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].

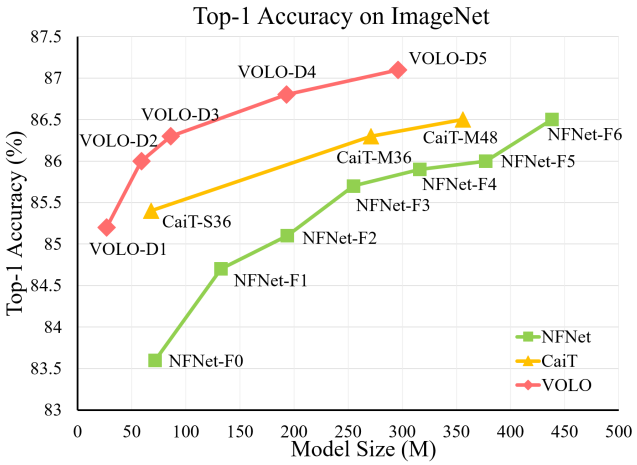

Figure 1. ImageNet top-1 accuracy of state-of-the-art CNN-based and Transformer-based models. All the results are obtained based on the best test resolutions, without using any extra training data. Our VOLO-D5 achieves the best accuracy, outperforming the latest NFNet-F6 w/ SAM [2, 15] and CaiT-M48 w/ 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.

FaceGuard: Proactive Deepfake Detection

Yuankun Yang[1]*, Chenyue Liang[2]*, Hongyu He[3], Xiaoyu Cao[3], Neil Zhenqiang Gong[3]

[1]Fudan University, 17307110068@fudan.edu.cn
[2]Chinese Academy of Sciences, lllcy_cheryl@outlook.com
[3]Duke University, {hongyu.he, xiaoyu.cao, neil.gong}@duke.edu

Abstract

Existing deepfake-detection methods focus on *passive* detection, i.e., they detect fake face images via exploiting the artifacts produced during deepfake manipulation. A key limitation of passive detection is that it cannot detect fake faces that are generated by new deepfake generation methods. In this work, we propose *FaceGuard*, a *proactive* deepfake-detection framework. FaceGuard embeds a watermark into a real face image before it is published on social media. Given a face image that claims to be an individual (e.g., Nicolas Cage), FaceGuard extracts a watermark from it and predicts the face image to be fake if the extracted watermark does not match well with the individual's ground truth one. A key component of FaceGuard is a new deep-learning-based watermarking method, which is 1) robust to normal image post-processing such as JPEG compression, Gaussian blurring, cropping, and resizing, but 2) fragile to deepfake manipulation. Our evaluation on multiple datasets shows that FaceGuard can detect deepfakes accurately and outperforms existing methods.

1 Introduction

As deep learning becomes more and more powerful, deep learning based *deepfake generation methods* can produce more and more realistic-looking deepfakes [8, 18, 19, 20, 30, 35, 41, 42, 51, 56]. In this work, we focus on fake faces because faces are key ingredients in human communications. Moreover, we focus on *manipulated* fake faces, in which a deepfake generation method replaces a target face as a source face (known as *face replacement*) or changes the facial expressions of a target face as those of a source face (known as *face reenactment*). For instance, in the well-known Trump-Cage deepfakes example [34], Trump's face (target face) is replaced as Cage's face (source face). Fake faces can be used to assist the propagation of fake news, rumors, and disinformation on social media (e.g., Facebook, Twitter, and Instagram). Therefore, fake faces pose growing concerns to the integrity of online information, highlighting the urgent needs for deepfake detection.

Existing deepfake detection mainly focuses on *passive* detection, which exploits the artifacts in fake faces to detect them after they have been generated. Specifically, given a face image, a passive detector extracts various features from it and classifies it to be real or fake based on the features. The features can be manually designed based on some heuristics [2, 14, 22, 23, 27, 50] or automatically extracted by a deep neural network based feature extractor [1, 6, 11, 14, 28, 29, 37, 38, 48, 54]. Passive detection faces a key limitation [7], i.e., it cannot detect fake faces that are generated by new deepfake generation methods that were not considered when training the passive detector. As new deepfake generation methods are continuously developed, this limitation poses significant challenges to passive deepfake detection.

