UNIVERSIDAD DE CHILE
FACULTAD DE CIENCIAS FÍSICAS Y MATEMÁTICAS
DEPARTAMENTO DE CIENCIAS DE LA COMPUTACIÓN

CATALOGACIÓN AUTOMÁTICA DE VIDEOS POR ENTRENAMIENTO DE MODELO MULTIMODAL AUTO-SUPERVISADO PARA MEDIO TELEVISIVOS

MEMORIA PARA OPTAR AL TÍTULO DE
INGENIERO CIVIL EN COMPUTACIÓN

BENJAMÍN ALONSO AVENDAÑO LAGOS

PROFESOR GUÍA:
JOSÉ SAAVEDRA RONDO

MIEMBROS DE LA COMISIÓN:
PATRICIO INOSTROZA FAJARDIN
ANDRÉS ABELIUK KIMELMAN

SANTIAGO DE CHILE
2024

Resumen

Este estudio se enfoca en la catalogación de videos en el contexto de los medios de comunicación, abordando un desafío fundamental en la gestión de contenido audiovisual. Comienza destacando la complejidad de catalogar manualmente grandes volúmenes de videos en un entorno mediático. Se subraya la importancia de adoptar enfoques de aprendizaje automático para mejorar la eficiencia y la precisión de este proceso.

Se exploran modelos de representación de videos, incluidos aquellos basados en CNN y Transformers, resaltando su capacidad para capturar características visuales y relaciones temporales en los videos. Se discuten modelos específicos como TimeSformer, Video Swin Transformer y VideoMAE, que han demostrado su efectividad en la extracción de características de videos.

El estudio profundiza en las metodologías desarrolladas para la catalogación de videos en medios de comunicación, abordando desafíos como la alta entropía de los datos y la falta de etiquetas precisas. Se presentan enfoques para la creación de clases y etiquetas a partir de la metadata de los videos, así como la utilización de modelos NER para procesar entidades en el texto.

Se realiza una evaluación exhaustiva de modelos de representación en diversos conjuntos de datos públicos y en el contexto del canal de televisión. Se utiliza el ARI para medir la calidad de las representaciones de videos y se identifican áreas de mejora. Se destacan las mejoras realizadas en la creación de clases y etiquetas, lo que conduce a un aumento significativo en el rendimiento de los modelos, con TimeSformer como el modelo más efectivo.

La discusión se centra en la importancia de elegir adecuadamente las clases y etiquetas y se sugieren direcciones futuras, como el análisis más profundo de los datos y la exploración de modelos multimodales.

En conclusión, este estudio ofrece una visión completa de la catalogación de videos en medios de comunicación, enfatizando la relevancia de los modelos de representación, la creación de clases y etiquetas, y la evaluación en contextos del mundo real. Reconoce limitaciones como el costo computacional y sugiere futuras investigaciones para mejorar aún más este proceso fundamental en la gestión de contenido audiovisual.

Agradecimientos

En este trabajo, deseo comenzar expresando mi profundo agradecimiento a todas las personas que han sido parte de mi vida, desde familiares hasta amigos cercanos. Han sido un apoyo inquebrantable a lo largo de este largo camino académico, brindándome su apoyo, paciencia y comprensión. Sus palabras de aliento y su presencia constante han sido pilares fundamentales que me han mantenido enfocado y motivado.

Mi gratitud se extiende de manera especial a mi amada familia. Mi madre, Tatiana, y mi padre, Edgardo, han sido ejemplos de amor incondicional y sabios consejos que han guiado mis pasos a lo largo de la vida. A mis dos hermanos, Matías y Tomás, y a mi hermana, Josefa, quienes siempre han estado presentes para alentarme y compartir momentos especiales juntos.

No puedo pasar por alto el papel esencial de mi tía Maritza, cuya sabiduría y aliento han sido un faro en mi camino. También agradezco a mi Mami Hilda, a mi Abu Teresa, a mi Tata Edgardo, a mi Tata Victor, a mis tías, tíos, primos y demás familiares que, de diversas formas, han contribuido a mi crecimiento personal y académico.

A mis amigos y seres queridos, a Jorge la pareja de mi hermana, al Visho, Are, Muni, Bosch, Aguayon, Camilon, Tomi y todos los amigos del colegio con quienes compartimos innumerables recuerdos. A Jojo, Ignacio, Frolopo, Seba y Tomi, quienes se cruzaron en mi camino en la universidad y se han convertido en amigos entrañables.

Finalmente, quiero expresar mi sincero agradecimiento al Profesor de Guía, José Saavedra R., y al equipo que trabajó incansablemente en este proyecto (Lukas y Cristobal). Su orientación experta, infinita paciencia y dedicación incansable fueron fundamentales para llevar este proyecto al éxito. Sus conocimientos y consejos resultaron invaluables en cada etapa de este trabajo.

Tabla de Contenido

Índice de Tablas

Índice de Ilustraciones

Capítulo 1

Introducción

La producción y distribución de contenidos audiovisuales en la industria de los medios televisivos han experimentado una expansión sin precedentes en las últimas décadas. Con la creciente disponibilidad de plataformas digitales y la multiplicidad de canales de transmisión, la cantidad de material de video generado diariamente es abrumadora. Sin embargo, este crecimiento exponencial en la producción de contenido se ha traducido en un desafío crítico para la gestión y catalogación de estos recursos audiovisuales.

La catalogación precisa de contenido de video es esencial para una serie de aplicaciones en la industria de los medios, que van desde la búsqueda y recuperación eficiente de contenido, hasta la recomendación de programas y la personalización de la experiencia del espectador. Sin embargo, el etiquetado manual de videos, en un panorama de crecimiento constante de datos, resulta prohibitivamente costoso y requiere una inversión de tiempo significativa. Esto ha impulsado la búsqueda de soluciones automatizadas que puedan asumir la tarea de catalogación de manera eficaz y precisa.

Por otra parte, en los últimos años, hemos sido testigos de avances extraordinarios en el campo de la inteligencia artificial, impulsados por la revolución del aprendizaje profundo. En particular, la visión por computadora, que se refiere al desarrollo de modelos computacionales capaces de interpretar y comprender el entorno a través de imágenes y videos, ha experimentado una transformación radical gracias a estas innovaciones en IA.

Dos enfoques destacados en esta revolución son las redes convolucionales y los modelos basados en *Transformers*. Las redes convolucionales, inspiradas en la arquitectura biológica de la corteza visual, han demostrado una capacidad excepcional para procesar y analizar imágenes de manera jerárquica, extrayendo características clave y patrones visuales. Por otro lado, los modelos basados en *Transformers*, inicialmente concebidos para el procesamiento de lenguaje natural, han revelado su versatilidad al aplicarse a tareas de visión por computadora, permitiendo una comprensión más profunda y contextual de la información visual.

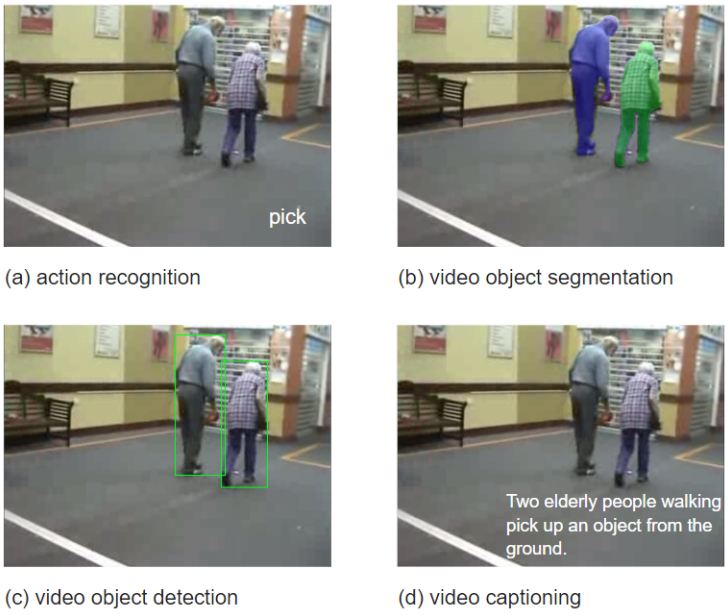


Figura 1.1: Diferentes tareas de visión por computadora para videos. (a) Reconocimiento de acciones. (b) Segmentación de objetos en video. (c) Detección de objetos en video. (d) Descripción de videos.

Esta convergencia entre la catalogación automática de videos y las tecnologías de aprendizaje profundo representa un punto de inflexión en la industria de los medios televisivos. La posibilidad de aprovechar modelos de representación de videos avanzados, basados en redes convolucionales y *Transformers*, ofrece una oportunidad única para abordar el desafío de gestionar y clasificar el creciente caudal de contenido audiovisual de manera eficiente y precisa.

Exploraremos cómo estas tecnologías pueden ser utilizadas para capturar patrones y características clave en los videos, permitiendo así la asignación automática de etiquetas y metadatos relevantes. Además, examinaremos las implicaciones prácticas de implementar estos modelos en un entorno de medios televisivos, considerando aspectos de escalabilidad, eficiencia y precisión, con una evaluación que se llevará a cabo en el contexto del medio televisivo para lograr un mejor resultado.

En este contexto, este trabajo se realiza bajo el supuesto de que la información visual contenida en los videos debe ser suficiente para generar descripciones y metadatos relevantes. Este supuesto se fundamenta en el potencial de las tecnologías de visión por computadora, especialmente aquellas basadas en redes neuronales convolucionales y *Transformers*, para extraer y comprender características visuales clave en el contenido audiovisual, en línea con la capacidad humana de interpretar y captar información relevante a través de la observación visual.

A través de esta investigación, aspiramos a proporcionar una visión integral de los avances y desafíos actuales en el campo de la catalogación automática de videos para medios televisivos, así como contribuir al desarrollo de soluciones que tengan un impacto significativo en la eficiencia y calidad de la gestión de contenido audiovisual en la era digital.

1.1. Relevancia del Problema

En esta memoria se abordará el análisis del impacto derivado de la aplicación de modelos de aprendizaje profundo en el contexto de un canal televisivo. El objetivo es sentar las bases para una posible automatización en la tarea de catalogación del contenido audiovisual, que implica la generación de metadatos, y que actualmente conlleva una carga de tiempo mayor.

1.2. Objetivos

1.2.1. Objetivo General

El objetivo del proyecto es llevar a cabo una evaluación exhaustiva de modelos unimodales (modelos de video que se basan únicamente en información visual) en el contexto del canal televisivo y de el aprendizaje en un enfoque tanto supervisado como el auto-supervisado. A través de esta evaluación, se pretende generar representaciones de videos de alta calidad que sienten una sólida base y representen un primer paso hacia la automatización del proceso de catalogación de medios audiovisuales en el ámbito de la industria televisiva.

