Autoencoder Transformer (AMAT) for Image Classification. Our approach introduces a novel adaptive masking mechanism within the training framework of AMAT, enabling the adaptive removal of irrelevant tokens based on the semantic context of image patches. This adaptive masking technique effectively reduces the computational complexity of ViT-based models. Additionally, leveraging the benefits of Masked Image Modeling (MIM), our enhanced model surpasses the performance of the original ViT model while maintaining the same network architecture.

The contributions of this work are documented as follows:

1. We proposed a novel method called AMAT for pre-training ViT-based models in image classification. AMAT utilizes masked image modeling, reducing the computational resources required during inference.
2. We develop an adaptive masking mechanism, which dynamically conceals extraneous tokens throughout the entirety of both the pre-training and fine-tuning stages, thereby substantially alleviating the computational overhead incurred by ViT-based models.
3. We have proposed a novel objective function to facilitate the training of the AMAT method, effectively bridging the chasm that exists between the pre-training and fine-tuning stages, which in turn enhances the final accuracy.
4. The experimental results on the ILSVRC-2012 and Places365 datasets demonstrate the efficiency of the proposed AMAT. It achieves an impressive reduction in FLOPs of up to 40%, while surpassing the accuracy of the original ViT model within the same network architecture.

The rest of this paper is organized as follows: Section 2 introduces the background and related work of this study. Subsequently, Section 3 presents the details of the proposed AMAT. In Section 4, we discuss the experimental configurations and highlight the promising performance of the proposed AMAT. Lastly, we provide a summary of our work in Section 5.

2. Background and Related Work

2.1. Vision Transformer

The Vision Transformer (ViT), as presented by Dosovitskiy et al. [7], has emerged as a promising paradigm within the field of vision tasks, excelling at the capture of intricate long-term dependencies. This evolution is underpinned by the resounding success of Transformer architectures in the realm of NLP [27], and ViT revolutionizes image analysis by conceptualizing images as a multitude of visual tokens. This paradigm shift allows ViT to leverage the power of attention mechanisms to encode intricate relationships between these visual tokens, ultimately facilitating robust representation learning.

The adaptation of vanilla Transformers for image processing poses a formidable challenge due to their inherent sequential input requirement. To overcome this limitation, the Vision Transformer (ViT) employs a multi-step preprocessing pipeline that involves splitting the input image into nonoverlapping patches. Then, the patches are subsequently subjected to a transformation process, wherein they are converted into patch embeddings through the application of a projection mechanism. In order to preserve essential spatial information, a 1-D positional encoding, which can be adjusted through learning, is thoughtfully integrated with the patch embeddings.

Touvron et al. [26] introduced DeiT, an innovative and efficient transformer-based model that leverages knowledge distillation from a well-trained CNN teacher network. This approach, which surpasses traditional methods such as data augmentation and regularization, employs a teacher-student distillation technique to enhance auxiliary representation learning. By incorporating a distinctive token, guided by labels provided by the teacher model, the student ViT gains its empowerment. Remarkably, extensive experiments conducted by Touvron and colleagues have revealed that CNNs outperform transformer-based models as teacher models. Furthermore, the distilled student Transformer model exhibits even higher performance than its teacher CNN model.

Building upon the foundation laid by the ViT, the CrossViT [3] was developed to learn the multi-scale feature representations in vision-based tasks. This novel method incorporates a dual-branch transformer block and a cross-attention fusion strategy, setting it apart from other existing techniques and establishing its superiority. The T2T-ViT [31] presents a pioneering paradigm in the realm of image processing, leveraging the transformative capabilities of a vision transformer model. This architectural marvel encompasses two pivotal components: the Tokens-to-Token (T2T) module and the T2T-ViT backbone, synergistically collaborating to fortify the model's proficiency in extracting profound and semantically rich features from images.

Han et al. [8] introduced a new architecture called Transformer iN Transformer (TNT) that explores the importance of attention within local patches for building high-performance visual transformers. Unlike the vanilla ViT model, TNT uniquely interprets local patches (e.g., 16x16) as "visual sentences", followed by a subsequent subdivision into smaller patches (e.g., 4x4) denominated as "visual words". Through the computation of attention interactions among these distinctive "words", the model effectively enhances its proficiency in extracting profound image features. The fusion of features stemming from both "words" and "sentences" profoundly bolsters TNT's representation prowess. Experiments on several benchmarks show that this architecture is effective. Chu et al. [4] propose a conditional positional encoding (CPE) scheme for vision Transformers to address the gap between ViT and higher-resolution images. Unlike previous positional encodings, CPE generates encodings based on the local neighborhood of input tokens, allowing for greater flexibility. CPE ensures the desired translation equivalence in vision tasks while improving the model's performance. By incorporating CPE seamlessly into the Transformer framework using a Position Encoding Generator (PEG), Conditional Position encoding Vision Transformer (CPVT) is created. CPVT exhibits attention maps similar to those with learned positional encodings and achieves superior performance. Liu et al. introduce Swim-Transformer, a hierarchical Transformer architecture that computes representations using shifted windows. This approach augments computational efficiency by constraining self-attention computations within non-overlapping, local windows, all the while fostering cross-window connections. The hierarchical structure of Swim-Transformer facilitates modeling at multiple scales, rendering it amenable to a wide spectrum of visual tasks. Moreover, Swim-Transformer sustains a linear computational complexity concerning image dimensions. DynamicViT, proposed by Rao et al. [20], incorporates a dynamic token sparsification framework for pruning redundant tokens in an input sequence of images. The framework includes a lightweight prediction module that estimates the importance score of each token and prunes tokens hierarchically across different layers. To refine the prediction module in an end-to-end fashion, an attention masking strategy has been introduced to differentially trim tokens by obstructing their interactions with others.

2.2. Masked Image Modeling

Masked Image Modeling (MIM) is a technique where specific regions or pixels in an image are concealed, and a model is trained to anticipate concealed regions through the utilization of the surrounding visual context. While Masked Language Modeling (MLM) has been extensively studied in natural language processing, MIM has received less attention in vision tasks until recently [18].

The groundbreaking work by He with the masked autoencoder (MAE) [10] has generated interest in MIM for image processing. MAE simplifies training ViT models by using a decoder with the same structure as the original ViT. During training, MAE randomly masks parts of input image patches, significantly improving computational efficiency compared to previous ViT methods. During the fine-tuning stage, MAE directs its attention toward training the encoder, while the decoder is intentionally set aside. This straightforward approach has garnered notable interest within the academic community.

Other notable contributions include MaskFeat [29], which explores different types of features and highlights the effectiveness of Histograms of Oriented Gradients (HOG). Bao et al. [2] introduced BEiT, a self-supervised vision representation model inspired by BERT. BEiT employs a masked image modeling task during pre-training to recover original visual tokens from corrupted image patches. The model can then be fine-tuned for specific tasks by adding task-specific layers on top of the pretrained encoder.

