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

Google Research, Brain Team

Yuxin Liu, Yijia Xue

Background

Vision Transformer (ViT) is an innovative image classification model proposed in 2021. It is breaking away from the dominant role of traditional CNNs in image processing.

CNNs: Relies on convolution operations.

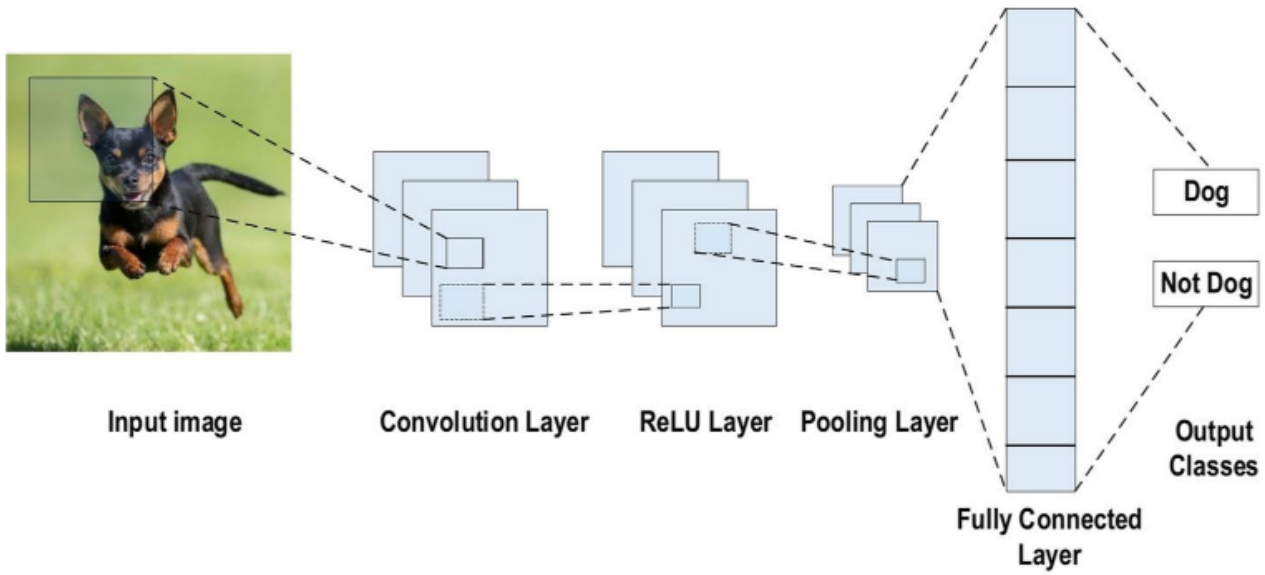


Figure: An example of CNNs architecture for image classification [1].