1.2.2. Objetivos Específicos

1. Definir conjuntos de datos públicos Kinetics [18], UCF101 [25], HMDB51 [19] y Something-Something v2 [16].
2. Analizar y definir de modelos unimodales SlowFast [15], TimeSformer [13], ViViT (FE) [12], Video Swin Transformer [23] y VideoMAE [27].
3. Entrenar y evaluar modelos en conjuntos de datos públicos.
4. Procesar metadata del canal televisivo, crear conjunto de clases a partir de la metadata brindada.
5. Seleccionar y evaluar modelos destacados en el conjunto de datos del canal.
6. Entrenar y evaluar modelos destacados en el conjunto de datos del canal.
7. Comparar resultados de modelos entrenados en el contexto del canal y de modelos no entrenados en el contexto del canal.
8. Implementar cambios y mejoras metodológicas.

1.3. Evaluación

La evaluación de los modelos se centrará en el uso del Índice de Rand Ajustado (ARI) como métrica principal. Esta medida será empleada para comparar y analizar la efectividad

de las metodologías empleadas en este estudio. Además del ARI, se emplearán otras métricas adicionales para realizar un análisis más detallado de los modelos y metodologías utilizadas. La evaluación se llevará a cabo mediante el *clustering* de las representaciones latentes del contenido visual de los videos, en conjunción con las representaciones de las descripciones asociadas a dichos videos.

El cálculo del ARI en los *clusters* generados a partir de las representaciones de video y texto permitirá evaluar la coherencia del espacio latente generado por el modelo de video. Se espera que descripciones similares queden cercanas entre sí, mientras que las descripciones diferentes se ubiquen distantes unas de otras. Por consiguiente, al realizar el *clustering* y comparar los *clusters* obtenidos por las representaciones de video y texto, se espera una correspondencia, donde las descripciones que coincidan en el espacio de texto también se agrupen en el mismo *cluster* en el espacio de video. Esta coherencia entre los *clusters* obtenidos de ambas representaciones validarán la calidad del espacio latente de los videos.

1.4. Estructura

La presente memoria posee la siguiente estructura:

- **Capítulo 2:** Estado del arte

  Se presenta el estado del arte, donde se describen los principales conceptos relacionados con el trabajo.

- **Capítulo 3:** Metodología

  Se presenta a la metodología utilizada para resolver el problema, preparación de datos, implementación y detalle de entrenamiento.

- **Capítulo 4:** Resultados Iniciales y Discusión

  Se presentan los resultados iniciales obtenidos de los modelos evaluados y se realiza la correspondiente discusión.

- **Capítulo 5:** Mejoras y Ajustes Metodológicos

  Se presentan y explican las mejoras y ajustes realizados en las metodologías en respuesta a los resultados iniciales.

- **Capítulo 6:** Resultados Finales y Discusión

  Se presentan los resultados finales obtenidos de los modelos y metodologías evaluadas y se realiza la correspondiente discusión.

- **Capítulo 7:** Conclusión

  Se presentan las conclusiones del trabajo realizado, junto con el planteamiento de las limitaciones y el trabajo futuro.

Capítulo 2

Estado del Arte

A lo largo del presente capítulo, se abordarán los conceptos esenciales necesarios para una comprensión completa del problema planteado y de la solución abordada. Este recorrido comenzará desde los fundamentos del aprendizaje automático y avanzará hacia una exploración detallada de las distintos componentes y arquitecturas de redes neuronales de vanguardia en el ámbito de la representación de videos.

2.1. Aprendizaje de Máquina

El aprendizaje de máquina o "*machine learning*" en inglés, es una rama fundamental de la inteligencia artificial que se centra en la capacidad de las máquinas para aprender de los datos. Se basa en algoritmos y modelos matemáticos que permiten a las máquinas adaptarse y mejorar su rendimiento a medida que se les proporciona más información.

En el *machine learning* existen diversas modalidades de aprendizaje, y para cada una de ellas existen diferentes modelos de implementación. La clasificación tradicional incluye el aprendizaje supervisado, el aprendizaje no supervisado, el aprendizaje reforzado y el aprendizaje auto supervisado.

- El aprendizaje supervisado se basa en que, para cada conjunto de datos que se introduce en el modelo, se proporciona una etiqueta que corresponde a la respuesta deseada. En este enfoque, el modelo se adapta para generar respuestas correctas al saber cuál debería ser la respuesta.
- El aprendizaje no supervisado se caracteriza por la falta de etiquetas en los datos de entrada. Aquí, el modelo analiza los datos en busca de relaciones inherentes y busca agruparlos de manera coherente, sin conocer de antemano cuál debería ser la salida.
- El aprendizaje reforzado implica la interacción de un agente que debe llevar a cabo una tarea específica y recibe recompensas, ya sean positivas o negativas, según las acciones que realiza. Este enfoque está relacionado con la toma de decisiones óptimas en cada paso para maximizar las recompensas a lo largo de una secuencia de acciones.

- El aprendizaje auto supervisado, que cobra cada vez más relevancia, se refiere a la capacidad de un modelo para aprender de datos no etiquetados, donde el propio modelo genera etiquetas o tareas auxiliares para el aprendizaje, lo que permite la extracción de representaciones significativas de los datos.

El aprendizaje de máquina se aplica en una amplia variedad de campos, desde la visión por computadora y el procesamiento de lenguaje natural hasta la predicción financiera y la medicina. Los modelos de *machine learning* pueden ser entrenados en grandes conjuntos de datos para realizar tareas específicas, y su capacidad para aprender y adaptarse a nuevos datos los hace herramientas poderosas en la toma de decisiones y la automatización de procesos.

2.2. *Clustering*

El *clustering* es una técnica de análisis de datos que agrupa un conjunto de elementos en subconjuntos homogéneos, donde los elementos dentro de cada grupo son más similares entre sí que con los elementos en otros grupos. Esta técnica es ampliamente utilizada en una variedad de aplicaciones, desde la segmentación de clientes en marketing hasta la detección de anomalías en la seguridad de la red.

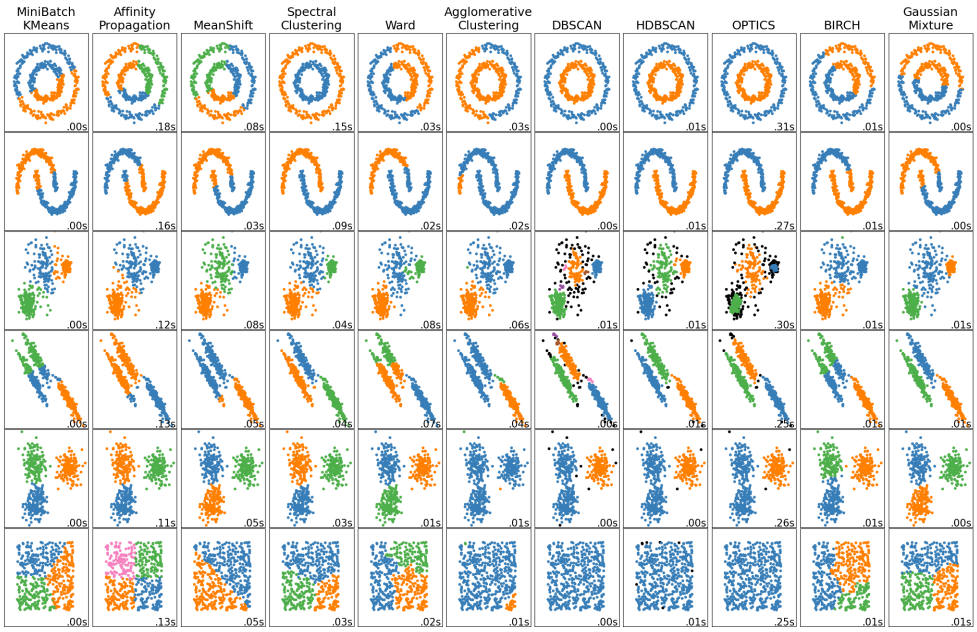


Figura 2.1: Diferentes técnicas de *clustering*. (1) K-Means *(Mini-Batch)*. (7) DBSCAN. (8) HDBSCAN. Fuente: [20]

2.2.1. K-Means

K-Means es un algoritmo de *clustering* que busca dividir un conjunto de datos en K grupos, con K predefinido. Inicialmente, asigna aleatoriamente K centroides y asigna cada

punto de datos al centroide más cercano. Luego, actualiza los centroides tomando el promedio de los puntos asignados a cada grupo y repite este proceso hasta la convergencia. Es eficiente para grandes conjuntos de datos, pero requiere la especificación previa de K y su rendimiento puede verse afectado por los valores iniciales de los centroides.

2.2.2. DBSCAN

DBSCAN (*Density-Based Spatial Clustering of Applications with Noise*) es un algoritmo de *clustering* basado en la densidad que identifica grupos de formas arbitrarias en conjuntos de datos con densidades variables. Comienza con un punto de datos y forma grupos alrededor de él al encontrar puntos cercanos dentro de un radio epsilon. Luego, se expande para incluir puntos vecinos que también tengan densidades adecuadas. No requiere especificar previamente el número de grupos, puede identificar grupos de formas irregulares y es resistente al ruido, pero puede ser menos eficaz en conjuntos de datos de alta dimensionalidad, y la elección de parámetros es crítica.

2.2.3. HDBSCAN

HDBSCAN (*Hierarchical Density-Based Spatial Clustering of Applications with Noise*) es un algoritmo de *clustering* que extiende DBSCAN al proporcionar una estructura jerárquica de *clusters*. En lugar de depender de un valor fijo de epsilon, HDBSCAN utiliza un enfoque basado en densidad que encuentra *clusters* de formas variadas en datos con densidades variables. Proporciona una representación más completa de la estructura de datos al permitir la identificación de *clusters* anidados y *subclusters* dentro de grupos más grandes. Además, HDBSCAN puede identificar automáticamente el número de *clusters*, lo que lo hace especialmente útil cuando no se conoce de antemano la cantidad de grupos en los datos. Este algoritmo es valioso para el análisis de datos con estructuras complejas y aplicaciones donde se requiere una comprensión detallada de la organización de los datos.

