Reconocimiento de acciones en video usando Inteligencia Artificial

Edgar Rangel

Visión por Computador

Universidad Industrial de Santander

Abr 01, 2020

BIVLab

Biomedical Imaging, Vision and Learning Laboratory

MACV

Motion Analysis and Computer Vision

Universidad Industrial de Santander

Reconocimiento de acciones

- El reconocimiento acciones humana basado en la visión por computador es el proceso de etiquetar secuencias de imágenes (video) con su respectiva acción.
- Una acción se puede ver como una secuencia de movimientos del cuerpo humano y puede envolver varias partes a la vez.
- **Su aplicabilidad es general, desde videovigilancia, análisis deportivo, indexación de videos, etc.**

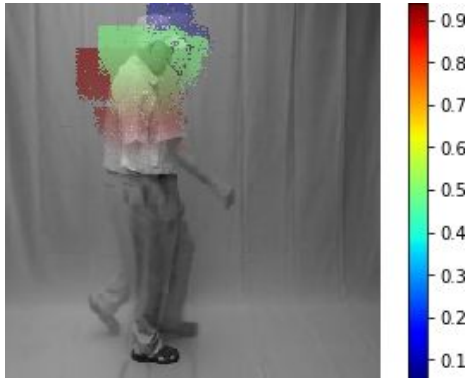
(a) c_walk (b) c_jump (c) side step (d) walk (e) jump

Motivación

El reconocimiento de acciones es usado en varios campos, como:

a. Traducción de lenguaje de señas.
b. Detección de enfermedades.
c. Videovigilancia.

(a)

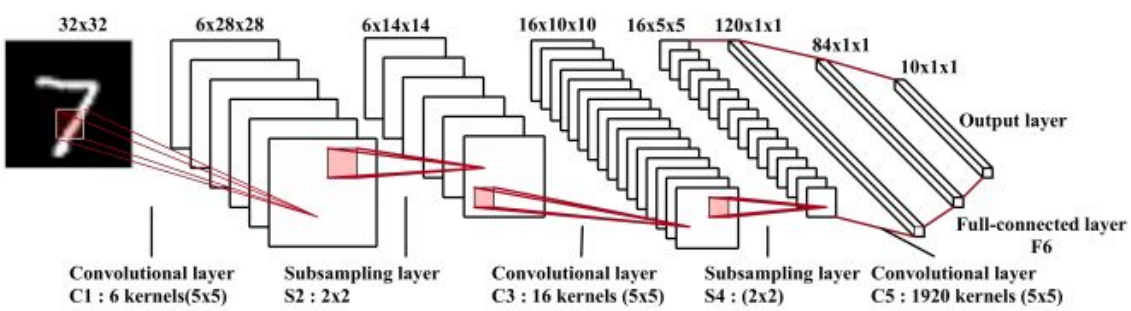


(b)

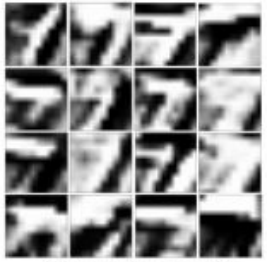
(c)

Redes neuronales convolucionales

- Están compuestas de múltiples capas, generalmente convoluciones, pooling y fully connected.
- Desde su aparición en 1990 por Yann LeCunn, han demostrado tener un buen desempeño para problemas de visión por computador.
- **Obtienen buenas representaciones de los datos originales para su clasificación.**



Convolutional layer C1
6x28x28



Convolutional layer C3
16x10x10

Problema

- La falta de grandes datos de entrenamiento limitan el desempeño de estas arquitecturas.
- Se han explorado ampliamente para problemas con imagenes (convoluciones en 2 dimensiones) pero relativamente poco con videos (convoluciones en 3 dimensiones).

Problema

- El modelamiento de los datos (vídeos) afectan en gran manera el aprendizaje del modelo. (Susceptibles de ruido).
- La complejidad de las mismas acciones al ser compuestas o repetitivas dificultan la tareas de aprender acciones.

