PONTIFICIA UNIVERSIDAD CATÓLICA DEL PERÚ

Escuela de Posgrado

Clasificación automática de eventos en videos de fútbol utilizando Redes Convolucionales Profundas

Tesis para obtener el grado académico de Maestro en Informática con mención en Ciencias de la Computación que presenta:

Alipio Laboriano Galindo

Asesor:

Dr. César Armando Beltrán Castañón

Lima, 2024

Informe de Similitud

Yo, César Armando BELTRÁN CASTAÑÓN, docente de la Escuela de Posgrado de la Pontificia Universidad Católica del Perú, asesor de la tesis titulado *Clasificación automática de eventos en videos de fútbol utilizando Redes Convolucionales Profundas* de el autor Alipio LABORIANO GALINDO, dejo constancia de lo siguiente:

- El mencionado documento tiene un índice de puntuación de similitud de 11%. Así lo consigna el reporte de similitud emitido por el software *Turnitin* el 15/04/2024.
- He revisado con detalle dicho reporte y la tesis, y no se advierte indicios de plagio.
- Las citas a otros autores y sus respectivas referencias cumplen con las pautas académicas.

Lugar y fecha:

San Miguel, 15 de Abril de 2024.

Apellidos y nombres del asesor / de la asesora: César Armando Beltrán Castañón	
DNI: 29561260	Firma 
ORCID: 0000-0002-0173-4140	

DEDICATORIA

A mis padres don Javier Laboriano y doña Marta Galindo, y a mi hermano mayor William Laboriano Galindo, que siempre han sido los pilares fundamentales para mi formación académica y personal.

AGRADECIMIENTOS

Al Dr. César Armando Beltrán Castañón, por haber aceptado el reto y compartir su experiencia y dedicación para lograr desarrollar este trabajo.

Al grupo de Inteligencia Artificial de la Pontificia Universidad Católica del Perú (IAPUCP), por haberme facilitado la infraestructura de su laboratorio de Inteligencia Artificial para desarrollar la experimentación durante el desarrollo de la presente tesis.

RESUMEN

La forma en que las nuevas generaciones consumen y experimentan el deporte especialmente el fútbol, ha generado oportunidades significativas en la difusión de contenidos deportivos en plataformas no tradicionales y en formatos más reducidos. Sin embargo, recuperar información con contenido semántico de eventos deportivos presentados en formato de video no es tarea sencilla y plantea diversos retos. En videos de partidos de fútbol entre otros retos tenemos: las posiciones de las cámaras de grabación, la superposición de eventos o jugadas y la ingente cantidad de fotogramas disponibles.

Para generar resúmenes de calidad y que sean interesantes para el aficionado, en esta investigación se desarrolló un sistema basado en Redes Convolucionales Profundas para clasificar automáticamente eventos o jugadas que ocurren durante un partido de fútbol.

Para ello se construyó una base de datos a partir de videos de fútbol descargados de SoccerNet, la cual contiene 1,959 videoclips de 5 eventos: saques de meta, tiros de esquina, faltas cometidas, tiros libres indirectos y remates al arco.

Para la experimentación se utilizó técnicas de preprocesamiento de video, una arquitectura convolucional propia y se aplicó transfer learning con modelos como ResNet50, EfficientNetb0, Visión Transformers y Video Visión Transformers.

El mejor resultado se obtuvo con una EfficentNetb0 modificada en su primera capa convolucional, con la cual se obtuvo un 91% accuracy, y una precisión de 100% para los saques de meta, 92% para los tiros de esquina, 90%

para las faltas cometidas, 88% para los tiros libres indirectos y 89% para los remates al arco.

Palabras claves: Clasificación de eventos en fútbol, Redes Convolucionales Profundas, Transformers, SoccerNet.

ABSTRACT

The way the new generations consume and experiment sports, especially soccer, has generated significant opportunities in the dissemination of sports content on non-traditional platforms and in smaller formats. However, retrieving information with semantic content of sporting events presented in video format is not an easy task and poses several challenges. In videos of soccer matches, among other challenges we have: the positions of the recording cameras, the overlapping of events or plays and the huge amount of frames available.

In order to generate quality summaries that are interesting for the fan, this research developed a system based on Deep Convolutional Networks to automatically classify events or plays that occur during a soccer match.

For this purpose, a database was built from soccer videos downloaded from SoccerNet, which contains 1,959 video clips of 5 events: goal kicks, corner kicks, fouls, indirect free kicks and shots on target.

For the experimentation, video preprocessing techniques were used, a proprietary convolutional architecture and transfer learning was applied with models such as ResNet50, EfficientNetb0, Vision Transformers and Video Vision Transformers.

The best result was obtained with a modified EfficentNetb0 in its first convolutional layer, with which 91% accuracy was obtained, and an accuracy of 100% for goal kicks, 92% for corner kicks, 90% for fouls committed, 88% for indirect free kicks and 89% for shots on target.

Key words: Football Event classification, Deep Convolucional Network, Transformers, SoccerNet.

CONTENIDO

ÍNDICE DE FIGURAS

ÍNDICE DE TABLAS

1. INTRODUCCIÓN

1.1. Problemática

Las nuevas generaciones viven el deporte de manera diferente, combinando la experiencia con dispositivos móviles, estadísticas y la interacción en plataformas como TikTok, Twitch, entre otras. Un estudio de Nielsen Sports (Empresa dedicada al análisis de patrocinio e inteligencia de fans), resalta las oportunidades en difusión de contenidos deportivos, especialmente entre generaciones jóvenes, donde aproximadamente el 44% de los aficionados de 16 a 29 años consumen contenido deportivo en canales digitales (García, 2022). Otro estudio de You First (Agencia global de deportes y entretenimiento creada para asesorar talentos, marcas y titulares de derechos deportivos), encontró que estos canales digitales también influyen en cómo se consume el fútbol y otros deportes, pues el 59% de la Generación Z (nacidos en el nuevo milenio) sigue los partidos de fútbol a través de resúmenes en redes sociales y muestra interés en formatos no tradicionales, como podcasts o documentales deportivos (ReasonWhy, 2023).

Esta nueva forma de vivir el deporte y especialmente el fútbol, supone un gran desafío al momento de identificar y recuperar los eventos más importantes que suceden durante un partido y que queda grabado en formato de video. En este tipo de videos se presentan diversas situaciones y desafíos que deben abordarse, como las variadas posiciones de las cámaras, las diferentes condiciones de iluminación, la dispersión o superposición de los eventos a lo largo del video, la duración no homogénea de los eventos (una falta cometida no dura lo mismo que

un tiro de esquina o un tiro libre) y la gran cantidad de frames o imágenes que se deben procesar.

Debido a la complejidad para obtener resúmenes de eventos deportivos en especial el fútbol, y a la creciente demanda durante los últimos años, en esta investigación se propone desarrollar un sistema basado en Redes Convolucionales Profundas que permita clasificar automáticamente cinco eventos que puedan suceder durante un partido de fútbol. Este sistema incluye procesos que van desde la generación o creación de un conjunto de datos de videoclips que contengan eventos como foul (falta), shots on target (remate al arco), corner (tiro de esquina), indirect free-kick (tiro libre indirecto) y clearance (despeje); hasta el entrenamiento mediante arquitecturas convolucionales implementadas desde cero y sobre todo utilizando transfer learning de modelos basados en Redes Convolucionales Profundas y Transformers.

Los resultados obtenidos con el sistema desarrollado pueden tener diversas aplicaciones, además de la creación de resúmenes con menos participación humana y mayor rapidez, también pueden utilizarse para recopilar datos que ayuden a los ojeadores de talento a tomar decisiones más informadas al hacer seguimiento de nuevas promesas del fútbol

1.2. Objetivos

1.2.1. Objetivo general

Clasificar automáticamente eventos en videos de partidos de fútbol, utilizando Redes Convolucionales Profundas.

1.2.2. Objetivos específicos

OE1. Construir un conjunto de datos de videoclips a partir de SocceNet, que se pueda utilizar para la clasificación automática de eventos en un partido de fútbol.

OE2. Utilizar diferentes Redes Convolucionales Profundas para clasificar eventos en un partido de fútbol.

OE3. Validar el rendimiento de los diferentes modelos con los que se experimentó.

OE4. Desplegar a nivel de prototipo el modelo que mejor desempeño tenga en la clasificación automática de eventos en partidos de fútbol.

1.3. Resultados esperados

RE1. Obtener como resultado una base de datos de videoclips con 5 tipo de eventos o jugadas de fútbol a clasificar (Foul = Falta, Shots on target = Remate al arco, Corner = Tiro de esquina, Indirect free-kick = Tiro libre indirecto y Clearance - Despeje) con sus respectivas etiquetas, a partir del cual se realiza la experimentación.

RE2. Obtener como resultado final un algoritmo funcional que permita clasificar automáticamente los eventos presentes en un video de fútbol.

RE3. Obtener y mostrar los resultados comparativos de los experimentos realizados (utilizando la métrica de rendimiento seleccionado), a partir de los cuales se seleccionó el modelo final que será implementado.

RE4. Mostrar a nivel de prototipo, el funcionamiento del modelo que mejor desempeño tenga en la fase de experimentación.

1.4. Métodos y procedimientos

La clasificación en videos es un poco diferente a la clasificación en imágenes, en esta tarea, se trabaja con una dimensión adicional: el Tiempo (fotogramas), por lo tanto, para que una Red Convolucional Profunda sea adecuada para la clasificación de videos, se requiere una modificación en las arquitecturas tradicionales, es ahí donde surgen enfoques como Early Fusion, Late Fusion, Slow Fusion, que permiten fusionar y ajustar la información antes que pase por la arquitectura de una red tradicional (Karpathy et al., 2014)

En una red convolucional tradicional, la primera capa convolucional tiene una entrada de tres canales, que representan los tres canales de color de la imagen, mientras que, con data de videos, la primera capa maneja 3 canales de color simultáneamente en cada frame. En consecuencia, en cada lote se tendrá como dimensión de entrada un tensor de dimensiones [B- tamaño del lote, F- número de fotogramas, C- número de canales, H- altura, W- ancho] [B, F, 3, H, W], el cual precisamente no es un tensor con las dimensiones requeridas por las arquitecturas entrenadas en ImageNet, que utilizan bloques convolucionales 2D. Además, como todos los fotogramas deben analizarse

simultáneamente, se requiere una remodelación, por lo que la primera capa de convolución se realizará en todos los marcos de la arquitectura, por lo tanto, las nuevas dimensiones del tamaño del lote son [B, 3F, H, W] (Karpathy et al., 2014).

