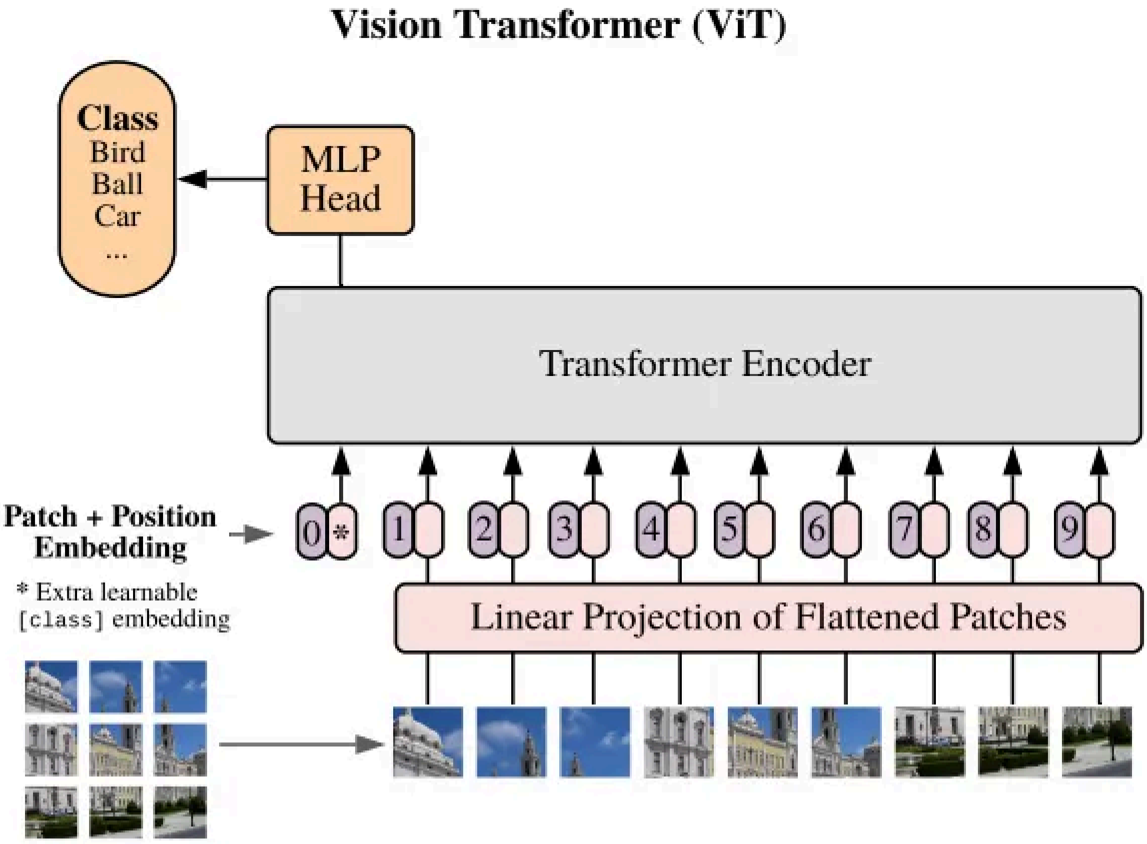
A Image is Worth Words The ViT Research Paper Notes

Summary

The Vision Transformer (ViT) replaces traditional Convolutional Neural Networks (CNNs) by splitting images into patches and processing them like words in a sentence, achieving state-of-the-art accuracy through massive-scale pre-training.


Vision Transformer (ViT)
Class
Bird
Ball
Car
...
MLP Head
Transformer Encoder
Patch + Position Embedding
* Extra learnable [class] embedding
0
*
1
2
3
4
5
6
7
8
9
Linear Projection of Flattened Patches

Concepts

- **Vision Transformer (ViT):** A model architecture that applies the standard Transformer (popular in Natural Language Processing) directly to images, with minimal modifications.
- **Image Patches:** The method of breaking an image down into small, fixed-size squares to create a sequence, similar to how a sentence is broken down into a sequence of words.
- **Inductive Bias:** Assumptions built into a model about the data (e.g., CNNs "know" that pixels near each other are related). Transformers lack this bias, meaning they don't inherently understand spatial relationships (left, right, up, down) until they are taught.

- **High Inductive Bias:** The model relies on strict, built-in rules (like a study guide) to learn quickly from small datasets, but this rigidity limits its ability to find novel patterns outside those rules.
- **Low Inductive Bias:** The model starts with minimal assumptions and needs massive amounts of data to learn relationships from scratch, but this freedom allows it to eventually discover more complex and accurate patterns than models with strict rules
- **Pre-training:** The process of training a model on a massive generic dataset before "fine-tuning" it for a specific task.