Objetivos

Objetivos

- Seleccionar un conjunto de datos académicos y públicos relacionados con el reconocimiento de acciones,
- Proponer un modelo de aprendizaje profundo sobre volúmenes de datos para codificar relaciones espacio-temporales.
- Clasificar las acciones dadas por los conjuntos de datos académicos seleccionados aplicando los métodos vistos en clase.

Datasets

UT

- Dataset que contiene 60 vídeos de 6 diferentes acciones:
  - Apretón de manos
  - Apuntar
  - Abrazar
  - Empujar
  - Patear
  - Golpear
- Cada video contiene una ejecución por interacción, con resolución de 720x480 grabados a 30 fps y la altura de la persona es de 200 pixeles.

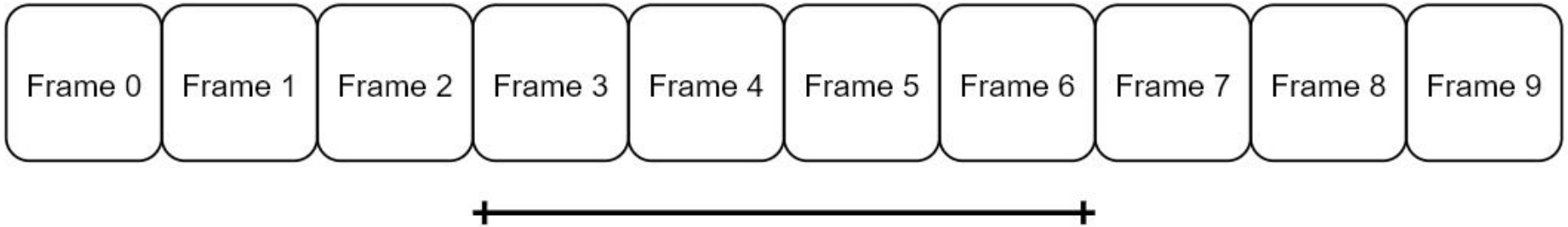
JHMDB

- Dataset que contiene 928 vídeos con 21 diferentes acciones provenientes del dataset HMDB con 51 acciones.
- Cada video fue tomado teniendo en cuenta acciones que no incluyeran movimientos faciales (sonreir), interacciones con otros (apretón de manos) y acciones que solo se puede realizar de una manera (voltereta lateral), además, no se dejan vídeos donde el actor no es obvio. Por último el primer y último frame del video corresponden respectivamente al inicio y final de la acción.

Aumento de Datos

Recortes temporales

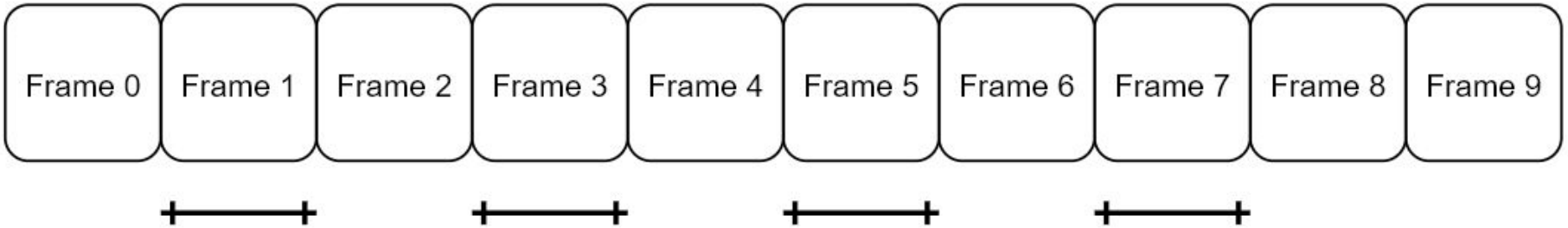
- Recorte centrado único.
- Recorte espaciado lineal único.
- Recorte espaciado lineal múltiple.