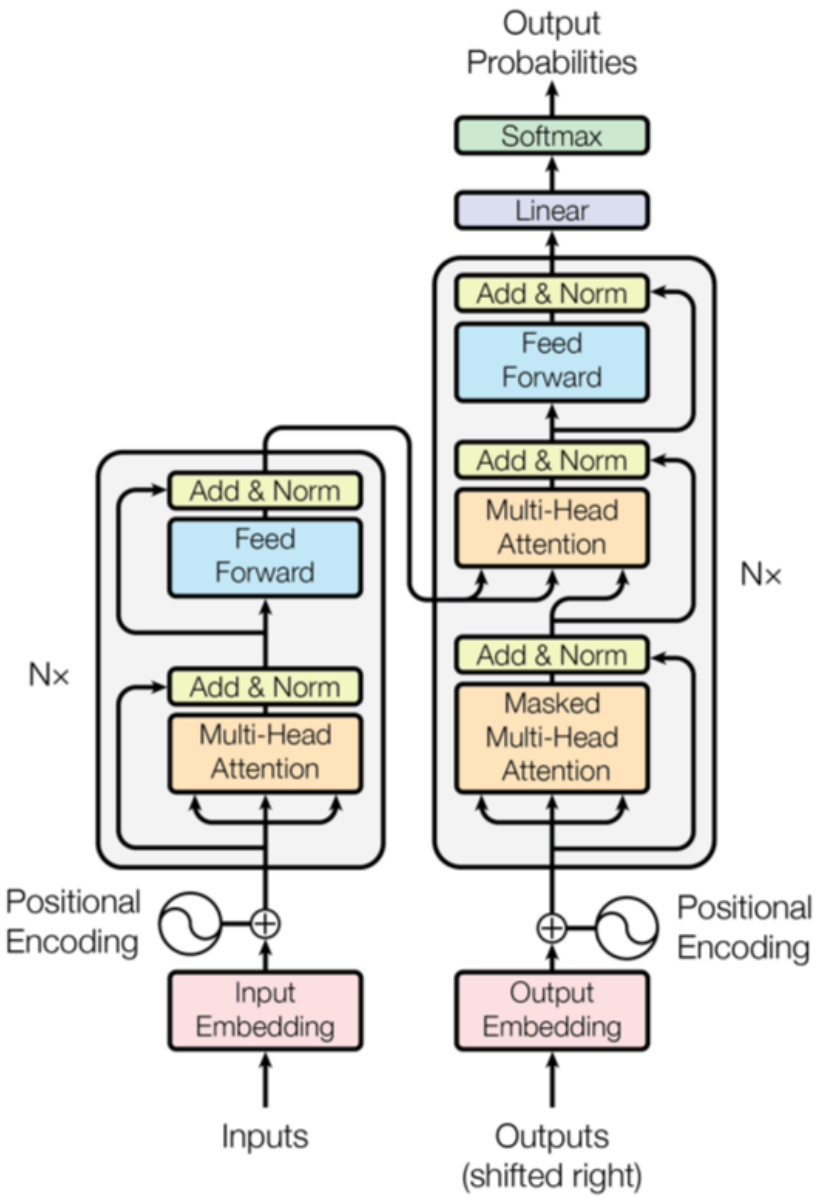
Vision Transformer (ViT): Transformer $\xrightarrow{\text{apply}}$ Image regconition

1. Inspired by the Transformer scaling successes in NLP, they experiment with applying a standard Transformer directly to images, with the fewest possible modifications.
2. Treat image patches the same way as tokens (words) in an NLP application.

A transformer is a deep learning architecture that was developed by researchers at Google and is based on the multi-head attention mechanism, which was proposed in the 2017 paper "Attention Is All You Need" [2].