Tomando en cuenta lo anterior, la solución propuesta se divide en cinco partes: En la primera se descarga los videos etiquetados de SoccerNet, en la segunda se extraen clips de 2 segundos de estos videos, en la tercera se obtienen los frames con los cuales se van entrenar los modelos, en la cuarta se entrenan los modelos utilizando arquitecturas basadas en redes convolucionales profundas y Transformers y en la quinta parte se obtienen los resultados que predice el evento que ocurre en un videoclip de dos segundos de duración dado (Ver Figura 1). Para construir la Figura 1 se utilizaron imágenes de (Karpathy et al., 2014), (Adhikari et al., 2020), (Andrade, 2021)

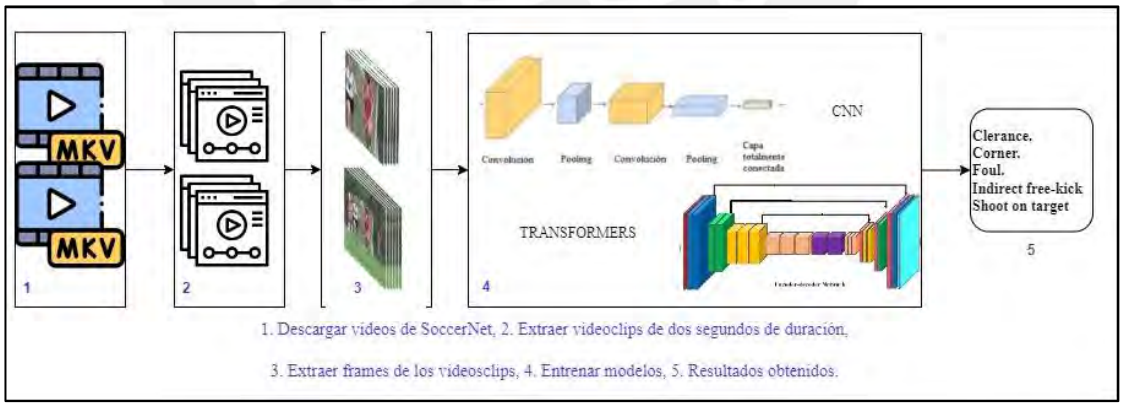

Figura 1: Diseño detallado de la solución propuesta

1.5. Alcances y limitaciones

El presente proyecto de investigación se centra principalmente en construir un conjunto de datos que contenga videoclips de 5 jugadas de

fútbol, para lo cual se utiliza la información disponible en SoccerNet, el cual es un conjunto de datos que contiene la información de 550 partidos de fútbol etiquetados; para luego realizar el entrenamiento mediante Redes Neuronales Profundas sobre este conjunto de datos, con el fin de obtener un algoritmo con el cual se pueda detectar automáticamente cada uno de estos 5 tipos de eventos; para luego implementar un prototipo piloto, que tendrá como entrada un videoclip de cualquiera de los 5 eventos considerados y devolverá como resultado el evento que está presente en el videoclip dado.

SoccerNet es una base de datos que contiene 550 partidos de fútbol con 17 eventos etiquetados, debido a las limitaciones en la capacidad de hardware que se tiene, procesar toda esta cantidad de información no es factible, por lo que en este proyecto se utilizará 53 partidos de fútbol (106 videos descargados) y se trabajará con cinco jugadas de fútbol, las cuales fueron seleccionadas tomando en cuenta la frecuencia de aparición en la mayoría de reportes estadísticos actuales de los partidos de fútbol, ya sea en televisión, web e incluso redes sociales.

2. MARCO CONCEPTUAL

En este capítulo se aborda de manera general los principales conceptos teóricos que sirven como base para el desarrollo de la presente investigación.

2.1. Redes Neuronales Convolucionales

Las redes convolucionales, también conocidas como Redes Neuronales Convolucionales o CNN por sus siglas en inglés, son un tipo especializado de red neuronal para procesar datos que tienen una estructura cuadriculada o de grilla conocida (Lecun Y. , 1989). Algunos ejemplos son los datos de series temporales, que pueden considerarse como una cuadrícula 1-D que toma muestras a intervalos de tiempo regulares, los datos de imágenes, que pueden considerarse como una cuadrícula 2-D de píxeles (Goodfellow et al., 2016) o fotogramas de video que pueden representarse como cuadriculas 3-D (Karpathy et al., 2014).

El nombre "Red Neuronal Convolucional" indica que la red emplea una operación matemática llamada convolución. La convolución es un tipo especializado de operación lineal. Las redes convolucionales son simplemente redes neuronales que utilizan la convolución en lugar de la multiplicación matricial general en al menos una de sus capas (Goodfellow et al., 2016).

En su forma más general, la convolución es una operación sobre dos funciones de un argumento de valor real, la cual puede definirse como la integral que expresa la cantidad de solapamiento de una función g, a medida que se desplaza sobre otra función f, es decir, "combina" una función con otra (Weisstein, 2002).

Finalmente, la arquitectura típica de una CNN está conformada por un conjunto de capas de convolución, capas de reducción o de submuestreo (Pooling) y una o más capas completamente conectadas (Fullyconected), tal como se muestra en la Figura 1 (Lecun et al., 1998).

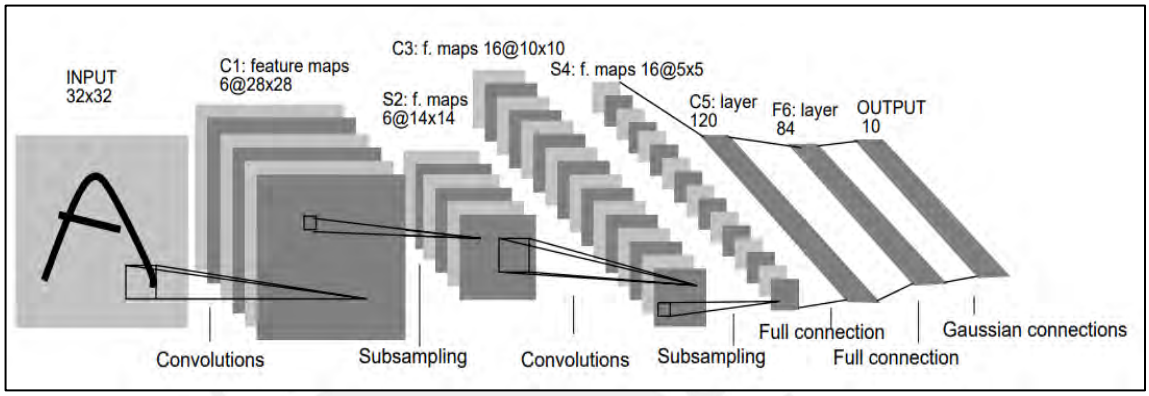

Figura 2: Arquitectura LeNet 5 (Lecun et al., 1998)

2.2. Clasificación en video utilizando Redes Neuronales Convolucionales

En los datos de entrada de videos se tiene una característica que de por sí no están presentes en las imágenes: el tiempo. Es por ello que esta característica puede ser de interés al momento de definir los modelos a trabajar. Karpathy en el 2014 propone cuatro métodos que se pueden utilizar cuando se trabaja con video (Karpathy et al., 2014).

- Frames individuales. Propone trabajar con cada uno de los frames del video, lo cual permite analizar la capacidad de la red para clasificar el video en base a información estática.
- Fusión tardía. Propone trabajar solamente con los frames inicial y final de un clip del video, para los cuales se utilizaría una CNN individual. Finalmente, estas dos CNN se unen en una capa totalmente conectada.

- Fusión temprana. Propone trabajar con una secuencia de frames consecutivos, los cuales son fusionados para entrar a una CNN, esto permitiría conocer la dirección y velocidad del movimiento.
- Fusión lenta. La propuesta es una mezcla de la fusión temprana y la fusión tardía. Consiste en mezclar secuencias de frames consecutivos en una ventana de tiempo móvil.

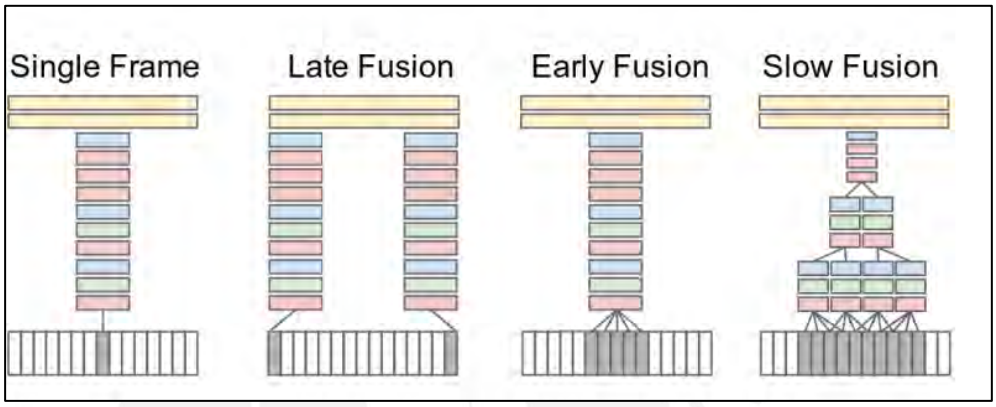

Figura 3: Métodos de fusión de información temporal en video

(Karpathy et al., 2014)

2.3. Principales Arquitecturas Clásicas de Redes Neuronales Convolucionales

ResNet (Residual Networks)

Las Redes Neuronales Residuales, o ResNets, son una arquitectura de redes neuronales convolucionales profundas que abordan el desafío de entrenar redes muy profundas. La característica distintiva de las ResNets es la inclusión de conexiones residuales que permiten que la información fluya directamente a través de las capas sin restricciones, lo que facilita el entrenamiento de redes profundas y evita el problema del desvanecimiento del gradiente. La arquitectura de una ResNet se basa en bloques residuales. Un bloque residual típico consta de dos capas

convolucionales con activaciones ReLU, seguidas de la operación de suma con la entrada original. Estos bloques residuales se apilan para formar la red. Además, las ResNets a menudo incluyen capas de normalización, como Batch Normalization, para estabilizar el entrenamiento (He et al., 2016).

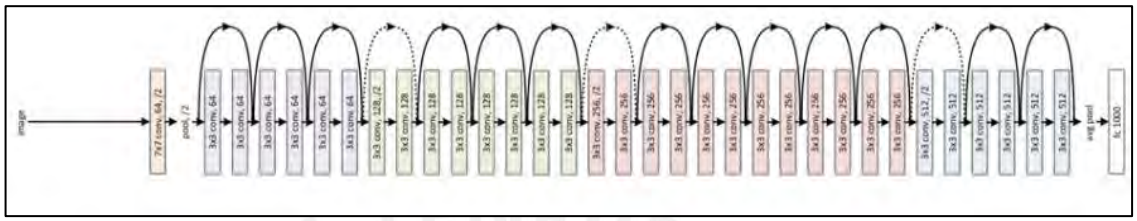

Figura 4: Detalle de la arquitectura de una ResNet (He et al., 2016)

Desde su lanzamiento en el 2016, han surgido diferentes tipos y variaciones de las ResNet, entre ellas se tiene:

- ResNet-18, ResNet-34.

