ViT

Research Problem/Objective

- While Transformers dominate NLP, Convolutional Neural Networks (CNNs) remain the standard for computer vision, largely due to their built-in inductive biases (locality, translation equivariance).
- Previous attempts to use attention in vision either combined it with CNNs or used specialized attention patterns that were hard to scale. Applying standard Transformers directly to pixels is computationally infeasible due to quadratic complexity.
- The objective is to demonstrate that a standard Transformer architecture, with minimal modifications, can be applied directly to images for classification and achieve state-of-the-art performance, challenging the necessity of CNN-specific inductive biases, provided it's pre-trained on sufficient data.

Core Idea/Hypothesis

- The central hypothesis is that the reliance on CNN-specific inductive biases is not necessary for large-scale image recognition. Scaling up model size and, crucially, pre-training dataset size can compensate for the lack of these biases in a standard Transformer.
- The core idea is to treat an image as a sequence of patches. The image is split into fixed-size patches, each patch is linearly embedded, position embeddings are added, and this resulting sequence of vectors is fed directly into a standard Transformer encoder.

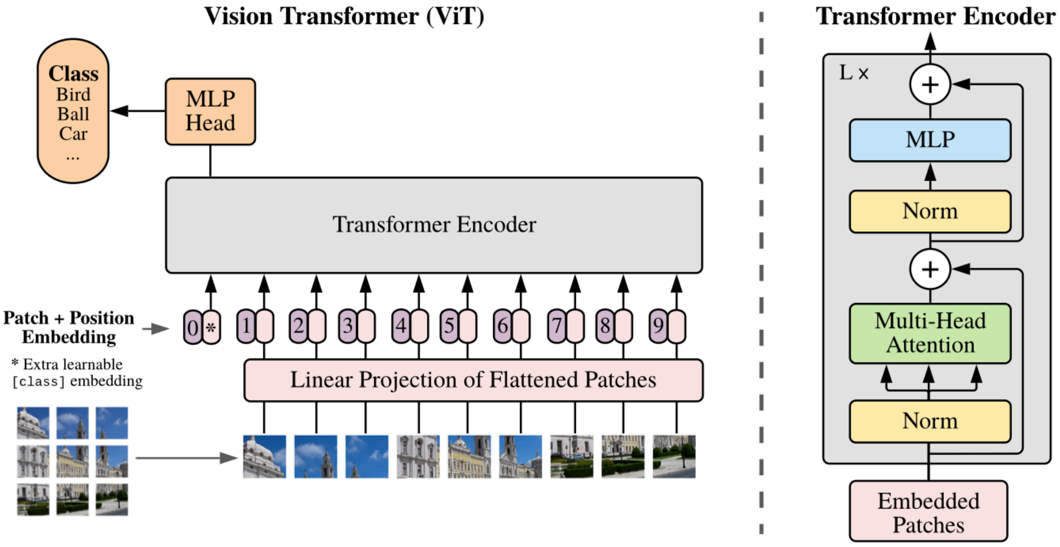
Model Architecture (Conceptual Overview)

1. **Patch Embedding:** An input image $x \in R^{(H \times W \times C)}$ is reshaped into a sequence of flattened 2D patches $xp \in R^{(N \times (P^2 \cdot C))}$, where (P, P) is the patch resolution and $N = HW/P^2$ is the number of patches (effective sequence length).
2. **Linear Projection:** These N patches are mapped to D dimensions using a trainable linear projection E, resulting in patch embeddings.
3. **Learnable Class Token:** Similar to BERT, a learnable [class] token embedding (x_{class}) is prepended to the sequence of patch embeddings. The final state of this token at the Transformer output (z^L_0) serves as the aggregate image representation.
4. **Position Embeddings:** Standard learnable 1D position embeddings (E_{pos}) are added to the patch embeddings (including the class token) to retain spatial information.
    - $z_0 = [x_{class}; x_p^1 E; x_p^2 E; ...; x_p^N E] + E_{pos}$
5. **Transformer Encoder:** A standard Transformer encoder composed of L layers. Each layer consists of Multi-Head Self-Attention (MSA) and MLP blocks. LayerNorm is applied

before each block (pre-LN), and residual connections are used after each block.