2.3. Redes Neuronales

Las redes neuronales tienen su origen en la simulación de las neuronas biológicas y su comunicación, donde se procesan impulsos de entrada, atraviesan neuronas intermedias y se obtiene un impulso de salida. Cada neurona se caracteriza por tener una o más entradas, que se ponderan mediante pesos y se les suma un valor llamado *bias*. Luego, esta suma se pasa a través de una función de activación. En una neurona con múltiples entradas, como la de la Figura 2.2 en donde se puede ver un ejemplo de una neurona con 3 entradas y una salida, la salida se calcula mediante la Ecuación 2.1.

$$y = F(\sum_i x_i * w_i) \tag{2.1}$$

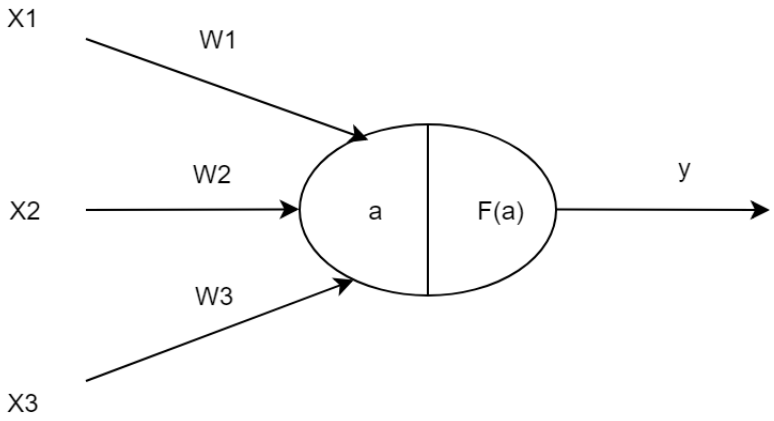


Figura 2.2: Ejemplo neurona con 3 entradas. Fuente: [11]

En una red neuronal, se compone de un conjunto de neuronas organizadas en capas, y las conexiones entre estas capas varían según la arquitectura específica del modelo. En la Figura 2.3 se ilustra un ejemplo de una red neuronal simple, donde los datos de entrada pasan a través de tres capas de neuronas antes de llegar a una neurona de salida.

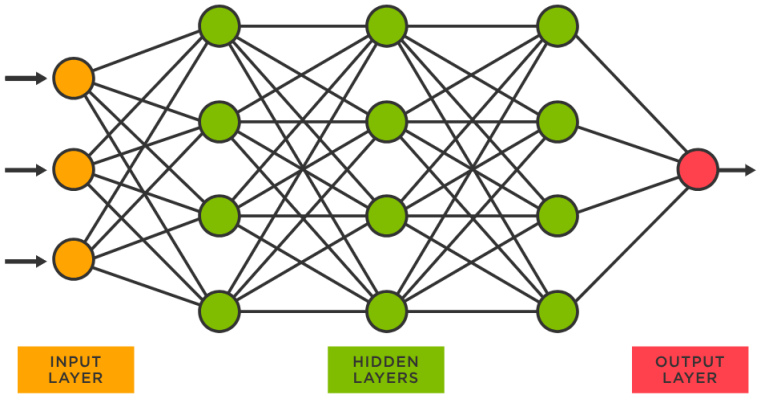


Figura 2.3: Ejemplo red neuronal de 5 capas, con 3 entradas y una salida. Fuente: [26]

En el proceso de aprendizaje de una red neuronal, la actualización de los pesos y *bias* constituye un elemento esencial. La modalidad de esta actualización se encuentra intrínsecamente ligada a la arquitectura específica del modelo propuesto. No obstante, otro aspecto de suma relevancia es la función de pérdida, comúnmente conocida como "*loss*". Esta función de pérdida representa una métrica que cuantifica la discrepancia entre las predicciones del modelo y los valores reales en el conjunto de datos de entrenamiento. La elección de la función de pérdida está íntimamente relacionada con la naturaleza del problema que se busca resolver y el tipo de modelo empleado. El objetivo central durante el proceso de entrenamiento radica en la minimización del valor asignado por la función de pérdida.

2.3.1. Funciones de *loss*

A continuación se describen algunas de las funciones de *loss* más populares utilizadas en el aprendizaje automático.

Error Cuadrático Medio

El Error Cuadrático Medio (*Mean Squared Error*, MSE), es una función de pérdida que se utiliza ampliamente en problemas de regresión. Su fórmula se expresa como la Ecuación 2.2.

$$MSE(y, \widehat{y}) = \frac{1}{N}\sum_{i=1}^{N}(y_i - \widehat{y}_i)^2 \tag{2.2}$$

Donde:

- y representa los valores reales u objetivos.
- $\widehat{y}$ representa las predicciones del modelo.
- N es el número total de muestras en el conjunto de datos.

El MSE calcula el promedio de las diferencias al cuadrado entre los valores reales y las predicciones del modelo. Cuanto menor sea el valor del MSE, mejor será el ajuste del modelo a los datos. Sin embargo, el MSE es sensible a los valores atípicos (*outliers*) porque eleva al cuadrado las diferencias. Esto significa que los errores grandes tienen un impacto significativamente mayor en el MSE que los errores más pequeños, lo que puede ser problemático en conjuntos de datos con *outliers*.

Error Absoluto Medio

La función de pérdida de Error Absoluto Medio (*Mean Absolute Error*, MAE) es una métrica comúnmente utilizada en problemas de regresión para evaluar la diferencia absoluta promedio entre las predicciones del modelo y los valores reales. Su fórmula general se expresa como la Ecuación 2.3.

$$MAE(y, \widehat{y}) = \frac{1}{N}\sum_{i=1}^{N}|y_i - \widehat{y}_i| \tag{2.3}$$

Donde:

- y representa los valores reales u objetivos.

- $\widehat{y}$ representa las predicciones del modelo.
- N es el número total de muestras en el conjunto de datos.

El MAE mide la magnitud promedio de los errores absolutos entre las predicciones y los valores reales. Es una métrica robusta que no penaliza los errores de manera exponencial, como lo hace el Error Cuadrático Medio (MSE), por lo que es menos sensible a valores atípicos (*outliers*). Si un modelo tiene un MAE más bajo, generalmente indica una mejor capacidad de predicción en términos de la magnitud de los errores.

Entropía Cruzada

La función de pérdida de Entropía Cruzada (*Cross-Entropy Loss*, CE *Loss*), es ampliamente utilizada en problemas de clasificación y es especialmente común en tareas de aprendizaje profundo. Su fórmula general se expresa como 2.4.

$$Loss_{CE}(y, \widehat{y}) = -\frac{1}{N}\sum_{i=1}^{N}(y_i * log(\widehat{y}_i) + (1 - y_i) * log(1 - \widehat{y}_i)) \tag{2.4}$$

Donde:

- y representa los valores reales u objetivos, usualmente codificados.
- $\widehat{y}$ representa las predicciones del modelo, generalmente son valores de probabilidad que se obtienen a través de funciones de activación como la función sigmoide o *softmax*.
- N es el número total de muestras en el conjunto de datos.

La función de *Cross-Entropy Loss* es útil en la optimización de modelos de clasificación, ya que penaliza eficazmente las predicciones incorrectas y fomenta la convergencia hacia una distribución de probabilidad que se asemeje a las etiquetas reales.

2.3.2. Redes Neuronales Convolucionales

Las Redes Neuronales Convolucionales (*Convolutional Neural Networks*, CNN) son modelos de redes neuronales que se destacan por su capacidad para llevar a cabo operaciones de convolución. En el ámbito de la visión por computador, si bien las CNN han demostrado ser altamente efectivas en tareas relacionadas con el análisis de imágenes, cuando se trata del análisis de videos, es crucial considerar tanto las características espaciales (aquello que sucede en cuadros individuales) como las características temporales (las relaciones y cambios a lo largo del tiempo).

En este contexto, se han desarrollado diversas arquitecturas de CNN específicamente diseñadas para abordar la complejidad de la información presente en los videos. En primer

lugar, se encuentran las CNN 2D, que están diseñadas para procesar cada cuadro de video de manera individual. En este enfoque, cada cuadro se considera como una imagen estática y se aplican capas convolucionales 2D para extraer características espaciales. Aunque este enfoque tiene la ventaja de ser computacionalmente eficiente, no logra capturar completamente las relaciones temporales entre los cuadros, lo que puede ser esencial en muchas aplicaciones de análisis de video. En la Figura 2.4 se observa una operación de convolución 3D de forma gráfica.

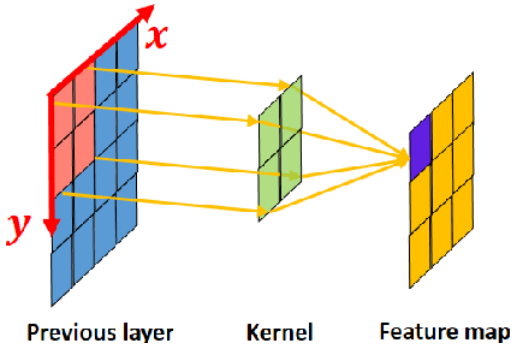


Figura 2.4: Operación de convolución con kernel 2x2. Fuente: [17]

Por otra parte se encuentran las CNN 3D. Estas arquitecturas amplían el concepto de las CNN 2D al agregar una dimensión temporal. En lugar de procesar cuadros de manera individual, las capas convolucionales 3D se utilizan para aprender patrones en secuencias de cuadros. Esta extensión permite la captura de características temporales de manera directa a través de la convolución 3D, lo que resulta en una representación más completa de la información en videos y es fundamental para tareas que requieren un entendimiento profundo de la evolución temporal en los datos visuales. En la Figura 2.5 se observa una operación de convolución 3D de forma gráfica.

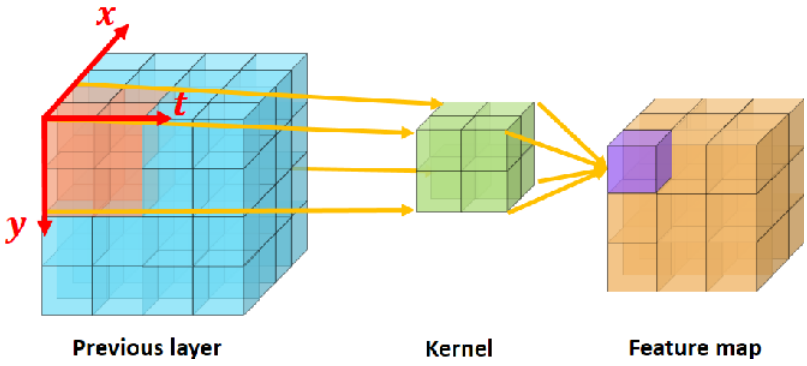


Figura 2.5: Operación de convolución con kernel 2x2x2. Fuente: [17]

El empleo de convoluciones en dos y tres dimensiones (2D y 3D) ha generado una interesante amalgama de enfoques híbridos. Uno de estos enfoques destacados es la fusión de convoluciones 2D y 1D, donde las convoluciones 2D son aplicadas de manera individual a cada cuadro, mientras que las convoluciones 1D operan a lo largo de la dimensión temporal. Este enfoque híbrido ofrece la capacidad de crear representaciones efectivas que abarcan tanto la información espacial como la temporal de manera simultánea.