Our work: In this work, we propose *FaceGuard*, a *proactive* deepfake-detection framework. FaceGuard addresses the limitation of passive detection via proactively embedding watermarks into real face images before they are manipulated by deepfake generation methods. Figure 1 illustrates the difference between passive detection and FaceGuard. Specifically, before posting an individual's real face image on social media, FaceGuard embeds a watermark (i.e., a binary vector in our work) into it. The watermark is human imperceptible, i.e., a face image and its watermarked version look visually the same to human eyes. For instance, the watermark can be embedded into an individual's face image using the individual's smartphone. Suppose a face image is claimed to be an individual, e.g., the manipulated

*The first two authors made equal contributions. They performed this research when they were remote interns in Gong's group.

Revisiting ResNets: Improved Training and Scaling Strategies

Irwan Bello [1] **William Fedus** [1] **Xianzhi Du** [1] **Ekin D. Cubuk** [1] **Aravind Srinivas** [2] **Tsung-Yi Lin** [1]
Jonathon Shlens [1] **Barret Zoph** [1]

Abstract

Novel computer vision architectures monopolize the spotlight, but the impact of the model architecture is often conflated with simultaneous changes to training methodology and scaling strategies. Our work revisits the canonical ResNet (He et al., 2015) and studies these three aspects in an effort to disentangle them. Perhaps surprisingly, we find that training and scaling strategies may matter more than architectural changes, and further, that the resulting ResNets match recent state-of-the-art models. We show that the best performing scaling strategy depends on the training regime and offer two new scaling strategies: (1) scale model depth in regimes where overfitting can occur (width scaling is preferable otherwise); (2) increase image resolution more slowly than previously recommended (Tan & Le, 2019). Using improved training and scaling strategies, we design a family of ResNet architectures, ResNet-RS, which are 1.7x - 2.7x faster than EfficientNets on TPUs, while achieving similar accuracies on ImageNet. In a large-scale semi-supervised learning setup, ResNet-RS achieves 86.2% top-1 ImageNet accuracy, while being 4.7x faster than EfficientNet-NoisyStudent. The training techniques improve transfer performance on a suite of downstream tasks (rivaling state-of-the-art self-supervised algorithms) and extend to video classification on Kinetics-400. We recommend practitioners use these simple revised ResNets as baselines for future research.

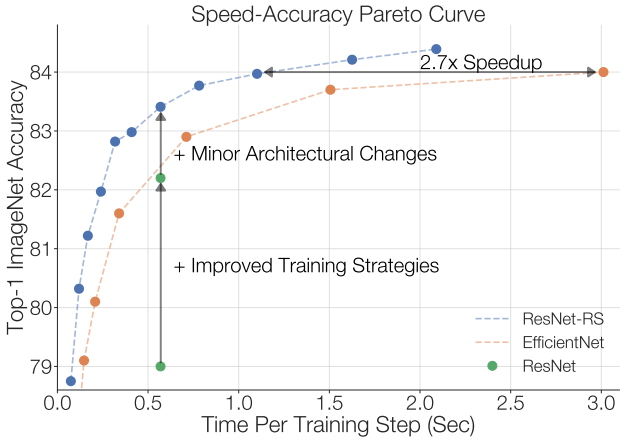

Figure 1. **Improving ResNets to state-of-the-art performance.** We improve on the canonical ResNet (He et al., 2015) with modern training methods (as also used in EfficientNets (Tan & Le, 2019)), minor architectural changes and improved scaling strategies. The resulting models, **ResNet-RS**, outperform EfficientNets on the speed-accuracy Pareto curve with speed-ups ranging from **1.7x - 2.7x** on TPUs and **2.1x - 3.3x** on GPUs. ResNet (•) is a ResNet-200 trained at 256×256 resolution. Training times reported on TPUs.