  - z'ℓ = MSA(LN(z{ℓ-1})) + z_{ℓ-1}
  - zℓ = MLP(LN(z'ℓ)) + z'_ℓ

6. **Classification Head:** A classification head (MLP during pre-training, single linear layer during fine-tuning) is attached to the output [class] token representation y = LN(z^L_0).
7. **Hybrid Architecture:** An alternative where the input patches are extracted from the feature maps of a CNN backbone instead of the raw image.



Key Innovations & Contributions

- **Pure Transformer for Vision:** Demonstrating that a standard Transformer, applied directly to image patches, can achieve SOTA image classification results without CNN backbones.
- **Scalability Trumps Inductive Bias:** Showing empirically that with sufficient pre-training data (14M-300M images), the performance of ViT surpasses that of highly optimized CNNs, overcoming its lack of built-in vision-specific inductive biases.
- **Patch-based Input:** Introducing the simple yet effective method of splitting images into patches and treating them as a sequence for a standard Transformer.
- **Competitive Performance with Lower Pre-training Cost:** Achieving SOTA or comparable results on multiple benchmarks while requiring significantly less computational resources for pre-training compared to equivalent SOTA CNNs (like BiT).

Research Methodology & Evaluation (High-Level)

- **Models:** ViT (ViT-Base, ViT-Large, ViT-Huge with varying patch sizes, e.g., ViT-L/16), ResNet (BiT variants), Hybrid (ResNet+ViT).
- **Pre-training Datasets:** ImageNet (1.3M), ImageNet-21k (14M), JFT-300M (303M).

- **Downstream Tasks:** ImageNet (incl. ReaL labels), CIFAR-10/100, Oxford-IIIT Pets, Oxford Flowers-102, VTAB suite (19 tasks).
- **Evaluation:** Fine-tuning accuracy on downstream tasks, few-shot linear accuracy. Comparison based on performance vs. pre-training compute cost (TPUv3-core-days).
- **Training Details:** Adam for pre-training, SGD w/ momentum for fine-tuning. Explored regularization (weight decay, dropout, label smoothing). Used higher resolution for fine-tuning.

Significant Findings & Results

- **Data Scaling is Key:** ViT models underperform comparable ResNets when pre-trained on mid-sized datasets (ImageNet) but excel when pre-trained on large datasets (ImageNet-21k, JFT-300M). Larger ViT models require larger datasets to show their benefit.
- **SOTA Performance:** ViT-H/14 pre-trained on JFT-300M achieved SOTA results on ImageNet (88.55%), ImageNet-ReaL (90.72%), CIFAR-100 (94.55%), and VTAB (77.63%), surpassing previous SOTA CNNs like BiT-L and Noisy Student at the time.
- **Pre-training Efficiency:** ViT models achieve better performance/compute trade-offs than ResNets. E.g., ViT-L/16 pre-trained on JFT outperformed BiT-L pre-trained on the same dataset, using ~2-4x less compute.
- **Hybrid Performance:** Hybrids slightly outperform pure ViT at smaller compute budgets, but this gap closes for larger models.
- **Learned Representations:**
    - The initial patch embedding layer learns basis functions similar to CNN filters.
    - Learned 1D position embeddings effectively encode 2D image structure (distance, row/column relationships), making explicit 2D embeddings unnecessary.
    - Self-attention integrates information globally even in early layers. Attention distance ("receptive field") increases with depth. Attention focuses on semantically relevant image regions.
- **Self-Supervision:** Preliminary experiments using masked patch prediction showed improvement over training from scratch but significantly lagged behind supervised pre-training on large datasets.

Conclusions & Implications

- The Vision Transformer (ViT) demonstrates that architectures based purely on self-attention can achieve excellent results on image recognition tasks, directly challenging the long-standing dominance of CNNs.
- The importance of image-specific inductive biases diminishes with large-scale pre-training data; relevant patterns can be learned directly.
- ViT provides a simple, scalable, and computationally efficient alternative for vision, benefiting from the advancements and infrastructure developed for Transformers in NLP.

- The work opens avenues for applying Transformers to other vision tasks (detection, segmentation) and further exploring self-supervised learning methods for vision Transformers.