  Estas variantes contienen 18 y 34 capas, respectivamente, son más livianas y se utilizan en aplicaciones con recursos limitados. Las ResNet-18 y ResNet-34 son populares para tareas de clasificación de imágenes y transfer learning (He et al., 2016).

- ResNet-50, ResNet-101, ResNet-152.

  Estas son versiones más profundas con 50, 101 y 152 capas, respectivamente, son adecuadas para aplicaciones donde se requiere una mayor profundidad y una representación más rica de los datos, como la detección de objetos y segmentación semántica (He et al., 2016).

- Wide ResNets:

  Estas variantes aumentan la cantidad de filtros en cada capa, lo que permite capturar características más ricas. Las Wide ResNets se

utilizan en tareas de clasificación de imágenes y pueden superar el rendimiento de las ResNets tradicionales (Zagoruyko & Komodakis, 2017)

- ResNeXt:

  Se basa en la idea de "cardinalidad", que representa el número de subcaminos paralelos en un bloque residual. Son eficaces en una variedad de tareas y se utilizan en aplicaciones de visión por computadora avanzadas (Xie et al., 2017)

EfficientNet

EfficientNet es una familia de arquitecturas de redes neuronales profundas diseñada por Tan & Le, presentada en el artículo "EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks" en 2019. La principal innovación de las EfficientNets es su enfoque en el equilibrio entre la profundidad, el ancho y la resolución de la red para lograr una mayor eficiencia en términos de rendimiento y uso de recursos computacionales. Esta red está compuesta por una familia de modelos, denominados EfficientNets, que logran una precisión y una eficiencia muy superiores a las ConvNets anteriores (Tan & Le, 2019).

La arquitectura EfficientNet se basa en un compuesto de tres dimensiones: profundidad, ancho y resolución. La red se escalona en estas tres dimensiones para crear variantes de diferentes tamaños. La profundidad se refiere al número de capas, el ancho se refiere al número de canales en cada capa, y la resolución se refiere al tamaño de las imágenes de entrada. La arquitectura utiliza una técnica llamada

"compound scaling" para ajustar estos tres parámetros de manera eficiente (Tan & Le, 2019).

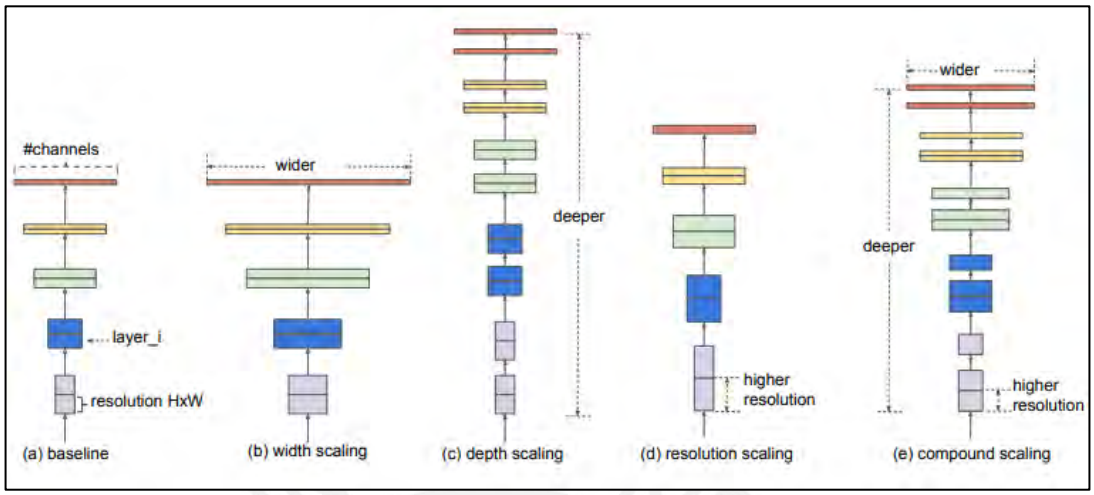

Figura 5: Detalle de la arquitectura de una EfficientNet (Tan & Le, 2019)

Las eEfficientNets tienen diferentes versiones, que van desde B0 a B7, EfficientNet B0 es el modelo base y más pequeño de la familia, mientras que B7 es el más grande. A medida que avanzas de B0 a B7, la profundidad, el ancho y la resolución de la red aumentan. B0 es adecuado para aplicaciones con recursos limitados, como dispositivos móviles, mientras que B7 es más poderoso y se utiliza en aplicaciones de alta gama (Tan & Le, 2019).

EfficientNet se ha utilizado en una variedad de aplicaciones de visión por computadora y más allá, gracias a su eficiencia y alto rendimiento, dentro de las aplicaciones en la que se utiliza, destacan (Tan & Le, 2019):

- Clasificación de Imágenes: Las EfficientNets son altamente efectivas en tareas de clasificación de imágenes y han logrado un alto rendimiento en conjuntos de datos desafiantes, como ImageNet.

- Detección de Objetos: Se aplican en sistemas de detección de objetos, como en aplicaciones de reconocimiento facial y en la detección de objetos en tiempo real.
- Segmentación Semántica: Las EfficientNets se han utilizado en aplicaciones de segmentación semántica para clasificar píxeles de una imagen en categorías específicas.
- Transferencia de Aprendizaje: Son ideales para la transferencia de aprendizaje en una amplia variedad de tareas, ya que las diferentes variantes pueden adaptarse a las necesidades de recursos computacionales.

2.4. Recientes Arquitecturas para Clasificación de Imágenes

Vision Transformers – ViT

Mientras que la arquitectura Transformer se ha convertido en el estándar de facto para las tareas de procesamiento del lenguaje natural, sus aplicaciones a la visión por ordenador han ido ganando un espacio considerable en los últimos años. En visión, la atención o bien se aplica junto con redes convolucionales, o bien se utiliza para sustituir ciertos componentes de las redes convolucionales manteniendo su estructura general. Con Visión Transformers, se demuestra que las Redes Neuronales Convolucionales no necesariamente son las que mejor funcionan en tareas de Visión Computacional, y que un Transformer puro aplicado directamente a secuencias de parches de imágenes puede rendir muy bien en tareas de clasificación de imágenes. Cuando se preentrena en grandes cantidades de datos y se transfiere a múltiples puntos de referencia de reconocimiento de imágenes de tamaño medio o

pequeño (ImageNet, CIFAR-100, VTAB, etc.), Vision Transformer (ViT) alcanza excelentes resultados en comparación con las redes convolucionales de última generación, al tiempo que requiere sustancialmente menos recursos computacionales para entrenarse. La arquitectura de una Vision Transformer consta de varios componentes clave (Dosovitskiy et al., 2021).

- Codificación de posición: Se utiliza codificación de posición para asignar ubicaciones relativas a cada fragmento de la imagen. Esto permite a la red conocer la disposición espacial de los fragmentos.
- Transformers: Los Transformers son el núcleo de la arquitectura. Consisten en capas de atención multi-head y capas de alimentación hacia adelante. La atención se utiliza para capturar relaciones entre segmentos de la imagen, mientras que las capas de alimentación hacia adelante procesan la información localmente.
- Clasificación: Al final de la red, se agrega una Perceptrón Multicapa, con el cual se busca clasificar o predecir las etiquetas de salida, como las clases de objetos que hay en una imagen.

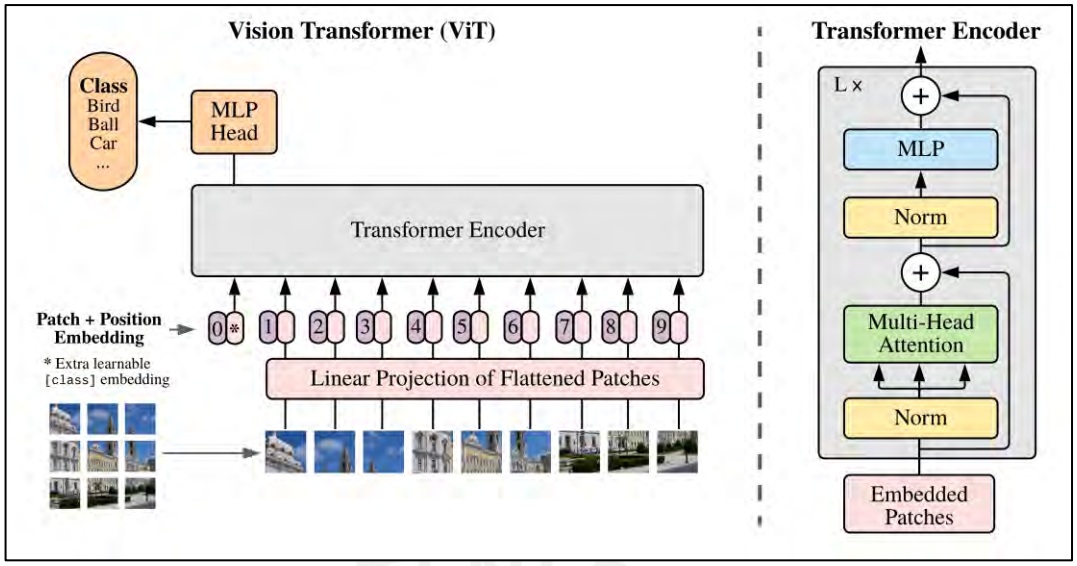

Figura 6: Detalle de la arquitectura de un Vision Transformer

(Dosovitskiy et al., 2021)

Video-Vision Transformers – ViViT

ViViT es un modelo de clasificación de video basado solamente en transformers que con el cual logra mejorar los resultados de múltiples benchmarks. ViViT es más eficiente computacionalmente que los modelos de clasificación de imagen basados en transformadores y puede manejar secuencias de tokens más largas de manera más eficiente. Además, ViViT utiliza técnicas de regularización que permiten que el modelo se entrene en conjuntos de datos más pequeños sin sobreajuste. La arquitectura de ViViT consta de dos etapas principales: extracción de tokens espaciotemporales y codificación de tokens mediante capas de transformadores. En la primera etapa, se extraen tokens espaciotemporales de la entrada de video utilizando una red de detección de objetos previamente entrenada. En la segunda etapa, se codifican los tokens utilizando capas de transformadores, que permiten que el modelo

capture relaciones espaciotemporales complejas en los datos de entrada (Arnab et al., 2021).

