An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale

Author: Dosovitskiy et al.

created by paper2slides

2020-10-22

https://arxiv.org/abs/2010.11929

Executive Summary

- The study explores the application of Transformers, typically used in NLP, to image recognition tasks.
- Proposes the **Vision Transformer (ViT)**, which processes input images as sequences of patches, achieving competitive results.
- Demonstrates that large-scale pre-training on datasets (like ImageNet-21k and JFT-300M) surpasses CNNs on several benchmarks.
- Highlights significant performance improvements on tasks such as ImageNet, CIFAR-100, VTAB, etc., with better compute efficiency.

Introduction

- Transformers have achieved state-of-the-art results in NLP tasks.
- Application to vision tasks has been limited, often combined with CNNs.
- Key idea: pure Transformers can be applied directly to sequences of image patches.
- Goal: evaluate the potential of **Vision Transformer (ViT)** for image classification.

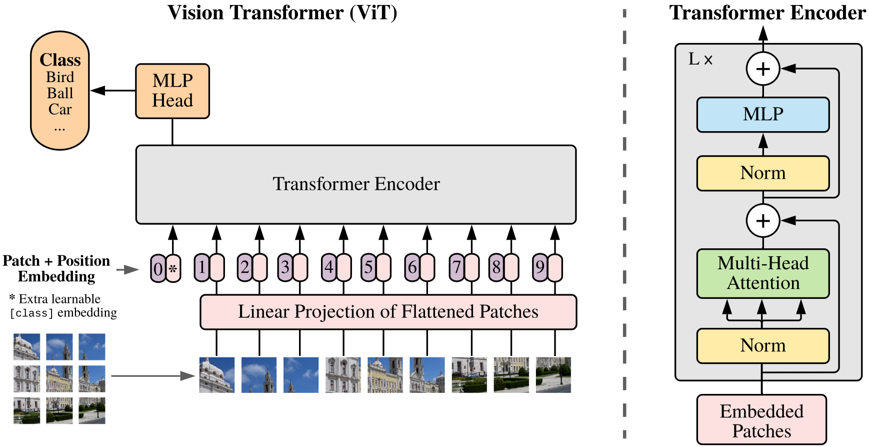

Figure: Overview of Vision Transformer (ViT): An image is split into fixed-size patches, linearly embedded, and processed by a standard Transformer encoder.

Proposed Method: Vision Transformer (ViT)

- An image is divided into non-overlapping patches.
- Each patch is linearly embedded into a fixed-size vector.
- Position embeddings are added to retain spatial information.
- The sequence of patch embeddings is input to a standard Transformer encoder.
- A classification token is added to the sequence for image classification tasks.

Design Choices in Vision Transformer

- Inspired by BERT architecture with slight modifications.
- Dimensions: Base, Large, and Huge variants for scalability.
- Patch size impacts computational efficiency and sequence length.
- Utilize large-scale pre-training datasets to leverage the model's potential.

Pre-training and Fine-tuning

- Pre-training on large datasets like ImageNet-21k and JFT-300M.
- Fine-tuning tailored to specific tasks and resolutions.
- Key benefit: Fine-tuning at higher resolutions leads to better performance.
- Position embeddings are interpolated for varying resolutions during fine-tuning.

Experimental Setup

- Datasets: ImageNet, ImageNet-21k, JFT-300M plus several benchmarks for transfer learning.
- Compared with state-of-the-art models like Big Transfer (BiT) and Noisy Student on several metrics.
- Evaluated both fine-tuning and linear few-shot performance.
- Use of Adam optimizer with specific hyperparameters for training.

Results: Comparison to State of the Art

- ViT-H/14 achieves 88.55% on ImageNet, 90.72% on ImageNet-ReaL, and 94.55% on CIFAR-100.
- Outperforms BiT-R152x4 in most datasets with lower pre-training cost.
- Significant improvement over existing models on VTAB benchmark.