Aumento de Datos

Recortes temporales

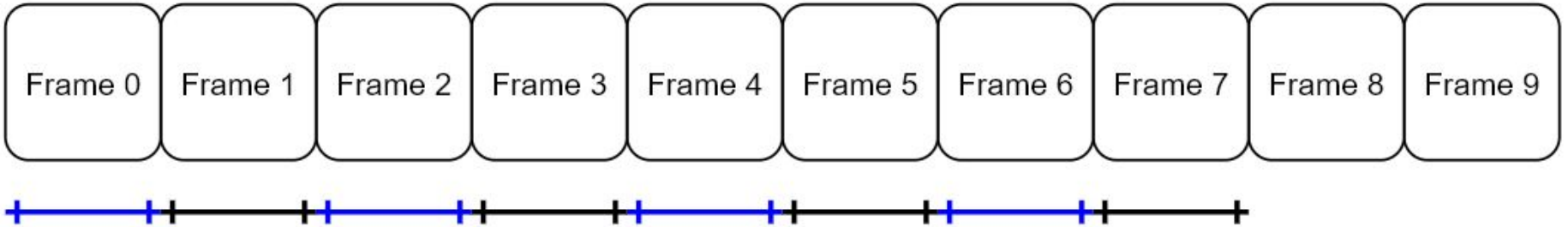
- Recorte centrado único.
- Recorte espaciado lineal único.
- Recorte espaciado lineal múltiple.