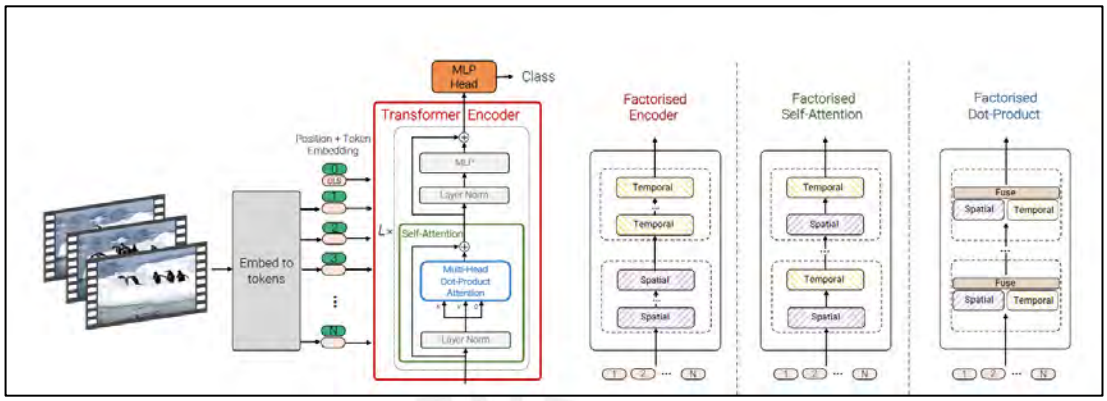

Figura 7: Detalle de la arquitectura de un Video Vision Transformer (Arnab et al., 2021)

Una de las principales ventajas de ViViT en comparación con los modelos de clasificación de video existentes es su eficiencia computacional. ViViT utiliza una arquitectura de transformador más eficiente que la utilizada en los modelos de clasificación de imagen basados en transformadores (VIT), lo que permite que ViViT maneje secuencias de tokens más largas de manera más eficiente. Además, ViViT utiliza una técnica de regularización que permite que el modelo se entrene en conjuntos de datos más pequeños sin sobreajuste (Arnab et al., 2021).

3. ESTADO DEL ARTE

Los avances sobre el estado del arte se obtuvieron de las bases de datos SCOPUS e IEEE, aunque también fue necesario realizar búsquedas apoyándose en Google Scholar. Los principales términos con los que se realizó las búsquedas fueron: “sport video classification”, “football video classification”, “soccer video classification”, “sport events classification”, “convolucional neural network”, “computer vision”, “deep learning”.

En la investigación “Football Game Video Analysis Method with Deep Learning” (Liu et al., 2022), se utilizaron métodos basados en aprendizaje profundo para desarrollar un modelo que detecte seis eventos contenidos en un video de fútbol. Este modelo se divide en dos etapas. La primera etapa se utiliza para generar fragmentos de eventos candidatos (que va desde el preprocesamiento del video, extracción de características y clasificación de los eventos). La segunda etapa se encarga de la detección de los eventos, la cual incluye tareas como, fijación de umbrales probabilísticos para obtener las posiciones inicial y final de los eventos y poder generar fragmentos de eventos completos.

Para la primera etapa utilizaron Redes Neuronales Convolucionales 3D agregando la dimensión del tiempo a una CNN tradicional y una Inception v2 (CNN – 2D) para extraer características y clasificar los eventos, logrando los mejores resultados con la CNN-3D. A nivel global obtuvieron una precisión de 77.2% y por clase se obtuvo 78.8% para los remates (shot), 90.1% para los tiros de esquina (corner kick), 78.9% para los tiros libres (free kick), 97.9% para las tarjetas amarillas (yellow card), 87.1% para las faltas (fouls) y 30.6% para los goles (goal)

En la investigación **"Fine-Grained Soccer Actions Classification Using Deep Neural Network"** (Sen et al., 2022), los investigadores propusieron un enfoque de aprendizaje profundo para clasificar de manera automática diez acciones de fútbol diferentes. Para realizar esta clasificación construyeron el conjunto de datos SoccerAct10 que consiste en clips con las 10 acciones a clasificar, estas son: tiro de esquina (corner), falta (foul), tiro libre (free kick), saque de meta (goal kick), pase largo (long pass), remate al arco (on target shoot), penal (penalty), pase corto (short pass), sustitución (substitution) y lateral (throw in).

La extracción de características se realizó utilizando transfer learning, aprovechando los parámetros aprendidos por los diferentes modelos de Redes Neuronales Convolucionales de última generación como DenseNet201, InceptionResNetV2, MobileNetV2, ResNet152V2 y Xception, los cuales fueron entrenados en ImageNet. En la etapa de clasificación utilizaron una LSTM (Long short-term memory), que recibió como input las características extraídas por las Redes Neuronales Convolucionales, lo cual permitió modelar los cambios temporales frame a frame de las jugadas de fútbol. En la capa final, la función de activación softmax permitió obtener las distribuciones de probabilidades de cada acción clasificada. A nivel global obtuvieron una precisión del 90% para la clasificación de estas 10 jugadas de fútbol.

En la investigación **"End-to-end soccer video scene and event classification with deep transfer learning"** (Hong et al., 2018), se introduce un nuevo conjunto de datos denominado Soccer Video Scene and Event Dataset (SVSED), el cual contiene 600 clips de video de 3 segundos de

duración cada uno, distribuidos en seis clases (100 clips por clase): tiro de esquina (corner), tiro libre (free-kick), gol (goal), penal (penalty), vista amplia sin eventos (long view with no events), vista en primer plano (close-up view). Utilizando este conjunto de datos, y aplicando técnicas de transfer learning, los investigadores experimentaron con diferentes modelos basados en Redes Neuronales Convolucionales como, VGG16, VGG19, ResNet50, Inception V3 y MobileNet, para lograr extraer las características de los fotogramas de cada clip y poder clasificarlos. Al final obtuvieron una precisión global de 76% y por cada clase obtuvieron una precisión de: tiros de esquina-54%, tiros libres-75%, goles-61%, vistas amplias sin eventos-92%, vistas de primer plano-94% y penales-100%.

En la investigación **"Multi-camera Temporal Grouping for Play/Break Event Detection in Soccer Games"** (Song & Rasmussen, 2019), se utilizan las I3D ConvNets, las cuales son una variante de las Redes Neuronales Convolucionales tradicionales, diseñadas específicamente para el procesamiento de videos en 3D (altura, ancho + tiempo). Utilizando estas ConvNets implementan el reconocimiento de seis diferentes acciones que suceden en un partido de fútbol en videos de larga duración y con tomas de múltiples cámaras fijas, estas acciones son: juego fluido (normal play), saques de puerta (plus breaks in play due to kick-offs), tiros libres (free kicks), laterales(throw-ins), goles (goal) y tiros de esquina (corner kicks), obteniendo una precisión global del 84,2% y una precisión individual para cada clase de: juego fluido – 91.6%, saques de puerta – 71.6%, tiros libres – 27.7%, laterales – 75.7% y tiro de esquina – 72.6%.

En la investigación **“Applying Convolutional Neural Network for Detecting Highlight Football Events”** (Le et al., 2021), se propone detectar y clasificar acciones o situaciones destacadas en vídeos de fútbol, teniendo como entrada un vídeo completo de cualquier partido fútbol y como salida un conjunto de escenas con las principales acciones (Córner, Falta, Gol). La arquitectura general del sistema propuesto incluye: preprocesamiento de vídeo (extracción de keyframes, redimensionarlos y recortarlos en la zona dónde se está dando la jugada), extracción y etiquetado de características, construcción de modelos de aprendizaje automático a partir de las características extraídas utilizando una arquitectura basada en Redes Neuronales Convolucionales, para detectar situaciones destacadas en el vídeo de entrada. A nivel global obtuvieron un accuracy de 95,8% para las tres acciones clasificadas.

En la investigación **“User-selectable event summarization in unedited raw soccer video via multimodal bidirectional LSTM”** (Haruyama et al., 2021), se introduce una arquitectura CNN-BiLSTM multimodal para analizar vídeos de fútbol sin editar. Esta arquitectura extrae escenas candidatas (clips) para el resumen de eventos a partir de vídeos de fútbol completos. Luego utilizando una VGG16 extrae tres grupos de características de estos video clips, el primer grupo consolida la información visual de los frames (Visual Features), el segundo grupo consolida la información de la ubicación de los jugadores (Player Features) y el tercer grupo se encarga de consolidar en espectrogramas el audio del clip (Audio Features). A continuación, clasifica estas escenas utilizando modelos basados en la arquitectura propuesta, en eventos típicos como remates (shot), tiro de esquina (corner

kick), tiro libre (free kick) y falta (Foul). Finalmente, generan resúmenes de eventos seleccionables por el usuario considerando simultáneamente la importancia de las escenas candidatas y los resultados de la clasificación de eventos. En la tarea de clasificación, lograron un f1-score global de 0.71 y por cada evento lograron obtener un f1-score de: remate-0.81, tiro de esquina-0.87, tiro libre-0.75 y falta 0.78.

En la investigación **"Automated Event Detection and Classification in Soccer: The Potential of Using Multiple Modalities"** (Nergård Rongved et al., 2021), se presenta y evalúa diferentes enfoques basados en redes neuronales en los que se combina características visuales con características de audio para detectar y clasificar eventos en vídeos de fútbol, los eventos que se clasificaron fueron: tarjeta (card), cambio de jugador (substitution), gol (goal) y no hay evento a detectar (background). Emplean la fusión de modelos para combinar las dos modalidades (vídeo y audio), y realizan experimentos utilizando técnicas de preprocesamiento de videos como early y late fusion y diferentes modelos basados en Redes Neuronales Convolucionales como ResNet 2-D y 3-D en el conjunto de datos SoccerNet. Los resultados muestran que un enfoque multimodal es beneficioso, obteniendo una accuracy global de 89% y una precisión por evento de: tarjeta-87%, cambio de jugador-94%, gol-95% y fondo o background-83%.

4. DISEÑO DE LA SOLUCIÓN

4.1. Base de datos

Para realizar la presente investigación se utilizó los videos de partidos de fútbol con sus respectivas etiquetas disponibles en SoccerNet.

SoccerNet es un conjunto de datos a gran escala para la comprensión de vídeos de fútbol, ha evolucionado a lo largo de los años para incluir diversas tareas, como la detección de acciones, la calibración de cámaras, la reidentificación y el seguimiento de jugadores. Se compone de 550 partidos completos de fútbol (550 x 2 tiempos = 1,100 videos) retransmitidos y 12 partidos (24 videos, 12 de cada tiempo del juego) de una sola cámara tomados de las principales ligas europeas, las características principales de estos videos son que tienen una duración promedio de 45 minutos, están a 25 frames por secundo y tienen un peso aproximado de 1gb cada uno. (Giancola et al., 2018).

