Transformer-Based Image Classification Network MultiFormer Postprint

Authors: Hu Jie, Chang Minjie, Xiong Zongquan, Boyuan Xu, Xie Lihao, Guo Di

Date: 2022-06-06T00:00:00+00:00

Abstract

To address the limitations of existing ViT models, which are unable to alter input patch size and where input patches contain only single-scale information, we propose MultiFormer, a Transformer-based image classification network. MultiFormer utilizes an AWS (Attention With Scale) module to embed multi-scale input small patches at each stage into large patches rich in semantic information; through a GLA-P (Global-Local Attention With Patch) module, it alternately captures local and global attention, thereby preserving both fine-grained and coarse-grained features during embedding. We design three variants of the MultiFormer architecture: MultiFormer-Tiny, -Small, and -Base, which achieve Top-1 accuracies of 81.1%, 82.2%, and 83.2% respectively on ImageNet image classification experiments. The latter two models demonstrate improvements of 3.1% and 3.4% over convolutional neural networks of comparable size, ResNet-50 and ResNet-101; compared to the Transformer-based classification model ViT, MultiFormer-Base achieves a 2.1% improvement while requiring significantly fewer parameters and computational resources than the ViT-Base/16 model and without the need for large-scale data pre-training.

Full Text

MultiFormer: An Image Classification Network Based on Transformer

Hu Jiea,b,c†, **Chang Minjie**a,b,c, **Xiong Zongquan**a,b,c, **Xu Boyuan**a,b,c, **Xie Lihao**a,b,c, **Guo Di**a,b,c

aHubei Key Laboratory of Advanced Technology for Automotive Components
bHubei Collaborative Innovation Center for Automotive Components Technology

cHubei Research Center for New Energy & Intelligent Connected Vehicle
Wuhan University of Technology, Wuhan 430070, China

Abstract

To address the limitations of existing Vision Transformer (ViT) models—namely their inability to change input patch size and their reliance on single-scale information—we propose MultiFormer, a novel image classification network based on Transformer. MultiFormer employs an AWS (Attention With Scale) module to embed small patches from different scales at each stage into larger patches rich in semantic information. Through the GLA-P (Global-Local Attention With Patch) module, it alternately captures local and global attention, preserving both fine-grained and coarse-grained features during embedding. We design three variants: MultiFormer-Tiny, -Small, and -Base. On ImageNet, these achieve Top-1 accuracies of 81.1%, 82.2%, and 83.2% respectively. The latter two models outperform convolutional networks of comparable size—ResNet-50 and ResNet-101—by 3.1% and 3.4%. Compared to ViT-Base/16, MultiFormer-Base achieves a 2.1% improvement with far fewer parameters and computational costs, and without requiring large-scale pre-training.

Keywords: machine vision; deep learning; image classification; self-attention; Transformer

0 Introduction

Computer vision tasks such as image classification [1], object detection [2], and semantic segmentation [3] have long been dominated by convolutional neural networks (CNNs). Since AlexNet [4] won the ImageNet competition, CNN architectures have evolved to become deeper, denser, and more complex [5–7]. ResNet [5] introduced residual connections to mitigate gradient vanishing when deepening networks. DenseNet [6] employed densely connected topologies linking each convolutional block to all previous ones. VGG [8] expanded receptive fields by stacking convolutional kernels, while GoogLeNet [9] approximated optimal sparse structures through dense block constructions to improve performance without increasing computation. EfficientNet [10] demonstrated that uniformly scaling all dimensions of a model with compound coefficients enhances performance.

Meanwhile, Transformers have been widely adopted in natural language processing due to their self-attention modules' ability to capture long-range dependencies [11]. Inspired by this, researchers have explored Transformer architectures for computer vision tasks. Studies [12–15] have integrated self-attention modules into CNNs for various vision applications. Vision Transformer (ViT) [16] applied Transformers to image classification by serializing images into sequences, sparking rapid improvements [17–20] and extensions to downstream tasks [21–24].