1. Introduction

The performance of a vision model is a product of the architecture, training methods and scaling strategy. However, research often emphasizes architectural changes. Novel architectures underlie many advances, but are often simultaneously introduced with other critical – and less publicized – changes in the details of the training methodology and hyperparameters. Additionally, new architectures enhanced by modern training methods are sometimes compared to older architectures with dated training methods (e.g. ResNet-50 with ImageNet Top-1 accuracy of 76.5% (He et al., 2015)). Our work addresses these issues and empirically studies the impact of *training methods* and *scaling strategies* on the popular ResNet architecture (He et al., 2015).

We survey the modern training and regularization techniques widely in use today and apply them to ResNets (Figure 1). In the process, we encounter interactions between

[1]Google Brain [2]UC Berkeley. Correspondence to: Irwan Bello and Barret Zoph <{ibello,barretzoph}@google.com>.

* Code and checkpoints available in TensorFlow:
https://github.com/tensorflow/models/tree/master/official/vision/beta
https://github.com/tensorflow/tpu/tree/master/models/official/resnet/resnet_rs

Plug-and-Play Methods Provably Converge with Properly Trained Denoisers

Ernest K. Ryu [1] **Jialin Liu** [1] **Sicheng Wang** [2] **Xiaohan Chen** [2] **Zhangyang Wang** [2] **Wotao Yin** [1]

Abstract

Plug-and-play (PnP) is a non-convex framework that integrates modern denoising priors, such as BM3D or deep learning-based denoisers, into ADMM or other proximal algorithms. An advantage of PnP is that one can use pre-trained denoisers when there is not sufficient data for end-to-end training. Although PnP has been recently studied extensively with great empirical success, theoretical analysis addressing even the most basic question of convergence has been insufficient. In this paper, we theoretically establish convergence of PnP-FBS and PnP-ADMM, without using diminishing stepsizes, under a certain Lipschitz condition on the denoisers. We then propose real spectral normalization, a technique for training deep learning-based denoisers to satisfy the proposed Lipschitz condition. Finally, we present experimental results validating the theory.

1. Introduction

Many modern image processing algorithms recover or denoise an image through the optimization problem

$$\underset{x\in\mathbb{R}^d}{\text{minimize}} \quad f(x)+\gamma g(x),$$

where the optimization variable $x\in\mathbb{R}^d$ represents the image, $f(x)$ measures data fidelity, $g(x)$ measures noisiness or complexity of the image, and $\gamma\geq 0$ is a parameter representing the relative importance between f and g. Total variation denoising, inpainting, and compressed sensing fall under this setup. *A priori* knowledge of the image, such as that the image should have small noise, is encoded in $g(x)$. So $g(x)$ is small if x has small noise or complexity. *A posteriori* knowledge of the image, such as noisy or partial measurements of the image, is encoded in $f(x)$. So $f(x)$ is small if x agrees with the measurements.

First-order iterative methods are often used to solve such optimization problems, and ADMM is one such method:

$$\begin{aligned}
x^{k+1} &= \underset{x\in\mathbb{R}^d}{\operatorname{argmin}} \left\{\sigma^2 g(x)+(1/2)\|x-(y^k-u^k)\|^2\right\}\\
y^{k+1} &= \underset{y\in\mathbb{R}^d}{\operatorname{argmin}} \left\{\alpha f(y)+(1/2)\|y-(x^{k+1}+u^k)\|^2\right\}\\
u^{k+1} &= u^k+x^{k+1}-y^{k+1}
\end{aligned}$$

with $\sigma^2=\alpha\gamma$. Given a function h on $\mathbb{R}^d$ and $\alpha>0$, define the proximal operator of h as

$$\operatorname{Prox}_{\alpha h}(z)=\underset{x\in\mathbb{R}^d}{\operatorname{argmin}} \left\{\alpha h(x)+(1/2)\|x-z\|^2\right\},$$

which is well-defined if h is proper, closed, and convex. Now we can equivalently write ADMM as

$$\begin{aligned}
x^{k+1} &= \operatorname{Prox}_{\sigma^2 g}(y^k-u^k)\\
y^{k+1} &= \operatorname{Prox}_{\alpha f}(x^{k+1}+u^k)\\
u^{k+1} &= u^k+x^{k+1}-y^{k+1}.
\end{aligned}$$

