MULTI-EXIT VISION TRANSFORMER WITH CUSTOM FINE-TUNING FOR FINE-GRAINED IMAGE RECOGNITION

Tianyi Shen, Chonghan Lee, and Vijaykrishnan Narayanan

The Pennsylvania State University
Dept. of Electrical Engineering and Computer Science
State Colege, USA

ABSTRACT

Capturing subtle visual differences between subordinate categories is crucial for improving the performance of Fine-grained Visual Classification (FGVC). Recent works proposed deep learning models based on Vision Transformer (ViT) to take advantage of its self-attention mechanism to locate important regions of the objects and extract global information. However, their large number of layers with self-attention mechanism requires intensive computational cost and makes them impractical to be deployed on resource-restricted hardware including internet of things (IoT) devices. In this work, we propose a novel Multi-exit Vision Transformer architecture (MEViT) for early exiting based on ViT, as well as a fine-tuning strategy that involves self-distillation to improve the accuracy of early exit branches on FGVC task compared to the baseline ViT model. The experiments on two standard FGVC benchmarks show our proposed model provides superior accuracy-efficiency trade-offs compared to the state-of-the-art (SOTA) ViT-based model and demonstrate that it is possible to accurately classify many subcategories with significantly less effort.

1. INTRODUCTION

Fine-grained Visual Classification (FGVC) aims to accurately identify discriminative local parts and features from visually similar subcategories, such as different bird species or various car models. Recently, Transformer model architecture, which has led to successes in Natural Language Processing (NLP) tasks, has also been applied to the computer vision domain and resulted in high performance in vision tasks. Vision Transformer (ViT) has demonstrated high performance in regular classification tasks [1]. In a similar manner in which Transformer divides a sentence into words and learns the correlation between each word, ViT splits an image into a series of ordered image patches and learns the association between the image patches. Specifically, a series of specialized ViT-based models are proposed and achieved better performance in FGVC tasks compared to the existing convolutional neural networks (CNNs) [2, 3, 4, 5, 6, 7, 8, 9]. TransFG has introduced a ViT-based framework that generates overlapping image patches with a sliding window to avoid any information loss among hard-split patches [10]. Another work proposed a Feature Fusion Vision Transformer (FFVT) which aggregates features of important patches from early ViT layers to extract low-level and middle-level information effectively [11]. However, they come with a high computational cost and the slow inference speed of these models could hinder their deployment on edge devices for real-world applications.

While Deep Neural Networks including Transformer architectures gain an advantage from a large number of layers, it is often found that fewer layers can still precisely identify a large number of classes in classification tasks. There have been numerous studies exploring the concept of an early exit in deep neural networks, which involves exiting the network in earlier layers than the normal exit point. Initially, [12] highlighted that classification difficulty varies widely across data classes in real-world datasets and only a small portion of classes requires the full compute power of the model, while the majority can be identified precisely with minimal effort. BranchyNet augmented the existing early CNN model architectures with additional side branch classifiers and achieved superior accuracy-efficiency trade-offs compared to the original models [13]. The early exit mechanism was also introduced in Transformer architectures as an effective dynamic inference framework for low-resolution image classification tasks with broad categories [14]. In this paper, we propose a novel multi-exit ViT model named MEViT that is able to adaptively perform the early exit based on a preset threshold and a fine-tuning strategy that utilize the inplace distillation to solve the knowledge degradation problem of the early-exit model in FGVC tasks.

The main contribution of this work includes:

- We introduce a Multi-exit Vision Transformer architecture (MEViT) with an uncertainty score predictor module to dynamically exit on earlier branches based on the predicted threshold value, which is determined by the input image's difficulty level.

This work was supported in part by Semiconductor Research Corporation and Center for Brain-inspired Computing.