Debido a la gran cantidad de información disponible en SoccerNet, se decidió seleccionar y descargar 53 partidos de fútbol (106 videos) de la Premier League de las temporadas 2014-2015, 2015-2016 y 2016-2017, los cuales se usaron como fuente primaria para construir el dataset de videoclips, con los cuales se realizó la experimentación (Ver el detalle de los partidos de fútbol en el Anexo A).

En los 106 videos descargados, como en todo el conjunto de datos hay 17 eventos, de los cuales para realizar esta investigación, solamente se consideraron los siguientes cinco eventos:

- Foul = Falta.
- Shots on target = Remate al arco.
- Corner = Tiro de esquina.
- Indirect free-kick = Tiro libre indirecto.
- Clearance - Despeje o saque de meta.

Una vez obtenidos los 106 videos de SoccerNet con sus respectivas etiquetas, se desarrolló un proceso para extraer videoclips de 2 segundos de duración centrados alrededor del tiempo en el que se da la acción de cada uno de los cinco eventos seleccionados, con los cuales se construyó una base de datos de videoclips personalizada.

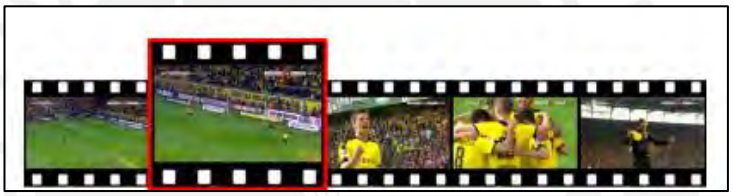

Figura 8: Secuencia de frames que ocurren durante un evento.

(Giancola et al., 2018)

Para extraer estos videoclips se probó con 3 opciones, para capturar de la manera más precisa el momento en el que ocurre el evento (Frame Central = etiqueta de SoccerNet), estas opciones fueron:

- Frame Central – 25, Frame Central, Frame Central + 25 = 51 frames = 2 segundos.
- Frame Central, Frame Central + 49 = 50 frames = 2 segundos.
- Frame Central-15, Frame Central, Frame Central + 34 = 50 frames = 2 segundos.

Después de verificar visualmente los clips extraídos de un video completo, se observó que la tercera opción es la que mejor captaba la secuencia completa de los cinco eventos seleccionados, por lo que se decidió construir el dataset de videoclips utilizando esta estructura de extracción.

Para realizar la experimentación se extrajo solamente 16 frames de manera aleatoria, dando prioridad a los que se encuentran entre el frame 20 y 45 de cada uno de los videoclips obtenidos anteriormente.

Para el entrenamiento de los modelos basados en redes convolucionales tradicionales, se seleccionó aleatoriamente 5 de estos 16 frames como imágenes de entrada y para modelos que incluyen además el tiempo (secuencia de frames) se entrenó con la secuencia completa de los 16 frames.

También se realizó una revisión manual de todos los frames extraídos, de manera que el dataset contenga solamente tomas de primer plano y tomas de un plano general o amplio.

Al final se obtuvo una base de 1,959 videoclips, de los cuales 1,579 se utilizaron para el entrenamiento (80%), 298 para la validación (15%) y 91 para el testeo (5%), y su distribución por clases se muestra en la Figura 10.

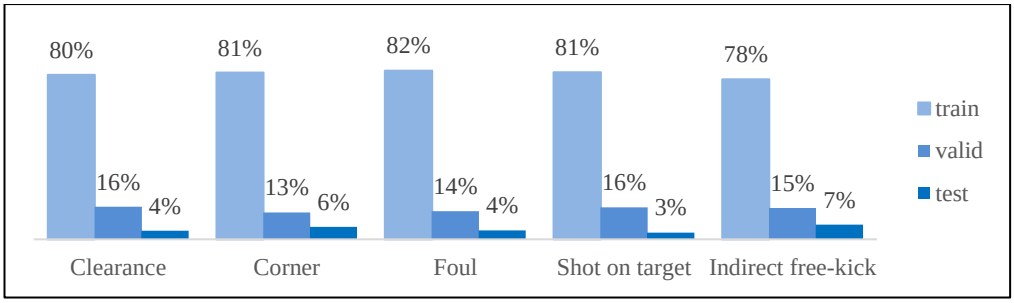

Figura 9: Distribución de eventos en cada uno de los datasets

4.2. Métricas de rendimiento

Se utilizó como métricas de rendimiento el Accuracy, la Precisión el Recall y la matriz de confusión para cada una de las clases.

El accuracy es una métrica de rendimiento comúnmente utilizada en tareas de clasificación en visión por computadora. Se define como la fracción de predicciones correctas realizadas por un modelo en comparación con el total de predicciones. El accuracy mide con qué frecuencia un modelo clasifica correctamente las muestras de datos. Es una métrica fácil de entender y comunicar, ya que se expresa en términos de un porcentaje, donde un 100% de precisión indica que el modelo no comete errores en sus predicciones (Swamynathan, 2017).

La Precisión es una métrica que se utiliza para evaluar qué porcentaje de valores que se han clasificado como positivos por el modelo son realmente positivos. Mientras que el Recall es una métrica que ayuda a medir la cantidad de casos positivos reales que han sido identificados por el modelo. (Swamynathan, 2017).

La matriz de confusión es una herramienta fundamental en tareas de clasificación en visión por computadora y aprendizaje automático, proporciona una forma de visualizar y evaluar el rendimiento de un modelo

de clasificación al mostrar cuántas instancias de cada clase se han clasificado correctamente o incorrectamente. Muestra la relación entre las predicciones del modelo y las clases reales en un conjunto de datos. La matriz de confusión se compone de cuatro valores (Swamynathan, 2017):

- Verdaderos Positivos (True Positives, TP): Representa la cantidad de muestras de la clase positiva que se han clasificado correctamente como positivas por el modelo.
- Verdaderos Negativos (True Negatives, TN): Representa la cantidad de muestras de la clase negativa que se han clasificado correctamente como negativas por el modelo.
- Falsos Positivos (False Positives, FP): Representa la cantidad de muestras de la clase negativa que se han clasificado incorrectamente como positivas por el modelo.
- Falsos Negativos (False Negatives, FN): Representa la cantidad de muestras de la clase positiva que se han clasificado incorrectamente como negativas por el modelo.

4.3. Experimentación

Para realizar la experimentación, se utilizó diferentes enfoques de preprocesamiento de video (Single frame, Early Fusion), además se aplicó técnicas de aumento de datos para evitar el sobreajuste, entre las que destacan técnicas para resaltar bordes, desenfocar las imágenes, agregar ruido gaussiano, desenfoque de movimiento y se rotación aleatoria la imagen en un rango de -35 a 35 grados.

Los modelos que se utilizaron durante el periodo de entrenamiento se pueden dividir en cuatro grandes segmentos:

- Modelos propios.

  Se desarrolló una red convolucional (customConvNet) propia, que sirviera como línea de base o punto de partida. Esta red puede recibir una imagen como entrada o una secuencia de imágenes.

  La arquitectura de la red neuronal construida, contiene principalmente:

  Cinco capas convolucionales de extracción de características, la primera capa convolucional recibe una imagen de 3 canales de entrada (correspondientes a los canales RGB), 96 filtros convolucionales con un tamaño de kernel de 11x11 y un stride de 3, luego se aplica una función de activación ReLU después de cada capa convolucional para introducir no linealidades en la red, luego se aplica una Normalización local o en el batch, luego se reduce a la mitad mediante una capa de Max Pooling. Este patron se repite para las siguientes capas convolucioanles.

  Finalmente, la capa de clasificación contine una función de AdaptiveAvgPool2d, de manera que pueda realizar un average pooling adaptativo para convertir las características en un tensor de tamaño fijo (1, 1), seguido de una capa de una capa de dropout que ayuda a prevenir el sobreajuste, desactivando aleatoriamente un porcentaje de neuronas durante el entrenamiento, y finalmente una capa completamente conectada que produce la salida final con 5 clases.

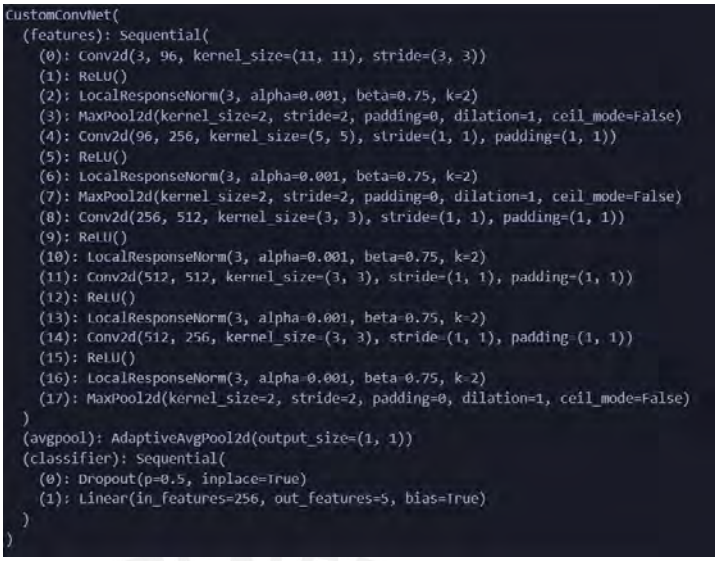

Figura 10: Arquitectura CustomConvNet

- Modelos convolucionales para imágenes.

  En este caso se realizó transfer learning utilizando un modelo ResNet50 y EfficientnetB0.

- Modelos convolucionales para secuencias.

  Para utilizar modelos convolucionales que admita secuencias, fue necesario modificar las primeras capas tanto de la CustomConvNet, ResNet50 y EfficientnetB0, pasando la estructura de la primera capa convolucional de [B, C, H, W] a [B, C*F, H, W].

- Modelos basados en Transformers.

  En este caso se entrenó usando transfer learning con arquitecturas basadas en Transformers como Vision Transformers y Video Vision Transformers.

Los experimentos se ejecutaron en los servidores proporcionados por IAPUCP, en el cual se tuvo acceso a 4 gpus, los cuales cuentan con una tarjeta gráfica NVIDIA RTX A5500 con 25 GB de memoria cada uno. El código fuente fue implementado en Python y PyTorch Lightning.

El entrenamiento de todos los modelos se realizó durante 100 épocas, los hiperparámetros de cada uno de ellos lo puede encontrar en el Anexo B.

4.4. Prototipo de interfaz web

Se construyó una interfaz web para realizar la clasificación automática del evento o jugada que está ocurriendo en un videoclip dado. Para la implementación de esta interfaz se utilizó FastAPI, las funciones de preprocesamiento y transformación de los videos en frames y el modelo entrenado (EfficientNetb0 modificado en su primera capa convolucional).

