Comparative Analysis of Transformer Based Models for Image Classification

Mohammad Ashraful Hoque
Computer Science and Engineering
Southeast University
Dhaka, Bangladesh
ashraful@seu.edu.bd

Md. Rafiqul Hasan
Computer Science and Engineering
Southeast University
Dhaka, Bangladesh
2018100011007@seu.edu.bd

Ummy Kolsum Mohona
Computer Science and Engineering
Southeast University
Dhaka, Bangladesh
2018100011003@seu.edu.bd

Khadijatul Kobra Sonia
Computer Science and Engineering
Southeast University
Dhaka, Bangladesh
2019200011003@seu.edu.bd

***Abstract*—This research article provides a thorough comparison of ten notable vision transformer models. In the field of computer vision, these models have drawn a lot of attention and significantly influenced the state-of-the-art in image categorization and related fields. While comparative research papers for CNN models are plentiful, comparable research for vision transformer models are sparse. This study tries to close the gap by evaluating and comparing transformers models for images across multiple dimensions. This study methodically breaks down every model's structure, explaining its architectural choices and defending their pre-training strategy, measuring their computing efficiency, and evaluating their performance on the ImageNet dataset. This paper provides an invaluable resource for the computer vision community by comparing their distinct qualities, capabilities, and limits, simplifying the selection of the most suitable vision transformer model for various real-world applications.**

***Index Terms*—Vision Transformer, Computer Vision, Image Classification, Transformer-based Models, Comparative Analysis, Self-Attention Mechanisms**

I. Introduction

The introduction of vision transformer models has radically altered the field of computer vision, resulting in a fundamental shift in image classification and related applications. With their distinctive architecture and pre-training techniques, vision transformers have become effective substitutes for traditional convolutional neural networks (CNNs), completely changing the landscape of computer vision [1], [2]. In particular, the vision community is curious to know if transformers can successfully compete with the widely used Convolutional Neural Network-based architectures (CNNs) in vision tasks, such as DETR [3], it isthe first study that use a transformer for object detection, ResNet [4] and EfficientNet [5], given the remarkable performance of transformers in natural language processing (NLP). As these models expand and vary, it becomes increasingly important to undertake a thorough comparison study that sheds light on their specific characteristics, advantages, and trade-offs.

Recently, Transformers, self-attention models for language modeling, have been applied to vision tasks such as detecting objects, image classification, noise reduction and super-resolution. [6]. Given the increasing growth of transformer based models, it is critical to undertake an in- depth analysis that sheds light on their the intricate detail and relative advantages. This article attempts to meet this demand by examining ten leading vision transformer models, including their architectural complexities, pre-training procedures, computational efficiency, and performance on major benchmark datasets, with a focus on ImageNet [7].

In the following sections, This paper will go through ten different models in detail. This study aims to provide an invaluable resource for the computer vision community by evaluating their strengths and drawbacks, making it easier to select the best vision transformer model for various real-world applications. The structure of the paper is as follows: The introduction will provide an overview of the study's aims. The following part, paper summary will go over the ten transformer models in detail, focusing on their distinguishing features and training configurations. Following that, a comparison analysis section will look at the differences between the models based on specific parameters. The key observations part will present our findings and conclusions after studying these ten models. Finally, the conclusions will summarize the findings and suggest future study possibilities.

II. Paper Summary

This section will provide a complete overview of ten transformer models technological improvements and strategies, with a focus on architectural changes and performance benefits.

A. Vision Transformer (ViT)

By utilizing a Transformer-like architecture over regions of the image, The Vision Transformer (ViT) presents a novel method for categorizing images. Beginning with fixed patch segmentation, an image is then linearly modified and enhanced with position embeddings. The typical Transformer encoder is then fed this series of embeddings. The tried-and-true method of adding an extra classification token that can be learned to the sequence is used to enable categorization [8]. Vision Transformer performs well on multiple small or medium-sized image recognition benchmarks (ImageNet, CIFAR-100, VTAB, etc.) with a significant reduction in CPU usage during training [8]. The best results come from supervised pre-training, however, this is not true for NLP. The authors also utilized masked patched prediction, a self-supervised pre-training aim inspired by masked language modeling, in a different experiment.Seven pre-training epochs were followed by 300 epochs of training for this model. Using the Adam optimizer, the batch size was 4096. Although this method improves accuracy over training from scratch, the smaller ViT-B/16 model still performs 4% less accurately on ImageNet (79.9%) than supervised pre-training [8]. The architecture of this model can be seen in Figure 1.

