International Journal of Computer Vision (2023) 131:3136–3151
https://doi.org/10.1007/s11263-023-01861-3

Super Vision Transformer

Mingbao Lin[1,2] · Mengzhao Chen[1] · Yuxin Zhang[1] · Chunhua Shen[3] · Rongrong Ji[1] · Liujuan Cao[1]

Received: 28 October 2022 / Accepted: 13 July 2023 / Published online: 2 August 2023

Abstract
We attempt to reduce the computational costs in vision transformers (ViTs), which increase quadratically in the token number. We present a novel training paradigm that trains only one ViT model at a time, but is capable of providing improved image recognition performance with various computational costs. Here, the trained ViT model, termed super vision transformer (SuperViT), is empowered with the versatile ability to solve incoming patches of multiple sizes as well as preserve informative tokens with multiple keeping rates (the ratio of keeping tokens) to achieve good hardware efficiency for inference, given that the available hardware resources often change from time to time. Experimental results on ImageNet demonstrate that our SuperViT can considerably reduce the computational costs of ViT models with even performance increase. For example, we reduce 2 × FLOPs of DeiT-S while increasing the Top-1 accuracy by 0.2% and 0.7% for 1.5 × reduction. Also, our SuperViT significantly outperforms existing studies on efficient vision transformers. For example, when consuming the same amount of FLOPs, our SuperViT surpasses the recent state-of-the-art EViT by 1.1% when using DeiT-S as their backbones. The project of this work is made publicly available at https://github.com/lmbxmu/SuperViT.

Keywords Hardware efficiency · Supernet · Vision transformer

1 Introduction

Vision transformers (ViTs) initially introduced in 2020 (Dosovitskiy et al., 2020) have spread widely in the field of computer vision and soon become one of the most pervasive and promising architectures in varieties of prevalent vision tasks, such as image classification (Dosovitskiy et al., 2020; Jiang et al., 2021; Graham et al., 2021), object detection (Carion et al., 2020; Zhu et al., 2022), video understanding (Bertasius et al., 2021; Arnab et al., 2021) and many others (Zheng et al., 2021; Xie et al., 2021; Liang et al., 2021; Zamir et al., 2022; Huang et al., 2020). The basic idea behind ViTs is to break down an image as a series of local patches and use a linear projection to tokenize these patches as inputs. In particular, ViTs merit in its property of capturing the long-range relationships between different portions of an image with the mechanism of multi-head self-attention (MHSA). Therefore, increasing attention has been paid to developing ViTs in various vision tasks.

Recent studies focus more on an efficient ViT (Liu et al., 2021; Chu et al., 2021a; Li et al., 2022; Graham et al., 2021; Chavan et al., 2022) since the excessive computational costs, which increase quadratically to the number of tokens, have severely barricaded the broader usage of ViTs in real-world applications. Note that the transformer's token sequence length is inversely proportional to the square of the patch size, which denotes that models with smaller patch sizes are computationally more expensive. The most intuitive way is to reduce the transformer's token number by enlarging the patch size. However, it has been an experimental consensus in the literature (Dosovitskiy et al., 2020) that a ViT model performs better with smaller-size patches as its inputs. For example, ViT-B (Dosovitskiy et al., 2020) observes 77.91% Top-1 accuracy on ImageNet when the patch size is 16 × 16 while only 73.38% is reached if the patch size is 32 × 32. Modern ViT structures simply accept a fixed patch size *w.r.t.* all input images when training a ViT model.

Communicated by Nikos KOMODAKIS.

Mingbao Lin and Mengzhao Chen have contributed equally to this work.

✉ Liujuan Cao
caoliujuan@xmu.edu.cn

[1] Key Laboratory of Multimedia Trusted Perception and Efficient Computing, Ministry of Education of China, School of Informatics, Xiamen, China

[2] Tencent Youtu Lab, Shanghai, China

[3] Zhejiang University, Hangzhou, China