SimMIM, as postulated by Xie et al. [30], elucidates a straightforward yet formidable framework for MIM. Their approach involves predicting pixel values of randomly concealed patches using a lightweight one-layer head and training the model with a straightforward L1 loss. SimMIM attains cutting-edge results on the ImageNet dataset, showcasing the efficacy of their training approach.

These advancements in MIM have deepened our understanding of visual representation learning and highlighted the potential of applying masked language models to the image domain.

3. Adaptive Masked Autoencoder Transformer

In this section, we will delve into the specifics of the Adaptive Masking Autoencoder Transformer (AMAT) we have developed. We will outline the training objective function for AMAT and explain how it facilitates the training phase of our variant of the ViT to successfully accomplish the image classification task.

3.1. Vision Transformer

When processing 2D images within the context of a vision transformer (ViT), we start with an input image $X \in \mathbb{R}^{h\times w\times c}$. To handle the image effectively, it is converted into a series of small image patches denoted as $X_p \in \mathbb{R}^{n\times(p^2\cdot c)}$, where c represents the number of channels. Here, (h, w) refers to the height and width of the input image, while (p, p) represents those of each small image patch.

To ensure the transformer operates efficiently on these patches, the length of input sequence is determined as $n = hw/p^2$. This allows the transformer to process the image patches coherently. Given that the transformer utilizes layers characterized by consistent widths, a trainable linear projection is implemented to transform each vectorized patch into the model dimension d. The resulting output is referred to as patch embeddings, encapsulating the essential information extracted from the image. The projected images patches can be represented as $X_e \in \mathbb{R}^{n\times d}$.

Positional Encoding. To account for the absence of positional information in the original input, a common practice is to incorporate positional encoding into the resulting embeddings. This additional positional encoding helps the transformer model effectively capture the sequential or spatial relationships between elements in the input sequence:

$$\begin{aligned} PE(pos, 2i) &= \sin\left(\frac{pos}{10000^{\frac{2i}{d}}}\right); \\ PE(pos, 2i+1) &= \cos\left(\frac{pos}{10000^{\frac{2i}{d}}}\right), \end{aligned} \tag{1}$$

Multi-head Self-Attention. The self-attention mechanism plays a crucial role as a fundamental component within transformer-based models. It requires three input variables: queries $Q \in \mathbb{R}^{n\times d}$, keys $K \in \mathbb{R}^{n\times d}$, and values $V \in \mathbb{R}^{n\times d}$. Importantly, these input variables are consistently derived from the same input sequence X. Multi-head self-attention splits the original attention into several sub-attentions, and then aggregates all the sub-attended values to form the final

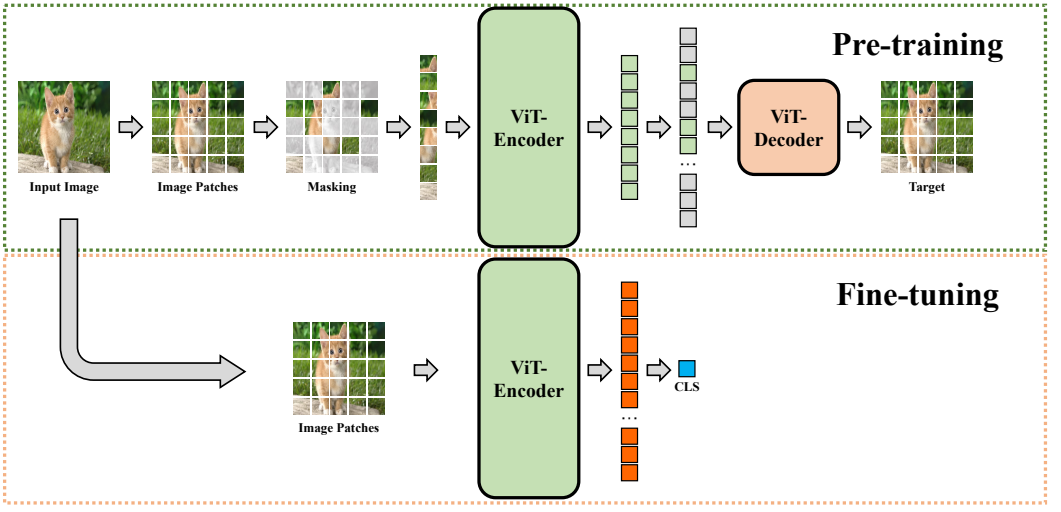

Figure 1: The masked autoencoder undergoes two stages of training. In the pre-training stage, the model aims to reconstruct image patches that have been masked during the pre-processing step. Once the pre-training stage is complete, only the encoder part of the model is retained for training the classifier.

attended value:

$$\begin{aligned} \mathbf{MHA}(Q,K,V) &= \left[\text{head}_1, \cdots, \text{head}_h\right] W^o, \\ \text{head}_i &= \text{softmax}\left(\frac{QW_i^Q(KW_i^K)^\dagger}{\sqrt{d_k}}\right) VW_i^V, \end{aligned} \tag{2}$$

where $W_i^Q, W_i^K, W_i^V \in \mathbb{R}^{d\times d_k}$ projects the inputs into the different subspaces.

3.2. Autoencoder Transformer

Masked Autoencoder (MAE) has greatly advanced the field of image processing. These models, based on vision transformers, effectively capture hidden structures and extract meaningful features from images using masking techniques. Their ability to capture intricate patterns and enhance image quality has gained significant attention in recent years. Like vanilla autoencoders, MAE consists of an encoder and a decoder component working together to complete the target vision tasks.

Encoder. Presume a stochastic mask $M^p \in \{0,1\}^n$ for the image patches X_e, the encoder will remove the masked tokens. The remaining tokens will be stacked and processed:

$$X_{vis} = \mathbf{ViTEncoder}\left(X_e \circ \overline{M^p}\right), \tag{3}$$

and the median features X_{vis} will be passed to the decoder.

Decoder. After obtaining the visual representation X_{vis}, to restore the original image, X_{vis} is expanded back to the initial sequence length. This expansion is achieved by incorporating a special mask token denoted as:

$$\hat{X}_{vis} = \left[X_{vis} + PE \circ \overline{M^p}, [\mathbf{MASK}] + PE \circ M^p\right], \tag{4}$$

where PE represents the corresponding location information that needs to be added for the supplementary token. For the output generated by the decoder, only the masked tokens need to be predicted.

$$X_{dec} = \mathbf{ViTDecoder}\left(\hat{X}_{vis}\right). \tag{5}$$

The training process of MAE encompasses two distinct stages. The initial stage is known as pre-training. In this stage, each input image undergoes a random masking procedure, incorporating a fixed masking ratio (i.e., 75%). The MAE is trained to predict the masked tokens and reconstruct the input image faithfully.