Aumento de Datos

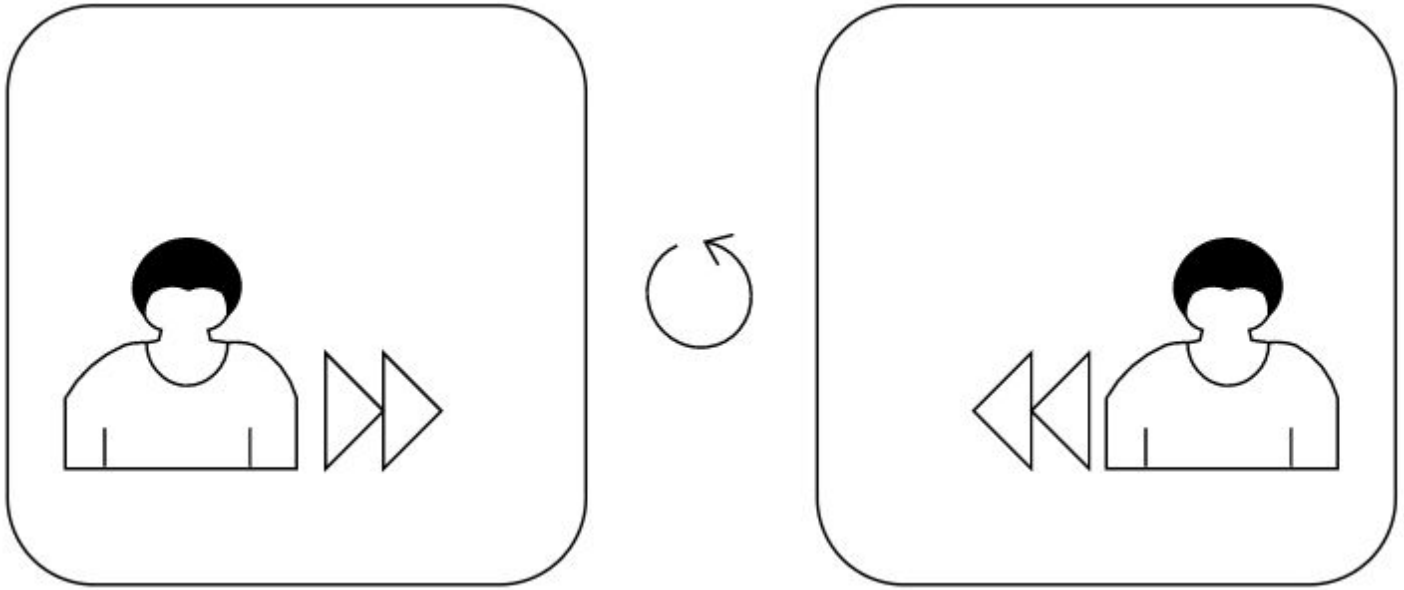
Recortes temporales

- Recorte centrado único.
- Recorte espaciado lineal único.
- Recorte espaciado lineal múltiple.



Aumento de Datos

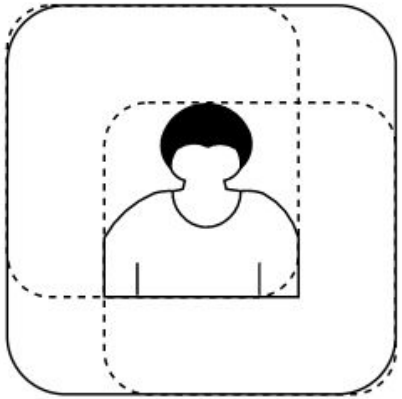
Rotación horizontal del video

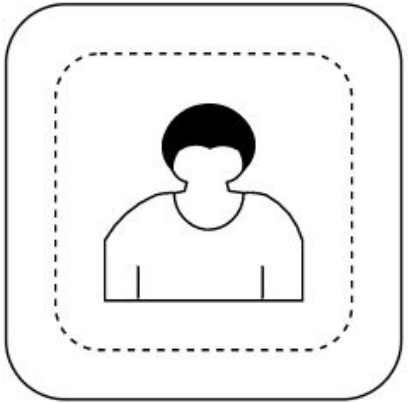


Aumento de Datos

Recortes espaciales

- Cortes en esquinas.
- Cortes centrados en el sujeto.