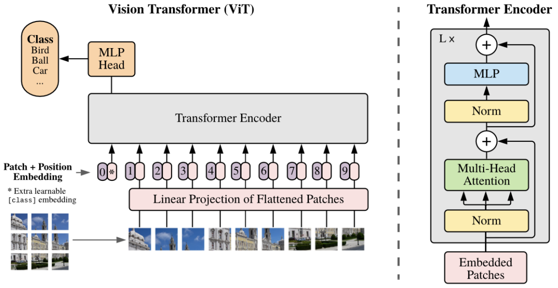

Fig. 1. Architecture of Vision Transformer [8].

B. Data-Efficient Image Transformer (DeiT)

For image classification tasks, a sort of Vision Transformer is known as a Data-Efficient Image Transformer. A transformer-specific teacher-student training method is used to train the model. It depends on a distillation token to make sure that the pupil pays attention to the teacher and learns from them [9]. There are no variations from the architectural design suggested by Dosovitskiy et al. (2020). The distillation token and training methods are the only variations. Additionally, they simply utilize a linear classifier for the pre-training instead of an MLP head. To avoid any mistake, the author prefixes the results with DeiT and refers to the findings of the earlier work by ViT. DeiT-B384, for instance, is the final operating resolution obtained after fine-tuning DeiT at a higher resolution. Finally, the parameters of ViT-B (and hence of DeiT-B) are fixed as follows: D = 768, h = 12, and d = D/h = 64 when using the distillation approach, which the authors refer to as DeiT [9]. The DeiT-S and DeiT-Ti are two further compact variations that they introduce. They maintain d constant while varying the number of heads. The AdamW optimizer was used to train DeiT over 300 epochs with a batch size of 1024 [9]. This reference vision transformer (86M parameters) attains top-1 accuracy of 83.1% (single-crop) on ImageNet without requiring external data [9]. Figure 2 shows the structure(distillation procedure) of this model.

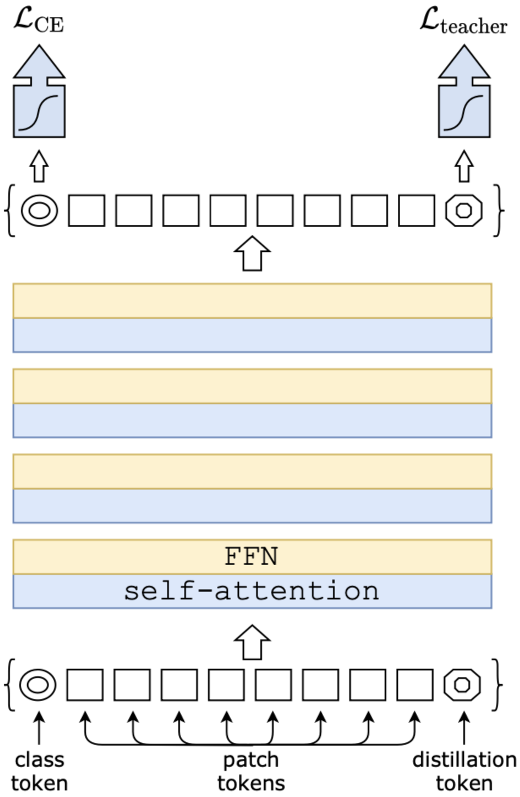

Fig. 2. Architecture of DeiT (distillation procedure) [9].

C. CrossViT

To classify images, CrossVit uses a dual-branch architecture to collect multi-scale information. In this method, various-sized image patches (tokens) are processed utilizing independent branches, which repeatedly fuse these tokens to increase their total strength [10]. This fusion process takes place using a cross-attention module that functions successfully. A non-patch token that functions as a messenger and encourages communication between the two branches is produced by each transformer branch. This process substantially accelerates the development of the attention map during fusion. The performance of the ViT is significantly impacted by the granularity of patch sizes. Although they need more processing resources, smaller patch sizes produce better results. Although they need more processing resources, smaller patch sizes produce better results. A patch size of 16 performs 6% better than a patch size of 32, but it requires four times as many computer processes. The suggested method seeks to take advantage of smaller patch sizes while retaining a manageable level of complexity. [10] The model created by the author comprises numerous multi-scale transformer encoders. Two branches are integrated by each encoder:

a. The L-Branch (large) uses coarse-grained patches that have more encoders and bigger embedding dimensions [10].

b. The S-Branch (small) makes use of fine-grained patches, which have fewer encoders and smaller embedding dimensions.