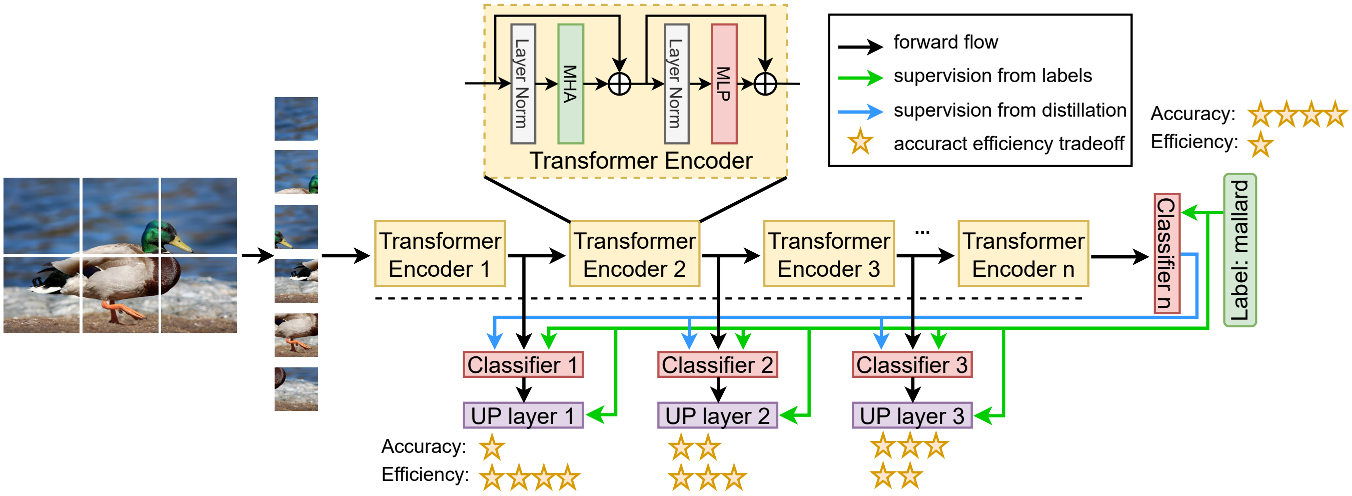

Fig. 1. The overview of our proposed Multi-exit Vision Transformer. The input sequence of image patches are fed to the model. The uncertainty predictor module (UP) is introduced to the exit branches to dynamically decide on which layer to exit based on the difficulty level of the input images.

- We apply layer-wise inplace distillation and sandwich rule technique during training to prevent the model from conflicting feedback from multiple branch classifiers while updating all the exit points without compromising the performance of the full model.

Empirical results show that MEViT is capable of exiting early during inference and remarkably cuts down computational cost while achieving superior performance on two FGVC benchmarks compared to the baseline ViT and a SOTA model. Our model can dynamically exit early based on the difficulty level of the input images.

2. METHOD

2.1. Multi-Exit Structure in ViT

We begin with the pre-trained ViT model as the backbone, which consists of 12 encoder layers with 12 attention heads in each layer. Then, we append auxiliary classifiers after each encoder layer to represent the early exit branches. Each branch classifier is a single-layer fully-connected network. To allow the early-exit mechanism to be adaptive and enable dynamic multi-level exit points for classes at different levels of difficulty, we introduce an uncertainty score predictor module during training similar to [15] along with the early branch classifiers to estimate the uncertainty level (the lower the value, the better the quality) to determine the prediction quality. The model will terminate on a certain branch when the uncertainty level falls below a pre-defined threshold.

2.2. Training Strategy of Multi-Exit ViT

In our preliminary experiments, we focused on training the model in an end-to-end manner where the loss signals of all exits are combined and backpropagated through the network at the same time. However, we found that the different signals from multiple exit branches with independently initialized weights greatly hinder optimization when they are simultaneously updating the model. We address this issue with layer-level inplace distillation technique, following the sandwich rule training approach [16]to make the early exit classifiers learn from the probability distribution of the final classifier to prevent conflicting feedback to the backbone during backpropagation.

We applied the layer-level sandwich rule training technique to effectively train our multi-exit model. First, we update the model with the upper bound which is the full model with the final classifier. Then we apply inplace distillation to update the model with the lower bound which includes the earliest exit branch classifier and other randomly sampled early exit branch classifiers to transfer the knowledge from the full model's final classifier to the sparse models with various exit points. In each iteration, both the full and spare models are optimized simultaneously to enable the model to be adaptable to arbitrary exit points.