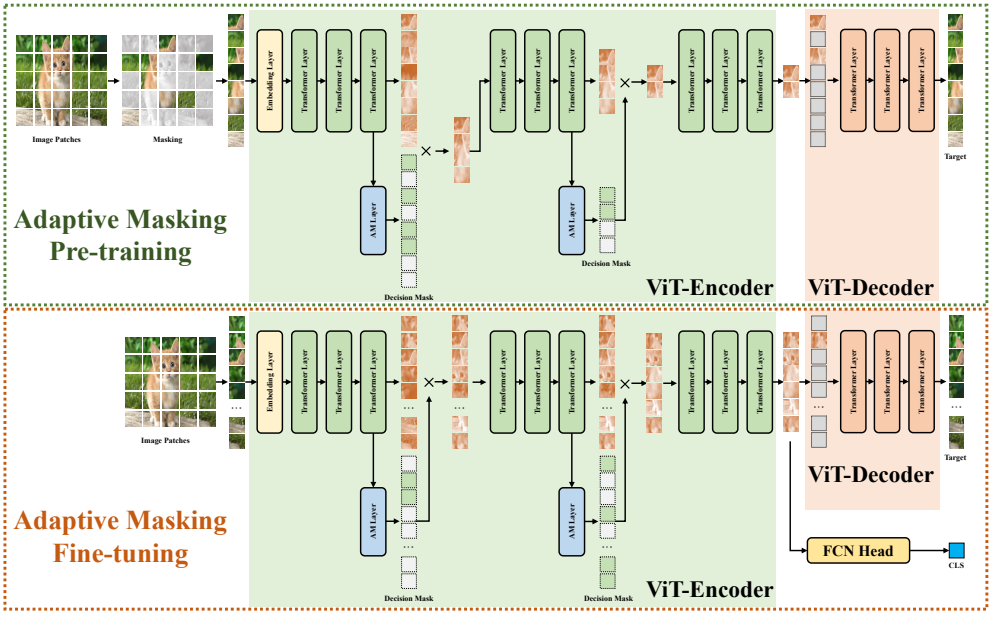

Figure 2: Within the framework of the proposed AMAT, we start by training the encoder and decoder together using masked image modeling and the developed **A**daptive **M**asking (AM) layers. During the pre-training stage, the AM layers/branches are also extensively trained to learn how to remove unnecessary tokens in the particular layers. In the next fine-tuning stage, the AM layers work to eliminate unimportant tokens that don't harm the final accuracy.

The subsequent stage is called fine-tuning. During this stage, the supervised signal is integrated, and classification tasks are performed. Notably, the decoder used in the pre-training stage is discarded, and a fully-connected network (FCN) is appended on top of the encoder. This linear layer is responsible for carrying out the classification process, leveraging the encoded representations generated by the MAE. The predicted classification vector can be computed by:

$$\hat{\boldsymbol{y}} = \text{softmax}\left(\text{FCN}\left(\text{avg}\left(\textbf{ViTEncoder}\left(X_e\right)\right)\right)\right). \tag{6}$$

The architecture of a MAE is depicted in Fig. 1. Both the pre-training and fine-tuning procedures are replicated within this diagram.

3.3. Adaptive Masking Mechanism

Despite the remarkable performance achieved by transformer-based models in vision tasks, the high computational burden remains a challenge when it comes to deploying them on resource-constrained devices. This challenge primarily stems from the inherent complexity of the original multi-head self-attention mechanism, which grows quadratically with the length of the input sequence. Furthermore, the impressive achievements of MAE in masked self-supervised learning have greatly influenced our research. It is astonishing to see that even when an original image is heavily damaged, with over 75% of its information missing, it can still be reconstructed successfully. This reconstruction leads to improved accuracy in subsequent classification tasks. These findings highlight the significant potential of MAE in advancing the field of image analysis and classification.

Our investigations have uncovered fascinating insights into this phenomenon. On one hand, we have observed that random masks show promise in enhancing the performance of neural networks. However, it is important to recognize that certain key patches may hold greater significance for discrimination in classification tasks. On the other hand, we have found that the semantic information used for discrimination in diverse images is typically concentrated in specific regions. To tackle the computational burden associated with ViT, we developed this adaptive masking method for self-supervised training and inference, aiming to optimize the efficiency of ViT while preserving its effectiveness. This adaptive mechanism allows us to customize distinct mask strategies for each image, dynamically adjusting to diverse inputs while effectively preserving discriminative features. By adopting this strategy, our goal is to enhance

both the efficiency and effectiveness of the ViT model, thereby pushing the boundaries of its performance in image classification.

To better exhibit how the adaptive masking mechanism work in the pre-training stage and fine-tuning stage, we show the overall AMAT framework in Figure 2. This figure illustrates the overall framework of AMAT. The encoder consists of nine transformer layers, while the decoder comprises three transformer layers. There are two adaptive masking (AM) layers after the 3-rd and 6-th layers, respectively. During the pre-training stage using masked self-supervised learning, images are initially masked according to a predefined ratio (75%). The remaining tokens are used to generate embedding patches. When encountering an AM layer, the current tokens are input into the AM layer to obtain the corresponding decision mask. Based on this decision mask, the less important tokens are removed and passed on to the subsequent transformer layers for further processing. Once the ViT-Encoder completes sequential processing of the token sequence, the ViT-Decoder utilizes the masked tokens to reconstruct the image. During fine-tuning stage, the image is directly input into the ViT-Encoder. Similarly, the AM layers remove corresponding output tokens and obtain intermediate representations. At this stage, aside from the ViT-Decoder used for generating reconstructed images, a classification branch (FCN Head) is employed to determine the category of the sample.

3.3.1. Adaptive Masking Layers

To dynamically decide which token to be removed for each input sequence, we use a small four-layer network to evaluate the importance of tokens and place them behind some specified layers. Assume the input sequence of a specific branch is $X \in \mathbb{R}^{n\times d}$, we want to obtain the decision mask $\boldsymbol{M}^d$ to decide which tokens should be eliminated. To accomplish this objective, the compact prediction network adeptly gathers both local and global features from the input sequence, thereby ensuring a robust estimation of the significance of each input token. The local features are obtained by:

$$X^l = \mathrm{FCN}(X) \in \mathbb{R}^{n\times d'}. \tag{7}$$

Here, it is noteworthy that d' may be judiciously chosen as a dimensionality that is reduced relative to the original. Furthermore, we astutely compute an all-encompassing global feature by invoking the aggregation process, which amalgamates information derived from the fully connected network (FCN), in concordance with the esteemed decision mask, $\boldsymbol{M}^d$. The resultant global feature is expressed as:

$$X^g = \mathrm{avg}\left(\mathrm{FCN}\left(X\right)\right) \in \mathbb{R}^{d'}. \tag{8}$$

In this context, the aggregation function "avg" plays a crucial role in consolidating knowledge from all existing tokens. During the training process, to ensure that masked tokens do not influence the results, we exclude them from the calculation.

