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

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, Neil Houlsby

Published: 12 Jan 2021, Last Modified: 26 May 2025 ICLR 2021 Oral Readers: Everyone Show Bibtex Show Revisions

Keywords: computer vision, image recognition, self-attention, transformer, large-scale training

Abstract: While the Transformer architecture has become the de-facto standard for natural language processing tasks, its applications to computer vision remain limited. In vision, attention is either applied in conjunction with convolutional networks, or used to replace certain components of convolutional networks while keeping their overall structure in place. We show that this reliance on CNNs is not necessary and a pure transformer applied directly to sequences of image patches can perform very well on image classification tasks. When pre-trained on large amounts of data and transferred to multiple mid-sized or small image recognition benchmarks (ImageNet, CIFAR-100, VTAB, etc.), Vision Transformer (ViT) attains excellent results compared to state-of-the-art convolutional networks while requiring substantially fewer computational resources to train.

One-sentence Summary: Transformers applied directly to image patches and pre-trained on large datasets work really well on image classification.

Code Of Ethics: I acknowledge that I and all co-authors of this work have read and commit to adhering to the ICLR Code of Ethics

Code: google-research/vision_transformer + 142 community implementations

Data: CIFAR-10, CIFAR-100, ImageNet, ImageNet-W, JFT-300M, MAFW, ObjectNet, OmniBenchmark, Oxford 102 Flower, Oxford-IIIT Pets, VizWiz-Classification

Community Implementations: 16 code implementations

ViT = {Vision Transformers}

An Image is Worth 16x16 Words → ViT Paper

Year = 2021

Original Transformer → NLP Tasks
(Vaswani et. al)

ViT Paper → Computer Vision Tasks
↓
apply transformers directly to images.
(No CNN Involved)

- Findings

→ Pretraining ViT on small/medium datasets don't perform as well as ResNet. But with large datasets (4-300M) outperforms.

Why??

→ Inductive Bias = Built in assumptions a model has about the structure of the data

like Translation Equivariance (if an obj moves accross image the detection also moves)

Locality (pxls close to each other are related), Heirarchical Composition (small features (edges) combine into larger ones (objects)), Weight Sharing (Same filter applied accross image)

In ViT,

→ No idea of heirarchy

→ Treat all image patches equally.

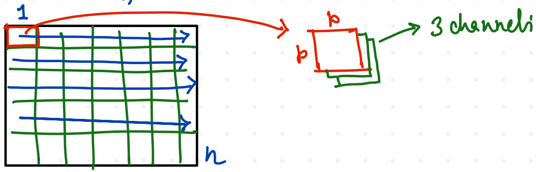
$\boxed{\text{Base Process}}$

$\boxed{\text{Step 1}}$ = Split the image into fixed image patches

if $H \times W$ = resolution, P - patch size, C = Channel size (3 for RGB

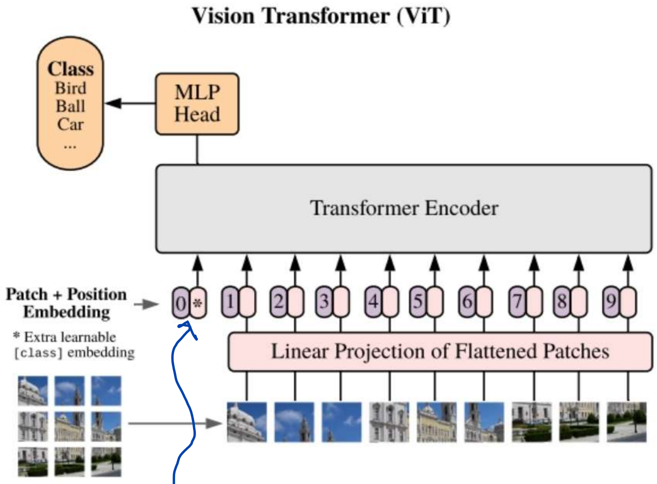
$\therefore$ for an image $x \in \mathbb{R}^{H \times W \times C}$

Resulting no. of patches, $N = HW/P^2$

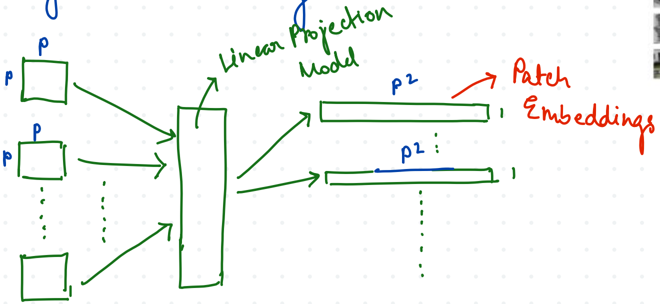


$\boxed{\text{Step 2}}$: Pass the patches through a linear Projection Layer

(Similar to the word embedding layer of the transformer architecture, where they embed the tokens)



Here each patch is converted into a $(1 \times p^2)$ long vector embedding.