FastAPI es un framework moderno y rápido (de alto rendimiento) para crear API´s con Python 3.8 o superior basado en sugerencias de tipo estándar de Python (FastAPI, s.f.).

Esta aplicación permite cargar un videoclip de 2 segundos desde una carpeta local, los procesa y devuelve las probabilidades de cada uno de los cinco eventos evaluados, las interfaces se muestran en el Anexo D.

5. RESULTADOS

Con los modelos entrenados en base a frames individuales extraídos de los videoclips se logró obtener resultados que superan el 70% de accuracy global, obteniéndose los resultados más bajos con la red convolucional implementada desde cero, el cual sirve como línea base, para establecer una comparativa de los siguientes modelos entrenados. Luego mediante el uso de transfer learning los modelos ResNet50 y EfficientNetb0, logran superar la barrera del 75% de accuracy global, resultados que son parecidos a los obtenidos por otros investigadores sobre el tema (Ver Figura 11).

Mientras que con los modelos basados en Transformes, se logra obtener resultados que superan el 80% de accuracy, la desventaja de estos modelos es que están sobre ajustando demasiado la data de entrenamiento, ya que hay una diferencia considerable entre el accuracy en la data del entrenamiento y validación o test (Figura 11).

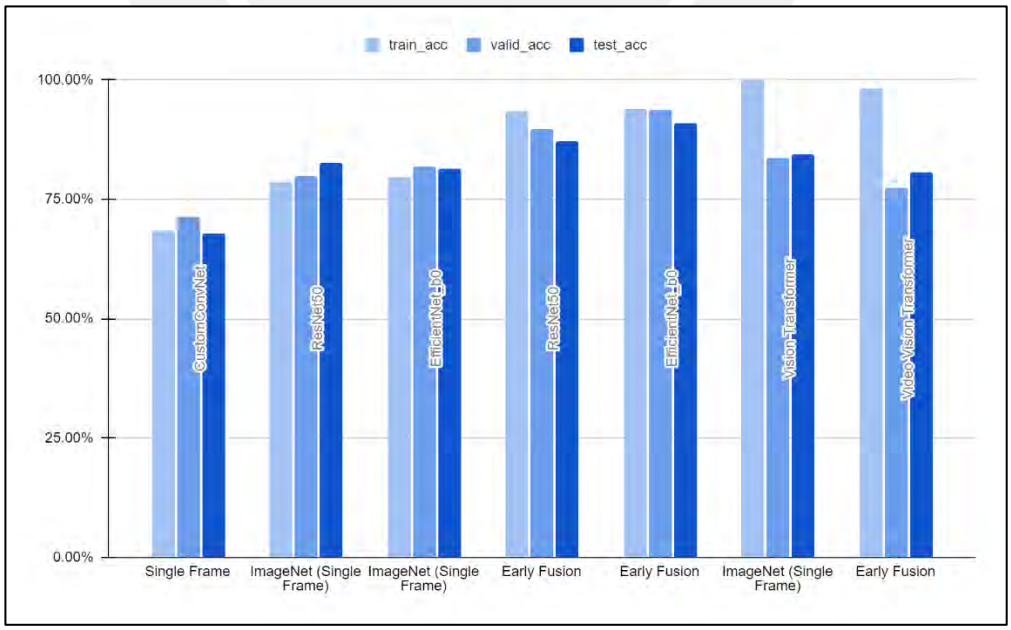

Figura 11: Accuracy global obtenida durante 100 épocas de entrenamiento

Los modelos convolucionales incluyendo el tiempo como secuencia de frames (Modificar la primera capa convolucional de entrada), son los que mejor resultados dieron. Se obtuvo un accuracy global de 91%, el cual está dentro de los mejores resultados obtenidos en este campo de análisis (Ver Figura 11).

El modelo que mejor resultados dio fue el EfficientNetb0 modificado en su primera capa para que se pueda trabajar con secuencias de frames, con este modelo se obtuvo un accuracy global de 91% y para cada uno de los eventos clasificados se obtuvieron los resultados que se muestran en la Tabla 1.

Evento	Precisión	Recall
Clearence	1.00	1.00
Corner	0.92	0.92
Foul	0.90	0.83
Indirect Free Kick	0.88	0.95
Shot on target	0.89	0.89

Tabla 1. Métricas por evento – Modelo EfficientNetb0 – Early Fusión

La matriz de confusión muestra que el modelo identifica muy bien todos eventos, aunque todavía hay espacio de mejora al momento de identificar jugadas que se dan una detrás de otra como una falta seguido de un tiro libre.

	Clearence	Corner	Foul	Indirect Free Kick	Shot on target
Clearence	14	0	0	0	0
Corner	0	12	0	0	1
Foul	0	0	19	3	1
Indirect Free Kick	0	0	1	21	0
Shot on target	0	1	1	0	17

Tabla 2. Matriz de confusión – Modelo EfficientNetb0 – Early Fusión

Finalmente, las curvas de aprendizaje muestran que este modelo se mantiene estable durante las 100 épocas de entrenamiento, garantizando que el modelo entrenado es robusto y estable. Las curvas de aprendizaje de los otros modelos con los que se experimentó se encuentran en el Anexo C.

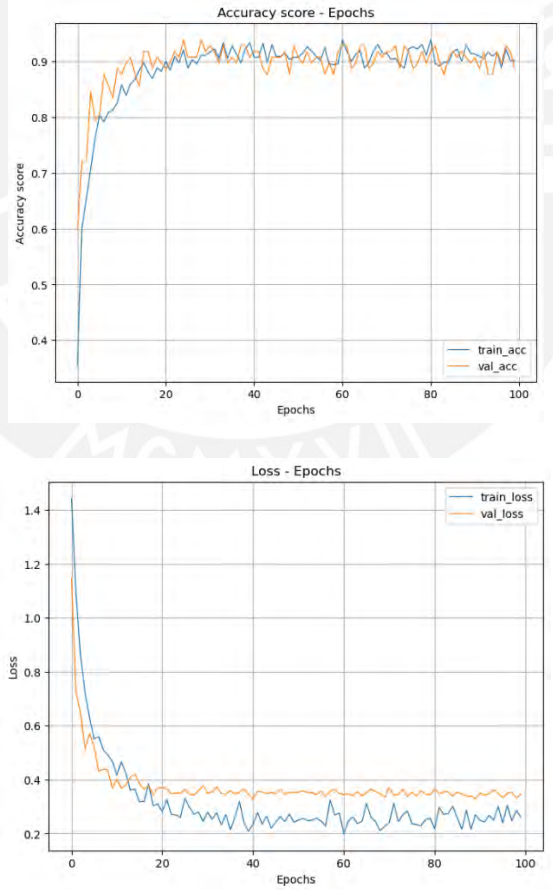

Figura 12: Curvas de accuracy y loss - Modelo EfficientNetb0 Early Fusión

6. CONCLUSIONES

Se logró construir un conjunto de datos de videoclips a partir de SocceNet, conformado por 1,959 clips de 2 segundos de duración de 5 eventos distribuidos de la siguiente manera: 334 clips de saques de meta (Clearence), 215 clips de tiros de esquina (corners), 521 clips de faltas cometidas (fouls), 310 clips de tiros libres indirectos (indirect free kick) y 579 clips de remates al arco (shoots on target), a partir del cual se entrenaron los modelos durante la etapa de experimentación.

Se diseñó una red convolucional desde cero para establecer una línea base y se empleó transfer learning para entrenar redes neuronales convolucionales como ResNet50, EfficientNetb0 (En su arquitectura original y modificando la primera capa de cada una de ellas para que acepten secuencias de imágenes como entrada). También se utilizaron modelos basados en Transformers como Visión Transformers o Video Vision Transformers.

Se utilizó el accuracy como medida principal para evaluar el rendimiento de los diferentes modelos con los que se experimentó, encontrando que el modelo EfficientNetb0 modificado en su primera capa convolucional es el que mejor resultados da, ya que es más preciso y también más estable.

Finalmente utilizando FastApi de Python, y el modelo entrenado (EfficientNetb0 modificado en su primera capa convolucional), se implementó un prototipo que proporciona la clasificación automática un evento de fútbol dado.

7. TRABAJOS FUTUROS

Durante el desarrollo de la investigación, surgieron diversos temas que pueden ayudar a mejorar aún más la precisión obtenida, como tener un dataset más amplio con diferentes posiciones de las cámaras, obtener otras características además de las visuales de los videos y probar con otros modelos de visión computacional.

También sería interesante, desarrollar una investigación con el objetivo de identificar en que parte del frame se está dando el evento (ubicación) lo cual puede ser un input interesante y que podría ayudar a mejorar el accuracy de la tarea de clasificación.

Finalmente se podría ampliar el estudio, para que la red en vez de recibir un clip de entrada y me diga que jugada es, esta red tenga la capacidad de analizar todo un partido entero de fútbol y devolver las principales jugadas.

8. BIBLIOGRAFÍA

Adhikari, S., Kim, G., & Kim, H. (2020). Deep Neural Network-Based System for Autonomous Navigation in Paddy Field. *IEEE Access*.

Andrade, H. S. (2021). Modelo para detectar el uso correcto de mascarillas en tiempo real utilizando redes neuronales convolucionales. *Revista de Investigación en Tecnologías de la Información*. https://doi.org/10.36825/RITI.09.17.011

Arnab, A., Dehghani, M., Heigold, G., Sun, C., Lučić, M., & Schmid, C. (2021). ViViT: A Video Vision Transformer. *arXiv:2103.15691*.

Dosovitskiy, A., Alexander, K., Weissenborn, D., Heigold, G., Uszkoreit, J., Beyer, L., . . . Zhai, X. (2021). An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. *ICLR*.

García, M. Á. (2022, Junio 05). *La forma de consumir deporte está cambiando*. Primera Plana - Diario Marca - España: https://www.marca.com/primera-plana/2022/06/05/629481ed268e3ee1118b4589.html

Giancola, S., Amine, M., Dghaily, T., & Ghanem, B. (2018). SoccerNet: A Scalable Dataset for Action Spotting in Soccer Videos. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*. https://doi.org/10.1109/CVPRW.2018.00223

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep Learnin.* MIT Press. http://www.deeplearningbook.org

Haruyama, T., Takahashi, S., Ogawa, T., & Haseyama, M. (2021). User-selectable Event Summarization in Unedited Raw Soccer Video via Multimodal Bidirectional LSTM. *ITE Transactions on Media Technology and Applications, 9*, 42-53.

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep Residual Learning for Image Recognition. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 770-778.