Además, las convoluciones 2D se han combinado con redes neuronales recurrentes (RNN) para construir modelos que aprovechan tanto las convoluciones espaciales como las secuencias de tiempo. Esta combinación se ha utilizado con éxito en tareas que requieren un entendimiento profundo de la dinámica temporal y la interacción entre cuadros en secuencias de video.

2.3.3. *Transformers*

Esta arquitectura, introducida por primera vez en el artículo "*Attention Is All You Need*" [29] de Vaswani et al. en 2017, ha redefinido la manera en que las máquinas comprenden y generan datos secuenciales. Los Transformers, en particular, han dejado una profunda huella en los campos del Procesamiento del Lenguaje Natural (NLP) y la Visión por Computadora (*Computer Vision*), impulsando el surgimiento de modelos de lenguaje de vanguardia y desencadenando avances significativos en diversas aplicaciones de inteligencia artificial.

En su esencia, los Transformers se destacan por su capacidad de modelar relaciones a larga distancia en datos secuenciales, superando las limitaciones de las arquitecturas recurrentes tradicionales. Esta innovación se logra a través de un mecanismo de atención (*attention mechanism*) que permite a la red asignar diferentes niveles de importancia a las partes relevantes de una secuencia, sin importar cuán lejanas estén entre sí. Esta atención flexible y contextual ha demostrado ser esencial en la comprensión y generación de lenguaje natural, así como en la resolución de tareas de visión por computadora, como la detección de objetos y el procesamiento de imágenes médicas.

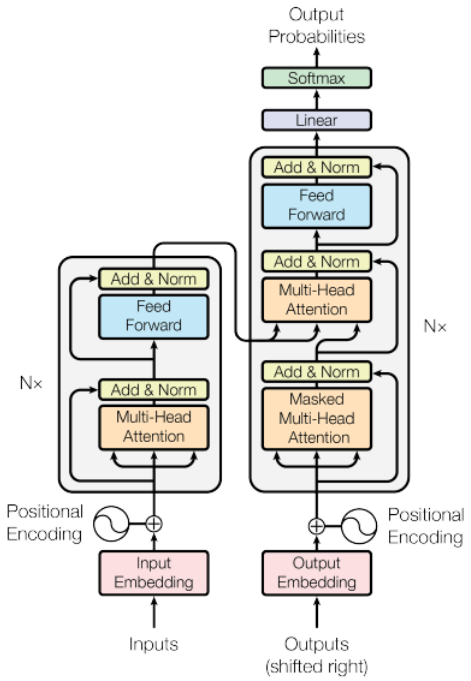


Figura 2.6: Arquitectura *Transformer*. Fuente: [29]

Como se observa en la Figura 2.6, la arquitectura Transformer sigue esta estructura ge-

neral utilizando capas apiladas de autoatención y capas completamente conectadas punto a punto tanto para el codificador como para el decodificador. El Transformer es una arquitectura de redes neuronales profundas ampliamente conocida y esencial en el campo del procesamiento de lenguaje natural y otras aplicaciones de aprendizaje automático. Para obtener una comprensión más detallada de esta arquitectura, se recomienda referirse al artículo original de Transformer [29].

2.4. Procesamiento del Lenguaje Natural

La Procesamiento del Lenguaje Natural (*Natural language processing*, NLP) es un campo multidisciplinario de la inteligencia artificial que se centra en la interacción entre las computadoras y el lenguaje humano. En la era digital actual, donde la información fluye en cantidades masivas a través de texto escrito, el NLP emerge como una disciplina esencial que permite a las máquinas entender, interpretar y generar lenguaje humano de manera efectiva.

2.4.1. Tareas

Las tareas en el Procesamiento del Lenguaje Natural (NLP) abarcan un amplio espectro de aplicaciones que se centran en la interacción entre las máquinas y el lenguaje humano. Desde el análisis de texto hasta la generación de texto, estas tareas se han convertido en un pilar esencial de la inteligencia artificial moderna. En este contexto, exploraremos algunas de las tareas más prominentes del NLP, incluyendo la traducción automática, el reconocimiento de entidades nombradas (NER), la corrección ortográfica y la representación de oraciones. Cada una de estas tareas desempeña un papel crucial en nuestra capacidad de comprender y utilizar el lenguaje de manera efectiva en el mundo digital actual.

Corrección ortográfica

La corrección ortográfica [5] (*Spell Checking*) es una tarea que va más allá de la simple identificación de errores tipográficos. Los sistemas modernos utilizan algoritmos avanzados de corrección que tienen en cuenta el contexto del texto y sugieren correcciones precisas, mejorando la calidad de la escritura y la comunicación en línea. En la Figura 2.7 se ilustra un ejemplo de la finalidad de esta tarea.

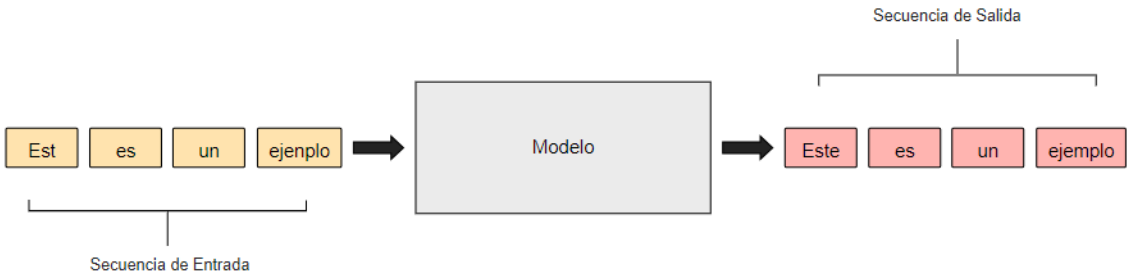


Figura 2.7: Ejemplo de corrección ortográfica.

Traducción automática

La traducción automática [2] (*Machine translation*, MT) es una de las tareas más impactantes del NLP, permitiendo la conversión instantánea de texto de un idioma a otro. Desde las primeras herramientas de traducción hasta los sistemas de traducción neuronal de última generación, esta tarea ha desempeñado un papel fundamental en la superación de barreras lingüísticas y en la promoción de la comunicación global. En la Figura 2.8 se ilustra un ejemplo de la finalidad de esta tarea.

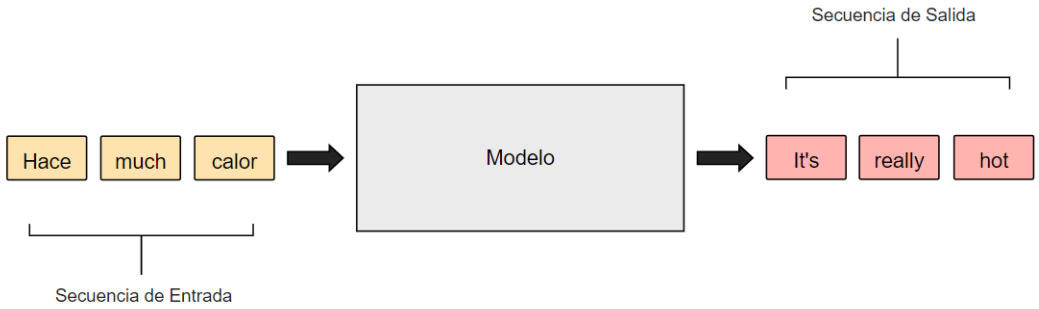


Figura 2.8: Ejemplo de traducción automática

Reconocimiento de Entidades Nombradas

El reconocimiento de entidades nombradas [3] (*Named Entity Recognition*, NER) es una tarea que implica la identificación y clasificación de nombres propios, como nombres de personas, organizaciones y ubicaciones, en un texto. Esto es esencial para la extracción de información y la comprensión de los contextos en los que se utilizan estas entidades, lo que facilita la búsqueda y la recuperación de información relevante. En la Figura 2.9 se ilustra un ejemplo de la finalidad de esta tarea.

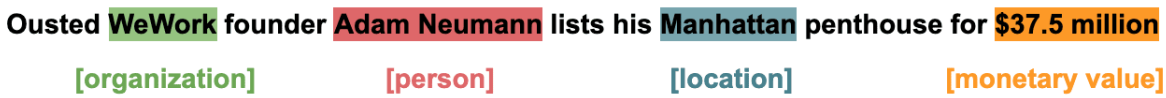


Figura 2.9: Ejemplo Reconocimiento de Entidades Nombradas. Fuente: [24]

Representación de Oraciones

La representación de oraciones [4] (*Sentence Representation*) se enfoca en la creación de representaciones vectoriales que capturan el significado semántico de una oración completa. Estos modelos avanzados permiten el análisis de similitud entre oraciones y la comprensión del contenido en un nivel más profundo, lo que es fundamental para aplicaciones como la recuperación de información y la respuesta automática a preguntas.

2.4.2. Modelos

El campo del Procesamiento del Lenguaje Natural (NLP) ha experimentado un avance espectacular en las últimas décadas gracias a la aplicación de modelos de aprendizaje profundo. Estos modelos se han convertido en la columna vertebral de muchas aplicaciones de NLP, permitiendo tareas que van desde la traducción automática hasta la generación de texto más coherente y significativa. En esta exploración, examinaremos algunos de los modelos más influyentes en el NLP contemporáneo, destacando a BERT, RoBERTa y otros que han revolucionado la forma en que las máquinas comprenden y generan lenguaje humano.

BERT

BERT (*Bidirectional Encoder Representations from Transformers*) [14] es uno de los modelos más destacados en la historia del NLP. Introducido por Google en 2018, BERT se basa en la arquitectura Transformer y se destaca por su capacidad para comprender el contexto bidireccional de las palabras en una oración. Esta característica revolucionaria ha mejorado significativamente el rendimiento en una variedad de tareas de procesamiento de lenguaje, como la respuesta a preguntas, la traducción automática y el resumen de texto.