We use p^2 for the dimension for not losing any information.

The final image size is $x_p \in \mathbb{R}^{N \times (p^2 \cdot c)}$

$\boxed{\text{Step 3}}$: Addition of Positional Embeddings & <CLS> token.

<CLS> Token is prepended to the sequence of the patch embeddings (Img. Representation)

(Acts as a placeholder to learn a representation of the entire image, gathering info for all patch tokens)

Positional Embedding: Provide info about each patch location of the entire image.

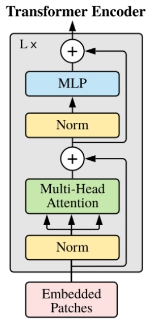
Significance of <class> or <cls> token

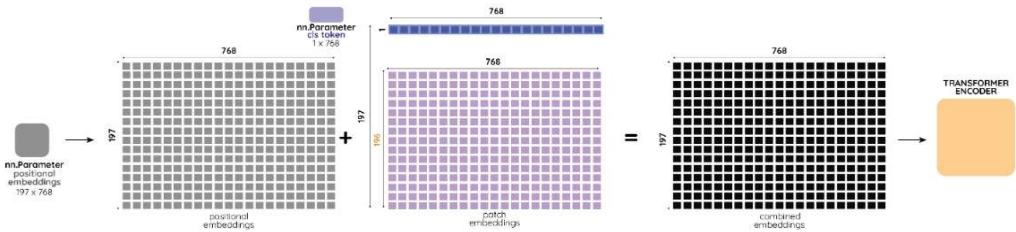
This token is prepended to the (Patch + Positional) Embedding which is then fed to the transformer.

The final value of this token becomes the class representation. which is fed to a classification head / MLP layer to make predictions.

<Remember the ViT paper was trying to attempt image classification using Transformer & hence classification head>

Step 4: These Embeddings are then passed through the Transformer Encoder which o/p's a $(1 \times p^2)$ vector → <CLS> embedding which is then passed through an MLP to get the predictions





* In ViT, only MLP layers are local (translation invariant) & the self attention layers are global

global ↓

capture long range dependencies

allows every patch to attend every other patch.

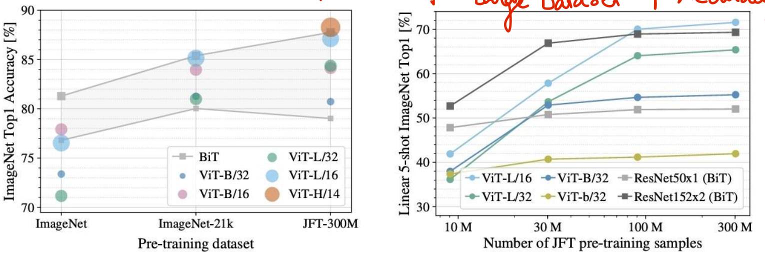
local ↓

affecting feature within a single token / operate independently on each patch

Training : Pretraining was done on ImageNet, ImageNet-21K & JFT 300M

→ Large Model ↑ Accuracy

Large Dataset ↑ Accuracy



B = Base, L = Large, H = Huge /16 → Patch Size

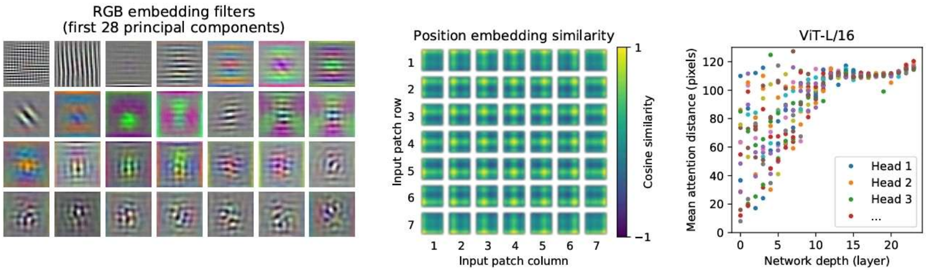


Figure 7: **Left:** Filters of the initial linear embedding of RGB values of ViT-L/32. **Center:** Similarity of position embeddings of ViT-L/32. Tiles show the cosine similarity between the position embedding of the patch with the indicated row and column and the position embeddings of all other patches. **Right:** Size of attended area by head and network depth. Each dot shows the mean attention distance across images for one of 16 heads at one layer. See Appendix D.7 for details.

→ The principle components of the projection layer resemble useful visual patterns like edges or textures.

→ Positional Embeddings are added to each patch to give spatial info about the image. Conc. : Patches closer together & patches in the same row & column have similar embeddings.

→ Attention Distance (Similar to Receptive field in CNN's) says how far each patch reaches out in terms of image distance. Some heads attend to all patch in the early layers itself, Others are more local focussing on nearby patches.

→ The model attends to image regions that are semantically relevant for classificatio