Introduction

Transformer - NLP



Vision Transformation - ViT

Model Overview

Split image into patches → Embedding → Transformer encoder

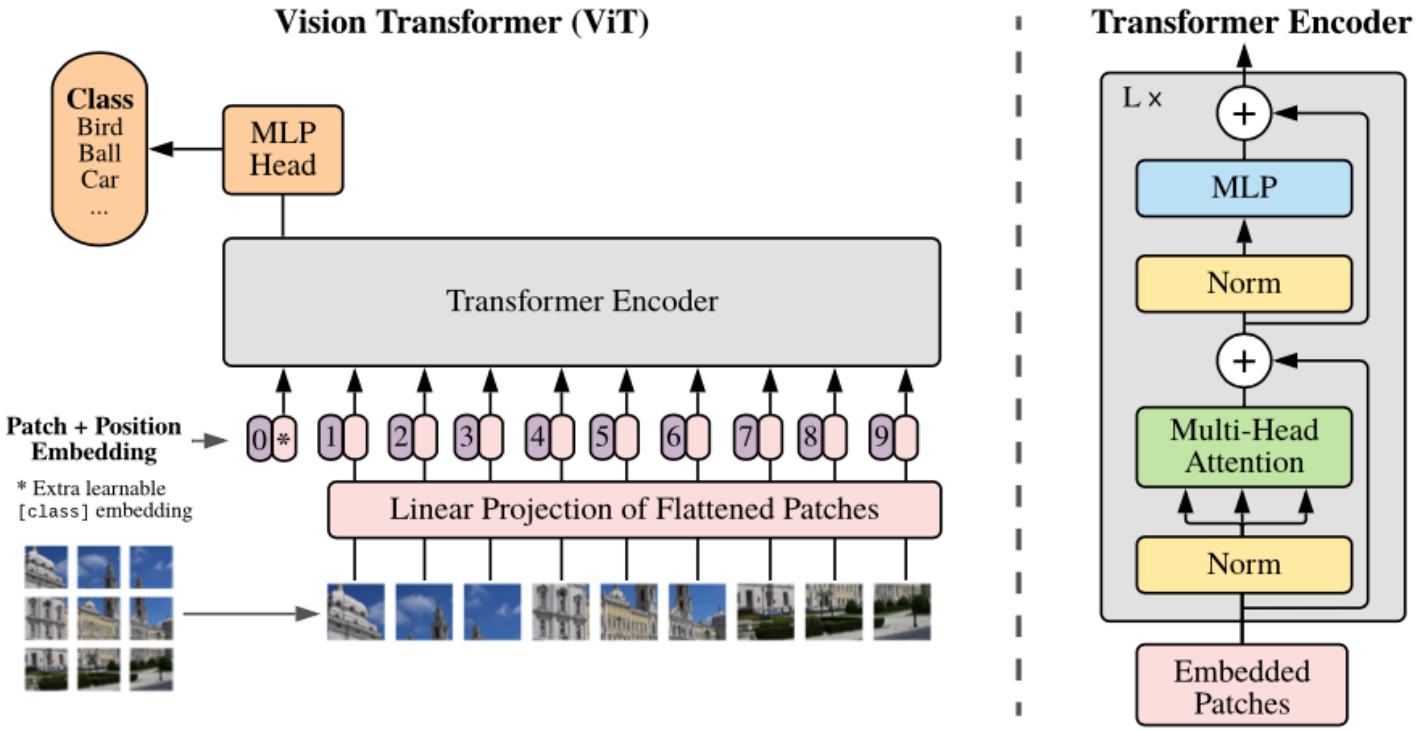


Figure: Model Overview

Vision Transformation - ViT

Method - Reshape Image

Reshape the image $x \in \mathbb{R}^{H \times W \times C}$ into a sequence of flattened 2D patches $x_p \in \mathbb{R}^{N \times (P^2 \cdot C)}$, where (H, W) is the resolution of the original image, C is the number of channels, (P, P) is the resolution of each image patch, and $N = HW/P^2$ is the resulting number of patches.

Figure: Image Reshaping

Vision Transformation - ViT

Method - Patch Embedding

Patch embedding:

Flatten the patches and map to D dimensions with a trainable linear projection.

$$z_0 = [x_{\text{class}};\ x_p^1\mathbf{E};\ x_p^2\mathbf{E};\ \ldots;\ x_p^N\mathbf{E}] + \mathbf{E}_{pos}$$

where $\mathbf{E} \in \mathbb{R}^{(P^2 \cdot C) \times D}$, $\mathbf{E}_{pos} \in \mathbb{R}^{(N+1) \times D}$.

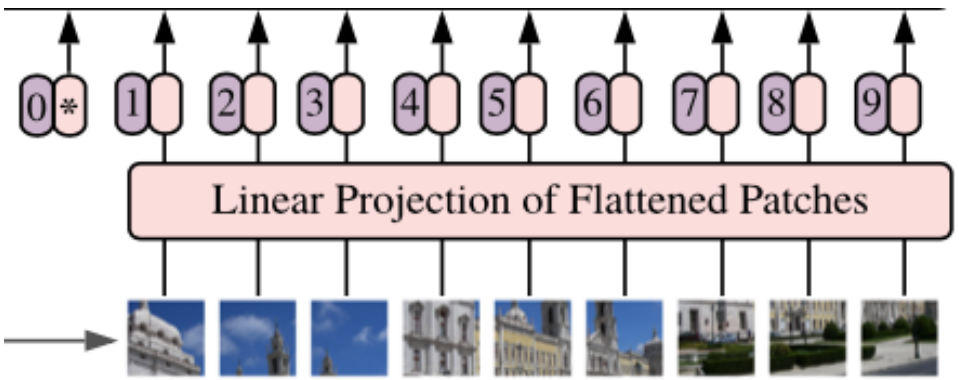


Figure: Patch Embedding

Vision Transformation - ViT

Method

- **Inductive Bias** Vision Transformer has much less image-specific inductive bias than CNNs. Only MLP layers are local and translationally equivariant, while the self-attention layers are global. Resolution adjustments and patch extraction are the only inductive biases introduced manually.
- **Hybrid Architecture** The input sequence can be formed from feature maps of a CNN. In this hybrid model, the patch embedding projection **E** is applied to patches extracted from a CNN feature map.