RoBERTa

RoBERTa (*A Robustly Optimized BERT Pretraining Approach*) [22] es una variante de BERT que se ha destacado por su enfoque en el pre-entrenamiento sin tareas (task-agnostic pretraining). Al ajustar varios hiperparámetros y entrenar en grandes conjuntos de datos, RoBERTa logró mejorar aún más el rendimiento en comparación con BERT en tareas de NLP. Esta adaptación refinada ha demostrado la importancia de la configuración del modelo y el pre-entrenamiento en el éxito de las aplicaciones de NLP.

Flair

Flair [9] es otro modelo destacado que se centra en la comprensión de contexto y el análisis de sentimientos. Utiliza una arquitectura basada en redes neuronales recurrentes bidireccionales (BiLSTM) para tareas como la identificación de entidades nombradas y el análisis de sentimientos. Flair se distingue por su capacidad de trabajar en varios idiomas y por su enfoque en tareas de procesamiento de lenguaje específicas.

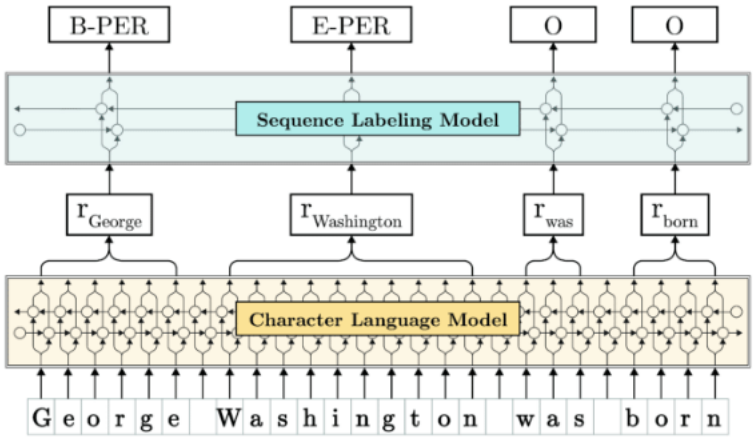


Figura 2.10: Arquitectura Flair. Fuente: [10]

2.5. Visión por Computador

La visión por computadora en el ámbito de los videos es una rama de la inteligencia artificial dedicada a enseñar a las máquinas a entender y procesar secuencias de imágenes en movimiento. Mientras que la visión por computadora tradicional se centra en imágenes fijas, este campo se enfoca en analizar cómo evolucionan las escenas a lo largo del tiempo. Desde la detección de objetos en movimiento hasta la generación de subtítulos automáticos, estas capacidades tienen aplicaciones en diversas industrias, desde la vigilancia hasta la medicina y el entretenimiento. Con el advenimiento del aprendizaje profundo y la disponibilidad de conjuntos de datos masivos de videos, este campo está experimentando un rápido progreso y desempeñará un papel cada vez más significativo en el mundo actual.

2.5.1. Tareas

La visión por computadora en el ámbito de videos aborda una serie de tareas que implican analizar y comprender secuencias de imágenes en movimiento. Estas tareas son fundamentales en una variedad de aplicaciones, desde la detección de objetos en videos de vigilancia hasta la generación de descripciones automáticas para contenido multimedia. A continuación, se describen algunas de las tareas clave en visión por computadora para videos.

Detección de Objetos en Videos

La detección de objetos en videos [7] (*Video Object Detection*) es similar a la detección de objetos en imágenes estáticas, esta tarea implica identificar y ubicar objetos específicos en secuencias de video en tiempo real. Es fundamental en aplicaciones como la monitorización de tráfico, la seguridad y la vigilancia, así como la navegación de vehículos autónomos. En la Figura 2.11 se ilustra un ejemplo de la finalidad de esta tarea.

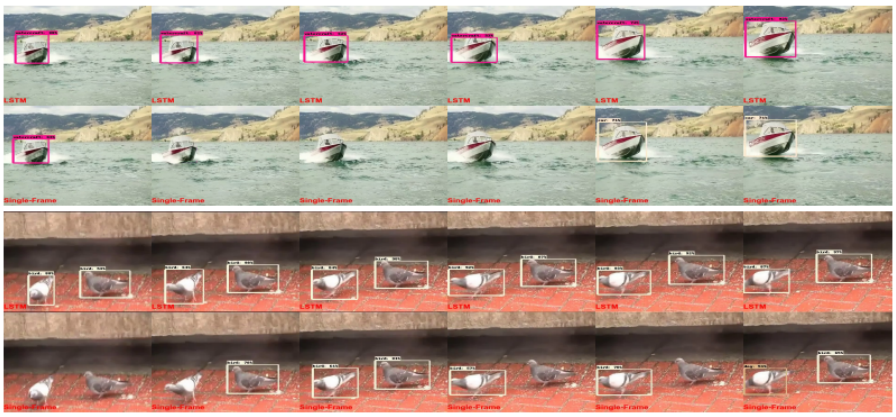


Figura 2.11: Ejemplo de detección de objetos en videos. Fuente: [21]

Seguimiento de Objetos en Videos

El seguimiento de objetos en videos [8] (*Video Object Tracking*) es una tarea que consiste en rastrear la trayectoria de un objeto específico a lo largo de un video. Puede utilizarse para el seguimiento de objetos en movimiento en aplicaciones de seguimiento de deportes, seguimiento de vehículos y seguimiento de personas. En la Figura 2.12 se ilustra un ejemplo de la finalidad de esta tarea.

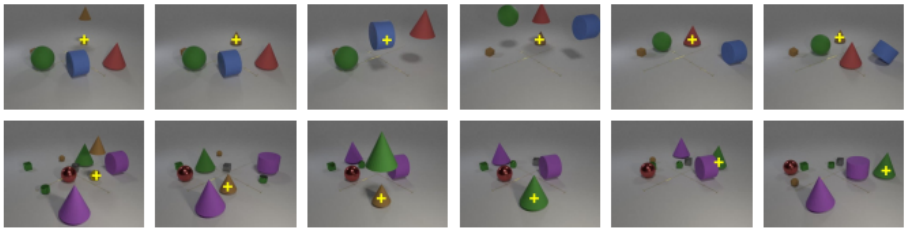


Figura 2.12: Ejemplo de seguimiento de objetos en videos. Fuente: [28]

Reconocimiento de Acciones en Videos

El reconocimiento de acciones en videos [1] (*Video Action Recognition*) es crucial para identificar y clasificar actividades presentes en un video, como caminar, correr, nadar u otras acciones. Esta tarea resulta fundamental en aplicaciones como el análisis de videos de vigilancia y la supervisión de comportamientos. Por ejemplo, puede utilizarse para detectar

actividades sospechosas o para analizar el movimiento de personas en áreas de seguridad. Además, esta labor contribuye significativamente a enriquecer la comprensión del contexto y la dinámica de un video.

Además de su aplicación directa, esta tarea también desempeña un papel fundamental en el entrenamiento de modelos para aprender un espacio de representaciones latentes de alta calidad. Al enfrentarse a una variedad de datos y escenarios, los modelos deben prestar atención a diversas acciones y contextos, lo que resulta en una comprensión más rica y detallada del contenido visual. Esto contribuye a mejorar la capacidad del modelo para generalizar y reconocer patrones en diferentes entornos, lo que es esencial para su desempeño en aplicaciones del mundo real. La Figura 2.13 ilustra un ejemplo que ejemplifica la aplicación de esta tarea..

Figura 2.13: Ejemplo de reconocimiento de acciones en videos. Fuente: [30]

Catalogación de Videos

La tarea de catalogación automática de videos en medios televisivos se refiere al proceso de asignar etiquetas o metadatos a los videos de una biblioteca o base de datos de contenido visual. Estos metadatos o etiquetas pueden incluir información como la descripción del video [6], el tema, los objetos o personas que aparecen en él, el género, las palabras clave y otros detalles relevantes. Bajo lo anterior, en este trabajo hacemos referencia a la tarea de catalogación de videos a este mismo trabajo de manera automatizada, sin la intervención manual de etiquetadores humanos

Catalogación de Videos - Contexto Canal

La catalogación de videos es un desafío significativo para los medios televisivos, especialmente considerando la vasta cantidad de contenido de video que se genera diariamente. El proceso de catalogación manual, que implica la visualización y etiquetado individual de cada video, se convierte en una tarea laboriosa y costosa en términos de tiempo y recursos humanos.

Esta catalogación manual, además de ser intensiva en tiempo, puede ser subjetiva y susceptible a errores humanos. Esto, a su vez, puede dar lugar a inconsistencias y falta de precisión en los metadatos asignados a los videos, lo que dificulta la búsqueda y recuperación

eficiente de contenido relevante en una extensa biblioteca de medios.

Un dato destacable es que, en este contexto, se ha observado que el proceso de catalogación manual por parte de los bibliotecarios puede llevar hasta aproximadamente dos meses desde la creación del video hasta su completa catalogación en el canal televisivo. Esta demora puede afectar significativamente la disponibilidad y accesibilidad del contenido para los espectadores, lo que resalta aún más la necesidad de soluciones eficaces para la catalogación automática de videos en medios televisivos.

2.5.2. Modelos

Los modelos de representación de videos son una categoría de modelos de aprendizaje automático diseñados específicamente para analizar, capturar y representar las características, patrones y relaciones presentes en los videos. Estos modelos se han convertido en herramientas fundamentales para una amplia gama de aplicaciones relacionadas con el procesamiento de contenido visual en movimiento. La elección de la arquitectura adecuada para un modelo de representación de video puede variar, y algunos de los enfoques y arquitecturas más destacados en este campo incluyen las redes neuronales convolucionales (CNN), que pueden ser 2D, 3D o híbridas, y los modelos basados en transformers. Estos enfoques permiten capturar tanto las características espaciales como las temporales presentes en los videos, lo que resulta en representaciones ricas y significativas para tareas de análisis y procesamiento de videos.

SlowFast

El modelo SlowFast [15] es un enfoque de representación de videos que fue publicado el 29 de octubre de 2019. Este modelo ha sido pre-entrenado en el conjunto de datos Kinetics-400 (K400). Su arquitectura se basa en el uso de una red neuronal convolucional en 3D (3D CNN) que procesa dos flujos diferentes, uno rápido y uno lento, de forma separada como se observa en la Figura 2.14. El flujo lento se enfoca en capturar detalles espaciales, mientras que el flujo rápido está diseñado para capturar información temporal rápidamente y también información espacial. Además, se aplican "*skip connections*" para mantener la información de alta resolución a través de las capas de *pooling*. Esta metodología puede aplicarse a distintos *backbones*, pero en [15] utilizan ResNet-50.