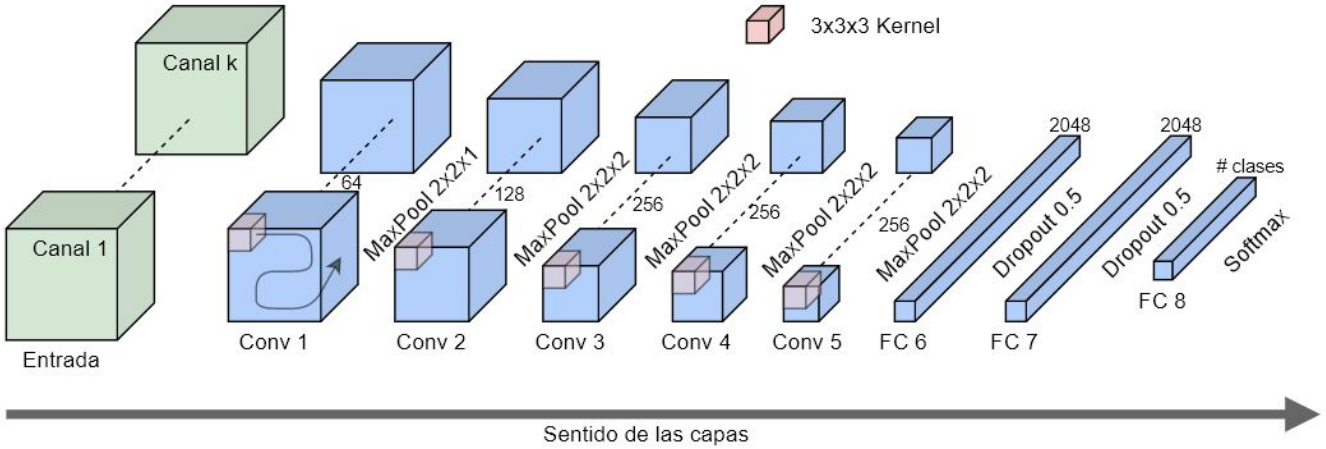




Trabajo realizado

Arquitectura convolucional base

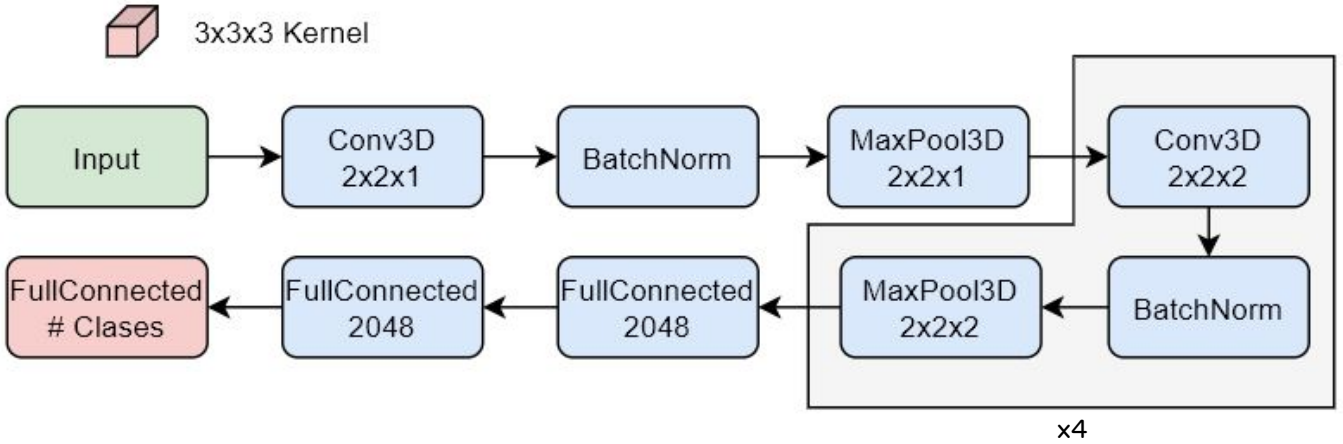
- LTC (Long-term temporal convolutions for action recognition)
- Compuesta de 5 convoluciones 3D y 3 capas fully connected para clasificar las acciones en los videos.
- Transfer learning con la red C3D y pesos de redes convolucionales 2D.



Modulo BatchNormalization

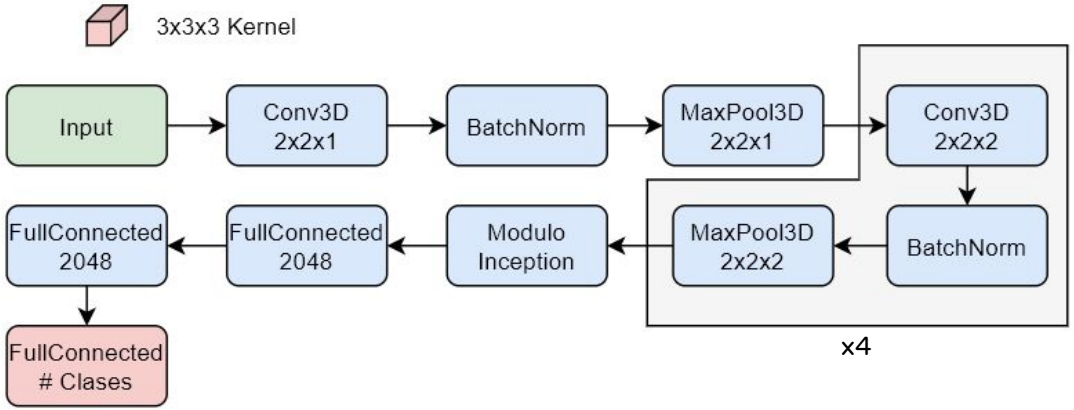
Esta capa corresponde a **normalizar** su entrada y realizar una transformación lineal sobre los datos para ayudar a evitar el desvanecimiento o explosión del gradiente.