The loss function for the i_{th} layer classifier is defined as

$$L_i(x,y) = L_{CE_i}(f_i(\mathbf{h}_i), y) + L_{MSE_i}(f_i(\mathbf{h}_i), \tilde{u}_i) \quad (1)$$

where $\mathbf{h}_i$ is the i_{th} layer hidden state corresponding to the CLS token. L_{CE_i} is cross-entropy loss between the i_{th} classifier f_i and the ground truth label y. L_{MSE_i} is mean squared error (MSE) from the i_{th} layer uncertainty predictor module defined as

$$L_{MSE_i}(f_i(\mathbf{h}_i), \tilde{u}_i) = \|u_i - \tilde{u}_i\|_2^2 \quad (2)$$

where

$$u_i = \sigma\left(\mathbf{w}^\top f_i(\mathbf{h}_i) + b\right) \quad (3)$$

σ is the sigmoid function. $\mathbf{w}$ is the weight vector and b is the bias of the uncertainty predictor module. u_i is the predicted uncertainty level and $\tilde{u}_i$ is the ground truth uncertainty level defined as

$$\tilde{u}_i = \tanh\left(\left|f_i\left(\mathbf{h}_i\right) - y\right|\right) \tag{4}$$

When the model is updated with the lower bound configuration to exit early, we additionally compute Kullback-Leibler divergence loss L_{KLDiv_i} and update the model as shown below.

$$L_{KLDiv_i} = f_n(\mathbf{h}_n) \ln \frac{f_n(\mathbf{h}_n)}{f_i(\mathbf{h}_i)} \tag{5}$$

where f_n and $\mathbf{h}_n$ are the final classifier and the final hidden state corresponding to the CLS token respectively. The additional inplace distillation loss during the fine-tuning of early exit classifiers prevents negative interference among multiple classifiers, leading to optimal performance.

3. EXPERIMENT

3.1. Experiment Setup

Our multi-exit model (MEViT) is based on the ViT model (ViT-B/16) that is pre-trained on the Imagenet21K dataset [17]. MEViT consists of an embedding layer, 12 encoder layers, and 6 corresponding classification heads starting from the 6_{th} encoder layer. All the classifiers are single-layer fully connected layers. We evaluate the performance of our model with two popular FGVC datasets including FGVC-Aircraft[18] and Stanford cars[19]. We followed the same data augmentation used by TransFG and randomly cropped training images to have the size of 448×448. During the fine-tuning stage, we used SGD optimizer with a momentum of 0.9 and applied the cosine annealing scheduler. The initial learning rate is set to 0.001 and 0.003 for FGVC Aircraft and Stanford cars datasets respectively. The training batch size is set to 32 for all the datasets. We trained the models on a single Nvidia Tesla V100 GPU and run inference on the GPU and an Nvidia Jetson Nano to validate their performance.

3.2. Evaluation metrics

We compare our model to the baseline ViT model and TransFG, which is a ViT-based SOTA model that outperforms CNN-based approaches in FGVC tasks. We fine-tuned the models on the two datasets mentioned earlier to evaluate their accuracy and efficiency. To evaluate the efficiency of our model on various hardware, we focused on two different metrics: the number of floating operations (FLOPs) required to run a single batch of input and the latency. Since FLOPs are independent of hardware, we used them as a proxy for efficiency. We measured the average latency on the GPU and Jetson Nano to test our model on hardware with different computational budgets. The latency on GPU was measured with the input of batch size 8. To run inference on Jetson Nano, We converted the models using ONNX [20] and measured the model size, latency, power consumption, and memory usage with a batch size of 1.