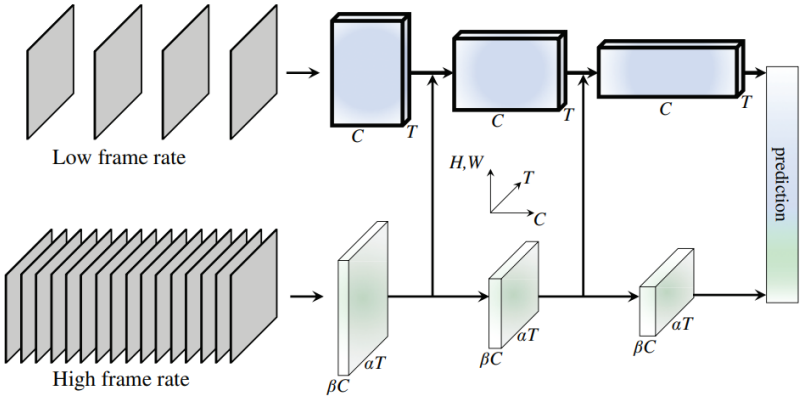


Figura 2.14: Arquitectura SlowFast. Fuente: [15]

TimeSformer

TimeSformer [13] es un modelo de representación de videos que fue publicado en la conferencia ICML 2021 el 9 de junio de 2021. Este modelo ha sido pre-entrenado en varios conjuntos de datos populares en la comunidad de visión por computadora, incluyendo Kinetics-400 (K400), Kinetics-600 (K600), Something-Something-V2 (SSv2) y HowTo100M. La arquitectura del TimeSformer se basa en un enfoque basado en Transformer libre de convoluciones, donde aprende directamente de los parches de los cuadros con un esquema de "atención dividida". Esto significa que la atención temporal y la atención espacial se aplican por separado dentro de cada bloque del modelo como se observa en la Figura 2.15, permitiendo una mejor captura de las relaciones temporales y espaciales en los videos.

La arquitectura consiste en *3D Patch Partition Layer*, una capa de *embedding*, una serie de bloques de (T+S) Transformers y una capa final MLP (*Multi Layer Perceptron*).

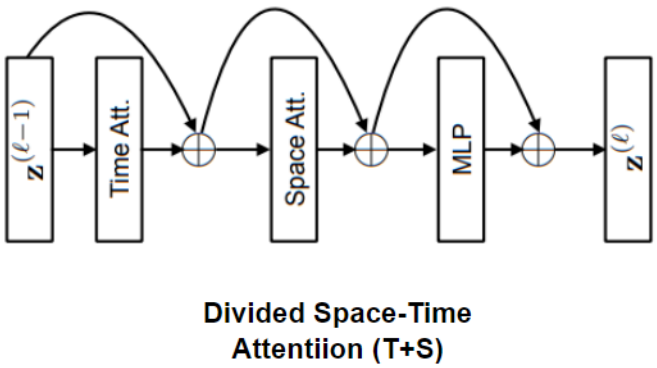


Figura 2.15: Arquitectura TimeSformer Transformer. Fuente: [13]

ViViT (FE)

ViViT *Factorised Encoder* (FE) [12] es un modelo de representación de videos que fue presentado en la conferencia ICCV 2021 el 1 de noviembre de 2021. Este modelo ha sido pre-entrenado en varios conjuntos de datos populares en la comunidad de visión por computadora, incluyendo Kinetics-400 (K400), Kinetics-600 (K600) y Epic Kitchens. La arquitectura del Vivit *Factorised Encoder* se basa en un enfoque basado en Transformer, donde se separa la codificación espacial y temporal en diferentes ramas de la red como se observa en la Figura 2.16. La entrada a la rama temporal es una representación latente de *tokens* extraidos del mismo indice temporal como se observa en la Figura 2.17.

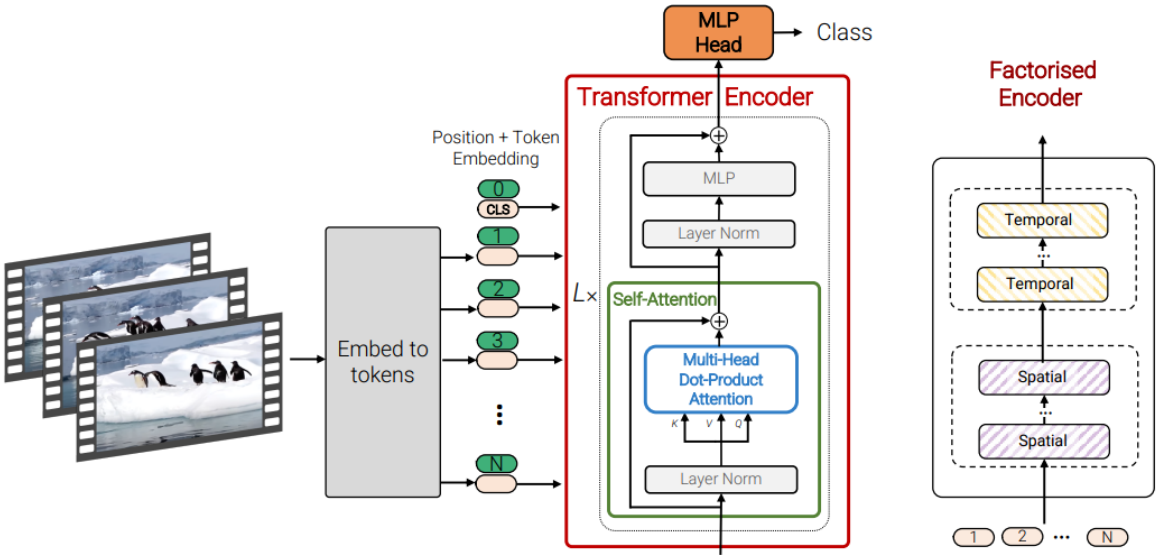


Figura 2.16: Arquitectura ViViT *Factorised Encoder*. Fuente: [12]

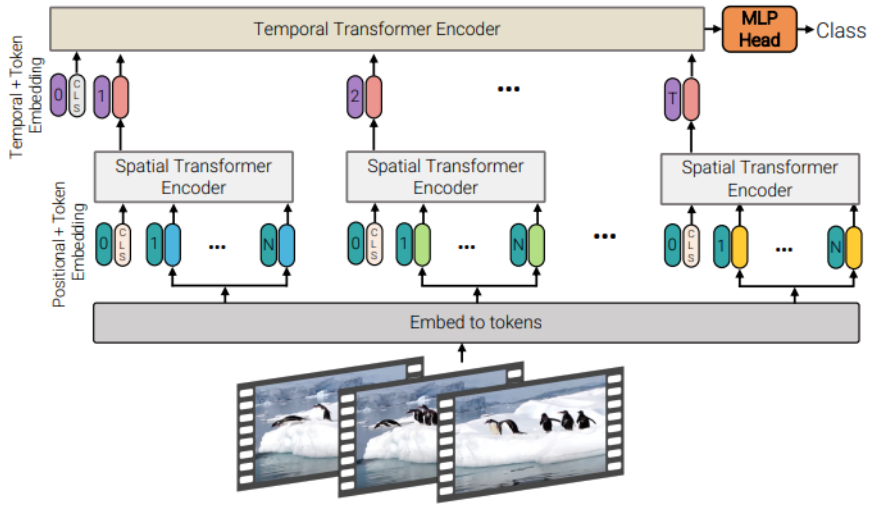


Figura 2.17: Arquitectura ViViT *Factorised Encoder*. Fuente: [12]

Video Swin Transformer

Video Swin Transformer [23] es un modelo de representación de videos que fue publicado el 24 de junio de 2021. Este modelo ha sido pre-entrenado en conjuntos de datos populares en la comunidad de visión por computadora, incluyendo Kinetics-400 (K400), Kinetics-600 (K600) y Something-Something-V2 (SSv2). La arquitectura del Video Swin Transformer se basa en un enfoque basado en Transformer, sin embargo, se diferencia de otros modelos en que reemplaza la "atención multi cabezal" por una "ventana deslizante". Además, también aplica convoluciones temporales y espaciales en la primera instancia del modelo, lo cual lo hace único en su enfoque de procesamiento de videos.

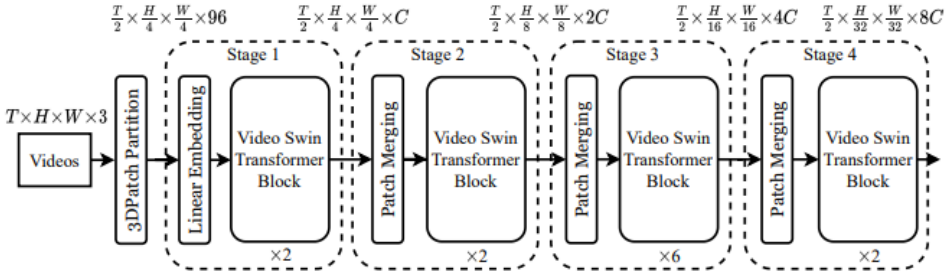


Figura 2.18: Arquitectura Video Swin Transformer. Fuente: [23]

En la Figura 2.18, se presenta una visualización de la arquitectura del Video Swin Transformer, un modelo diseñado específicamente para abordar la tarea de procesar contenido de video. En esta arquitectura, T representa los cuadros individuales, y cada cuadro contiene HxWx3 píxeles, donde H denota la altura, W el ancho, y 3 los canales de color (rojo, verde y azul). La arquitectura consiste en *3D Patch Partition Layer*, una capa de *embedding*, una serie de bloques de Video Swin Transformer (Figura 2.18) y capas adicionales de *patch merging*.

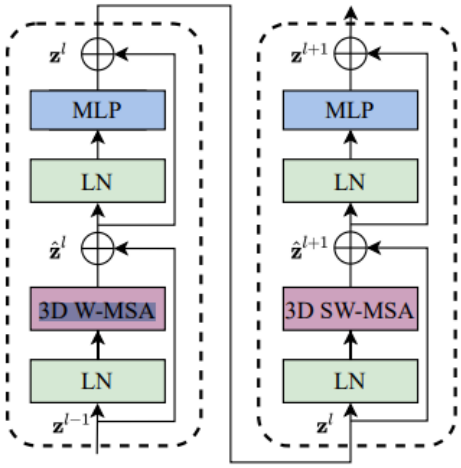


Figura 2.19: 2 bloques de Video Swin Transformer consecutivos. Fuente: [23]

VideoMAE

VideoMAE [27] es un modelo de representación de videos que fue publicado por primera vez en la versión de la cámara lista para NeurIPS 2022 el 18 de octubre de 2022. Este modelo ha sido pre-entrenado en varios conjuntos de datos ampliamente utilizados en la comunidad de visión por computadora, incluyendo Kinetics-400 (K400), Something-Something-V2 (SSv2) y UCF101. La arquitectura del VideoMAE utiliza como *backbone* ViT (Vision Transformer clásico) con una modalidad de Espacio-Tiempo unida (*Joint Space-Time*) como se observa en la Figura 2.20 y un enfoque generativo para aprender representaciones a partir de la reconstrucción de videos enmascarados (*Masked Auto Encoders*, MAE) como ilustra la Figura 2.21.