Vision Transformation - ViT

Method - Fine-Tuning and Higher Resolution

- Pre-training is done on large datasets; fine-tuning adapts ViT for downstream tasks.
- The pre-trained prediction head is removed, replaced with a new $D \times K$ feedforward layer.
- When fine-tuning at a higher resolution:
  - Keep the patch size constant, increasing the sequence length.
  - Perform 2D interpolation of position embeddings to adapt to the new resolution.
- ViT can handle arbitrary sequence lengths, but pre-trained position embeddings may lose meaning.

Experiments

Setup

Datasets: ILSVRC-2012 ImageNet; ImageNet-21; JFT
Benchmark tasks: original validation labels; ReaL labels; CIFAR-10/100; Oxford-IIIT Pets; Oxford Flowers-102; 19-task VTAB classification suite.
ViT: Base, Large, Huge Models. (Ex. ViT-L/16 $\sim$ "Large" variant with 16 $\times$ 16 input patch size) Configurations based on BERT. Smaller patch size are more computational expensive.
Baseline CNNs: ResNet; replaced Batch Normalization layer with Group Normalization.
Hybrid: Intermediate feature maps from CNN to ViT with patch size $= 1$.

Experiments

Setup

Training

Optimizer: Adam ($\beta_1 = 0.9$, $\beta_2 = 0.999$)
Batch Size: 4096
Weight Decay: 0.1 (for improved transfer learning)
Learning Rate: Linear warmup & decay

Fine-Tuning

Optimizer: SGD with momentum
Batch Size: 512
Averaging: Polyak Averaging (factor = 0.9999) for improved generalization.

Metrics

Few-Shot Accuracy: Regularized least-squares regression that maps training image to $\{-1, 1\}^K$ target vectors
Fine-Tuning Accuracy: Performance after fine-tuning on downstream tasks.

Comparison

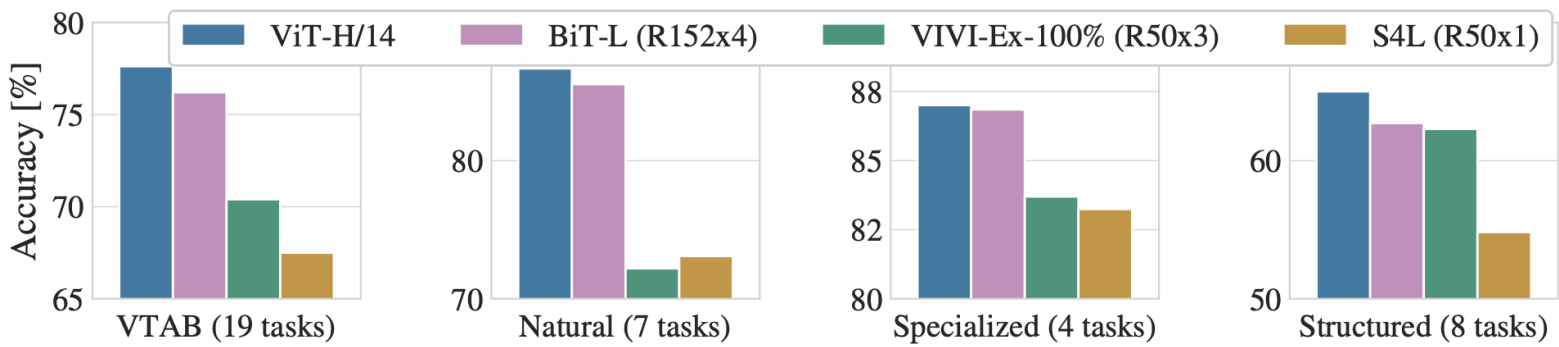
ViT-H/14 & ViT-L/16 vs.Big Transfer & Noisy Student

Result Overview

- **ViT-L/16** (pre-trained on JFT-300M) **outperforms BiT-L** on all tasks while using fewer computational resources.
- **ViT-H/14** (pre-trained on JFT-300M) shows improved performance, especially on challenging datasets like ImageNet, but also requires less compute than previous SOTA.
- **ViT-L/16** (pre-trained on ImageNet-21k) delivers strong performance across tasks and is resource-efficient. (as a controlled study for different architectures)

