Métodos de Análisis de Secuencias basados en Aprendizaje Profundo en problemas de Visión y Procesamiento de Imágenes

Estimación de Pose y Clasificación de Imágenes

Óscar Esaú Peralta Rosales[1] Dr. Mariano Rivera Meraz[1]

[1]Centro de Investigación en Matemáticas A.C.

Avance de Tesis

Tabla de Contenido

Centro de Investigación en Matemáticas, A.C.
CIMAT

Table of Contents

Centro de Investigación en Matemáticas, A.C.
CIMAT

Motivación de la Tesis

Con el auge de los Transformes como modelos de procesamiento de información secuencial, el trabajo de esta tesis ha sido dirigido en explorar dichos modelos en áreas fuera del Procesamiento del Lenguaje Natural.

Finalmente se propone una variante enfocado en aumentar la capacidad receptiva de las cabezas de atención permitiendo mayor flexibilidad al no estar ligada al tamaño de embedding predefinidos.

Las experimentaciones del funcionamiento del modelo se realizan en los siguientes problemas:

- Predicción de Pose 2D en humanos sobre imágenes
- Predicción de Pose 3D en humanos (Monocular, Desacoplado)
- ViT y Clasificación de Enfermedades Comunes de Tórax (INAOEP, CIMAT, IMSS)

CIMAT Centro de Investigación en Matemáticas, A.C.

Table of Contents

Centro de Investigación en Matemáticas, A.C.
CIMAT

Estimación de Pose 2D y 3D en Humanos

- 2D: Dada una imágen estimar las posiciones de las articulaciones de la persona en cuestión sobre la imagen.
- 3D: Dada una imagen estimar las posiciones de las articulaciones dentro de un marco de referencia que mejor ajuste la posición espacial de la persona en cuestión.

Figure: Estimación de Pose 2D

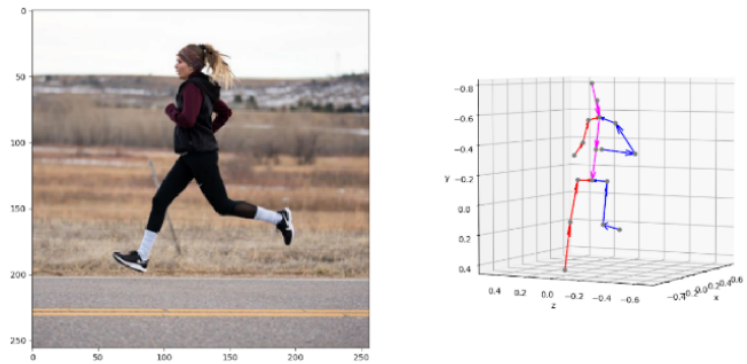

Figure: Estimación de Pose 3D

Centro de Investigación en Matemáticas, A.C.
CIMAT

Estimación de Pose 2D y 3D en Humanos

Esqueleto

- 17 articulaciones
- 16 huesos

Datasets

- Human 3.6M
- COCO 2017

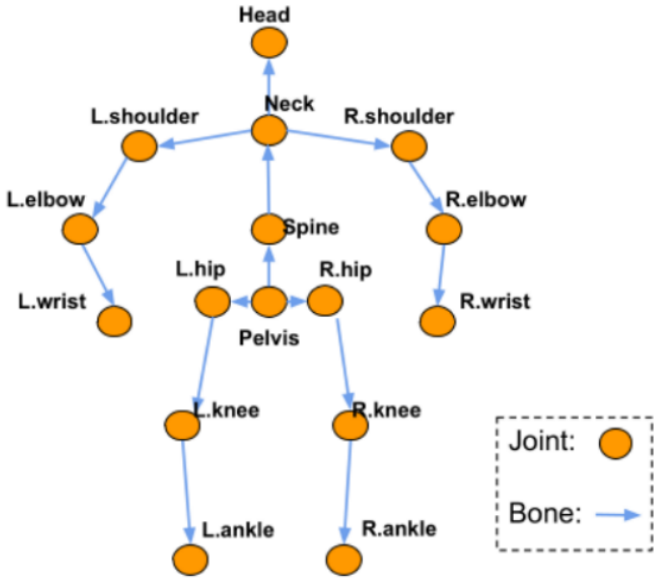

Figure: Articulaciones usadas en tareas de Estimación de Pose en Humanos. Imagen tomada del artículo "Anatomy-Aware 3D Human Pose Estimación en Videos"

CIMAT Centro de Investigación en Matemáticas, A.C.