The final CLS tokens of both branches are used for prediction throughout these branches' many iterations of fusion. Before entering the multi-scale transformer encoder, each token in both branches includes a learnable position embedding. Acquiring multi-scale feature representations relies heavily on effective feature fusion. AdamW optimizer were used in CrossViT's 300 period training process with a batch size of 4096. Another 30 epochs were added to its fine tuning [10].

It's interesting to note that, In comparison to VIt-B crossVit-18 requires 50% less floating point operations per second (FLOPs) and parameters and succeeds 4.9% more accurately. This technique, however with some architectural alterations, significantly outperforms the newest DeiT by 2% on the ImageNet1K dataset, with minimal to moderate increases in FLOPs and model parameters [10]. Figure 3 shows the architecture of CrossViT model.

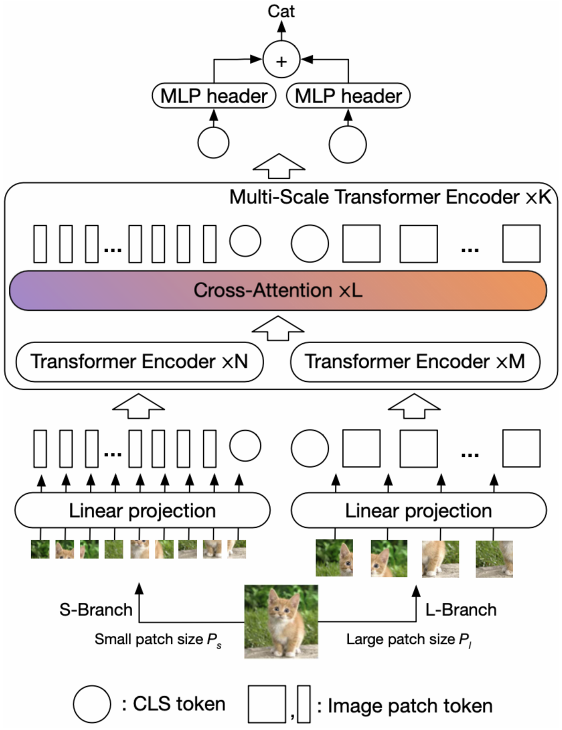

Fig. 3. Architecture of CrossViT [10].

D. Tokens-to-Token ViT (T2T-ViT)

T2T-ViT adopts a layer-wise Tokens-to-Token transformation strategy to capture local structures by amalgamating neighboring tokens. The two parts of the T2T-ViT are the Tokens-to-Token (T2T) module and the T2T-ViT backbone. The T2T module can be designed in a variety of ways. Since n is set to 2, There will be a n = 2 reorganization and an n+1 = 3 soft split of the T2T module. The three soft splits are P = [7, 3, 3] and S = [3, 1, 1] in terms of patch sizes, which reduces the size of the input image from 224*224 to 14*14. While the T2T-ViT backbone functions similarly to ViT in that it receives tokens from the T2T module of a predetermined length, it is designed with a deep-narrow structure that has less hidden dimensions (256-512) and MLP size (512-1536) than ViT. T2T-ViT-14 contains 14 transformer layers in its T2T-ViT backbone and 384 hidden dimensions, whereas ViT-B/16 has 12 transformer layers, 768 hidden dimensions, and is three times larger than T2T-ViT-14 in terms of parameters and Multiply-Accumulate Operations (MACs) [11].T2TViT was trained using the AdamW optimizer for 310 epochs with a batch size of 1024.

This T2T-Vit improves performance on ImageNet by 3.0% by reducing parameters and computation by 50% when paired with an effective deep-narrow backbone. T2T-ViT competes with ResNets [4] and MobileNets [12], reaching 83.3% accuracy on ImageNet at 384384 resolution and ResNet50-like size [11]. The T2T-ViT model's architecture is shown in Figure 4.