Comparison

	Ours-JFT (ViT-H/14)	Ours-JFT (ViT-L/16)	Ours-I21k (ViT-L/16)	BiT-L (ResNet152x4)	Noisy Student (EfficientNet-L2)
ImageNet	**88.55** ± 0.04	87.76 ± 0.03	85.30 ± 0.02	87.54 ± 0.02	88.4/88.5*
ImageNet ReaL	**90.72** ± 0.05	90.54 ± 0.03	88.62 ± 0.05	90.54	90.55
CIFAR-10	**99.50** ± 0.06	99.42 ± 0.03	99.15 ± 0.03	99.37 ± 0.06	–
CIFAR-100	**94.55** ± 0.04	93.90 ± 0.05	93.25 ± 0.05	93.51 ± 0.08	–
Oxford-IIIT Pets	**97.56** ± 0.03	97.32 ± 0.11	94.67 ± 0.15	96.62 ± 0.23	–
Oxford Flowers-102	99.68 ± 0.02	**99.74** ± 0.00	99.61 ± 0.02	99.63 ± 0.03	–
VTAB (19 tasks)	**77.63** ± 0.23	76.28 ± 0.46	72.72 ± 0.21	76.29 ± 1.70	–
TPUv3-core-days	2.5k	0.68k	0.23k	9.9k	12.3k



Scaling Study

Goal

- Evaluate transfer performance from JFT-300M vs. pre-training cost for different models.

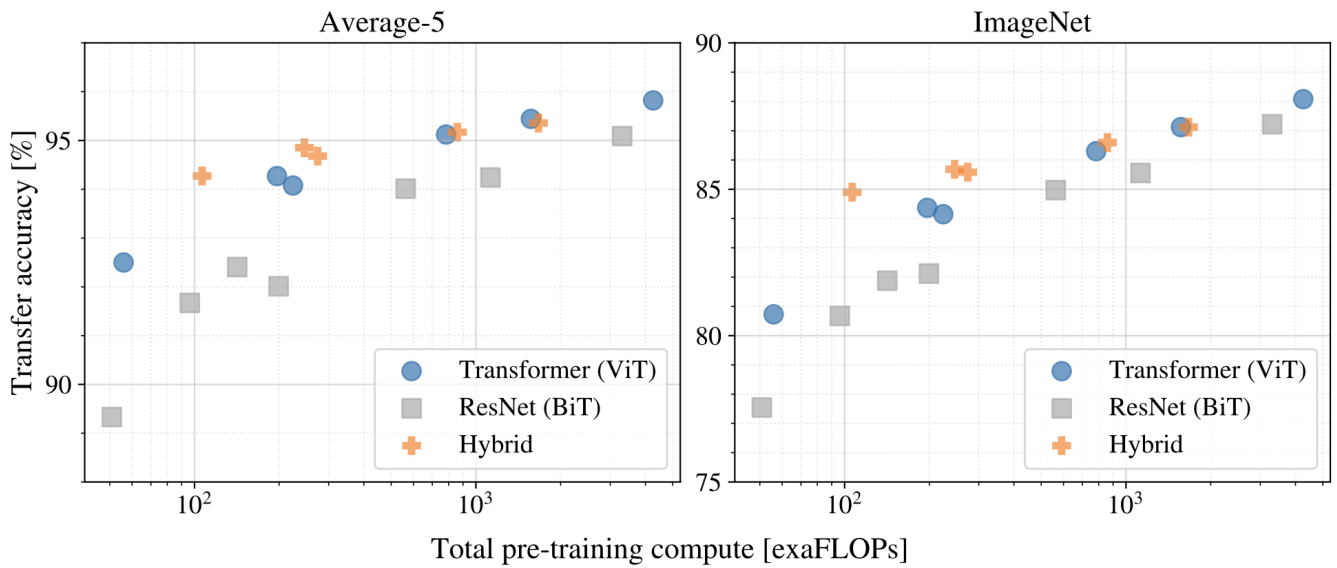
Model Set

- 7 ResNets: R50x1, R50x2, R101x1, R152x1, R152x2 (7 epochs), R152x2, R200x3 (14 epochs)
- 6 ViTs: ViT-B/32, B/16, L/32, L/16 (7 epochs), L/16, H/14 (14 epochs)
- 5 Hybrids: Combos of ResNet + ViT models (7 epochs for smaller, 14 for larger models)

Scaling Study

A controlled scaling study - performance on JFT-300M

- **ViTs outperform ResNets:** ViTs use 2-4× less compute to achieve the same performance across 5 datasets.
- **Hybrids outperform ViTs at small budgets:** Hybrids are better at lower computational budgets, but this advantage diminishes as model size increases.
- **ViTs don't saturate:** Performance continues to improve without significant saturation in the tested range.