Detección y Clasificación de Enfermedades Comunes de Tórax

- Trabajo colaborativo entre CIMAT, INAOE e IMSS.
- Modelo clasificador para la detección de 15 padecimientos incluyendo COVID-19.
- Se realiza la comparativa de un modelo basado en ViT con las modificaciones antes mencionadas.

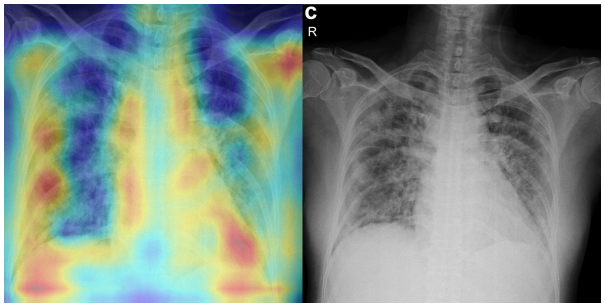

Figure: Áreas Afectadas por COVID-2019 detectadas por el modelo usando GradCam.

Centro de Investigación en Matemáticas, A.C.
CIMAT

Table of Contents

Centro de Investigación en Matemáticas, A.C.
CIMAT

Modelo Estimación de Pose 2D

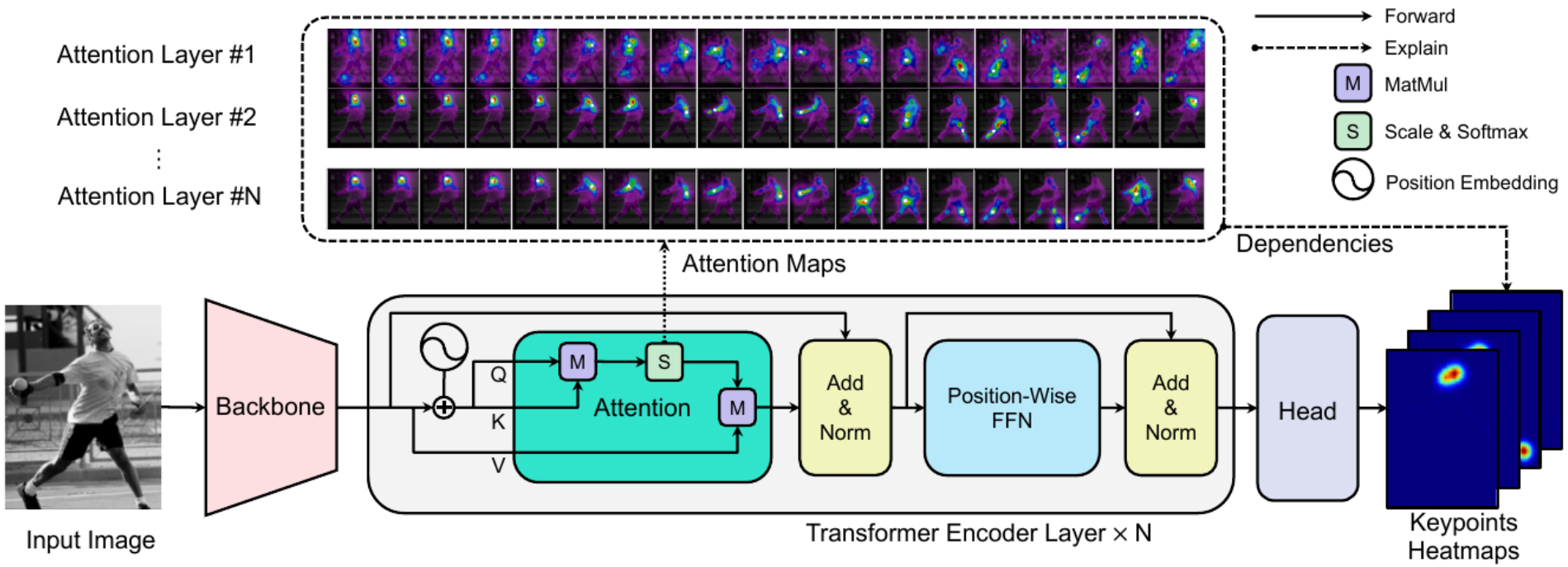

Figure: Modelo Predicción de Pose 2D. Al igual que ViT usa capas con Decoders con entrada las características obtenidas por un modelo convolucional usado como Backbone

Modelo Estimación de Pose 3D

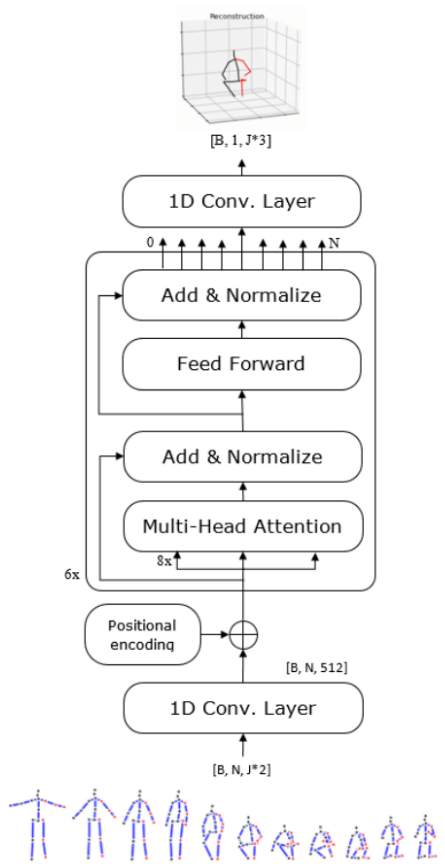

Figure: Modelo Estimación de Pose 3D. Al igual de ViT usa solo capas con Decoders. Las entradas son las estimaciones 2D de algún otro predictor o los GT.