The local feature, which captures the inherent characteristics of each token, and the global feature, which encompasses the overall contextual information of the entire image, hold immense importance. Therefore, we combine these unique features, the local and global, to create a seamless integration known as local-global embeddings:

$$X^d_{feat} = \left[X^l, \mathrm{expand}(X^g)\right] \in \mathbb{R}^{n\times d}. \tag{9}$$

These carefully designed embeddings are smoothly fed into another FCN (fully connected network), giving us the invaluable capability to predict the probabilities associated with the decision of keeping or discarding the tokens.

$$\boldsymbol{\pi} = \mathrm{softmax}(\mathrm{FCN}(X^d_{feat})) \in \mathbb{R}^{n\times 2}, \tag{10}$$

where the entry $\boldsymbol{\pi}_{i,0}$ signifies the probability of discarding the token indexed by i, whereas $\boldsymbol{\pi}_{i,1}$ represents the probability of retaining the specific token. Subsequently, the adaptive masking mechanism will invoke the process of sampling from the probability distribution $\boldsymbol{\pi}$ to obtain the current decision mask $\boldsymbol{M}^d$ for the current layer.

3.3.2. Training Phase

The adaptive masking layers are trained simultaneously with the encoder and decoder. However, since the token selection is not differentiable in deep neural networks, adaptive masking behaves slightly differently during training and testing. In the pre-training stage and the fine-tuning stage of the proposed AMAT, we employ a reparameterization trick to transform the token selection process into a differentiable operation. This allows the prediction branch to obtain gradients from the backward process. The reparameterization trick is implemented as follows:

$$M^d = \mathrm{GS}\left(\boldsymbol{\pi}\right)_{*,1} \in \{0,1\}^n, \tag{11}$$

where the term "GS" denotes the gumbel-softmax.

Furthermore, during the training process, aside from the initial random masking performed by the masked autoencoder to reduce image patches, direct reduction of input tokens does not occur through adaptive masking. Instead, a modified version of softmax is utilized to completely block unimportant tokens, thereby eliminating any influence from irrelevant tokens. Thus, the attended values in the Eqn. 2 should be revised accordingly:

$$\begin{aligned} \mathrm{head}_i &= \frac{\exp\left(\mathrm{sim}_i\right) \odot M^d}{\sum_{i=1}^{n} \exp\left(\mathrm{sim}_i\right) \odot M^d} V W_i^V, \\ \text{where } \ \mathrm{sim}_i &= \frac{Q W_i^Q (K W_i^K)^\top}{\sqrt{d_k}}. \end{aligned} \tag{12}$$

3.3.3. Testing Phase

In the testing phase, after going through the adaptive masking branch, certain less important tokens are simply removed based on the current probabilities $\boldsymbol{\pi}$. These removed tokens are then passed to the next layer for further processing.

Assume the pruning ratio for the i-th reduction branch is ρ_i. Firstly, we sort the prediction scores along the last dimension and obtain the indices of values from high to low. From this sorted set, we select the first $(1-\rho_i)n$ indices. This post-processing step in the reduction module can be expressed as:

$$M^d = \mathrm{argsort}\left(\mathrm{argsort}\left(\boldsymbol{\pi}_{:,1}\right)\right) > \rho_i n. \tag{13}$$

Here, M^d is a binary mask representing the current input tokens X during the test phase. The set of pruned tokens X' can be obtained as follows:

$$X'_i = \begin{cases} X_i, M_i^d = 1 \\ \emptyset, M_i^d = 0 \end{cases}, \quad \forall i \in [1..n]. \tag{14}$$

In the above equation, X'_i denotes the pruned tokens based on a binary mask M_i^d. If M_i^d is 1, it indicates that the token is retained in X', while if M_i^d is 0, it signifies that the token is not included ($\emptyset$) in X'.

3.4. Adaptive Masking Loss

The training of MAE consists of two stages: pre-training stage and fine-tuning stage. In both stages, we introduce a new objective function for training the ViT-based autoencoder. This objective function enables hierarchical dynamic token masking after each specific layer. The design of the loss function allows the AM layer to accurately identify which tokens should be preserved, while ensuring that the network can converge and achieve high classification accuracy on the target dataset.

3.4.1. Loss for Pre-training

Considering our previous discussion, let's assume that the embedded image patches are denoted as X_e, and the corresponding label is $\boldsymbol{y}$. We have a randomly sampled mask, M_p, with a pre-defined masking ratio. To pre-train a ViT-based autoencoder, we can define the basic objective function as follows:

$$\mathcal{L}_{rec} = \|X_e \circ M^p - X_{dec}\|^2, \tag{15}$$

where X_{dec} represents the reconstructed image obtained from the autoencoder. This objective function employs a simple Mean Squared Error (MSE) loss, which can be easily implemented.

Nevertheless, in the course of this training process, the tiny branch responsible for adaptive mask prediction lacks control. To ensure that the auxiliary network produces the desired pruning ratio for the current layer's input sequence, we introduce an additional regularized term. This term guides the overall training process, promoting the adaptive masking branch to output an appropriate masking ratio. For the i-th adaptive mask, we have:

$$\mathcal{L}_i^m = \left(\frac{\sum_{j=1}^{n} M_{i,j}^d}{n} - \rho_i \right)^2. \tag{16}$$

This term encourages the keeping ratio of the input tokens to approximate a preset compression rate.

Hence, we can summarize the overall loss function $\mathcal{L}_{ptr}$ for the pre-training stage as follows:

$$\mathcal{L}_{ptr} = \mathcal{L}_{rec} + \sum_{i=1}^{n_a} \frac{\mathcal{L}_i^m}{n_a}, \tag{17}$$

where n_a represents the total count of adaptive masking layers incorporated into the ViT model.

3.4.2. Loss for Fine-tuning

During the fine-tuning stage of standard MIM-based method, a fully-connected layer is added on top of the encoder. The decoder is discarded, and only the encoder is fine-tuned for the classification tasks. The original loss function for the fine-tuning stage can be derived as follows:

$$\mathcal{L}_{ftn} = -\sum_{c=1}^{C} y_c \log(\hat{y}_c) + \sum_{i=1}^{n_a} \frac{\mathcal{L}_i^m}{n_a}, \tag{18}$$

where the left term represents the cross-entropy loss for classification, while the right term corresponds to the regularization term for the pruning ratios of each AM layer.