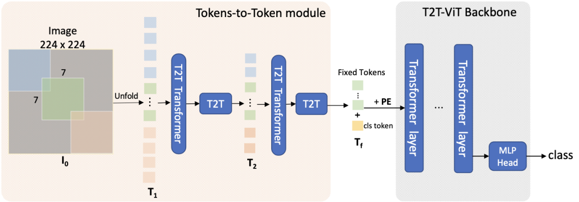

Fig. 4. Architecture of Tokens-to-Token ViT [11].

E. Masked Autoencoders (MAE)

MAE are extensible self-supervised learning for computer vision. This approach is straightforward: image patches with random input are masked, and the pixels that are missing are replaced. It relies on two primary concepts. They provide a lightweight encoder-decoder architecture in the first place, with an encoder that only processes the visible region of patches (i.e., without mask tokens) and a lightweight decoder that uses mask tokens and the implicit representation in order to recreate the original image. Second, they demonstrate that obscuring a significant percentage of the input image—say, 75%—produces a relevant and challenging assignment for self-supervision. They can train large models rapidly and effectively while also enhancing accuracy (by a factor of three or more) by combining these two designs. A basic ViT-Huge model obtains its highest accuracy (87.8%) among methods that only employ ImageNet-1K data because of its scalable methodology. The models are trained for 800 epochs using batch size 4096 [13]. Figure 5 illustrates the MAE model's architecture.

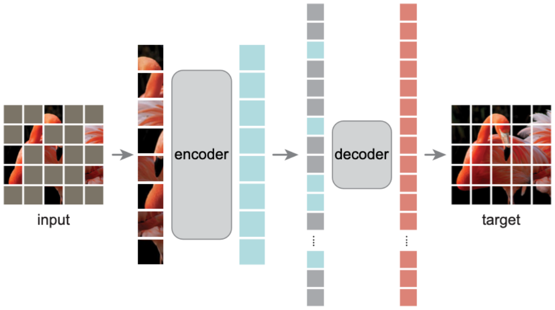

Fig. 5. Architecture of MAE [13].

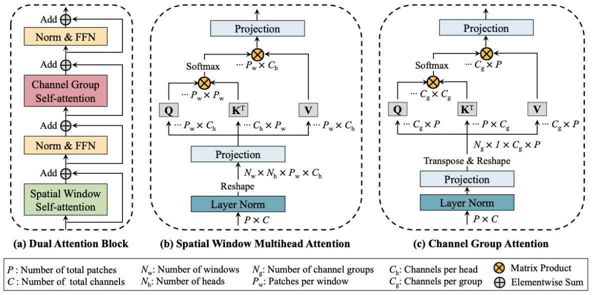

Fig. 6. Architecture of DaViT [14].

F. Dual Attention Vision Transformers (DaViT)

A vision transformer design called Dual Attention Vision Transformers (DaViT) successfully captures the global context while preserving computing efficiency. The strategy makes use of spatial tokens and channel tokens to harness self-attention mechanisms. Channel dimensions are used for token attributes in channel tokens, whereas spatial dimensions are used for scope in spatial tokens. For both categories, tokens are grouped by the direction of the sequence to maintain linear complexity. Together, these attentions capture global interactions by accounting for all spatial placements when computing scores, while spatial attention sharpens local representations to support the modeling of global information [14]. Window inside Multihead self-attention splits the spatial dimension into local windows, with many spatial tokens included in each of the windows. Furthermore, every token has several heads [14]. Single-head self-attention groups channel tokens into many groups. Using an entire image-level channel as a token that transmits globally relevant data, each channel group pays attention. By taking into account all spatial positions while calculating attention scores between channels, the channel tokens, which include abstract representations of the complete image, channel attention to naturally record global representations and interactions. Second, fine-grained interactions across spatial places are made easier by the spatial attention mechanism, which improves local perceptions. As a result, the channel attention mechanism's modeling of global information is improved [14]. DaViT was trained using the AdamW optimizer for 300 epochs with a batch size of 2048 [14]. DaViT efficiently completes four tasks with cutting-edge results. On ImageNet-1K, DaViT models of various sizes reach up to 84.6 [14]. Figure 6 shows the architecture of the DaViT model.