We can interpret the subroutine $\operatorname{Prox}_{\sigma^2 g}:\mathbb{R}^d\to\mathbb{R}^d$ as a denoiser, i.e.,

$$\operatorname{Prox}_{\sigma^2 g}: \text{noisy image}\mapsto\text{less noisy image}$$

(For example, if σ is the noise level and $g(x)$ is the total variation (TV) norm, then $\operatorname{Prox}_{\sigma^2 g}$ is the standard Rudin–Osher–Fatemi (ROF) model (Rudin et al., 1992).) We can think of $\operatorname{Prox}_{\alpha f}:\mathbb{R}^d\to\mathbb{R}^d$ as a mapping enforcing consistency with measured data, i.e.,

$$\operatorname{Prox}_{\alpha f}: \text{less consistent}\mapsto\text{more consistent with data}$$

More precisely speaking, for any $x\in\mathbb{R}^d$ we have

$$g(\operatorname{Prox}_{\sigma^2 g}(x))\leq g(x), \qquad f(\operatorname{Prox}_{\alpha f}(x))\leq f(x).$$

However, some state-of-the-art image denoisers with great empirical performance do not originate from optimization problems. Such examples include non-local means (NLM) (Buades et al., 2005), Block-matching and 3D filtering (BM3D) (Dabov et al., 2007), and convolutional neural

[1]Department of Mathematics, University of California, Los Angeles, USA [2]Department of Computer Science and Engineering, Texas A&M University, USA. Correspondence to: Wotao Yin <wotaoyinmath.ucla.edu>.

Proceedings of the 36^{th} International Conference on Machine Learning, Long Beach, California, PMLR 97, 2019.

Model Learning: Primal Dual Networks for Fast MR imaging

Jing Cheng[1], Haifeng Wang[1], Leslie Ying[3], Dong Liang[1, 2(✉)]

[1] Paul C. Lauterbur Research Center for Biomedical Imaging, Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences, Shenzhen, Guangdong, China
[2] Research center for Medical AI, Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences, Shenzhen, Guangdong, China
Dong.Liang@siat.ac.cn
[3] Departments of Biomedical Engineering and Electrical Engineering, University at Buffalo, the State University of New York, Buffalo, NY 14260 USA

Abstract. Magnetic resonance imaging (MRI) is known to be a slow imaging modality and undersampling in k-space has been used to increase the imaging speed. However, image reconstruction from undersampled k-space data is an ill-posed inverse problem. Iterative algorithms based on compressed sensing have been used to address the issue. In this work, we unroll the iterations of the primal-dual hybrid gradient algorithm to a learnable deep network architecture, and gradually relax the constraints to reconstruct MR images from highly undersampled k-space data. The proposed method combines the theoretical convergence guarantee of optimization methods with the powerful learning capability of deep networks. As the constraints are gradually relaxed, the reconstruction model is finally learned from the training data by updating in k-space and image domain alternatively. Experiments on in vivo MR data demonstrate that the proposed method achieves superior MR reconstructions from highly undersampled k-space data over other state-of-the-art image reconstruction methods.

Keywords: MR reconstruction, Primal dual, Deep learning.

1 Introduction

Accelerating magnetic resonance imaging (MRI) has been an ongoing research topic since its invention in the 1970s. Among a variety of acceleration techniques, compressed sensing (CS) has become an important strategy during the past decades [1]. In general, the imaging model of CS-based methods can be written as

$$\min_{m} \frac{1}{2}\|Am - f\|_2^2 + \lambda\|\Psi m\|_1 \tag{1}$$

where the first term is the data consistency and the second term is the sparse prior. Ψ is a sparse transform, such as wavelet transform or total variation, m is the image to be reconstructed, A is the encoding matrix, f denotes the acquired k-space data.