However, since self-attention operates on entire input sequences, treating each

pixel as a token would create sequences far longer than word sequences, incurring prohibitive memory and computational costs. ViT adopts a compromise by embedding multiple pixels as patches (tokens) for self-attention computation, yet computational complexity remains high and input images are restricted to fixed sizes. Improvements to ViT fall into three categories: (a) enhancing ViT' s design itself—DeiT [19] introduced effective training strategies to eliminate the need for large-scale pre-training and used distillation for better learning; T2T-ViT [20] progressively structured images into patches while preserving local structural information; DynamicViT [23] leveraged the unstructured nature of Transformer tokens to design a token sparsification pruning method that reduces computation by removing less informative tokens. (b) incorporating convolutions into ViT—using convolutions for positional encoding [17] or replacing linear projection layers [25]; CoAtNet [26] introduced CNNs to capture local attention and complement local features. (c) designing new backbones and attention modules—PVT [21] created a pyramid-structured backbone with progressive downsampling and spatial-reduction attention to balance efficiency and accuracy; CAT [24] designed cross-block attention combining intra-patch and inter-patch attention for local-global interaction; Swin [22] partitioned feature maps into fixed-size local windows, computing self-attention within each window to reduce costs; DPT [27] adaptively segmented images into variable-sized pixel blocks to preserve semantic information; FPT [28] encoded cross-scale non-local features and could integrate into other backbones.

These approaches still have limitations: using single-scale patches in self-attention loses significant semantic information, requiring cross-scale attention mechanisms to establish connections. Image classification also demands interaction between coarse and fine-grained features to capture target information.

Based on these observations, our contributions are: (a) We propose AWS (Attention With Scale), a multi-scale embedding module that addresses ViT' s single-scale limitation. AWS samples images with different convolutional kernels, fusing them into cross-scale attention patches with human-like receptive fields for each stage, ensuring multi-scale patch inputs throughout. (b) We design GLA-P (Global-Local Attention With Patch), a novel self-attention module that alternately captures global and local attention to aggregate coarse and fine-grained features, compensating for semantic information loss in patches. Using attention packing operations, it reduces computation without affecting performance. (c) We design three model variants—MultiFormer-Tiny, -Small, and -Base—and evaluate them on ImageNet, CIFAR-10, and CIFAR-100. Results demonstrate MultiFormer' s superiority over comparable networks, with ablation studies validating each module' s effectiveness.

1 Model Architecture

1.1 Overall Framework

The overall MultiFormer architecture is shown in Figure 1: see original paper. The backbone follows PVT [21] with a 4-stage pyramid structure, where each stage comprises an AWS multi-scale embedding module followed by multiple MultiFormer blocks. As shown in Figure 1: see original paper, each MultiFormer block consists of a GLA-P self-attention module and a multi-layer perceptron (MLP) that introduces non-linearity and transforms dimensions.

First, input images pass through the AWS module to generate multi-scale feature maps partitioned into patches containing multi-scale information. Except for Stage 1, outputs are downsampled to reduce patch count to one-quarter while doubling dimensions, forming a pyramid structure. The generated multi-scale patches then enter the GLA-P self-attention module within MultiFormer blocks, where GAP (Global Attention with Patch) and LAP (Local Attention with Patch) alternate. This creates a GLA-P module that simultaneously aggregates global and local features. Finally, a task-specific head such as a classification head is attached for image classification. We now detail each module' s principles and functions.

1.2 AWS Multi-Scale Embedding Module

Before feeding images into self-attention, pixels must be partitioned into equal-sized patches and serialized into 2D matrices. As shown in [Figure 2: see original paper], ViT simply partitions adjacent pixels into fixed-size patches, maintaining constant patch counts per stage to facilitate absolute position encoding. This approach causes self-attention computation to grow quadratically and requires massive datasets for pre-training while being difficult to converge.

Unlike ViT, the AWS module first partitions images into 4$\times$4$small patches, then uses down sampling to merge the$ larger patches while doubling dimensions. By reducing patch count and expanding dimensions, it forms a pyramid structure that lowers computational complexity and eliminates ViT' s fixed-size input restriction.

PVT [21] and Swin [22] also partition input images into equal-sized patches, but neglecting feature map scale effects results in single-scale patches that lose multi-scale semantic information, degrading performance. Our AWS module applies convolutions with different kernel sizes before patch generation to create semantically rich multi-scale feature maps that enhance patch information. As shown in [Figure 2: see original paper], AWS receives an RGB image input and samples it with three different convolutional kernels while maintaining consistent stride, ensuring each sampling window shares the same center but different scales. In Stage 1, AWS uses kernels of 2$\times$2, 4$\times$4, and 8$\times$8; $subsequent stages use$ 2$\times$2 and 4$\times$4 $kernels, all with$ 4$\times$4 $stride. Channel dimensions are set to D$ $scale feature maps are then fused into semantically rich embeddings inspired by human visual characteristics$[29]. sizes with doubled dimensions, forming a pyramid structure.