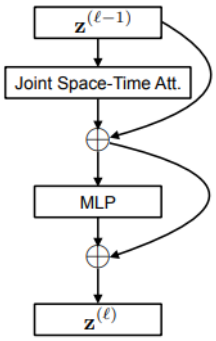


Figura 2.20: Arquitectura Transformer modalidad *Joint Space-Time*. Fuente: [13]

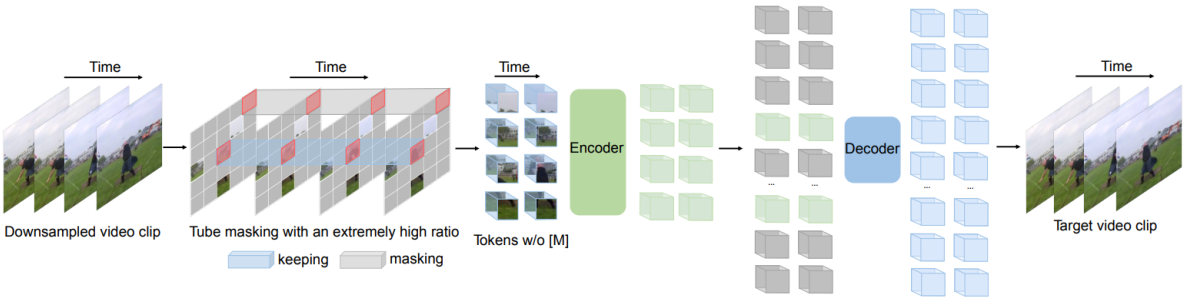


Figura 2.21: Arquitectura VideoMAE. Fuente: [27]

2.6. Métricas

En cuanto a la evaluación del rendimiento de cada red, existen diversas métricas disponibles. En primer lugar, es fundamental establecer las métricas básicas para evaluar si una muestra se clasifica correctamente en relación con una clase. En este contexto, existen cuatro posibles resultados que se deben considerar:

- Verdadero Positivo (TP): Cuando la muestra se clasifica correctamente como la clase esperada.
- Verdadero Negativo (TN): Cuando la muestra no se clasifica como la clase esperada, y esto es correcto, es decir, la muestra no pertenecía a esa clase.
- Falso Positivo (FP): Cuando la muestra se clasifica como la clase esperada, pero esto es incorrecto, ya que la muestra no pertenecía a esa clase.
- Falso Negativo (FN): Cuando la muestra no se clasifica como la clase esperada, pero debería haberlo hecho, es decir, la muestra pertenecía a esa clase.

2.6.1. Índice de Rand Ajustado

El índice de Rand ajustado (*Adjusted Rand Index*, ARI) es una métrica que se utiliza para evaluar la similitud entre los *clusters* obtenidos por un algoritmo de *clustering* y los *clusters* de referencia o *ground truth*. Se define en función de los TP, TN, FP y FN como se muestra en la Ecuación 2.5:

$$ARI = \frac{(RI - Expected\_RI)}{(max(RI) - Expected\_RI)} \tag{2.5}$$

Donde:

- RI (*Rand Index*) viene dado por la Ecuación 2.6:

$$RI = \frac{(TP + TN)}{(TP + TN + FP + FN)} \tag{2.6}$$

- Expected_RI es el valor esperado del índice de Rand para una asignación aleatoria de los datos a los *clusters*, que se calcula teniendo en cuenta las distribuciones marginales de los datos y se utiliza para tener en cuenta la posibilidad de coincidencia aleatoria.

2.6.2. Precisión

La precisión (*Accuracy*) es una métrica de evaluación de modelos que mide la proporción de predicciones correctas en relación con el total de predicciones realizadas. Su fórmula es la Ecuación 2.7.

$$Accuracy = \frac{(TP + FP)}{(N)} \tag{2.7}$$

Donde:

- N: Total de Muestras.

2.6.3. Similitud de Coseno

La similitud de coseno (*Cosine Similarity*) es una métrica que mide la similitud entre dos vectores, comúnmente utilizada en tareas de procesamiento de lenguaje natural y recuperación de información. Para calcular la similitud de coseno entre dos vectores se realiza por medio de la Ecuación 2.8.

$$Similar = Cos(\theta) = \frac{A \cdot B}{\|A\| * \|B\|} \tag{2.8}$$

- A·B: Producto punto entre A y B.
- ‖A‖‖B‖: Son las normas euclidianas de A y B.

Capítulo 3

Metodología

En este capítulo, se aborda en detalle el desarrollo inicial llevado a cabo para abordar la cuestión planteada. Se comienza por describir los conjuntos de datos empleados, los modelos evaluados y el procesamiento de los datos brindados por el canal para el posterior entrenamiento con estos.

3.1. *Datasets*

La elección de los conjuntos de datos se basa en su diversidad y amplitud, independientemente de la tarea específica para la que hayan sido ampliamente utilizados. Esta selección se fundamenta en la intención de aprovechar conjuntos de datos públicos para permitir que los modelos seleccionados aprendan un espacio de representaciones latentes de alta calidad. La idea subyacente es facilitar el aprendizaje transferido, permitiendo que estos modelos funcionen adecuadamente en conjuntos de datos externos a aquellos en los que fueron inicialmente entrenados. Esta estrategia nos permite comenzar con un espacio de representación ya bien aprendido desde el inicio, gracias a la variedad y riqueza de los datos utilizados.

3.1.1. *Kinetics*

Kinetics [18] es conocido por ser utilizado para pre-entrenar modelos unimodales con un enfoque auto-supervisado. Consta de 650.000 videos que cubren 400, 600 o 700 clases de acciones humanas, con una duración aproximada de 10 segundos por video. Cada video está etiquetado con una única clase de acción. Kinetics es particularmente relevante para la transferencia de aprendizaje en tareas de reconocimiento de acciones, ya que permite entrenar modelos en un gran corpus de datos antes de ajustarlos a tareas específicas con conjuntos de datos más pequeños.

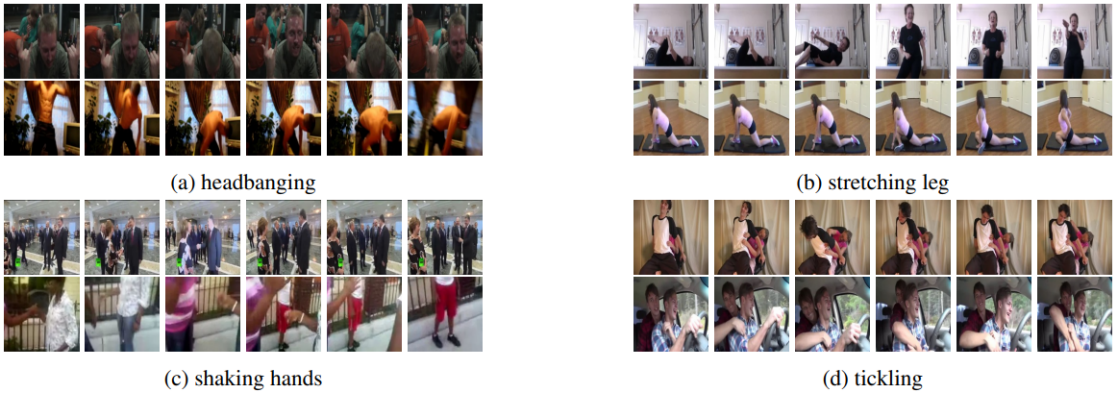


Figura 3.1: Muestra de datos del *dataset* Kinetics. Fuente: [18]

En la Figura 3.2 se muestra la cantidad de instancias por categoría en el *dataset* Kinetics.

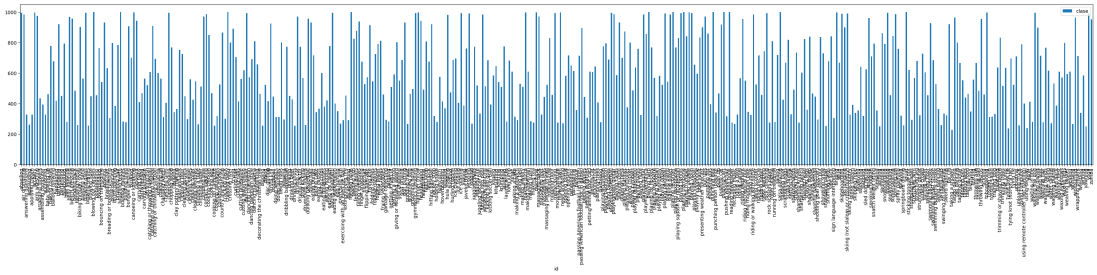


Figura 3.2: Número de instancias anotadas por clase en *dataset* Kinetics.

3.1.2. *UCF101*

UCF101 [25] es ampliamente utilizado en la comunidad de reconocimiento de acciones, ya que contiene 13.320 videos categorizados en 101 categorías distintas. Estos videos son generados por usuarios y recolectados de YouTube, con una resolución normalizada de 320x240 y una velocidad de cuadros de 25 FPS (*Frames Per Second*). Debido a su tamaño reducido, el *dataset* UCF101 es popular para el desarrollo y evaluación de algoritmos de reconocimiento de acciones en entornos con recursos limitados.

Figura 3.3: Muestra de datos del *dataset* UCF101. Fuente: [25]

En la Figura 3.4 se muestra la cantidad de instancias por categoría en el *dataset* UCF101.

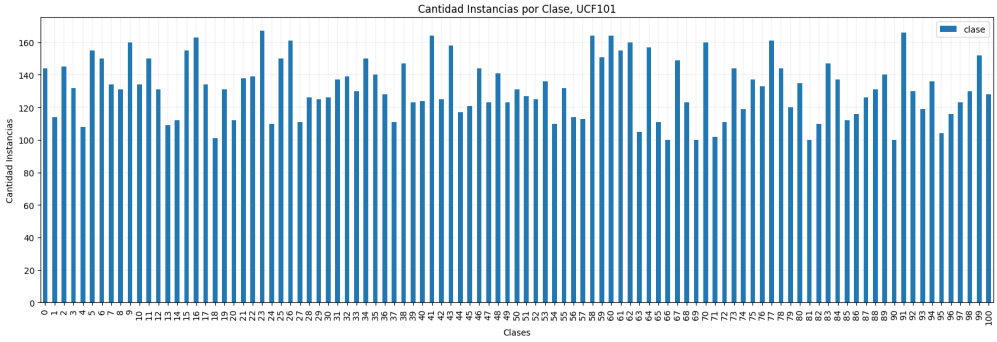


Figura 3.4: Número de instancias anotadas por clase en *dataset* UCF101.