G. PoolFormer

Transformers have shown a great deal of potential for computer vision applications. Most experts agree that their attention-based token mixer module contributes the most to their proficiency. According to recent research, transformers can still provide models that perform rather well even when the attention-based module is replaced by spatial Multi-Layer Perceptrons (MLPs). This finding prompts them to postulate that the performance of the model depends more on the overall transformer architecture than it does on the specific token mixer module. To illustrate this point, they replace an embarrassingly basic spatial pooling operator for the attention module on transformers in order to simplify the most basic token mixing. Through competitive performance on different computer vision tasks, the developed PoolFormer model is established. For instance, this architecture achieves 0.3%/1.1% accuracy on ImageNet-1K compared to the well-tuned vision transformer/MLP-like baselines DeiT-B/ResMLP-B24, although with 35%/52% less parameters and 48%/60% fewer MACs required. This model additionally achieves 82.1% top-1 accuracy on ImageNet-1K. Batch size 4096 is used to train the models for 300 epochs [15]. Figure 7 depicts the PoolFormer model's architecture.

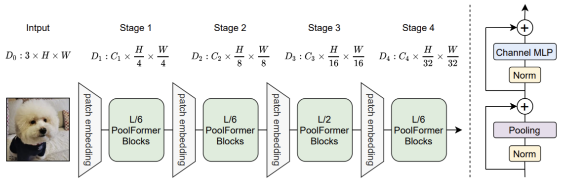
Intput
Stage 1
Stage 2
Stage 3
Stage 4
$D_0 : 3 \times H \times W$
$D_1 : C_1 \times \frac{H}{4} \times \frac{W}{4}$
$D_2 : C_2 \times \frac{H}{8} \times \frac{W}{8}$
$D_3 : C_3 \times \frac{H}{16} \times \frac{W}{16}$
$D_4 : C_4 \times \frac{H}{32} \times \frac{W}{32}$
patch embedding
L/6 PoolFormer Blocks
patch embedding
L/6 PoolFormer Blocks
patch embedding
L/2 PoolFormer Blocks
patch embedding
L/6 PoolFormer Blocks
Channel MLP
Norm
Pooling
Norm

Fig. 7. Architecture of PoolFormer [15].

H. Bidirectional Encoder representation from Image Transformers (BEiT)

BEiT is a self-supervised vision representation approach. They suggest using ”masked image modelling,” which is modeled after Bidirectional Encoder Representations from Transformers (BERT) [16], an innovator in natural language processing, for pre-training vision transformers. As an illustration, each image in their pre-training has two views: image patches (16x16 pixels) and visual tokens (i.e., discrete tokens). To generate visual tokens, they first ”tokenize” the original image. Then, a number of randomly applied image patches were supplied into the backbone transformer. Recovering the original visual tokens from the deteriorated image patches is one of the pre-training targets. After pre-training BEiT, they swiftly adjust the model parameters in downstream positions by adding task layers to the encoder.

The model outperforms past pre-training strategies, according to experimental results on semantic segmentation and image classification. On ImageNet-1K, for instance, base-size BEiT achieves 83.2% top-1 accuracy, significantly exceeding from-scratch DeiT training. Additionally, large-size BEiT surpasses ViT-L when pre-trained under supervision on ImageNet-22K (85.2%), achieving 86.3% while using only ImageNet-1K. The pre-training runs for 800 epochs with a 2k batch size [17]. The BEiT model's architecture is displayed in Figure 8.

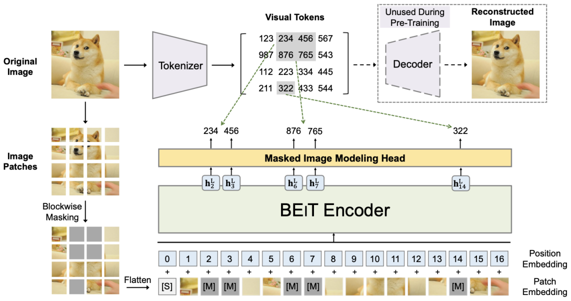

Fig. 8. Architecture of BEiT [17].