The right side of [Figure 2: see original paper] compares patch scales across Transformers. PVT and Swin crudely partition original feature maps into patches limited to 4$×4$pixelregions.Iftargetscalesexceedthisregion,themodellosessemanticinformationato$ through multi-scale convolutions, generating feature-rich patches that compensate for the insufficient multi-scale information in Swin and PVT, thereby improving performance.

1.3 GLA-P Self-Attention Module

After AWS generates multi-scale patches, they are fed into the GLA-P self-attention module within MultiFormer blocks. As shown in Figure 1: see original paper, image classification requires capturing both fine-grained and coarse-grained features. Therefore, MultiFormer blocks alternate GAP and LAP to form GLA-P, capturing global and local attention to preserve both feature granularities.

As illustrated in [Figure 3: see original paper], GLA-P receives multi-scale feature maps embedded by AWS. For LAP, every 4$×4$adjacentpixelblockisgroupedforLocalAttention.ForGAP,$ pixel blocks with stride 4 are grouped for Global Attention. The widely distributed non-adjacent blocks provide sufficient contextual information, making global attention more effective.

Let denote the input feature map and the generated matrices of corresponding dimensions. We add learnable relative position biases [30–32] to each attention head. denotes the convolution operation. With layer normalization [33] as , the self-attention operation computes:

[Figure 3: see original paper] shows the GLA-P module. Unlike CoAtNet [26], which uses CNNs for local features and self-attention for global features, GLA-P alternately captures global and local attention by modeling short-range and long-range pixel blocks without relying on CNNs. To visualize GLA-P' s behavior, we trained a MultiFormer-Base model and mapped the final layer' s pixel scores back to the original image after activation, as shown in [Figure 4: see original paper]. The bright regions indicate attention focus, demonstrating effective global information capture.

Compared to other Transformer attention modules, Swin partitions feature maps into non-overlapping windows with independent self-attention, using a shifted window strategy to exchange information between windows. However, Swin still restricts attention to adjacent blocks, preventing long-range modeling. Our GLA-P module combines patches generated by Global Attention on widely distributed blocks with those from Local Attention on adjacent blocks, preserving both global and local information. This enables simultaneous focus on coarse and fine-grained features, excelling at image classification.

To preserve semantic information, our large patches maintain relatively high resolution (e.g., 28$×$28 after GLA-P in Stage 1), which would still incur high computation when serialized. Therefore, we employ a convolutional packing

method instead of traditional multi-head attention [11]. Similar to MHA, it receives Query, Key, and Value to produce enhanced features.

Unlike Swin, which processes Query, Key, and Value simultaneously, we downsample Key-Value pairs to reduce computation by factor P without affecting accuracy. Parameter reduction after Attention Patch packing is shown in .

1.4 Model Variants

Following ResNet [5] design principles, we construct three model variants—MultiFormer-Tiny, -Small, and -Base—with size and complexity ratios of 1:1.5:3. Their computational costs and parameters are comparable to ResNet-18, ResNet-50, and ResNet-101 respectively. Key hyperparameters are:

MultiFormer-Tiny:
MultiFormer-Small:
MultiFormer-Base:

where D is the hidden dimension in Stage 1, Depth is the number of MultiFormer blocks per stage, and Heads is the multi-head attention dimension. During convolutional attention packing, Patch size P is set to . Detailed hyperparameters are in .

2 Experiments

We evaluate MultiFormer on ImageNet-1K, CIFAR-10, and CIFAR-100 for image classification, comparing against representative CNN backbones (ResNet [5]) and mainstream Transformer models, followed by ablation studies validating each module.

2.1 Image Classification on ImageNet

ImageNet-1K [34] contains 1.28 million training images and 50,000 validation images across 1,000 categories. We train on the training set and report Top-1 accuracy on the validation set. Images are randomly cropped to 224$×$224. We use AdamW optimizer with momentum 0.9, weight decay 0.05, cosine decay, batch size 128, and initial learning rate 0.001. All models are trained from scratch for 300 epochs on four 2080Ti GPUs. Results are in .

2.2 Results on CIFAR Datasets

We further validate MultiFormer on CIFAR-10 and CIFAR-100, containing 10 and 100 classes respectively, each with 50,000 training and 10,000 test images. To prevent overfitting on these smaller datasets, we follow ViT' s fine-tuning strategy: load weights from ImageNet experiments into MultiFormer-Tiny, -Small, and -Base, replace the classification head, and fine-tune with SGD (momentum 0.9) for 300 epochs at batch size 64. Results are in .