By utilizing the aforementioned equation, we can successfully accomplish the fine-tuning training of the ViT model with adaptive masking mechanism. Nonetheless, while discarding the decoder can enhance training efficiency, it may inadvertently compromise certain semantic features, resulting in a decrease in overall classification accuracy. Consequently, we propose that during the fine-tuning process, the decoder's reconstruction loss should still be preserved. This approach allows the encoder to receive signals from the decoder throughout the reconstruction process, thereby facilitating the training of the encoder. As a result, Eqn. 18 will be modified as follows:

$$\mathcal{L}_{ftn} = -\sum_{c=1}^{C} y_c \log(\hat{y}_c) + \sum_{i=1}^{n_a} \frac{\mathcal{L}_i^m}{n_a} + \hat{\mathcal{L}}_{rec}, \tag{19}$$

where the term

$$\hat{\mathcal{L}}_{rec} = \left\| X_e \circ \overline{\prod_{i=1}^{n_a} M_i^d} - \hat{X}_{dec} \right\|^2. \tag{20}$$

Hence, Eqn. 19 represents the final objective function $\mathcal{L}_{ftn}$ for the fine-tuning stage. The subsequent experimental section will showcase the performance enhancement resulting from the inclusion of the reconstruction term in the final fine-tuning objective function.

4. Experiments

4.1. Dataset & Setup

We evaluate the proposed AMAT on ILSVRC-2012 [22], which consists of approximately 1.2M real-world images for the training phase and 50k images for the validation. We measure the accuracy of our model through the Top-1

Table 1
Comparisons with the state-of-the-arts on ImageNet-1k. We compare the proposed model with state-of-the-art image classification models with comparable FLOPs. We employ the ViT-Base model as the reference model and employ the symbol "$\#\rho$" to denote the retention ratio of the input tokens for each selected layer, and use the "-()" to indicate the positions of the pruning branches. We also train the original masked autoencoder and report its results as another baseline model.

Model/Method	Params (M)	FLOPs (G)	Input Size	Top-1 Acc (%)
ViT-Base/16	86.6	17.6	224	77.9
DeiT-Base/16	86.6	17.6	224	81.8
MAE-Base/16	86.6	17.6	224	82.9
CrossViT-B	104.7	21.2	224	82.2
T2T-ViT-24	64.1	14.1	224	82.3
TNT-B	66.0	14.1	224	82.8
CPVT-B	88.0	17.6	224	82.3
RegNetY-16G	84.0	16.0	224	82.9
Swin-B	88.0	15.4	224	83.3
DynamicViT-B	-	11.2	224	81.3
AMAT#0.7-(3-6-9)	89.3	13.9	224	82.5
AMAT#0.6-(0-5-10)	89.3	10.7	224	81.4
AMAT#0.75-(0)	87.4	13.2	224	82.2

Table 2
Comparisons on the Places365-standard dataset, evaluating the performance of the proposed AMAT method against the original ViT model and the MAE model. Specifically, the AMAT method was trained using weights transferred from the ImageNet-1K dataset, while the original ViT model was trained from scratch. The MAE model, on the other hand, was trained using weights transferred from the ImageNet-1K dataset.

Model/Method	Params (M)	FLOPs (G)	Input Size	Top-1 Acc (%)
ViT-Base/16	86.6	17.6	224	49.1
MAE-Base/16	86.6	17.6	224	58.1
AMAT#0.7-(3-6-9)	89.3	13.9	224	57.8

classification metric. Additionally, we evaluate the computational efficiency of our model by reporting the Floating-Point Operations per Second (FLOPs) required per image.

Furthermore, to test the generalizability of the method, we also conducted experiments on the Places365-Standard dataset [33]. The dataset comprises 365 distinct scene categories and includes 1.8 million training images and 36,000 validation images, making it a comprehensive large-scale scene recognition dataset.

We use the base ViT model as the encoder architecture, which consists of 12 transformer layers. The patch size is 16x16. We pre-trained the AMAT on the ImageNet-1k for 400 epochs. The masking ratio is set to 75%, and the adamw optimizer is employed. During pre-training, the positions of the adaptive masking layers are set to (3-6-9), the base keeping ratio of tokens is set to 90%. After obtaining the weights, we fine-tuned the models with 100 epochs for supervised training under customer configurations of the adaptive masking mechanism. In this case, we only need to pre-train the model once, and subsequent fine-tuning processes with different configurations can directly use the pre-trained model for further supervised training. All the experiments are conducted on the DGX-V100 workstation with eight NVIDIA Tesla V100 GPUs.

4.2. Results on ImageNet

We conducted a comprehensive comparison between AMAT and other SOTA methods on the ImageNet-1k (ILSVRC-2012) dataset, focusing on computational complexity and classification accuracy. Additionally, to facilitate comparison, we presented the parameter number of different models together.

Table 3
Ablation experiments of the pre-training stage. For the method without adapive pre-training, we remove the token prediction branch during the pre-training stage. After the pre-training, the adaptive masking layers were stacked into the encoder and simultaneously trained with other components. The method without pre-training denote the model trained from scratch with the adaptive masking mechanism. The symbol "N/A" is short for not applicable, and "NaN" denotes that the model is unstable during training.

Model Configuration	pre-training Mode	Top-1 Acc (%)	Top-5 Acc (%)
MAE-base/16	N/A	82.9	96.3
AMAT#0.9	w/o pre-training	70.5	90.2
AMAT#0.9	w/o adaptive pre-training	79.4	94.7
AMAT#0.8	w/o adaptive pre-training	72.5	91.3
AMAT#0.7	w/o adaptive pre-training	65.4	87.1

To elaborate further, we employed diverse adaptive masking configurations for the AMAT method. Among the current mainstream ViT-based models, we selected base ViT [7], Deit [26], MAE [10], CrossViT [3], T2T-ViT [31], TNT [8], CPVT [4], Swim Transformer [16], and DynamicViT [20]. As for convolutional neural networks, we considered RegNet [19] as the most comparable competitor. As depicted in Table 1, it is evident that under the identical input resolution of 224x224, AMAT demonstrates competitive performance in terms of both FLOPs and accuracy.

In comparison to the original ViT, AMAT exhibits a marginal growth in the parameter count due to the inclusion of additional auxiliary network layers. However, in terms of FLOPs, AMAT achieves up to a 40% reduction compared to ViT. Moreover, at this computational scale, AMAT demonstrates an improvement of nearly 4% in accuracy. Furthermore, when compared to DeiT, which incorporates Knowledge Distillation (KD), AMAT achieves higher accuracy at lower FLOPs. This highlights the advantages of the MIM method in the fine-tuning procedure. CrossViT, on the other hand, possesses a higher parameter count and computational intricacy. In contrast, "AMAT#0.7-(3-6-9)" exhibits a 0.3% accuracy improvement with a complexity lower than 34% when compared to CrossViT. TNT, another relatively efficient transformer model, surpasses AMAT in terms of accuracy, although AMAT has lower complexity than TNT. Compared to DynamicViT, AMAT maintains higher accuracy at a lower complexity of 10.7 GFLOPs. It is worth noting that the DynamicViT method incorporates the utilization of a well-trained CNN model for knowledge distillation, which significantly contributes to improving the final accuracy.