I. CrossFormer++

Even though characteristics of different sizes are perceptually important to visual inputs, contemporary vision transformers do not yet explicitly exploit them. They initially proposed CrossFormer, a cross-scale vision transformer, to do this. The cross-scale embedding layer (CEL) and long-short distance attention (LSDA) are introduced by this paper. CEL provides cross-scale attributes to the self-attention module involves using a single token to combine several patches that have different scales. LSDA maintains the tokens' small-scale and large-scale properties while splitting the self-attention module into a short-distance and long-distance equivalent [18]. They present the progressive group size (PGS) paradigm and the amplitude cooling layer (ACL). CrossFormer++ is the CrossFormer which incorporates ACL and PGS. Batch size 1024 was employed to train the models for 300 epochs. A number of experiments suggest that CrossFormer++ works better than the other visual transformers in tasks including semantic segmentation, object detection, instance segmentation, and image classification [18]. Figure 9 represents the CrossFormer++ model's architecture.

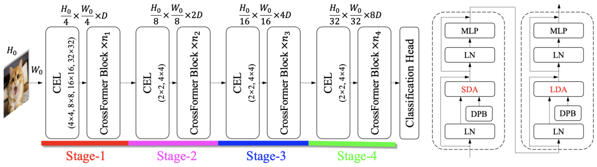

Fig. 9. Architecture of CrossFormer++ [18].

J. FastViT

FastViT is a hybrid vision transformer design that achieves the fastest accuracy/latency trade-off. By eliminating skip-connections from the network, they introduce the RepMixer building block of FastViT, a unique token mixing operator that makes use of structural reparameterization to lower memory access costs. They use huge kernel convolutions and train time over parametrization to boost accuracy, and research results suggest this strategy has very little effect on latency. On the ImageNet dataset, they demonstrate that, with the same accuracy, this model outperforms ConvNeXt [19] by 1.9 times, EfficientNet [5] by 4.9 times, and Cascaded Memory Transformer (CMT), a recently developed state-of-the-art hybrid transformer architecture, by 3.5 times. This model performs 4.2% better at the same latency than MobileOne's Top-1 accuracy on ImageNet. many different applications, such as 3D mesh regression, image segmentation, detection, and classification. The models pre-training length is 300 epochs with batch size 1024 [20]. Figure 10 exhibits the fastViT model's architecture.

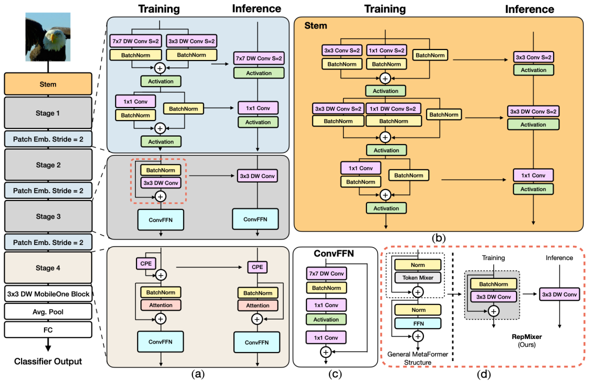

Fig. 10. Architecture of FastViT [20].

III. Comparative Analysis

The development of vision transformer models has significantly contributed to the area of computer vision, offering new perspectives on the processing and comprehension of images. This comparative study focuses on ten advanced vision transformer models. Each of these models brings to Table 1 distinct architectural innovations, pre-training methods, and computational efficiency when it comes to applying transformer architectures to complex visual issues. By carefully examining crucial elements such as architecture, pre-training methods, ImageNet accuracy, and computational efficiency, this analysis aims to provide a comprehensive understanding of the benefits and drawbacks of each model, helping practitioners and researchers select the most appropriate model for their specific use cases. The columns in the comparative table were designed to provide an exhaustive outline of each vision transformer model's essential elements and attributes, making it easier to examine and compare them. The reasoning behind each column is as follows:

- **Model:** The name of each vision transformer model is specified in this column, and it serves as the primary identifier for reference. With variations in design, architecture, and training strategies, each model provides a distinct method of applying transformer architectures for computer vision tasks.
- **Architecture:** This column emphasizes each model's essential architectural style, allowing readers to grasp the underlying structure, such as whether it is based on the Transformer architecture or adds extra components.
- **Pre-Training Strategy:** This column describes the approach used for pre-training the model. This includes pre-training datasets (e.g., ImageNet) data augmentation methods, learning rate schedules, optimization algorithms, and other training approaches designed to improve model performance prior to training. Understanding the pre-training strategy is crucial because it impacts the model's performance and ability to learn meaningful representations from data. BEiT and MAE are trained for 800 epochs; the rest of the eight models are trained for 300 epochs using AdamW optimizer, with a batch size of 1024 to 4096.
- **ImageNet Accuracy:** This column comprises a performance statistic that indicates the model's accuracy when tested against the ImageNet 1k dataset. A widely accepted standard for image classification tasks is ImageNet accuracy, which is used to assess the model's overall performance. To provide an in-depth comparison, variations in accuracy across various model sizes, training datasets, and pre-training processes are demonstrated.
- **Computational Efficiency:** This column provides information on the model's computing requirements, such as parameter count, memory usage, and processing speed. Computational efficiency is critical for practical deployment, particularly in resource-constrained situations.
- **Additional Notes:** This column contains brief additional information about each model. It may include distinguishing characteristics, benefits, or distinctions that set the model apart from others. These notes provide context and assist readers in understanding what distinguishes each model.

The table I provides a comprehensive perspective of each vision transformer model, which encompass architectural, performance, and practical factors. This enables readers to make educated conclusions about which model best meets their individual needs and limits.

A. Key Observations

Diverse Architectural Approaches: The study emphasizes the variety of architectural designs found in vision transformer models. While all of them use the Transformer framework, each model incorporates distinct architectural aspects to meet specific issues in image classification and associated activities.

Pre-Training Strategies: The research emphasizes the significance of vision transformer pre-training procedures. Models such as ViT and BEiT use supervised and unsupervised pre-training, demonstrating the importance of these methodologies in achieving high accuracy on image classification benchmarks. In contrast to other models trained for 300 epochs, the BEiT model, which used a batch size of 2k, and the MAE model, which used a batch size of 4,096 and was trained for 800 epochs, converging more slowly. This indicates variations in training dynamics and optimization tactics.

ImageNet Accuracy: The comparative research demonstrates that the vision transformer models have various levels of ImageNet 1k accuracy. This metric is used to evaluate their overall performance, with certain models obtaining top-tier results.

Computational Efficiency: The study delves into the computational efficiency of each model, taking into account parameters like parameter count, memory utilization, and processing speed. These findings are critical for real-world application, particularly in resource-constrained settings.

Innovative Features: Several models incorporate creative approach to improve their performance. CrossFormer++, for example, emphasizes the importance of cross-scale properties in visual inputs, whereas PoolFormer substitutes attention-based modules with spatial MLPs.

This comparison analysis gives a high-level overview of the essential characteristics of these ten vision transformer models, allowing beneficiaries to make educated decisions about their viability in various computer vision applications. Models can be chosen by researchers and practitioners based on their individual criteria for accuracy, efficiency, and architectural preferences.

IV. Conclusion

This investigation compared ten significant vision transformer models, offering insight into their architecture, meticulous pre-training processes, computational effectiveness, and ImageNet 1k benchmark dataset performance. This research offers scholars and professionals the necessary understanding to make well-informed selections of models by contrasting their advantages and disadvantages. The dynamic field of vision transformer models emphasizes how important comparison evaluations like these are in shaping the direction of computer vision research.

References

[1] Kai Han, Yunhe Wang, Hanting Chen, Xinghao Chen, Jianyuan Guo, Zhenhua Liu, Yehui Tang, An Xiao, Chunjing Xu, Yixing Xu, Zhaohui Yang, Yiman Zhang, and Dacheng Tao. A survey on vision transformer. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(1):87–110, 2023.

[2] Qizhe Xie, Eduard H. Hovy, Minh-Thang Luong, and Quoc V. Le. Self-training with noisy student improves imagenet classification. *CoRR*, abs/1911.04252, 2019.

[3] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. *CoRR*, abs/2005.12872, 2020.