MultiFormer-Base achieves 0.4% and 1.7% higher Top-1 accuracy on CIFAR-10 and CIFAR-100 respectively than EfficientNetV2-L, with 50% fewer parameters. With only one-tenth the parameters of ViT-H/16, it still improves by 0.3% on both datasets. MultiFormer-Tiny and -Small outperform LeViT-256 and -384 by 1.0% and 1.3% on CIFAR-10, further validating MultiFormer' s effectiveness.

2.3 Ablation Studies

We conduct ablation studies on ImageNet using MultiFormer-Tiny to validate AWS and GLA-P modules:

a) Removing AWS multi-scale embedding: Using single-scale embedding with a single 4$\times$4$kernelin Stage1 and single2\times$2 kernels in other stages reduces Top-1 accuracy by 0.6%, demonstrating AWS' s significant contribution.

b) Replacing GLA-P with attention modules from Swin [22], PVT [21], and CoAtNet [26] shows accuracy drops of 0.3%, 0.5%, and 0.8% respectively. Swin' s sliding window restricts attention locally, neglecting global connections. PVT' s simple Key-Value downsampling discards fine-grained semantics. CoAtNet' s early-stage CNN dependence loses global information. These results confirm that alternating local and global attention effectively boosts performance.

All experiments use identical hyperparameters and training settings: four 2080Ti GPUs, 300 training epochs.

3 Conclusion

This paper proposes MultiFormer, a Transformer-based image classification network featuring AWS (Attention With Scale) multi-scale embedding and GLA-P (Global-Local Attention With Patch) modules. Experiments demonstrate significant improvements over CNNs and other Transformer-based methods at comparable parameter counts and computational costs. Multi-scale embedding and alternating local-global attention substantially enhance self-attention' s ability to learn semantic features, making our backbone a promising universal feature extractor for downstream vision tasks. As Transformers rapidly evolve in computer vision, we hope this work inspires future research.

References

[1] Huang Kaiqi, Ren Weiqiang, Tan Tieniu. A review on image object classification and detection [J]. Chinese Journal of Computers, 2014, 378(06): 1225-1240.

[2] Li Xudong, Ye Mao, Li Tao. Review of object detection based on convolutional neural networks [J]. Application Research of Computers, 2017, 312(10): 2881-2886, 2891.

[3] Tian Xuan, Wang Liang, Ding Lei. Review of image semantic segmentation based on deep learning [J]. Journal of Software, 2019, 30(02): 440-468.

[4] Krizhevsky A, Sutskever I, Hinton G E. Imagenet classification with deep convolutional neural networks [C]//Advances in Neural Information Processing Systems. 2012: 1097–1105.

[5] He Kaiming, Zhang Xiangyu, Ren Shaoqing, et al. Deep residual learning for image recognition [C]//Proc of the IEEE Conference on Computer Vision and Pattern Recognition. 2016: 770–778.

[6] Huang Gao, Liu Zhuang, Van Der Maaten L, et al. Densely connected convolutional networks [C]//Proc of the IEEE Conference on Computer Vision and Pattern Recognition. 2017: 4700–4708.

[7] Xie Saining, Girshick R, Dollár P, et al. Aggregated residual transformations for deep neural networks [C]//Proc of the IEEE Conference on Computer Vision and Pattern Recognition. 2017: 1492–1500.

[8] Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition [EB/OL]. (2015-04-10) [2022-03-28]. https://arxiv.org/pdf/1409.1556.

[9] Szegedy C, Liu Wei, Jia Yangqing, et al. Going deeper with convolutions [C]//Proc of the IEEE conference on Computer Vision and Pattern Recognition. 2015: 1–9.

[10] Tan Mingxing, Le Q. Efficientnet: Rethinking model scaling for convolutional neural networks [C]//Proc of International Conference on Machine Learning. 2019: 6105–6114.

[11] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need [C]//Advances in Neural Information Processing Systems. 2017: 5998–6008.

[12] Wang Xiaolong, Girshick R, Gupta A, et al. Non-local neural networks [C]//Proc of the IEEE Conference on Computer Vision and Pattern Recognition. 2018: 7794-7803.

[13] Zhao Hengshuang, Jia Jiaya, Koltun V. Exploring self-attention for image recognition [C]//Proc of the IEEE Conference on Computer Vision and Pattern Recognition. 2020: 10076–10085.

[14] Ramachandran P, Parmar N, Vaswani A, et al. Studying standalone self-attention in vision models [EB/OL]. (2019-06-13) [2022-03-28]. https://arxiv.org/pdf/1906.05909.

[15] Carion N, Massa F, Synnaeve G, et al. End-to-End object detection with transformers [C]//Proc of European Conference on Computer Vision. 2020: 213-229.