Centro de Investigación en Matemáticas, A.C.
CIMAT

Modelo Estimación de Pose 3D

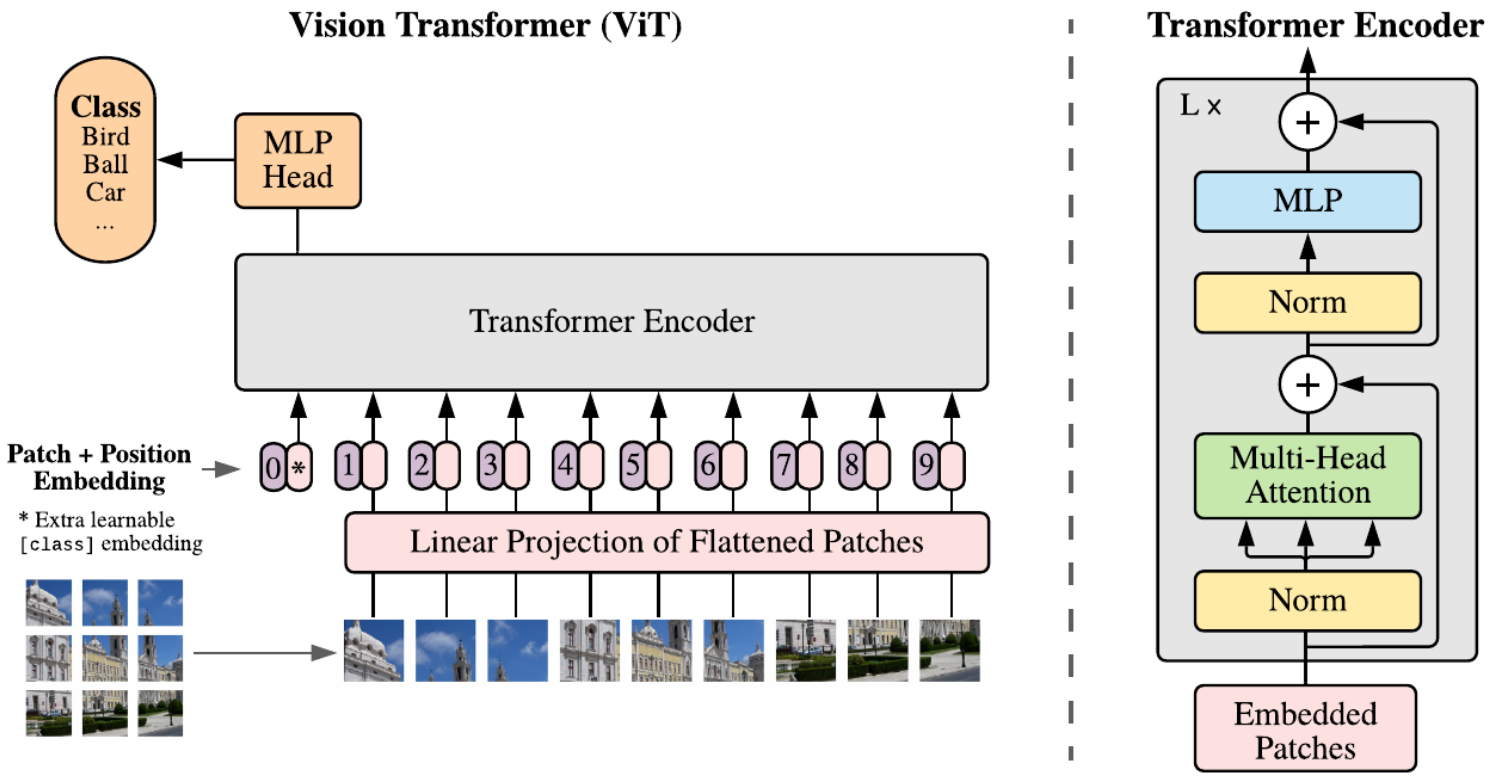

Figure: Modelo ViT usado en las tareas de clasificación de enfermedades pulmonares. La entrada es una secuencia obtenida al dividir la imagen en pequeños parches.

Centro de Investigación en Matemáticas, A.C.
CIMAT

Table of Contents

Centro de Investigación en Matemáticas, A.C.
CIMAT

MultiHead-Self-Attention

El Transformer está basado en la idea de Multihead-Self-Attention (MHA), permitiendo al modelo conjuntamente atender a información en diferentes posiciones desde diferentes subespacios de representación.

$$mha(Q, K, V) = Concat(head_1, head_2, head_3, ..., head_h)W^O$$

donde $Q, K, V \in \mathbb{R}^{nxd_m}$ son embeddings de entrada, n es el tamaño de la secuencia, d_m es el tamaño del embedding y h el número de cabezas de atención. Cada cabeza es definida como sigue:

$$head_i = Attention(QW_i^Q, KW_i^K, VW_i^V) = sofmax\left(\frac{QW_i^Q(KW_i^K)^T}{\sqrt{d_k}}\right)VW_i^V$$

donde W_i^Q, $W_i^K \in \mathbb{R}^{d_m x d_k}$, $W_i^V \in \mathbb{R}^{d_m x d_v}$, $W^O \in \mathbb{R}^{hd_v x d_m}$ y $d_k = d_v = d_m/h$

CIMAT Centro de Investigación en Matemáticas, A.C.

MultiHead-Self-Attention

- El tamaño de la cabeza es dependiente de la dimensión del embedding y el número de cabezas de atención.
- Mientras más cabezas de atención los embeddings son proyectados a dimensiones cada vez más bajas, lo que implica un compresión y perdida de información.
- Escalar el modelo un poco más costoso en memoría y costo computacional.
- Redefiniendo

Centro de Investigación en Matemáticas, A.C.
CIMAT