# Curso de aprendizaje profundo

*PCIC, UNAM*

## Tarea 3: redes basadas en atención

Resuelve el siguientes ejercicio en una libreta que sea replicable.

## 1. Clasificación de rostros por grupo etario mediante ViT

Entrena y evalúa modelos de clasificación de rostros por grupo etario basados en la arquitectura ViT[1], usando el conjunto de datos FairFace[2]. Agrega un bloque ResNet después de la capa convolucional que procesa los parches en ViT, entrena un modelo con la arquitectura modificada y compara su desempeño con ViT. Discute los resultados que obtuviste con las diferentes configuraciones. Para reducir el tiempo de entrenamiento puedes usar resoluciones de imagen más pequeñas, como $64 \times 64$ o incluso $32 \times 32$.

## 2. Reconocimiento de comandos de voz (2 pts. extra)

Entrena una red neuronal basada en bloques Transformer para el reconocimiento de comandos de voz usando el conjunto Speech Commands[3]. Se deberá representar los comandos de voz mediante espectogramas Mel, Mel-Log o MFCCs y procesar estos como una secuencia de vectores columna.

[1]Puedes usar el código visto en `https://github.com/gibranfp/CursoAprendizajeProfundo/blob/2024-1/notebooks/4b_vit_cifar10.ipynb`.

[2]Disponible en `https://github.com/joojs/fairface`

[3]Puedes usar el código para la carga de datos de la libreta `https://github.com/gibranfp/CursoAprendizajeProfundo/blob/2024-1/notebooks/1e_procesamiento_audio.ipynb`