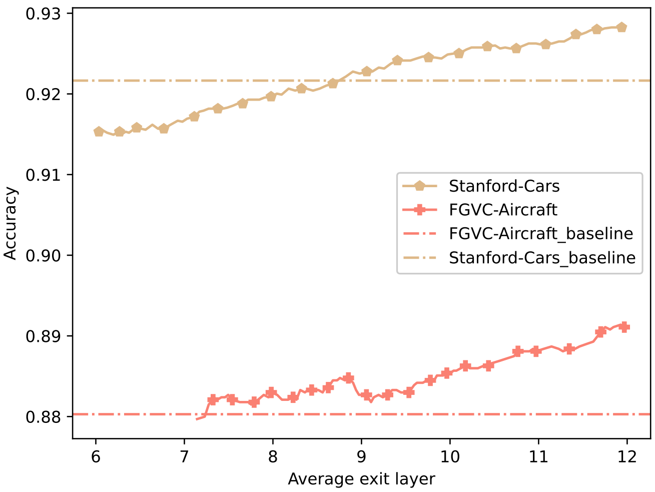

Fig. 2. Layer-wise trade-off curves of accuracy to the average number of exit layers on Stanford Cars and FGVC Aircraft.

3.3. Results

The results of our experiments are presented in Tables 1 to 3 and Fig. 2 and 4. Fig. 2 shows the layer-wise accuracy curves of Stanford Cars and FGVC Aircraft datasets. The horizontal lines represent the accuracy of the baseline ViT model for both datasets. It is noticeable in both curves that the early exit branch classifiers achieve even higher accuracy compared to the baseline ViT model. As we exit from the earlier branches, the accuracy drops consequently. However, the earliest exit point leverages only half of the full model while maintaining performance within 1 percent of the baseline model.

Method	Acc.(%)	latency(ms)	FLOPs
TransFG	92.4	671	5.70x
VIT-B	92.2	134	1.00x
MEViT$_{12}$	92.8	135	1.00x
MEViT$_6$	91.5	67	0.50x

Table 1. Comparison of different methods on Stanford Car.

Table 1 and 2 show the accuracy and efficiency of the models on the two benchmarks. MEViT$_6$ and MEViT$_{12}$ denotes our early exit model with 6_{th} classifier and full model with the final classifier. On Stanford Car benchmark, our full model achieves higher accuracy compared to the baseline ViT (ViT-B) and TransFG while it is 5× faster than TransFG. Our early exit model reduces the compute cost (FLOPs) by 50% resulting in 2× faster on GPU. In table 2, our early exit model is 2× faster while maintaining accuracy within 1 percent compared to ViT-B. With the same compute budget, our full model achieves 1.1% higher accuracy compared to ViT-B. This validates our layer-wise adaptive training scheme is effective to improve the overall performance of the model.

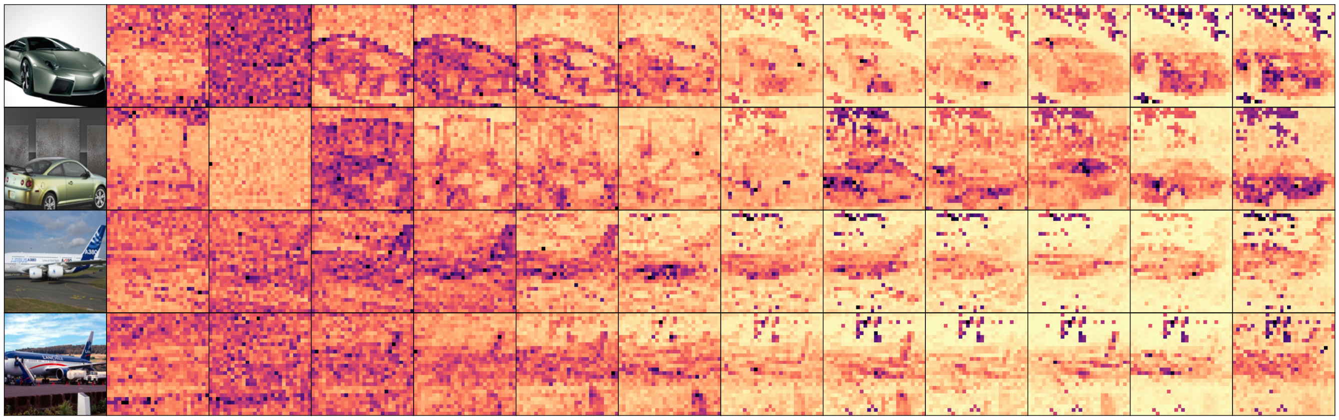

Fig. 3. Visualization of attention probability distribution on each layer. 6 to 11 layers generate similar attention patterns compared to that of the final layer due to the distillation supervision during training.