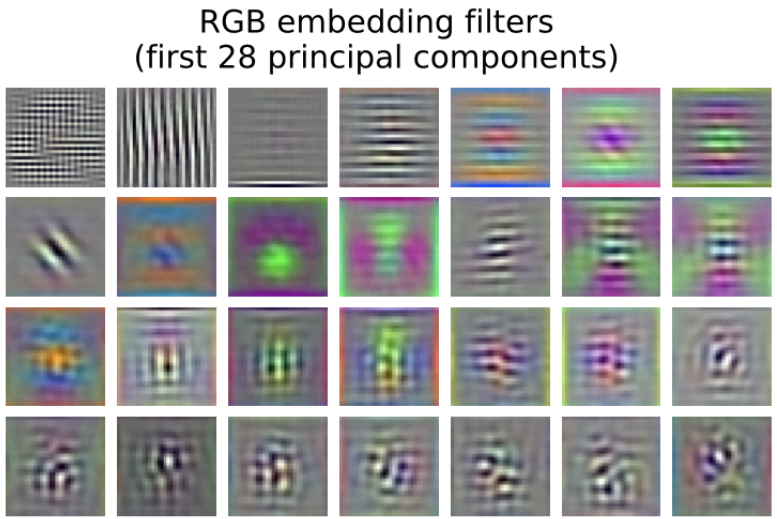
Internal Representation

1. **Projection and Embedding**

- **First Layer:** The Vision Transformer (ViT) linearly projects flattened image patches into a lower-dimensional space.

$$z_0 = [X_{class}\,; x_p^1 E;\ x_p^2 E; ...;\ x_p^N E] + E_{pos},$$
$$\mathrm{E} \in \mathbb{R}^{(P^2 \cdot C) \times D},\ E_{pos} \in \mathbb{R}^{(N+1) \times D}$$

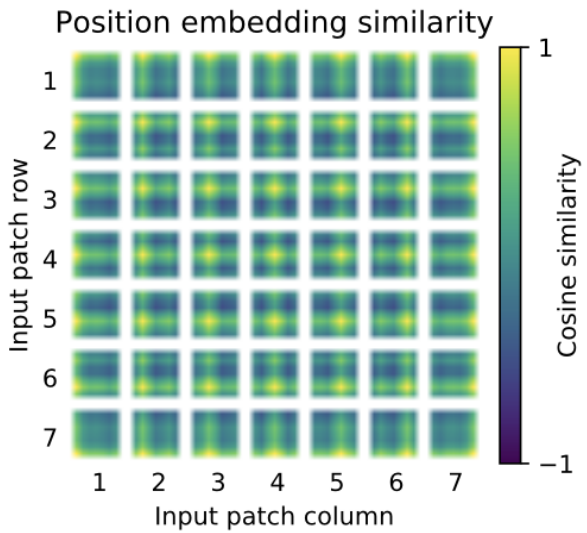
- **Embedding Filters:** The top principal components of the learned embeddings resemble basis functions for capturing fine details within each image patch.



Internal Representation

2. **Learned Position Embedding**

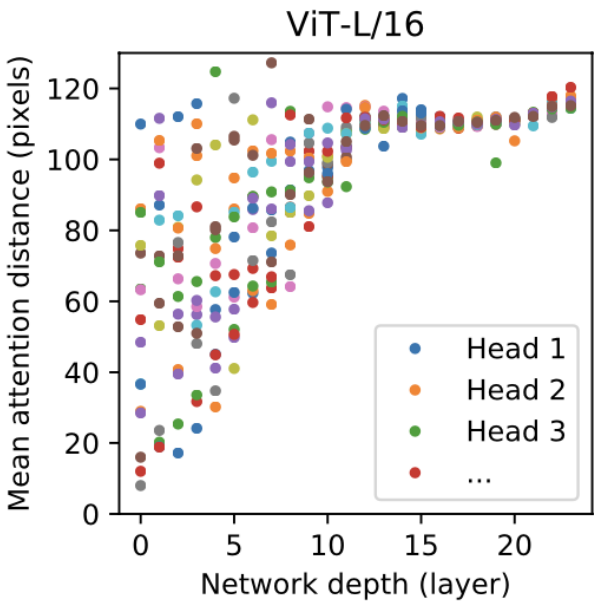
- **Spatial Distance:** Patches closer in space tend to have more similar position embeddings.
- **Row/Column Structure:** Patches in the same row or column have similar embeddings.
- **Sinusoidal Structure:** Apparent in larger grids, explaining why 2D-aware hand-crafted embeddings don't improve performance