Hong, Y., Ling, C., & Ye, Z. (2018). End-to-end soccer video scene and event classification with deep transfer learning. *2018 International Conference on Intelligent Systems and Computer Vision (ISCV)*. https://doi.org/10.1109/ISACV.2018.8369043

Karpathy, A., Toderici, G., Shetty, S., Leung, T., Sukthankar, R., & Fei-Fei, L. (2014). Large-Scale Video Classification with Convolutional Neural Networks. *IEEE Conference on Computer Vision and Pattern Recognition*, 1725-1732.

Le, T. H., Van, H. T., Tran, H. S., Nguyen, P. K., Nguyen, T. T., & Le, T. H. (2021). Applying Convolutional Neural Network for Detecting Highlight Football Events. *ICCASA 2021. Lecture Notes of the Institute for Computer Sciences, Social Informatics and Telecommunications Engineering*.

Lecun, Y. (1989). Generalization and network design strategies. (Z. S. R. Pfeifer, Ed.) *Connectionism in perspective Elsevier*.

Lecun, Y., Bottou, L., Bengio, Y., & Haffner, P. (1998). Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 2278-2324.

Liu, N., Liu, L., & Sun, Z. (2022). Football Game Video Analysis Method with Deep Learning. *Computational Intelligence and Neuroscience, 2022*, 12. https://doi.org/https://doi.org/10.1155/2022/3284156

Nergård Rongved, O., Stige, M., Hicks, S. A., Thambawita, V. L., Midoglu, C., Zouganeli, E., . . . Halvorsen, P. (2021). Automated Event Detection and Classification in Soccer: The Potential of Using Multiple Modalities. *Machine Learning and Knowledge Extraction*, 1030-1054. https://doi.org/https://doi.org/10.3390/make3040051

ReasonWhy. (2023, Enero 23). *Deporte, marcas y nuevas formas de consumo: adaptación a un terreno de juego cada vez más digital*. ReasonWhy: https://www.reasonwhy.es/actualidad/deporte-marcas-formas-consumo-terreno-digital-you-first

Sen, A., Minhaz Hossain, S. M., Ashraf Russo, M., Deb, K., & Jo, K.-H. (2022). Fine-Grained Soccer Actions Classification Using Deep Neural Network. *2022 15th International Conference on Human System Interaction (HSI)*. https://doi.org/10.1109/HSI55341.2022.9869480

Song, C., & Rasmussen, C. (2019). Multi-camera Temporal Grouping for Play/Break Event Detection in Soccer Games. *Advances in Visual Computing: 14th International Symposium on Visual Computing, ISVC*.

Swamynathan, M. (2017). *Mastering Machine Learning with Python in Six Steps.* https://doi.org/10.1007/978-1-4842-2866-1

Tan, M., & Le, Q. (2019). EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks. *Proceedings of the 36th International Conference on Machine Learning, Long Beach, 9-15 June 2019*.

Weisstein, E. W. (2002). *CRC Concise Encyclopedia of Mathematics.* New York: Chapman and Hall/CRC.

Xie, S., Girshick, R., Dollár, P., Tu, Z., & He, K. (2017). Aggregated Residual Transformations for Deep Neural Networks. *UC San Diego, Facebook AI Research, arXiv:1611.05431*.

Zagoruyko, S., & Komodakis, N. (2017). Wide Residual Networks. *Université Paris-Est, École des Ponts. arXiv :1605.07146*.

9. ANEXOS

Anexo A: Lista de partidos de fútbol descargados de SoccerNet, con sus respectivas labels (2 videos por partido, más archivo .json con sus respectivas etiquetas.)

id	league	season	game
1	england_epl	2014-2015	2015-02-21 - 18-00 Chelsea 1 - 1 Burnley
2	england_epl	2014-2015	2015-02-21 - 18-00 Crystal Palace 1 - 2 Arsenal
3	england_epl	2014-2015	2015-02-21 - 18-00 Swansea 2 - 1 Manchester United
4	england_epl	2014-2015	2015-02-22 - 19-15 Southampton 0 - 2 Liverpool
5	england_epl	2015-2016	2015-08-08 - 19-30 Chelsea 2 - 2 Swansea
6	england_epl	2015-2016	2015-08-29 - 17-00 Chelsea 1 - 2 Crystal Palace
7	england_epl	2015-2016	2015-08-29 - 17-00 Manchester City 2 - 0 Watford
8	england_epl	2015-2016	2015-09-12 - 14-45 Everton 3 - 1 Chelsea
9	england_epl	2015-2016	2015-09-12 - 17-00 Crystal Palace 0 - 1 Manchester City
10	england_epl	2015-2016	2015-09-19 - 19-30 Manchester City 1 - 2 West Ham
11	england_epl	2015-2016	2015-09-26 - 17-00 Liverpool 3 - 2 Aston Villa
12	england_epl	2015-2016	2015-10-17 - 17-00 Chelsea 2 - 0 Aston Villa
13	england_epl	2015-2016	2015-10-31 - 15-45 Chelsea 1 - 3 Liverpool
14	england_epl	2015-2016	2015-11-07 - 18-00 Manchester United 2 - 0 West Brom
15	england_epl	2015-2016	2015-11-21 - 20-30 Manchester City 1 - 4 Liverpool
16	england_epl	2015-2016	2015-11-29 - 15-00 Tottenham 0 - 0 Chelsea
17	england_epl	2015-2016	2015-12-05 - 20-30 Chelsea 0 - 1 Bournemouth
18	england_epl	2015-2016	2015-12-19 - 18-00 Chelsea 3 - 1 Sunderland
19	england_epl	2015-2016	2015-12-26 - 18-00 Manchester City 4 - 1 Sunderland
20	england_epl	2015-2016	2016-01-03 - 16-30 Crystal Palace 0 - 3 Chelsea
21	england_epl	2015-2016	2016-01-13 - 22-45 Chelsea 2 - 2 West Brom
22	england_epl	2015-2016	2016-02-07 - 19-00 Chelsea 1 - 1 Manchester United
23	england_epl	2015-2016	2016-02-14 - 19-15 Manchester City 1 - 2 Tottenham

24	england_epl	2015-2016	2016-03-02 - 23-00 Liverpool 3 - 0 Manchester City
25	england_epl	2015-2016	2016-03-05 - 18-00 Chelsea 1 - 1 Stoke City
26	england_epl	2015-2016	2016-03-19 - 18-00 Chelsea 2 - 2 West Ham
27	england_epl	2015-2016	2016-04-09 - 17-00 Swansea 1 - 0 Chelsea
28	england_epl	2015-2016	2016-04-09 - 19-30 Manchester City 2 - 1 West Brom
29	england_epl	2015-2016	2016-05-07 - 17-00 Sunderland 3 - 2 Chelsea
30	england_epl	2016-2017	2016-08-14 - 18-00 Arsenal 3 - 4 Liverpool
31	england_epl	2016-2017	2016-08-20 - 17-00 Burnley 2 - 0 Liverpool
32	england_epl	2016-2017	2016-08-20 - 19-30 Leicester 0 - 0 Arsenal
33	england_epl	2016-2017	2016-09-10 - 17-00 Arsenal 2 - 1 Southampton
34	england_epl	2016-2017	2016-09-16 - 22-00 Chelsea 1 - 2 Liverpool
35	england_epl	2016-2017	2016-09-17 - 17-00 Hull City 1 - 4 Arsenal
36	england_epl	2016-2017	2016-09-24 - 19-30 Arsenal 3 - 0 Chelsea
37	england_epl	2016-2017	2016-10-17 - 22-00 Liverpool 0 - 0 Manchester United
38	england_epl	2016-2017	2016-10-22 - 19-30 Liverpool 2 - 1 West Brom
39	england_epl	2016-2017	2016-10-29 - 14-30 Sunderland 1 - 4 Arsenal
40	england_epl	2016-2017	2016-10-29 - 17-00 Tottenham 1 - 1 Leicester
41	england_epl	2016-2017	2016-11-06 - 17-15 Liverpool 6 - 1 Watford
42	england_epl	2016-2017	2016-11-06 - 19-30 Leicester 1 - 2 West Brom
43	england_epl	2016-2017	2016-11-19 - 18-00 Southampton 0 - 0 Liverpool
44	england_epl	2016-2017	2016-11-26 - 18-00 Liverpool 2 - 0 Sunderland
45	england_epl	2016-2017	2016-12-10 - 20-30 Leicester 4 - 2 Manchester City
46	england_epl	2016-2017	2016-12-11 - 19-30 Liverpool 2 - 2 West Ham
47	england_epl	2016-2017	2016-12-14 - 22-45 Middlesbrough 0 - 3 Liverpool
48	england_epl	2016-2017	2016-12-19 - 23-00 Everton 0 - 1 Liverpool
49	england_epl	2016-2017	2016-12-27 - 20-15 Liverpool 4 - 1 Stoke City
50	england_epl	2016-2017	2016-12-31 - 20-30 Liverpool 1 - 0 Manchester City
51	england_epl	2016-2017	2017-01-02 - 15-30 Middlesbrough 0 - 0 Leicester
52	england_epl	2016-2017	2017-01-02 - 18-00 Sunderland 2 - 2 Liverpool
53	england_epl	2016-2017	2017-01-14 - 20-30 Leicester 0 - 3 Chelsea

Tabla 3. Listado de videos descargados de SoccerNet

Anexo B: Configuración de hiper parámetros por modelo

Hiperparámetro	CustomConvNet (Single Frame)	ResNet50 (Single Frame)	EfficientNet_b0 (Single Frame)	ResNet50 (Early Fusion)	EfficientNet_b0 (Early Fusion)	Vision Transformer (Single Frame)	Video Vision Transformer (Early Fusion)
seed	42	42	42	42	42	42	42
num_frames	5	5	5	16	16	5	32
num_classes	5	5	5	5	5	5	5
batch_size	32	32	32	8	8	64	1
num_workers	2	2	2	2	2	2	2
learning_rate	0.0001	0.0001	0.0001	0.0005	0.0004	0.0006	0.0006
num_epochs	100	100	100	100	100	100	100
drop_prob	0.2	0.2	0.2	0.4	0.4	0.2	0.4
pretrained	True	True	True	True	True	True	True
height	1024	1024	1024	1024	1024	224	224
width	1024	1024	1024	1024	1024	224	224