Overall, the AMAT method achieves comparable accuracy to the state-of-the-art models while reducing the complexity of the ViT model.

4.3. Results on Places365

We perform the transfer task on the Places365-standard dataset in order to obtain the classification accuracy on the target dataset. Specifically, the weights of the AMAT model are inherited from the pre-trained weights on the ImageNet-1k dataset, which are then fine-tuned using the target Places365 dataset. As for the vanilla ViT model, we train it from scratch, whereas for MAE, we transfer the weights from the ImageNet-1k dataset. The detailed results are presented in Table 2.

The results indicate that Masked Image Modeling significantly enhances the final classification results. The MAE model outperforms the original ViT model by a substantial margin. Although the proposed AMAT model exhibits a slight decrease in final accuracy compared to the MAE model, specifically 0.3% worse, it offers a remarkable complexity reduction of up to 21%. This reduction underscores the effectiveness of the proposed adaptive token removal strategy. Additionally, the improved results compared to the original ViT model serve as evidence of the proposed work's generalizability and robustness on the other dataset.

4.4. Ablation Study

To effectively demonstrate the advantages of adaptive masking pre-training and adaptive fine-tuning in the AMAT framework for ViT-based autoencoders, we conducted ablation experiments on the training phase of both stages in AMAT. The objective was to assess the individual contributions of each module and their collective impact on classification performance.

Table 4
The ablation experiments were conducted for the fine-tuning procedure to examine the influence of the reconstruction regularized term. The performance of the model was compared with and without the regularized term across various keeping ratios.

Keep Ratio	Masking Location	Reconstruction	Params (M)	GFLOPs	Top-1 (%)	Top-5 (%)
0.7	3-6-9	×	89.3	13.9	82.3	96.0
0.7	3-6-9	✓	89.3	13.9	82.5	96.2
0.75	0	×	87.4	13.2	82.0	95.9
0.75	0	✓	87.4	13.2	82.2	95.9
0.25	0	×	87.4	4.5	68.5	87.4
0.25	0	✓	87.4	4.5	70.2	88.5

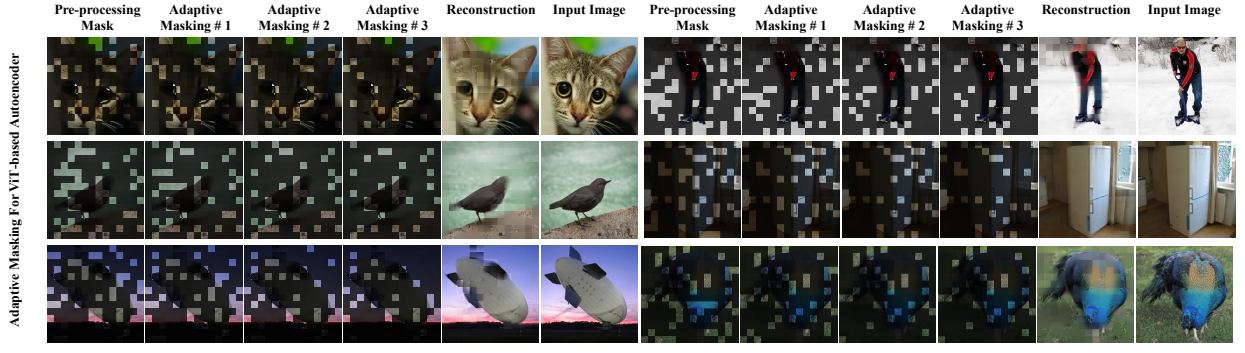

Figure 3: In the pre-training stage of the proposed AMAT, we engage in the visualization process to gain insights into its inner workings. We meticulously examine the pre-processing mask and meticulously observe the adaptive masks across all predefined locations.

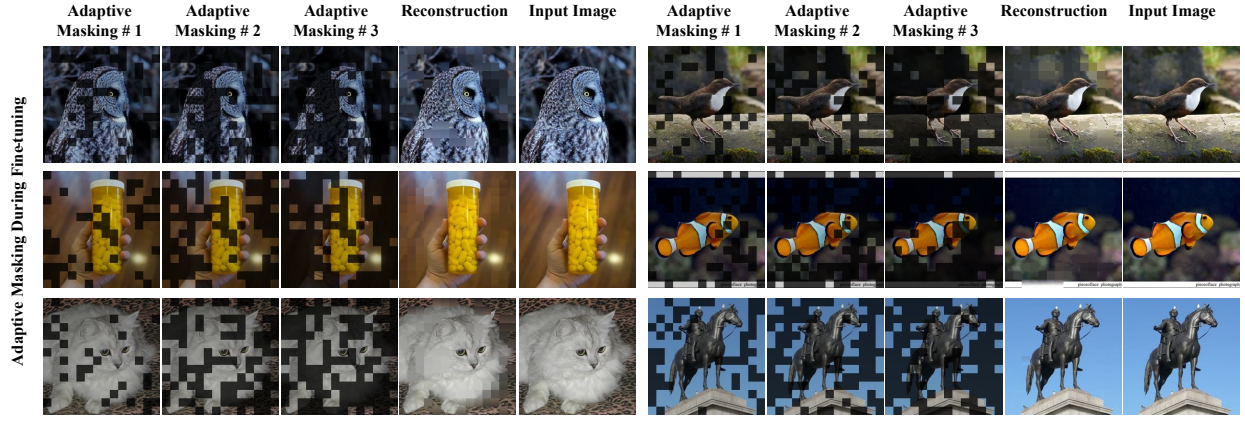

Figure 4: Visualization of the AMAT#0.7-(3-6-9) in the fine-tuning process. In addition, the adaptive masks in the locations (3-6-9) are depicted.

In order to demonstrate the importance of the pre-training stage in the proposed AMAT framework, we directly fine-tuned the base ViT model with the adaptive masking mechanism, i.e., train from scratch. Also, we conducted other experiments that did not employ the adaptive masking but used the original Masked Autoregressive strategy in the pre-training stage. We present the results in Table 3.

Table 5
The results of different locations of the prediction branches in AMAT. The reconstruction regularized term is disabled during fine-tuning.

Base Keeping Ratio	Masking Location	Params (M)	GFLOPs	Top-1 Acc (%)	Top-5 Acc (%)
0.7	3-6-9	89.3	13.9	82.3	96.0
0.7	0-5-10	89.3	12.6	82.0	95.8
0.6	3-6-9	89.3	12.5	81.7	95.8
0.6	0-5-10	89.3	10.7	81.0	95.3

As can be seen from Table 3, under the same settings, the original MAE has a Top-1 accuracy of 83% after pre-training procedure and fine-tuning procedure. However, AMAT without pre-training only has a Top-1 accuracy of about 70% even under a small sparsity. This proves the importance of pre-training for AMAT.

