INSTITUTO TECNOLÓGICO Y DE ESTUDIOS SUPERIORES DE MONTERREY

Inteligencia artificial avanzada para la ciencia de datos II

TC3007C - Gpo 501

Deep Learning

Integrantes:

Jorge Ignacio Reyes Pérez - A00573981

Fecha de entrega:
16 de Noviembre del 2025

Introducción

El reconocimiento automático de acciones humanas en video es una de las tareas más relevantes en visión por computadora moderna. Sus aplicaciones abarcan vigilancia inteligente, análisis deportivo, interacción persona-máquina, salud, y sistemas autónomos. En este proyecto se implementó un modelo de *deep learning* capaz de clasificar acciones humanas utilizando el conjunto de datos UCF101, un benchmark ampliamente utilizado que contiene más de 13,000 videos distribuidos en 101 clases diferentes.

A diferencia de los modelos basados en video crudo (RGB), este proyecto utiliza representaciones de esqueletos 2D, extraídas previamente como coordenadas de 17 articulaciones por frame. Este tipo de representación reduce drásticamente el tamaño de los datos y permite trabajar incluso sin GPU. Para este proyecto se utilizó la totalidad de las 101 clases.

El objetivo es demostrar el flujo completo: preprocesamiento, construcción del modelo, entrenamiento, evaluación y predicción sobre nuevos ejemplos.

Estrategia de trabajo

1. Revisión del dataset UCF101 Skeleton y del formato de anotaciones
2. Organización del proyecto en Github, con scripts separados para entrenamiento e inferencia
3. Carga y exploración del dataset
4. Implementación de una arquitectura deep learning LSTM, capaz de procesar secuencias temporales de keypoints
5. Entrenamiento del modelo durante 20 épocas usando el split oficial train1/test1
6. Evaluación del desempeño y comparación contra un baseline aleatorio
7. Aplicación de técnicas de regularización
8. Documentación de resultados y predicciones por inferencia

Pipeline implementado

- Preprocesamiento
    - Normalización de coordenadas
    - Concatenacion
    - Padding hasta 120 frames por secuencia con máscara temporal

- Modelo ActionLSTM
- Entrenamiento
- Evaluacion

Arquitectura del modelo

El modelo implementado es un LSTM bidireccional para capturar dependencias temporales tanto hacia adelante como hacia atrás en la secuencia de movimiento

Se usa un LSTM porque:

- Los keypoints forman secuencias temporales naturales
- Es ligero y eficiente sin GPU
- Ha sido ampliamente utilizado como baseline para reconocimiento de acciones basado en esqueletos

La salida final del LSTM se obtiene concatenando los estados finales forward y backward, lo que mejora el contexto temporal. La red concluye con una capa totalmente conectada que produce logits de tamaño 101

Resultados

Durante 20 épocas de entrenamiento el modelo pasó de un accuracy de 7% en la primera época a 31% en test en la mejor época que fue la época 17. Además el accuracy alcanzó el 50% en el train, mostrando un poco de overfitting, lo cual es esperado para un modelo LSTM sencillo con 101 clases

Métricas	Resultados
Clases utilizadas	101
Accuracy baseline (aleatorio)	0.99%
Mejor accuracy del modelo en test	31%
Mejor accuracy del modelo en train	50%
Número de parámetros	1.7M

El modelo supera ampliamente al baseline aleatorio lo cual valida que aprende patrones reales de movimiento

Inferencia y ejemplos reales

Se probó el modelo con el script “inference.py” obteniendo los siguientes resultados

```
None
Video idx: 0
Ground truth label (entero): 0
Top-5 predicciones:
  1) clase 77  prob=0.309
  2) clase 1  prob=0.275
  3) clase 0  prob=0.092
  4) clase 17  prob=0.073
  5) clase 19  prob=0.068

Video idx: 10
Ground truth label (entero): 0
Top-5 predicciones:
  1) clase 1  prob=0.277
  2) clase 77  prob=0.233
  3) clase 0  prob=0.197
  4) clase 19  prob=0.083
  5) clase 61  prob=0.046

Video idx: 25
Ground truth label (entero): 0
Top-5 predicciones:
  1) clase 1  prob=0.265
  2) clase 77  prob=0.241
  3) clase 0  prob=0.215
  4) clase 19  prob=0.107
  5) clase 61  prob=0.045
```

Aunque la clase correcta aparece solo en top3 esto es razonable dado que es un modelo baseline sencillo sin técnicas avanzadas, además no usa video, sino únicamente keypoints 2D y se debe de tomar en cuenta que se usan las 101 clases totales.

Regularizacion

Dropout 0.3 en las capas LSTM
Bidireccional LSTM
Padding y masking
Normalizacion simple de los keypoints

Conclusiones

En este proyecto se implementó un sistema funcional de clasificación de acciones humanas basado en coordenadas 2D de esqueletos. A pesar de no utilizar los datos de video completos, el modelo logró obtener un desempeño significativamente superior al baseline aleatorio, con una accuracy de aproximadamente 31% en el conjunto de prueba.

Este proyecto demuestra:

- La capacidad de los LSTM para procesar secuencias temporales de poses.
- La viabilidad de trabajar con keypoints en lugar de video completo.
- La eficiencia de PyTorch y modelos ligeros en entornos sin GPU.

Trabajo a futuro

Para mejorar el desempeño se podrían implementar otros modelos como Transformers, usar GCN para poder modelar la estructura del esqueleto, además de aplicar data augmentation y un fine-tuning específico por acción