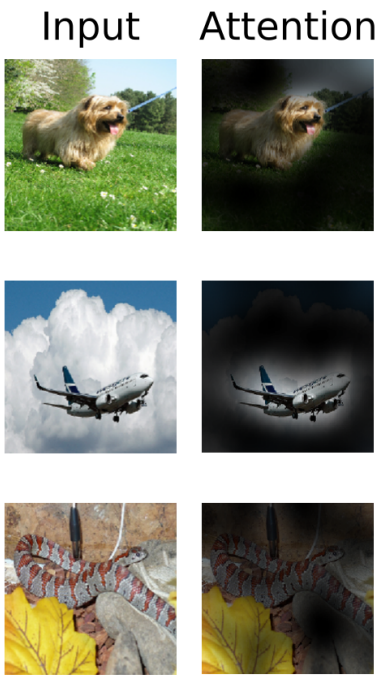


Internal Representation

3. **Self Attention**

- **Global Integration via Self-Attention:** ViT uses self-attention to integrate information across the entire image, even in the lowest layers.
- **Attention Distance:** The average distance across which information is integrated (similar to receptive field in CNNs)





Self-Attention

A popular building block for neural architectures.

For each element in an input sequence $\boldsymbol{z} \in \mathbb{R}^{N \times D}$, compute a weighted sum over all values $\boldsymbol{v}$ in the sequence. The attention weights A_{ij} are based on the pairwise similarity between two elements of the sequence and their respective query $\boldsymbol{q}^i$ and key $\boldsymbol{k}^j$ representations.

$$[\boldsymbol{q}, \boldsymbol{k}, \boldsymbol{v}] = \boldsymbol{z}\boldsymbol{U}_{qkv} \qquad \boldsymbol{U}_{qkv} \in \mathbb{R}^{D \times 3D_h}$$

$$A = \text{softmax}(\boldsymbol{q}\boldsymbol{k}^\top / \sqrt{D_h}) \qquad A \in \mathbb{R}^{N \times N}$$

$$SA(\boldsymbol{z}) = A\boldsymbol{v}$$

Conclusion

- **Direct application of Transformer**: Apply standard Transformer encoders to images by treating images as a sequence of patches with minimal image-specific inductive biases.
- **Pre-training Success:** This strategy works surprisingly well when coupled with pre-training on large datasets.
- **Computational Efficiency:** ViT offers strong performance with relatively low pre-training cost.

Quiz

Quiz 1: Match each feature with the correct model (CNN or ViT).

Feature	Model
(A) Uses Self-attention mechanism (Transformers) to process images	_____
(B) Use convolutional filter slides across the image and extracts features	_____
(C) Treat images as sequences of patches	_____
(D) Works well even with small datasets	_____
(E) Needs large datasets	_____
(F) Requires positional encodings to retain spatial information	_____

Quiz 2: In patch embedding process, if the flattened 2D patches is $x_p = \begin{pmatrix} 1 & 2 \\ 1 & 1 \end{pmatrix}$ and the linear projection matrix is $\mathbf{E} = \begin{pmatrix} 1 & 1 & 1 \\ 1 & 4 & 4 \end{pmatrix}$, what is $x_p^2\mathbf{E}$ after patch embedding?

$$\text{Hint:} \quad z_0 = [x_{\text{class}};\ x_p^1\mathbf{E};\ x_p^2\mathbf{E}; \ldots;\ x_p^N\mathbf{E}] + \mathbf{E}_{pos}$$

References

L. Alzubaidi, J. Zhang, A. J. Humaidi, A. Al-dujaili, Y. Duan, O. Al-Shamma, J. I. Santamaría, M. A. Fadhel, M. Al-Amidie, and L. Farhan.
Review of deep learning: concepts, cnn architectures, challenges, applications, future directions.
Journal of Big Data, 8, 2021.

A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin.
Attention is all you need, 2023.