In addition, for the model inherited weight from the original MAE pre-training, under the same sparsity conditions, the Top-1 accuracy is significantly improved compared to the model without pre-training. However, as the number of retained tokens decreases, the classification accuracy of AMAT drops significantly. Especially at the keeping ratio of 0.6, AMAT encountered a gradient explosion and failed to complete the training process. On one side, the table also demonstrates that the pre-training process is critical for the subsequent fine-tuning process in AMAT. On the other hand, only using the original masked autoregressive learning strategy for fine-tuning of AMAT is also a sub-optimal choice, since they have significant accuracy drop, which also verifies the necessity and effectiveness of the adaptive masking pre-training we proposed.

Furthermore, in the fine-tuning stage, to verify the effectiveness of introducing the decoder, we conducted an ablation experiment on the reconstruction loss in that process. For the sake of fairness, we enable/disable reconstruction loss at the same keeping ratio and the same mask location to assess the impact on the final accuracy. The results of this experiment are presented in Table 4.

From Table 4, it can be observed that the proposed AMAT, with the utilization of reconstruction loss at a keeping ratio of 0.7, exhibits a 0.2% improvement in Top-1 accuracy compared to the method without reconstruction loss. In addition, in order to test the impact of different masking locations, a masking branch located in front of the first layer was set. Under a keeping ratio of 0.75, the Top-1 accuracy also improved by 0.2% after enabling reconstruction. Nevertheless, the corresponding Top-5 accuracy did not exhibit significant changes.

It is worth noting that at a relatively high masking ratio, the method with enabled reconstruction loss demonstrates an improvement of nearly 2% in Top-1 accuracy, compared to the method without reconstruction. This proves that it is meaningful to introduce the decoder at the same time for reconstruction training in the fine-tuning stage. Although this will introduce a fresh level of computational complexity during the training process, given that the decoder network is comparatively more streamlined than the encoder network, it can accommodate additional processing during the fine-tuning phase. Therefore, the cost of the increase training time is acceptable. Because the decoder was not used in the actual testing phase, the final computational complexity of inference is completely consistent with the model without the decoder.

4.5. Visualization

In order to better demonstrate the importance of the adaptive masking mechanism in AMAT, we visualized the masks and the corresponding reconstruction images obtained by AMAT in the training phase, and plotted them in Figures 3, 4. All displayed image were sampled from ImageNet validation set.

Figure 3 shows the visual results obtained by the pre-processed mask of "AMAT#0.9-(3-6-9)" and the corresponding adaptive masking layers for the selected image samples. In order to facilitate reading, the dark part of the image obtained after masking represents that the patch has been removed, and the bright part represents that the patch has been retained. It can be clearly seen from the figure that even if the adaptive masking strategy is used to further remove 75% tokens based on the MIM, AMAT can still restore the removed parts of the test images.

Figure 4 shows the visual results of AMAT#0.7-(3-6-9) in the fine-tuning stage. It can be seen from the figure that adaptive masking gradually removes irrelevant information from the input patches. Most of removed patches revolve around the target classification objects. For the objective of reconstruction, AMAT can also relatively easily restore the original image information from the masked image.

4.6. Various Masking Locations

In this subsection, we will investigate the impact of the positioning of adaptive masking layers in the proposed AMAT method. To explore this, we maintained a consistent ratio of masked tokens but varied their placement when fine-tuning the model. The outcomes of these experiments are reported in Table 5, which provides insights into the effect of different adaptive masking layer positions on the classification accuracy of the AMAT approach.

In Table 5, when using the same token masking ratio, different positions of masking still have a significant impact on the ViT model. Models with more masking towards the beginning show a greater reduction in computational complexity (Lower FLOPs), but they also experience a significant decrease in Top-1 accuracy. On one hand, having the pruned layer closer to the front allows subsequent layers to benefit from the efficiency gained by masking input early on. On the other hand, removing tokens at shallow layers can lead to noticeable accuracy degradation.

5. Conclusion

In this work, we have introduced a novel method called Adaptive Masking Autoencoder Transformer (AMAT) for image classification. The AMAT method effectively tackles the computational complexity of the ViT model by dynamically sparsing input image patches in a hierarchical manner during both the pre-training and fine-tuning stages. By leveraging the inherent benefits of masking image modeling, the AMAT, when combined with the original ViT architecture, achieves competitive results compared to other state-of-the-art ViT-based models while maintaining a relatively lower computational complexity.

Our research carries both theoretical and practical implications. The AMAT method offers an innovative solution to reduce the computational demands of ViT models, making them more efficient and viable for real-world applications. By dynamically sparsing image patches, AMAT optimizes the utilization of computational resources while preserving the classification performance of the ViT. This advancement opens up new possibilities for deploying ViT models in resource-constrained environments, such as edge devices and embedded systems. The experimental results robustly demonstrate the capabilities of AMAT and validate its potential as an efficient alternative to existing ViT-based models.

However, it is crucial to acknowledge the limitations of our research. While AMAT significantly reduces the computational complexity of ViT models, there may exist trade-offs concerning certain aspects of classification accuracy and model complexity. For example, the token pruning/keeping ratio is currently predefined through human intervention, which may not be optimal in terms of model complexity and accuracy. Further investigation is necessary to comprehensively evaluate these trade-offs and identify any potential limitations in specific scenarios or datasets. Additionally, the two-stage training process employed by AMAT might increase the overall training time, and finding ways to alleviate this issue requires further improvement of the proposed method.

Looking ahead, there are several promising avenues for future research in this domain. Firstly, our focus will be on enhancing the efficiency of the Vision Transformer (ViT) by exploring additional techniques and architectures. One promising direction is the development of automated methods for searching and configuring adaptive layers in the proposed AMAT. The first challenge in designing such a search method is effectively solving the dual-objective optimization problem, which balances the conflicting objectives of computational complexity and predictive accuracy. The second challenge is mitigating the high search cost incurred by evaluating hundreds of candidate configurations in the optimization of adaptive masking branches. Furthermore, it is worthwhile to investigate the applicability of the proposed AMAT in other computer vision tasks beyond image classification. Exploring its potential in object detection and video analysis could lead to significant advancements in the field.

Acknowledgements

This work was supported by the National Natural Science Foundation of China (62372315), Sichuan Science and Technology Program (24ZDZX0007, 2023ZYD0143). This work was also financially supported by the China Scholarship Council (202106240095, 202106240096). This work was also supported by the Agency for Science, Technology, and Research (A*STAR) under its AME Programmatic Funds A1892b0026, and its MTC Programmatic Funds M23L7b0021.

References

[1] Arif, Z.H., Mahmoud, M.A., Abdulkareem, K.H., Kadry, S., Mohammed, M.A., Al Mhiqani, M.N., Al Waisy, A.S., Nedoma, J., 2022. Adaptive deep learning detection model for multi-foggy images. IJIMAI 7, 26–37.

