

# Visual Transformers

Sebastian López ·, Santiago Pineda, · Rafael Mejia , · Andrés  
Álvarez

Digital Signal Processing and Control Group - (GCPDS)  
Universidad Nacional de Colombia  
Manizales, Colombia  
July 2023

# Contenido



UNIVERSIDAD  
**NACIONAL**  
DE COLOMBIA

**1** Input Image Processing

**2** Visual Transformer Architecture

# Split the image in Patches

## Input Image Processing

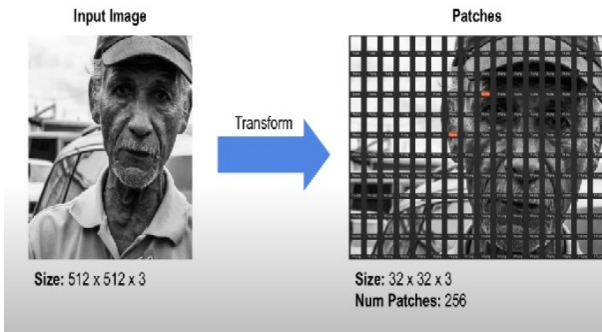


Figure: Split the image into patches

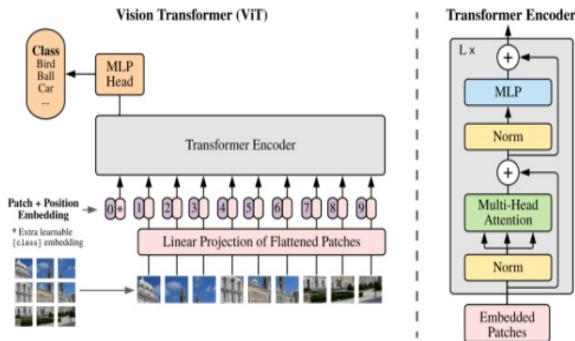
# Split the image in Patches

## Input Image Processing

THEORY	EXAMPLE
<b>Input Image:</b> $H \times W \times C$ .	<b>Input Image:</b> $200 \times 200 \times 3$
<b>Patch Size:</b> $P_h \times P_w$	<b>Patch Size:</b> $25 \times 25$
<b>Number of patches (N):</b> $(H \times W) / (P_h \times P_w)$	<b>Number of patches (N):</b> $= (200 \times 200) / (25 \times 25)$ $= 64$
<b>Transformed Input:</b> $(N, P_h \times P_w \times C)$	<b>Transformed Input:</b> $= (64, 25 \times 25 \times 3)$ $= (256, 1875)$
$H \times W$ = Image height x width	
$C$ = Image channels	
$P_h \times P_w$ = Patch height x width	
$N$ = Number of patches	

Figure: Split the image into patches

# Visual Transformer Architecture



The architecture of the proposed Vision Transformer (ViT)

Figure: Visual Transformer Architecture

# Positional Embedding



Figure: Positional Embedding

# Class Token

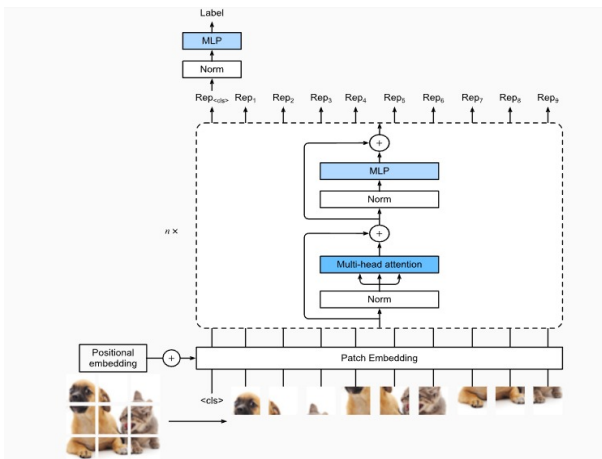
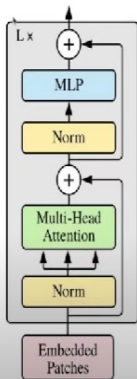


Figure: Class Token

# Transformer Encoder

## Transformer Encoder



**Norm:** Layer Normalization.

**MLP:** Uses GELU activation function.

Figure: Transformer Encoder



# Classics ViTs

Model	Layers	Hidden size $D$	MLP size	Heads	Params
ViT-Base	12	768	3072	12	86M
ViT-Large	24	1024	4096	16	307M
ViT-Huge	32	1280	5120	16	632M

Figure: Classic ViTs



# Thanks!