[16] Dosovitskiy A, Beyer L, Kolesnikov A, et al. An image is worth 16x16 words: Transformers for image recognition at scale [EB/OL]. (2021-01-03) [2022-03-28]. https://arxiv.org/pdf/2010.11929.

[17] Chu Xiangxiang, Tian Zhi, Zhang Bo, et al. Conditional positional encodings for vision transformers [EB/OL]. (2021-05-18) [2022-03-28]. https://arxiv.org/pdf/2102.10882.

[18] Han Kai, Xiao An, Wu Enhua, et al. Transformer in Transformer [EB/OL]. (2021-10-26) [2022-03-28]. https://arxiv.org/pdf/2103.00112.

[19] Touvron H, Cord M, Douze M, et al. Training data-efficient image transformers & distillation through attention [C]//Proc of International Conference on Machine Learning. 2021, 139: 10347-10357.

[20] Yuan Li, Chen Yunpeng, Wang Tao, et al. Tokens-to-token vit: Training vision transformers from scratch on imagenet [C]//Proc of the IEEE/CVF International Conference on Computer Vision. 2021: 558-567.

[21] Wang Wenhai, Xie Enze, Li Xiang, et al. Pyramid vision Transformer: A versatile backbone for dense prediction without convolutions [C]//Proc of the IEEE/CVF International Conference on Computer Vision. 2021: 568-578.

[22] Liu Ze, Lin Yutong, Cao Yue, et al. Swin Transformer: Hierarchical vision Transformer using shifted windows [C]//Proc of the IEEE/CVF International Conference on Computer Vision. 2021: 10012-10022.

[23] Rao Yongming, Zhao Wenliang, Liu Benlin, et al. Dynamicvit: Efficient vision transformers with dynamic token sparsification [EB/OL]. (2021-10-26) [2022-03-28]. https://arxiv.org/pdf/2106.02034.

[24] Lin Hezheng, Cheng Xing, Wu Xiangyu, et al. CAT: Cross attention in vision Transformer [EB/OL]. (2021-06-10) [2022-03-28]. https://arxiv.org/pdf/2106.05786.

[25] Wu Haiping, Xiao Bin, Codella N, et al. Cvt: Introducing convolutions to vision transformers [C]//Proc of the IEEE/CVF International Conference on Computer Vision. 2021: 22-31.

[26] Dai Zihang, Liu Hanxiao, Le Q V, et al. Coatnet: Marrying convolution and attention for all data sizes [C]//Advances in Neural Information Processing Systems. 2021, 34: 3965-3977.

[27] Chen Zhiyang, Zhu Yousong, Zhao Chaoyang, et al. Dpt: Deformable patch-based Transformer for visual recognition [C]//Proc of the 29th ACM International Conference on Multimedia. 2021: 2899-2907.

[28] Zhang Dong, Zhang Hanwang, Tang Jinhui, et al. Feature pyramid Transformer [C]//Proc of the European Conference on Computer Vision. 2020: 323-339.

[29] Liu Songtao, Huang Di, Wang Yunhong. Receptive field block net for accurate and fast object detection [C]//Proc of the European Conference on Computer Vision. 2018: 385-400.

[30] Bao Hangbo, Li Dong, Wei Furu, et al. Unilmv2: Pseudo-masked language models for unified language model pre-training [C]//Proc of International Con-

ference on Machine Learning. 2020: 642-652.

[31] Hu Han, Gu Jiayuan, Zhang Zheng, et al. Relation networks for object detection [C]//Proc of the IEEE Conference on Computer Vision and Pattern Recognition. 2018: 3588-3597.

[32] Raffel C, Shazeer N, Roberts A, et al. Exploring the limits of transfer learning with a unified text-to-text Transformer [J]. Journal of Machine Learning Research, 2020, 21(140): 1-67.

[33] Ba J L, Kiros J R, Hinton G E. Layer normalization [EB/OL]. (2016-07-21) [2022-03-28]. https://arxiv.org/pdf/1607.06450.

[34] Deng Jia, Dong Wei, Socher R, et al. Imagenet: A large-scale hierarchical image database [C]//Proc of the IEEE Conference on Computer Vision and Pattern Recognition. 2009: 248-255.

[35] Graham B, El-Nouby A, Touvron H, et al. LeViT: A vision Transformer in ConvNet' s clothing for faster inference [C]//Proc of the IEEE/CVF International Conference on Computer Vision. 2021: 12259-12269.

Note: Figure translations are in progress. See original paper for figures.