Aprendizaje profundo

Arquitecturas tipo Transformer para visión

Gibran Fuentes-Pineda

Octubre 2023

Transformers en imágenes: ViT (1)

- Comúnmente la imagen $\mathbf{I}^{(i)}$ de tamaño $H \times W \times C$ se divide en M partes $\mathbf{x}_k^{(i)}, k = 1, \ldots, M$ de tamaño $P \times P \times C$), donde $M = \frac{H \cdot W}{P^2}$.

- Cada parte se aplana y se proyecta con una capa densa o convolucional y se representa con un vector de d dimensiones.

- Los vectores asociados a las partes de la imagen se procesan por una red tipo Transformer directamente

- Se concatena un *embedding* de clase a la secuencia de vectores de cada imagen

Transformers en imágenes: ViT (2)

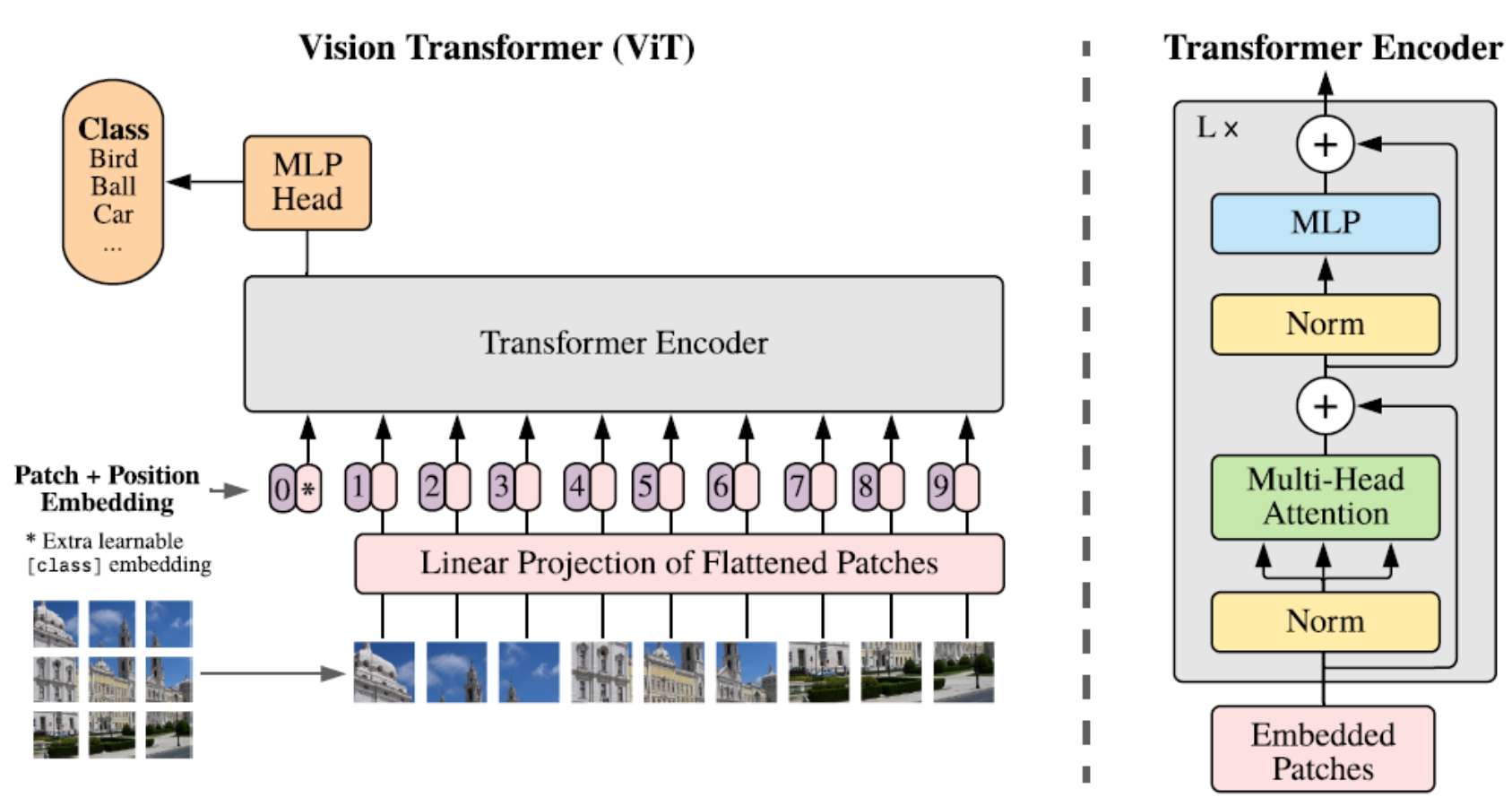

Imagen tomada de Dosovitski et al. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale, 2021.