Esta capa es entrenable, lo cual permite ir variando sus parámetros para normalizar mejor.



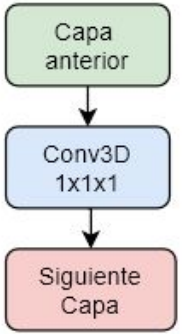
Módulos de la Inception

Estos módulos son usados en redes convolucionales para permitir una mayor eficiencia computacional y reducir la dimensionalidad de los canales con convoluciones 1x1. Estos módulos fueron diseñados para resolver el problema del costo computacional, el sobre aprendizaje, y reducción de dimensionalidad coherente.

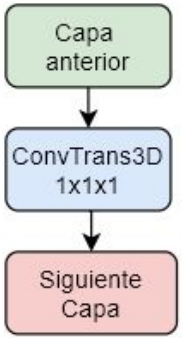


Módulos de la Inception (Contribución)

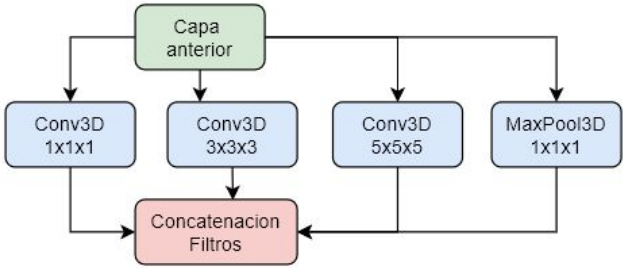
a. Módulo simple
b. Módulo con Convolución Transpuesta
c. Módulo Naive de Inception
d. Módulo Mejorado de Inception



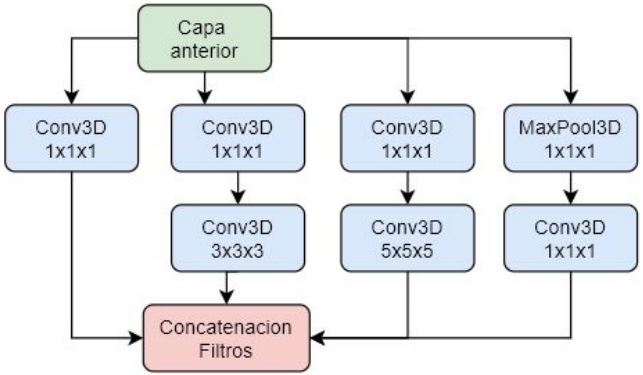
(a)



(b)



(c)

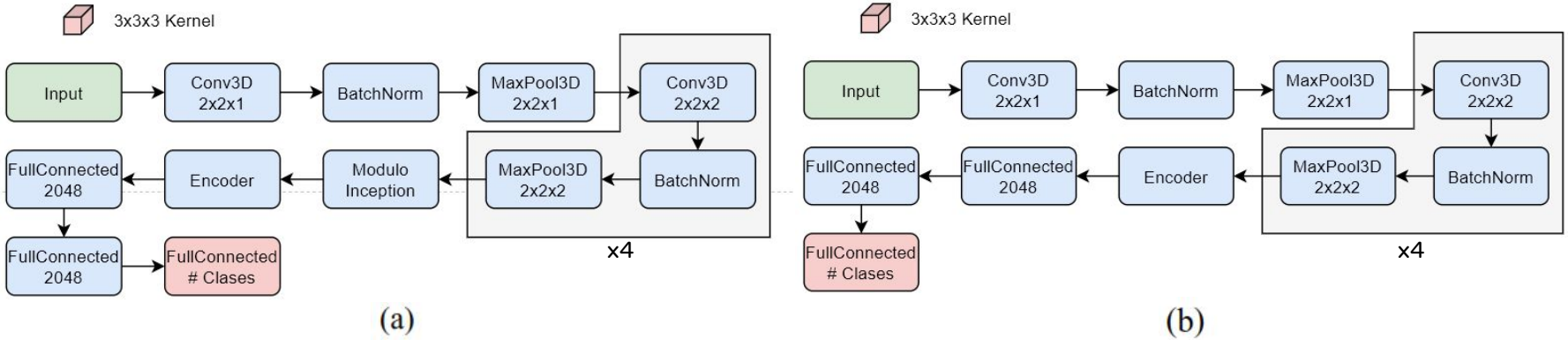


(d)

