VOLO: Vision Outlooker for Visual Recognition

Li Yuan[1,2*] Qibin Hou[2*] Zihang Jiang[2] Jiashi Feng[1,2] Shuicheng Yan[1]
[1]Sea AI Lab [2]National University of Singapore
{ylustcnus,andrewhoux,jzh0103}@gmail.com, {fengjs, yansc}@sea.com

Abstract

Visual recognition has been dominated by convolutional neural networks (CNNs) for years. Though recently the prevailing vision transformers (ViTs) have shown great potential of self-attention based models in ImageNet classification, their performance is still inferior to that of the latest SOTA CNNs if no extra data are provided. In this work, we try to close the performance gap and demonstrate that attention-based models are indeed able to outperform CNNs. We find a major factor limiting the performance of ViTs for ImageNet classification is their low efficacy in encoding fine-level features into the token representations. To resolve this, we introduce a novel outlook attention *and present a simple and general architecture, termed Vision Outlooker (VOLO). Unlike self-attention that focuses on global dependency modeling at a coarse level, the outlook attention efficiently encodes finer-level features and contexts into tokens, which is shown to be critically beneficial to recognition performance but largely ignored by the self-attention. Experiments show that our VOLO achieves 87.1% top-1 accuracy on ImageNet-1K classification, which is the first model exceeding 87% accuracy on this competitive benchmark, without using any extra training data. In addition, the pre-trained VOLO transfers well to downstream tasks, such as semantic segmentation. We achieve 84.3% mIoU score on the cityscapes validation set and 54.3% on the ADE20K validation set. Code is available at https://github.com/sail-sg/volo.*

1. Introduction

Modeling in visual recognition, which was long dominated by convolutional neural networks (CNNs), has recently been revolutionized by Vision Transformers (ViTs) [14, 51, 68]. Different from CNNs that aggregate and transform features via local and dense convolutional kernels, ViTs directly model long-range dependencies of local patches (*a.k.a.* tokens) through the self-attention mechanism which is with greater flexibility in modeling visual contents. Despite the remarkable effectiveness on visual recognition [37, 32, 52, 79], the performance of ViT models still lags behind that of the state-of-the-art CNN models. For instance, as shown in Table 1, the state-of-the-art transformer-based CaiT [52] attains 86.5% top-1 accuracy on ImageNet, which however is still 0.3% lower compared with the 86.8% top-1 accuracy achieved by the CNN-based NFNet-F5 [2] with SAM and augmult [15, 16].

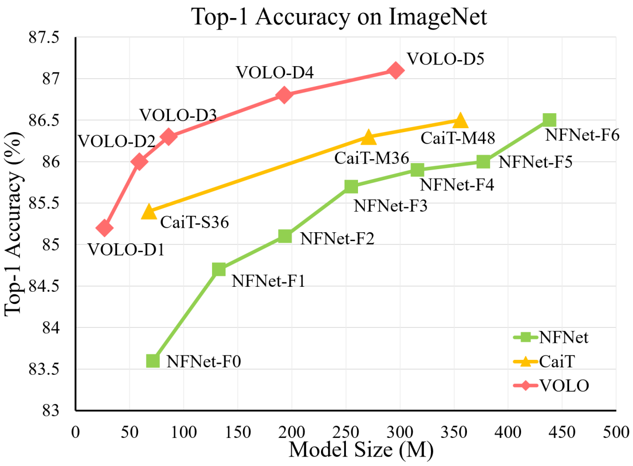

Figure 1. ImageNet top-1 accuracy of state-of-the-art CNN-based and Transformer-based models. All the results are obtained based on the best test resolutions, without using any extra training data. Our VOLO-D5 achieves the best accuracy, outperforming the latest NFNet-F6 w/ SAM [2, 15] and CaiT-M48 w/ KD [22, 69], while using much less training parameters. To our best knowledge, VOLO-D5 is the first model exceeding 87% top-1 accuracy on ImageNet.

In this work we try to close such performance gap. We find one major factor limiting ViTs from outperforming CNNs is their low efficacy in encoding fine-level features and contexts into token representations, which are critical for achieving compelling visual recognition performance. Fine-level information can be encoded into tokens by finer-grained image tokenization, which however would lead to a token sequence of greater length that increases quadratically the complexity of the self-attention mechanism of ViTs.

[*] Equal contribution.