Transformers en imágenes: Swin (1)

- Arquitectura tipo Transformer con ventanas y partes de distinta granularidad espacial que se mezclan de forma jerárquica

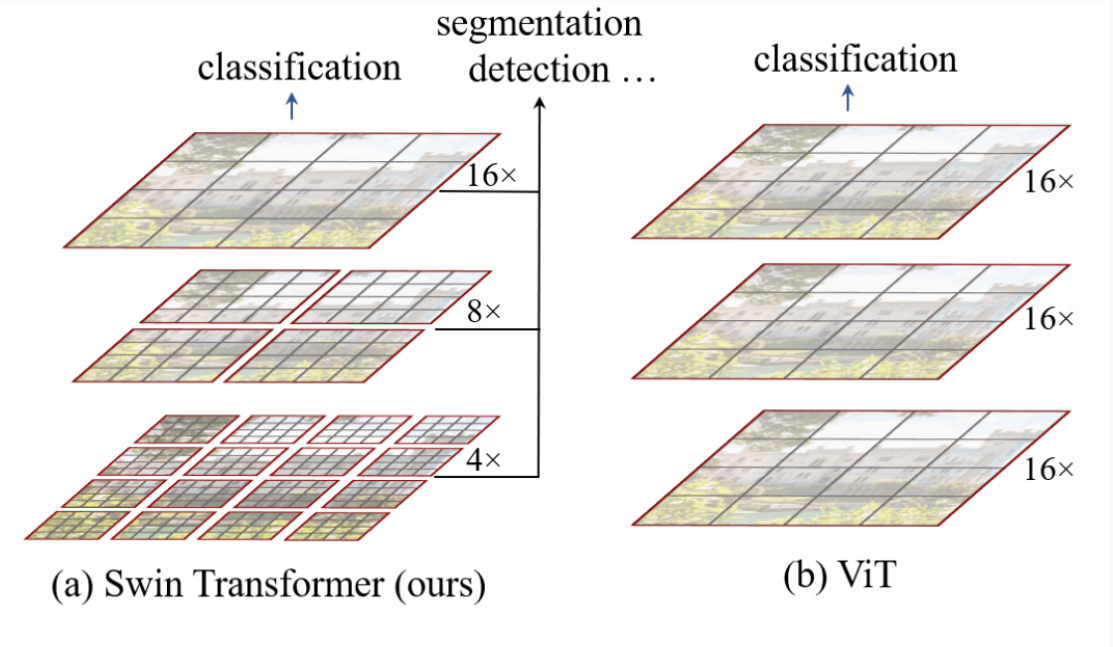

Imagen tomada de Swin Transformer: Hierarchical Vision Transformer using Shifted Windows, 2021.

Transformers en imágenes: Swin (2)

- La operación de autoatención se realiza solo entre las partes de una misma ventana, la cual se puede desplazar o mezclar en capas subsecuentes