Encoder - Decoder

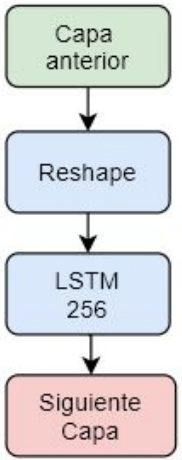
Estas redes consiste en su totalidad de 2 redes, una red neuronal que actúa como **codificador** y toma la entrada cruda para procesar unas características. El **decodificador** toma como entrada estas características y nos retorna una salida. Generalmente este tipo de redes siempre viene acompañado con **LSTM**.

Con estos modulos se busca modelar el eje temporal al final de las convoluciones.

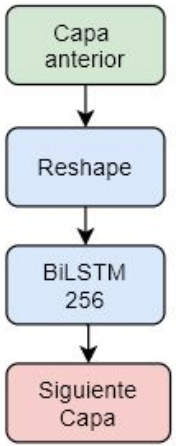


Encoder - Decoder (Contribución)

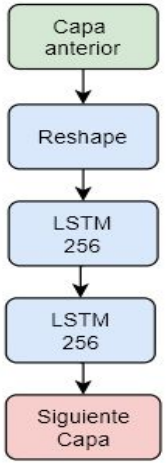
a. Módulo con LSTM de 256 neuronas
b. Módulo con una LSTM Bi Direccional de 256 neuronas
c. 2 Módulos de LSTM concatenados de 256 neuronas



(a)



(b)



(c)

Resultados

Resultados UT

Metodología de experimentos:

- 1 epoch.
- Decrecimiento controlado del learning rate con el optimizador del descenso del gradiente estocástico.
- Aumento de datos,inserción de módulos en la red y transfer learning.

Tipo Experimento	Train Accuracy	Test Accuracy
Línea base con espaciado temporal lineal múltiple	0,18	0,21
Módulo de BatchNormalization e Inception tipo D, junto a VideoFlip, Recorte espacial centrado y espaciado temporal múltiple	**0,67**	**0,75**
Módulo de BatchNormalization, Inception tipo D y Encoder tipo A, junto a VideoFlip, Recorte espacial centrado y espaciado temporal múltiple	0,63	0,64

Resultados JHMDB

Metodología de experimentos:

- 10 epoch.
- Decrecimiento controlado del learning rate con el optimizador del descenso del gradiente estocástico.
- Aumento de datos, inserción de módulos en la red y transfer learning.

Tipo Experimento	Train Accuracy	Test Accuracy
Línea base con espaciado temporal lineal múltiple y VideoFlip	0,69	0,19
Módulo de BatchNormalization e Inception tipo D, junto a VideoFlip, Recorte espacial centrado y espaciado temporal múltiple	**0,99**	**0,36**
Módulo de BatchNormalization, Inception tipo D y Encoder tipo C, junto a VideoFlip, Recorte espacial centrado y espaciado temporal múltiple	0,99	0,29

Conclusiones

Conclusiones

- Se seleccionó dos conjuntos de datos con características peculiares para comprobar el aprendizaje del modelo en diferentes escenarios.
- Se propuso un modelo de aprendizaje profundo basado en los conocimientos de visión por computador, usando nuevos módulos, transfer learning y modelamiento de datos.
- Se clasificaron los conjuntos de datos obteniendo como Score 75% y 36% respectivamente en UT y JHMDB.

¡Muchas gracias por su atencion!
¿Dudas?