TABLE I
COMPARATIVE ANALYSIS OF VISION TRANSFORMER MODELS

Model	Architecture	Pre-Training Strategy	ImageNet Accuracy	Number of Epochs	Computational Efficiency	Additional Notes
ViT	Transformer-based	Supervised/Self-Supervised	79.9%	300	Efficient Training	Patch-based, Transformer Encoder, Learnable Tokens
BEiT	Self-Supervised	Masked Image Modeling	86.3%	800	Competitive Efficiency	BERT-inspired, Improved Pre-Training
ViTMAE	Self-Supervised	Masked Autoencoders	87.8%	800	Scalable and Efficient	Masked Autoencoders, Scalable Training
CrossFormer++	Cross-Scale Vision	Cross-Scale Attention	84.9%	300	Versatile Performance	Long-short Distance Attention, Multi-Task Support
PoolFormer	Attention Replacement	Spatial MLPs	82.1%	300	Efficient and Lightweight	Spatial MLPs, Token Mixing Operator, Performance
FastVIT	Hybrid Architecture	RepMixer Building Block	84.9%	300	Balanced Latency-Accuracy	Hybrid Architecture, Reparameterization
CrossViT	Dual-Branch	Multi-Scale Feature Capture	82.8%	300	Improved Classification	Multi-Scale Feature Fusion, Efficient Architecture
T2T-ViT	Tokens-to-Token	Layer-Wise Transformation	83.3%	310	Parameter Reduction	Tokens-to-Token Transformation, Efficient Backbone
DeiT	Transformer-Based	Teacher-Student Strategy	85.2%	300	Knowledge Distillation	Distillation Token, Teacher-Student Training Strategy
DaViT	Dual Attention	Global Context Capture	84.6%	300	Computational Efficiency	Dual Attention Mechanisms, Strong Global Context

[4] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016.

[5] Mingxing Tan and Quoc Le. EfficientNet: Rethinking model scaling for convolutional neural networks. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 6105–6114. PMLR, 09–15 Jun 2019.

[6] Hanting Chen, Yunhe Wang, Tianyu Guo, Chang Xu, Yiping Deng, Zhenhua Liu, Siwei Ma, Chunjing Xu, Chao Xu, and Wen Gao. Pre-trained image processing transformer. *CoRR*, abs/2012.00364, 2020.

[7] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, 2009.

[8] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale, 2021.

[9] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training data-efficient image transformers distillation through attention, 2021.

[10] Chun-Fu Chen, Quanfu Fan, and Rameswar Panda. Crossvit: Cross-attention multi-scale vision transformer for image classification, 2021.

[11] Li Yuan, Yunpeng Chen, Tao Wang, Weihao Yu, Yujun Shi, Zihang Jiang, Francis EH Tay, Jiashi Feng, and Shuicheng Yan. Tokens-to-token vit: Training vision transformers from scratch on imagenet, 2021.

[12] Andrew G. Howard, Menglong Zhu, Bo Chen, Dmitry Kalenichenko, Weijun Wang, Tobias Weyand, Marco Andreetto, and Hartwig Adam. Mobilenets: Efficient convolutional neural networks for mobile vision applications. *CoRR*, abs/1704.04861, 2017.

[13] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners, 2021.

[14] Mingyu Ding, Bin Xiao, Noel Codella, Ping Luo, Jingdong Wang, and Lu Yuan. Davit: Dual attention vision transformers, 2022.

[15] Weihao Yu, Mi Luo, Pan Zhou, Chenyang Si, Yichen Zhou, Xinchao Wang, Jiashi Feng, and Shuicheng Yan. Metaformer is actually what you need for vision, 2022.

[16] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina N. Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. 2018.

[17] Hangbo Bao, Li Dong, and Furu Wei. Beit: BERT pre-training of image transformers. *CoRR*, abs/2106.08254, 2021.

[18] Wenxiao Wang, Lu Yao, Long Chen, Binbin Lin, Deng Cai, Xiaofei He, and Wei Liu. Crossformer: A versatile vision transformer hinging on cross-scale attention, 2021.

[19] Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A convnet for the 2020s. *CoRR*, abs/2201.03545, 2022.

[20] Pavan Kumar Anasosalu Vasu, James Gabriel, Jeff Zhu, Oncel Tuzel, and Anurag Ranjan. Fastvit: A fast hybrid vision transformer using structural reparameterization, 2023.