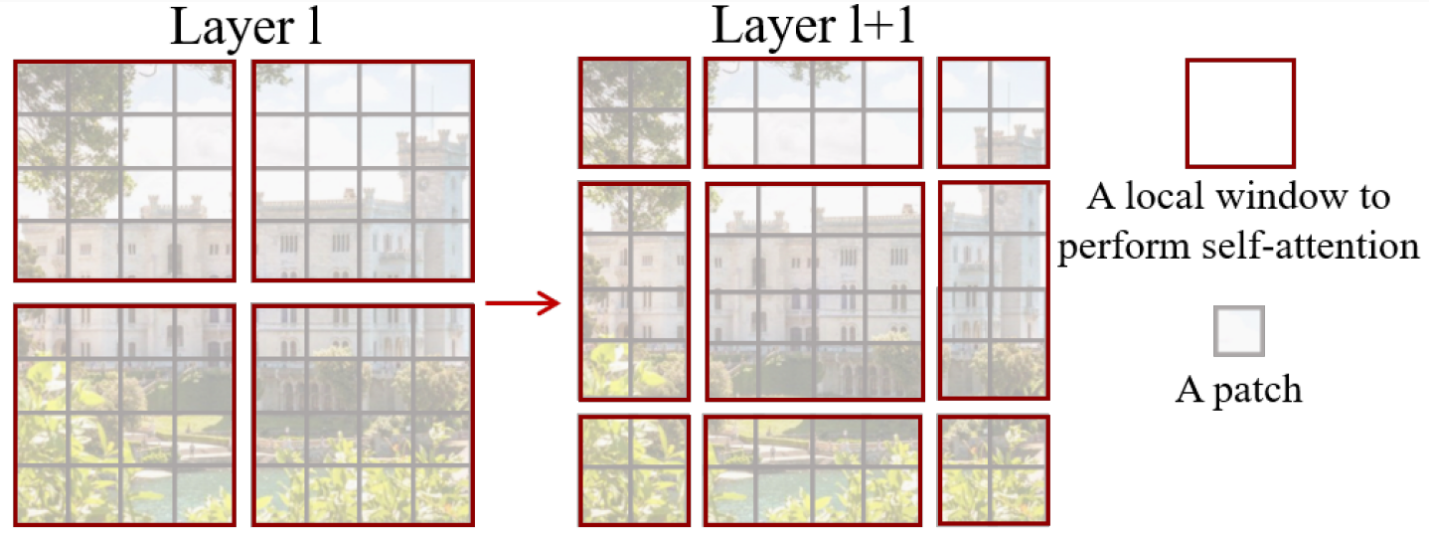

Imagen tomada de Swin Transformer: Hierarchical Vision Transformer using Shifted Windows, 2021.

Transformers en imágenes: Swin (3)

- La operación de autoatención se realiza solo entre las partes de una misma ventana, la cual se puede desplazar o mezclar en capas subsecuentes

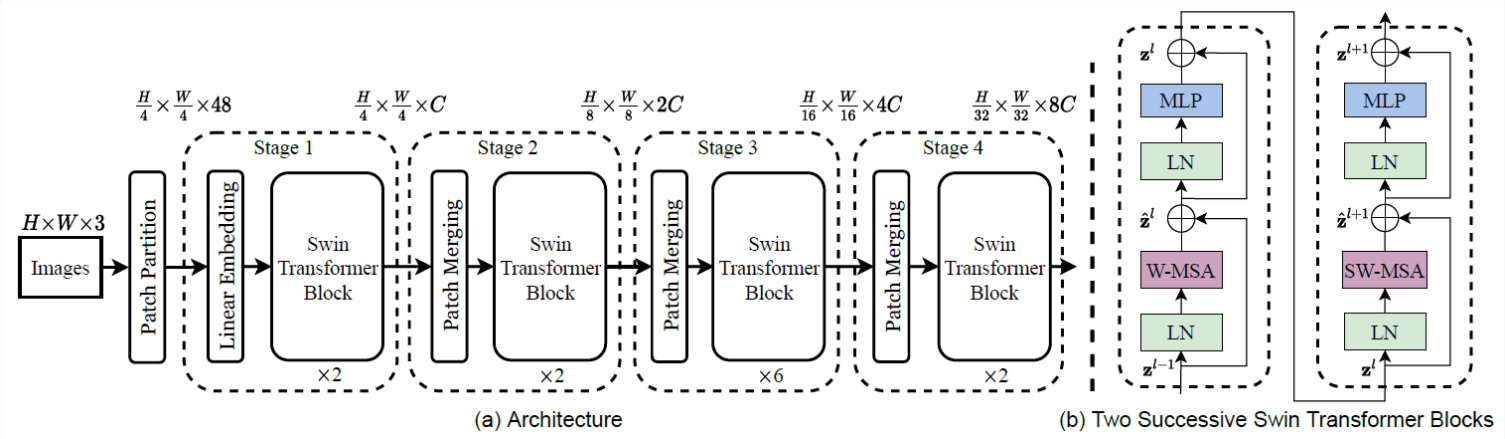

Imagen tomada de Swin Transformer: Hierarchical Vision Transformer using Shifted Windows, 2021.