Methodology / How it Works

The authors proposed a shift from processing pixels (CNN approach) to processing sequences (Transformer approach).

1. **Patch Partitioning:** The input image is split into fixed-size square patches (e.g., 16×16 pixels).
2. **Linear Embedding:** Each patch is flattened into a vector (a list of numbers).
3. **Position Embeddings:** Because the Transformer doesn't know the order of the patches (it doesn't know the top-left patch belongs in the top-left), "position tags" are added to the data so the model understands the spatial structure.
4. **Transformer Encoder:** The sequence of patches is fed into a standard Transformer Encoder.
5. **Massive Pre-training:** Since the model lacks "Inductive Bias," the authors pre-trained it on enormous datasets (14 million to 300 million images). This volume of data forces the model to learn spatial relationships from scratch.

> **Key Takeaway:** ViT treats an image exactly like a text document, where "patches" equal "words."

Key Results

- **Performance:** When pre-trained on large scales (JFT-300M dataset), ViT matches or exceeds the performance of the best CNNs (like ResNet) on standard benchmarks such as ImageNet and CIFAR-100.
- **Efficiency:** Despite being a massive model, ViT is actually *more* computationally efficient to train than state-of-the-art CNNs.
- **Scalability:** The research proves that Transformers do not need to rely on the complex hardware-specific architecture of CNNs to solve computer vision tasks.

Inductive Bias & Attention in Vision Transformers (ViT)

1. Removing the Guardrails (CNN vs. ViT)

- **CNN Approach:** CNNs bake in "Locality" and "Translation Equivariance" into every layer. They use sliding windows (kernels) that only look at small neighborhoods of pixels at a time.
- **ViT Approach:** ViT removes these hard-coded rules. It treats the image as a sequence of patches. The only manual structures are:
  1. Cutting the image into patches.
  2. Adding position embeddings (to tell the model where the patches go).

2. The Data Trade-off

Because ViT lacks the "template" (bias) that CNNs have, it struggles on smaller datasets (like ImageNet). It doesn't know that neighboring pixels are important yet.

- **Mid-sized Data:** ViT fails to generalize and performs worse than ResNet.
- **Massive Data (e.g., JFT-300M):** ViT eventually learns these spatial rules directly from the data.8 Once it learns them, it outperforms CNNs because it isn't restricted by the hard-coded architecture.

3. Managing Computation

Calculating attention for every single pixel is too expensive (N^2 complexity). ViT solves this by looking at **patches** (e.g., 16×16 zones) rather than individual pixels, making global attention computationally feasible.

Conclusion

1. Global Integration

Unlike CNNs, which need to stack many layers to "see" the whole image, ViT can integrate information across the entire image in a single layer due to global self-attention.9

2. Attention Distance Findings

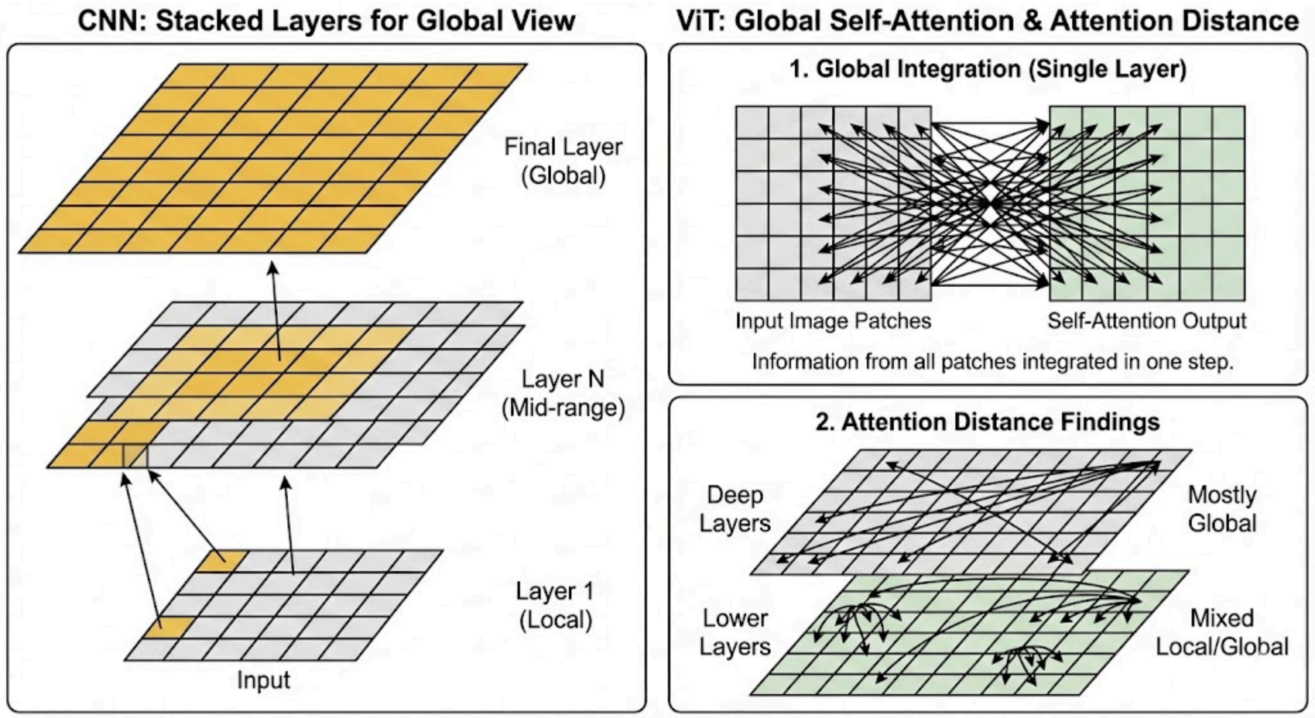
By analyzing how far the model "looks" (Attention Distance), the researchers found:

- **Lower Layers:** These behave like a mix. Some heads look globally, while others mimic CNNs by focusing on local neighborhoods (learning "Locality" from scratch).10
- **Deep Layers:** As the network gets deeper, the attention becomes almost entirely global, integrating information from everywhere.

3. Semantic Focus

Visualizing the attention maps shows that ViT automatically learns to focus on semantically relevant objects in the image (e.g., the dog) and ignore the background, without being explicitly told to do so.

4. Visual Representation


CNN: Stacked Layers for Global View
Final Layer
(Global)
Layer N
(Mid-range)
Layer 1
(Local)
Input
ViT: Global Self-Attention & Attention Distance
1. Global Integration (Single Layer)
Input Image Patches
Self-Attention Output
Information from all patches integrated in one step.
2. Attention Distance Findings
Deep
Layers
Mostly
Global
Lower
Layers
Mixed
Local/Global

Future Challenges & Directions for Vision Transformers (ViT)

The future of ViT revolves on applying the model to complex tasks like object detection, closing the performance gap between supervised and self-supervised learning, and continuing to scale model depth to unlock unsaturated performance gains.

- **Self-Supervised Learning:** A training method where the model learns from the data itself without human-provided labels. The goal is to stop relying on expensive, manually labeled datasets.
  - Real World Example:_ A student learning a language by listening to the radio and guessing the missing words (Self-Supervised) versus a student learning by memorizing a dictionary with a teacher grading every word (Supervised).
- **Masked Patch Prediction:** An experimental self-supervised technique used in ViT. The researchers corrupt (hide) 50% of the image patches and ask the model to guess the color of the missing parts.

- **Saturation:** The point where adding more data or making the model bigger stops improving performance. ViT has *not* reached this point yet.
- **Axial Attention:** A proposed variation of attention intended to be more efficient than standard attention, though currently difficult to implement efficiently on hardware (TPUs).

The researchers outline four specific directions to evolve the ViT architecture:

1. **Expanding Application Scope:**
   - Move beyond simple "Image Classification" (what is this image?) to complex spatial tasks like **Object Detection** (where is the object?) and **Segmentation** (what is the exact outline of the object?).
2. **Improving Self-Supervised Pre-training:**
   - Currently, there is a performance gap. Self-supervised ViT is 4% less accurate than Supervised ViT.
   - *Current Attempt:* They tried "Masked Patch Prediction" (mimicking BERT in NLP). It improved the model by 2% compared to training from scratch, but is still behind supervised methods.
   - *Next Step:* Explore **Contrastive Pre-training** to close this gap.
3. **Scaling Strategies:**
   - The model gets better the bigger it gets.
   - **Depth over Width:** Studies show that making the network *deeper* (more layers) is more effective than making it *wider* (more neurons per layer).
4. **Rethinking Architecture:**
   - The authors speculate that if datasets get big enough, they might remove even the last few "Inductive Biases," such as **Skip Connections** (Residual connections), making the model purely data-driven.