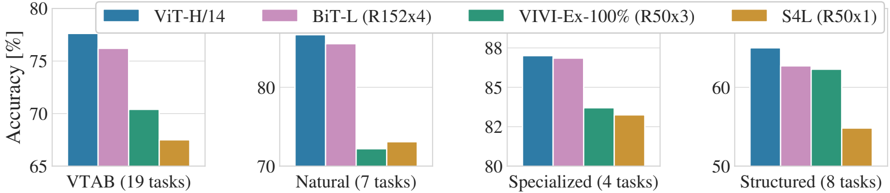

Figure: Breakdown of VTAB performance in *Natural*, *Specialized*, and *Structured* task groups.

Results: Pre-training Data Requirements

- ViT shows better performance with increased data size for pre-training.
- ViT-L/16 performs comparably to BiT with smaller datasets but excels with larger data like JFT-300M.
- Larger ViT models benefit significantly from larger pre-training datasets.

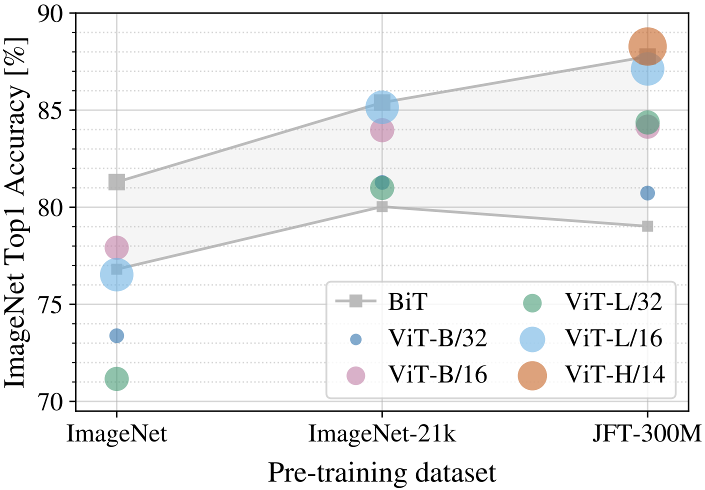

Figure: Transfer to ImageNet: Larger models show better performance with increased dataset size during pre-training.

Scaling Study and Comparison with ResNets and Hybrids

- Performance versus pre-training compute shows ViT's superior efficiency.
- ViT outperforms ResNets in performance-to-compute ratio.
- Hybrids show advantages at smaller scales but converge with ViT at larger scales.
- Larger ViT models (ViT-H/14) show no saturation, indicating potential for further scaling.

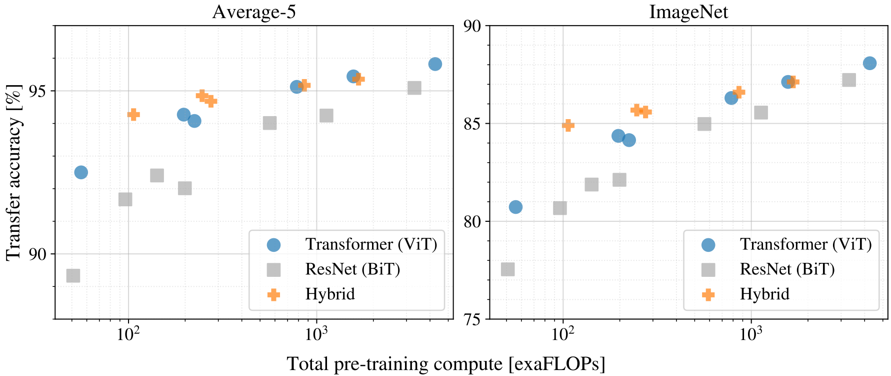

Figure: Performance versus pre-training compute for different architectures: Vision Transformers, ResNets, and hybrids.

Insights from Vision Transformer

- Attention patterns learned by ViT attend to image regions relevant for classification.
- Position embeddings capture 2D image topology, indicating learning beyond initial patch embedding.
- Large attention distances in some heads demonstrate global information integration.
- Performance improvements via self-supervised Masked Patch Prediction.

Conclusion and Future Directions

- Vision Transformer demonstrates the capability to outperform CNNs with appropriate pre-training.
- Future work includes applying ViT to other tasks such as detection and segmentation.
- Further exploration of self-supervised pre-training methods could unlock additional potential.
- Investigating more efficient Transformer variants for practical deployment.