Method	**Acc.(%)**	**latency(ms)**	**FLOPs**
VIT-B	88.0	135	1.00x
MEViT$_{12}$	89.1	136	1.00x
MEViT$_6$	87.2	68	0.50x

Table 2. Comparison of different methods on FGVC-Aircraft.

To further analyze the model performance on resource-constrained hardware, we measured additional efficiency metrics on Nvidia Jetson Nano. Table 3 shows the model size, latency, power consumption, and memory usage during the inference of the model. The efficiency metrics show linear correlations with FLOPs. Our MEViT$_6$ improves the efficiency by 2× compared to ViT-B.

Metrics	Size	Latency	Power	Mem usage
ViT-B	330MB	1446ms	11.0W	1.3Gb
MEViT$_6$	167MB	784ms	5.8W	0.73Gb

Table 3. Efficiency evaluation with the baseline model ViT-B and our MEViT$_6$ on Stanford Car.

Additionally, we analyzed the proportion of classes in each dataset that only require minimal computational effort from our model. As shown in Fig. 4, 107 out of 196 classes and 49 out of 100 classes from Stanford Car and FGVC Aircraft respectively could be accurately identified with only 6 encoder layers without accuracy loss compared to the full model. There is a noteworthy proportion of classes that exhibit higher accuracy while requiring significantly less compute effort compared to the full model on both benchmarks.

In Fig. 3, we visualize the attention probability distribution of each layer. We observe that the attention patterns of layers 6 to 11 are remarkably similar to that of the final layer. We attribute this similarity to the inplace distillation technique used during training, which facilitates knowledge transfer from the full model to the early exit branches.

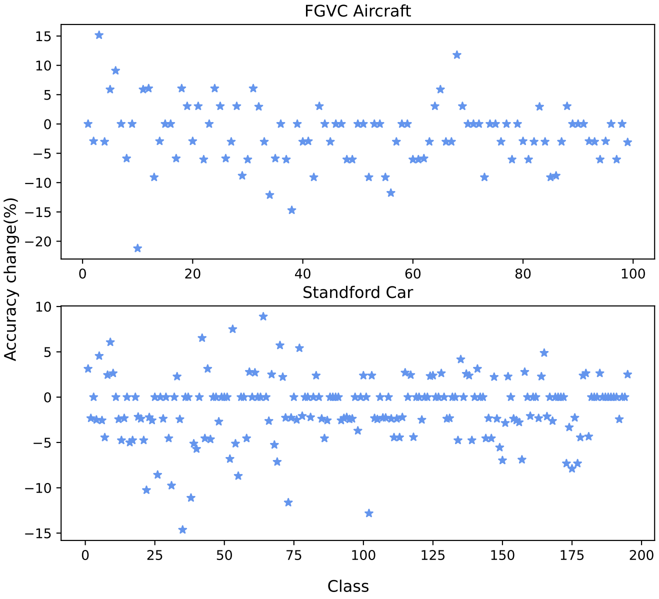

Fig. 4. Class-wise accuracy change of MEViT$_6$ compared to MEViT$_{12}$ on Stanford Cars and FGVC Aircraft.

4. CONCLUSION

In this paper, we propose a MEViT for efficient inference on FGVC. The uncertainty score predictor module enables the model to dynamically exit on earlier branches determined by the difficulty of the input images. Extensive experiments are conducted on several FGVC benchmarks, and the results show the superior accuracy-efficiency trade-offs of our model compared to ViT and a SOTA ViT-based model. Our early exit model drastically reduces the overall compute cost and memory usage on various systems with different computational budgets.

5. REFERENCES

[1] Alexey Dosovitskiy et al., "An image is worth 16x16 words: Transformers for image recognition at scale," in *International Conference on Learning Representations*, 2021.