[2] Bao, H., Dong, L., Piao, S., Wei, F., 2022. BEit: BERT pre-training of image transformers, in: International Conference on Learning Representations. URL: https://openreview.net/forum?id=p-BhZSz59o4.

[3] Chen, C.F.R., Fan, Q., Panda, R., 2021. Crossvit: Cross-attention multi-scale vision transformer for image classification, in: Proceedings of the IEEE/CVF international conference on computer vision, pp. 357–366.

[4] Chu, X., Tian, Z., Zhang, B., Wang, X., Shen, C., 2023. Conditional positional encodings for vision transformers, in: The Eleventh International Conference on Learning Representations. URL: https://openreview.net/forum?id=3KWnuT-R1bh.

[5] Datta, D., Mallick, P.K., Reddy, A.V., Mohammed, M.A., Jaber, M.M., Alghawli, A.S., Al-qaness, M.A., 2022. A hybrid classification of imbalanced hyperspectral images using adasyn and enhanced deep subsampled multi-grained cascaded forest. Remote Sensing 14, 4853.

[6] Devlin, J., Chang, M.W., Lee, K., Toutanova, K., 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805 .

[7] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N., 2021. An image is worth 16x16 words: Transformers for image recognition at scale, in: International Conference on Learning Representations. URL: https://openreview.net/forum?id=YicbFdNTTy.

[8] Han, K., Xiao, A., Wu, E., Guo, J., Xu, C., Wang, Y., 2021. Transformer in transformer. Advances in Neural Information Processing Systems 34, 15908–15919.

[9] Han, X., Wang, Y.T., Feng, J.L., Deng, C., Chen, Z.H., Huang, Y.A., Su, H., Hu, L., Hu, P.W., 2023. A survey of transformer-based multimodal pre-trained modals. Neurocomputing 515, 89–106.

[10] He, K., Chen, X., Xie, S., Li, Y., Dollár, P., Girshick, R., 2022. Masked autoencoders are scalable vision learners, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 16000–16009.

[11] He, K., Gkioxari, G., Dollár, P., Girshick, R., 2017. Mask r-cnn, in: Proceedings of the IEEE international conference on computer vision, pp. 2961–2969.

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

[13] Liu, C., Chen, X., Lin, J., Hu, P., Wang, J., Geng, X., 2024. Evolving masked low-rank transformer for long text understanding. Applied Soft Computing 152, 111207.

[14] Liu, C., Wang, J., Chen, X., 2022. Threat intelligence att&ck extraction based on the attention transformer hierarchical recurrent neural network. Applied Soft Computing 122, 108826.

[15] Liu, Y., Zhang, Y., Wang, Y., Hou, F., Yuan, J., Tian, J., Zhang, Y., Shi, Z., Fan, J., He, Z., 2023. A survey of visual transformers. IEEE Transactions on Neural Networks and Learning Systems .

[16] Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B., 2021. Swin transformer: Hierarchical vision transformer using shifted windows, in: Proceedings of the IEEE/CVF international conference on computer vision, pp. 10012–10022.

[17] Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., et al., 2022. Training language models to follow instructions with human feedback. Advances in Neural Information Processing Systems 35, 27730–27744.

[18] Park, N., Kim, W., Heo, B., Kim, T., Yun, S., 2023. What do self-supervised vision transformers learn?, in: The Eleventh International Conference on Learning Representations. URL: https://openreview.net/forum?id=azCKuYyS74.

[19] Radosavovic, I., Kosaraju, R.P., Girshick, R., He, K., Dollár, P., 2020. Designing network design spaces, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 10428–10436.

[20] Rao, Y., Zhao, W., Liu, B., Lu, J., Zhou, J., Hsieh, C.J., 2021. Dynamicvit: Efficient vision transformers with dynamic token sparsification. Advances in neural information processing systems 34, 13937–13949.

[21] Ren, S., He, K., Girshick, R., Sun, J., 2015. Faster r-cnn: Towards real-time object detection with region proposal networks. Advances in neural information processing systems 28.

[22] Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A.C., Fei-Fei, L., 2015. ImageNet Large Scale Visual Recognition Challenge. International Journal of Computer Vision (IJCV) 115, 211–252. doi:10.1007/s11263-015-0816-y.

[23] Shamrat, F.J.M., Akter, S., Azam, S., Karim, A., Ghosh, P., Tasnim, Z., Hasib, K.M., De Boer, F., Ahmed, K., 2023a. Alzheimernet: An effective deep learning based proposition for alzheimer's disease stages classification from functional brain changes in magnetic resonance images. IEEE Access 11, 16376–16395.

[24] Shamrat, F.J.M., Azam, S., Karim, A., Ahmed, K., Bui, F.M., De Boer, F., 2023b. High-precision multiclass classification of lung disease through customized mobilenetv2 from chest x-ray images. Computers in Biology and Medicine 155, 106646.

[25] Simonyan, K., Zisserman, A., 2015. Very deep convolutional networks for large-scale image recognition, in: Bengio, Y., LeCun, Y. (Eds.), 3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings. URL: http://arxiv.org/abs/1409.1556.

[26] Touvron, H., Cord, M., Douze, M., Massa, F., Sablayrolles, A., Jégou, H., 2021. Training data-efficient image transformers & distillation through attention, in: International conference on machine learning, PMLR. pp. 10347–10357.

[27] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I., 2017. Attention is all you need. Advances in neural information processing systems 30.

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

[29] Wei, C., Fan, H., Xie, S., Wu, C.Y., Yuille, A., Feichtenhofer, C., 2022. Masked feature prediction for self-supervised visual pre-training, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 14668–14678.

[30] Xie, Z., Zhang, Z., Cao, Y., Lin, Y., Bao, J., Yao, Z., Dai, Q., Hu, H., 2022. Simmim: A simple framework for masked image modeling, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 9653–9663.

[31] Yuan, L., Chen, Y., Wang, T., Yu, W., Shi, Y., Jiang, Z.H., Tay, F.E., Feng, J., Yan, S., 2021. Tokens-to-token vit: Training vision transformers from scratch on imagenet, in: Proceedings of the IEEE/CVF international conference on computer vision, pp. 558–567.

[32] Zaheer, M., Guruganesh, G., Dubey, K.A., Ainslie, J., Alberti, C., Ontanon, S., Pham, P., Ravula, A., Wang, Q., Yang, L., et al., 2020. Big bird: Transformers for longer sequences. Advances in neural information processing systems 33, 17283–17297.

[33] Zhou, B., Lapedriza, A., Khosla, A., Oliva, A., Torralba, A., 2017. Places: A 10 million image database for scene recognition. IEEE transactions on pattern analysis and machine intelligence 40, 1452–1464.