Tabla 4. Configuración de hiperparámetros

Anexo C: Curvas de Rendimiento

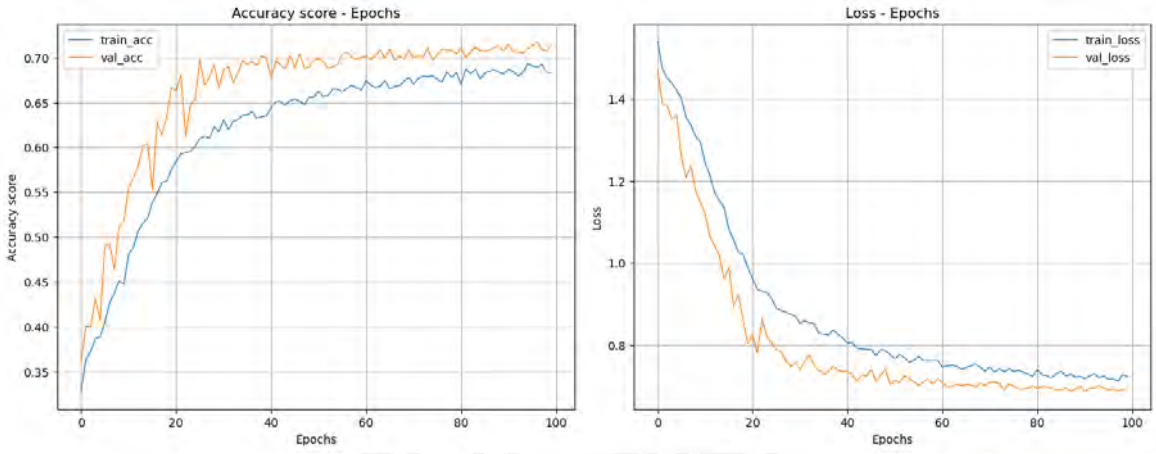

Figura 13: Curvas de accuracy y loss –CustomConvNet– Single Frame

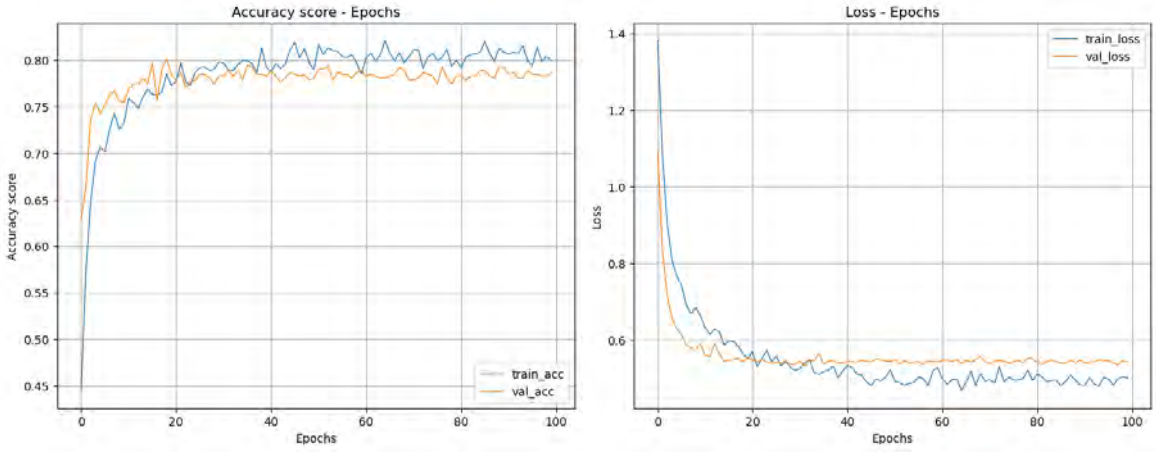

Figura 14: Curvas de accuracy y loss –ResNet50– Single Frame

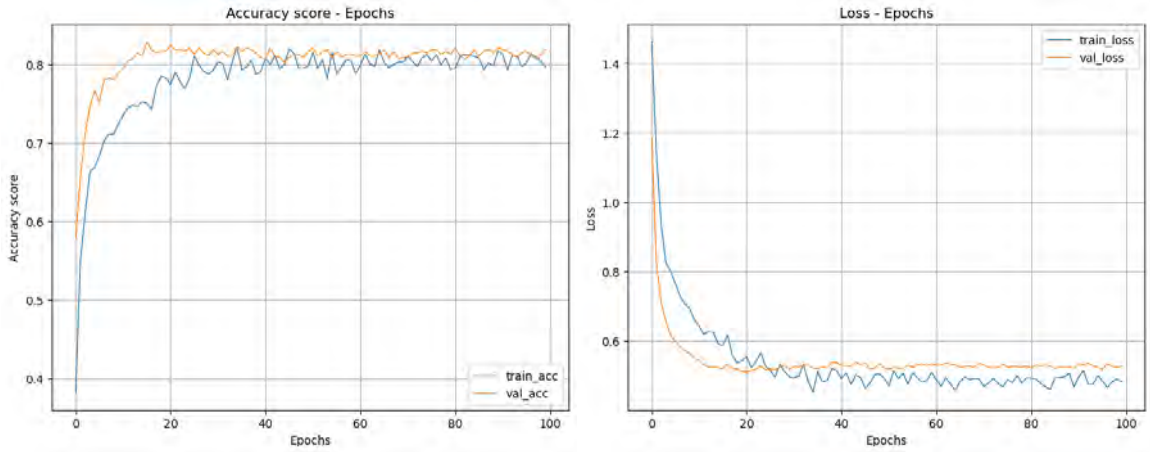

Figura 15: Curvas de accuracy y loss –EfficientNetB0– Single Frame

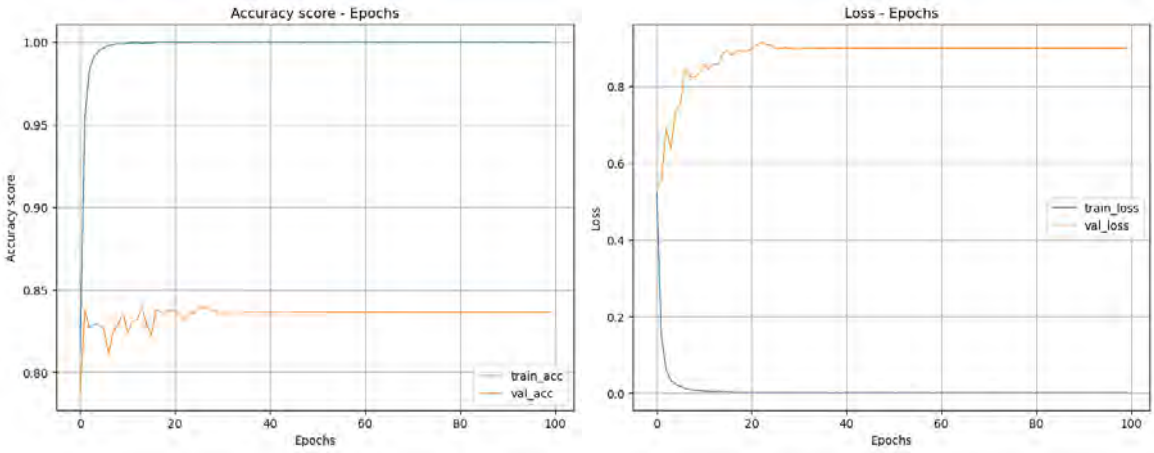

Figura 16: Curvas de accuracy y loss –ViT– Single Frame

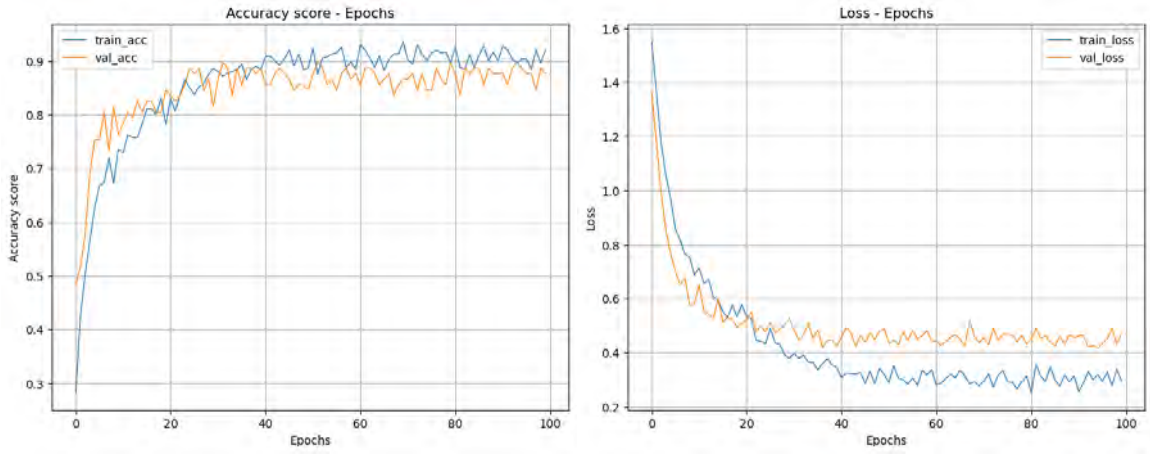

Figura 17: Curvas de accuracy y loss –ResNet50–Early Fusion

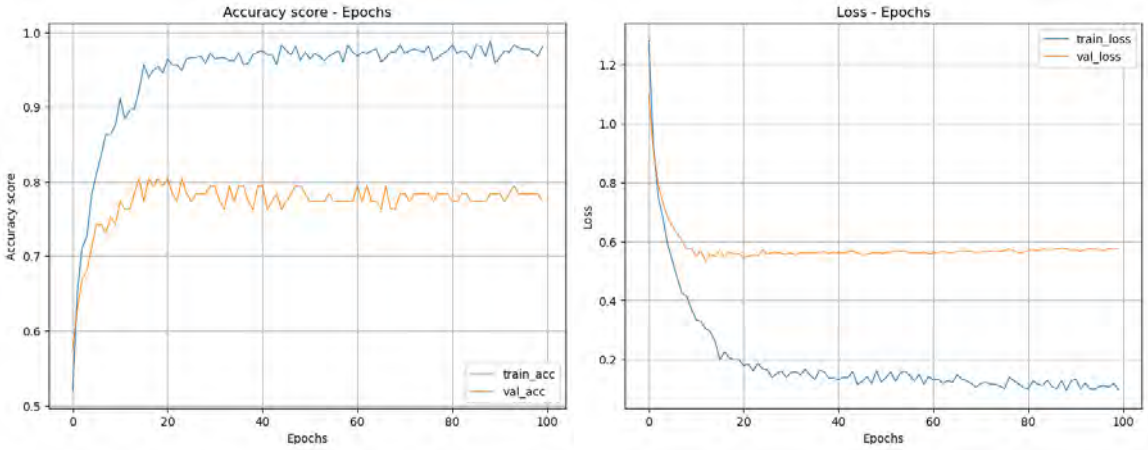

Figura 18: Curvas de accuracy y loss –ViViT–Early Fusion

Anexo D: Interfaz Web para predecir la jugada que aparece en un videoclip dado

Figura 19: Secuencia del prototipo - interfaz web