[2] Yutao Hu, Xuhui Liu, Baochang Zhang, Jungong Han, and Xianbin Cao, "Alignment enhancement network for fine-grained visual categorization," *ACM Trans. Multimedia Comput. Commun. Appl.*, vol. 17, no. 1s, mar 2021.

[3] Peiqin Zhuang, Yali Wang, and Yu Qiao, "Learning attentive pairwise interaction for fine-grained classification," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, pp. 13130–13137, 04 2020.

[4] Hao Li, Xiaopeng Zhang, Hongkai Xiong, and Qi Tian, "Attribute mix: Semantic data augmentation for fine grained recognition," 2020.

[5] Dimitri Korsch, Paul Bodesheim, and Joachim Denzler, "End-to-end learning of fisher vector encodings for part features in fine-grained recognition," in *Pattern Recognition*, Christian Bauckhage, Juergen Gall, and Alexander Schwing, Eds., Cham, 2021, pp. 142–158, Springer International Publishing.

[6] H. Touvron, A. Sablayrolles, M. Douze, M. Cord, and H. Jegou, "Grafit: Learning fine-grained image representations with coarse labels," in *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, Los Alamitos, CA, USA, oct 2021, pp. 854–864, IEEE Computer Society.

[7] Tian Zhang, Dongliang Chang, Zhanyu Ma, and Jun Guo, "Progressive co-attention network for fine-grained visual classification," *2021 International Conference on Visual Communications and Image Processing (VCIP)*, pp. 1–5, 2021.

[8] Yao Ding, Yanzhao Zhou, Yi Zhu, Qixiang Ye, and Jianbin Jiao, "Selective sparse sampling for fine-grained image recognition," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2019.

[9] Shaokang Yang, Shuai Liu, Cheng Yang, and Changhu Wang, "Re-rank coarse classification with local region enhanced features for fine-grained image recognition," *ArXiv*, vol. abs/2102.09875, 2021.

[10] Ju He et al., "Transfg: A transformer architecture for fine-grained recognition," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2022, vol. 36, pp. 852–860.

[11] Jun Wang, Xiaohan Yu, and Yongsheng Gao, "Feature fusion vision transformer for fine-grained visual categorization," in *BMVC*, 2021.

[12] Priyadarshini Panda, Abhronil Sengupta, and Kaushik Roy, "Conditional deep learning for energy-efficient and enhanced pattern recognition," in *2016 Design, Automation & Test in Europe Conference & Exhibition (DATE)*. IEEE, 2016, pp. 475–480.

[13] Surat Teerapittayanon, Bradley McDanel, and Hsiang-Tsung Kung, "Branchynet: Fast inference via early exiting from deep neural networks," in *2016 23rd International Conference on Pattern Recognition (ICPR)*. IEEE, 2016, pp. 2464–2469.

[14] Arian Bakhtiarnia, Qi Zhang, and Alexandros Iosifidis, "Multi-exit vision transformer for dynamic inference," *arXiv preprint arXiv:2106.15183*, 2021.

[15] Ji Xin, Raphael Tang, Yaoliang Yu, and Jimmy Lin, "BERxiT: Early exiting for BERT with better fine-tuning and extension to regression," in *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, Online, Apr. 2021, pp. 91–104, Association for Computational Linguistics.

[16] Jiahui Yu and Thomas S Huang, "Universally slimmable networks and improved training techniques," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 1803–1811.

[17] Tal Ridnik, Emanuel Ben-Baruch, Asaf Noy, and Lihi Zelnik-Manor, "Imagenet-21k pretraining for the masses," *arXiv preprint arXiv:2104.10972*, 2021.

[18] S. Maji, J. Kannala, E. Rahtu, M. Blaschko, and A. Vedaldi, "Fine-grained visual classification of aircraft," Tech. Rep., 2013.

[19] Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei, "3d object representations for fine-grained categorization," in *4th International IEEE Workshop on 3D Representation and Recognition (3dRR-13)*, Sydney, Australia, 2013.

[20] Junjie Bai, Fang Lu, Ke Zhang, et al., "Onnx: Open neural network exchange," https://github.com/onnx/onnx, 2019.