3.1.3. *HMDB51*

HMDB51 [19] es otro *dataset* popular utilizado en el campo del reconocimiento de acciones. Con aproximadamente 7.000 videos distribuidos en 51 clases de acciones distintas, y alrededor de 101 videos de 10 segundos cada uno, HMDB51 ofrece una mayor variabilidad en términos de acciones y escenarios representados en comparación con UCF101. Los videos en HMDB51 son recopilados de diversas fuentes, lo que lo hace adecuado para aplicaciones más realistas y diversas en términos de condiciones de grabación.

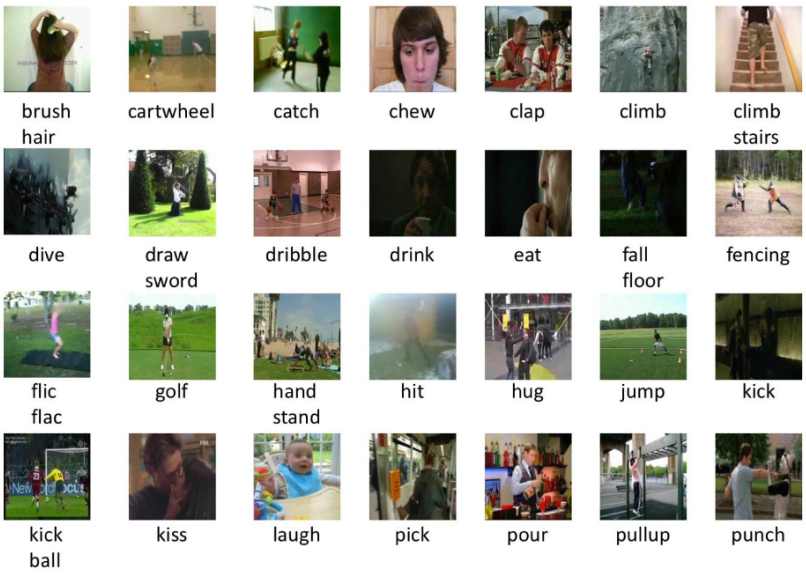


Figura 3.5: Muestra de datos del *dataset* HMDB51. Fuente: [19]

En la Figura 3.6 se muestra la cantidad de instancias por categoría en el *dataset* HMDB51.

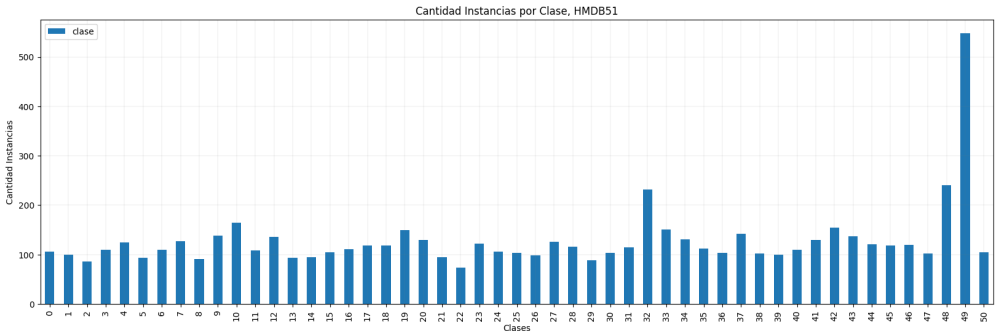


Figura 3.6: Número de instancias anotadas por clase en *dataset* HMDB51.

3.1.4. *Something-Something v2*

Something-Something v2 [16] es un conjunto de datos de videos utilizado en el campo del reconocimiento de acciones que contiene más de 220,000 videos etiquetados de acciones cotidianas. Con una duración de 2-6 segundos por video, capturados en fondos neutros, y más de 200 clases de acciones diferentes, este conjunto de datos es uno de los más grandes y completos en términos de diversidad de acciones representadas. Something-Something v2 ha sido ampliamente utilizado en la investigación de reconocimiento de acciones y ha servido como referencia en la comunidad académica y en la industria, permitiendo el desarrollo y evaluación de modelos de aprendizaje profundo para una amplia gama de aplicaciones en visión por computadora y procesamiento de videos.

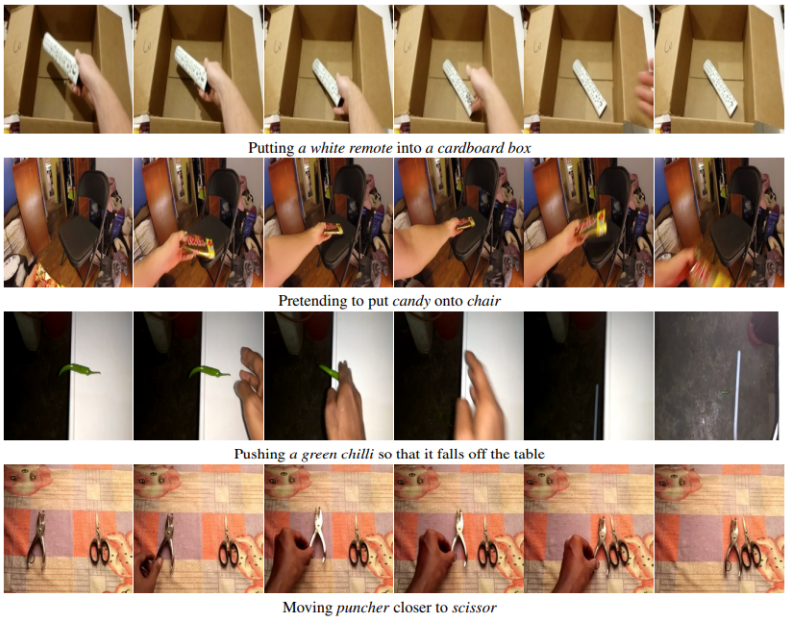


Figura 3.7: Muestra de datos del *dataset* Something-Something v2. Fuente: [16]

En la Figura 3.8 se muestra la cantidad de instancias por categoría en el *dataset* Something-Something v2.

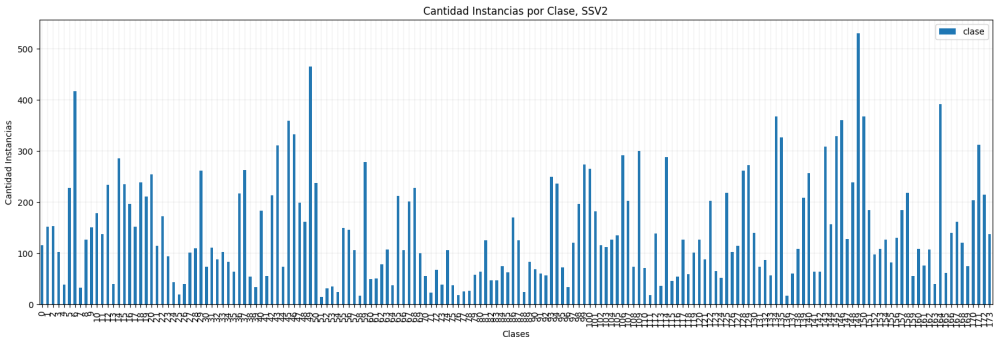


Figura 3.8: Número de instancias anotadas por clase en *dataset* Something-Something v2.

3.1.5. Videos y Metadata Canal Televisivo

Los datos proporcionados por el canal se componen de 266 videos del programa "Tu Día" del año 2022 que tienen una duración aproximada de 4 horas por capítulo, en la Figura 3.9 de ilustra una muestra de estos. Además de los videos, se suministra un archivo JSON asociado a cada video con distintas entidades "entry" que corresponden a distintas secciones del capitulo en cuestión. Cada "entry" contienen las siguientes propiedades:

1. **duration**: Duración de la sección en segundos ("entry").
2. **start**: Inicio de la sección en segundos ("entry").

3. **c13.thematicDescriptors**: Descriptores tematicos de la sección (“entry”), corresponde a una lista que puede tener cero o más descriptores con los temas de la sección del programa y los cuales no se encuentran predeterminados de ninguna forma, estos dependerán del bibliotecario encargado y lo que este vea correspondiente en su momento.
4. **c13.description**: Descripción del video de la sección (“entry”).
5. **c13.assetTitle**: Titulo de la sección (“entry”).

Figura 3.9: Muestra de datos del canal

Un ejemplo de este archivo JSON de datos se puede apreciar en la Figura 3.10.


```
{
"entry0": {
    "duration": "2652.95295295295",
    "start": "0",
    "c13.thematicDescriptors": [],
    "c13.description": "REVISAN VIDEO ASALTO EN AVENIDA EL GOLF, LAS
    CONDES.\nSE PRODUCE TIROTEO, SE VE JAIME ARELLANO  TRABAJADOR ,
     ENFRENTANDO A DELINCUENTES Y LANZANDO SILLAS.\nMOVIL  MARILYN
   PEREZ,  ENTREVISTA NIKOLE ROSENBERG,  (HIJA DE SARIKA RODNIK)
   VICTIMAS DE LA DELINCUENCIA Y VIOLENCIA. JAIME, TRABAJADOR QUE SE
    VE EN EL VIDEO.\n\n19.18 CONTACTO ZOOM DANIELA PE\u00d1ALOZA,
   ALCALDESA DE LAS CONDES. SE REFIEREN AL TEMA.",
    "c13.assetTitle": "CONDUCE: ANGELES ARAYA, MIRNA SCHINDLER."},
"entry1": {
    "duration": "698.264931598265",
    "start": "2676.10944277611",
    "c13.thematicDescriptors": [],
    "c13.description": "47.28 MOVIL ANA MARIA SILVA , AFUERAS
   ESTACION DE METRO. ENTREVISTA PERSONAS ABRIGADAS.",
    "c13.assetTitle": "INFORME DEL TIEMPO GIANFRANCO MARCONE"}
}
```


Figura 3.10: Objeto JSON de ejemplo que hace referencia los datos entregados por el canal.

Observando los metadatos disponibles, se ha decidido emplear los títulos asociados a cada sección como *ground-truth* inicial. Para ello, se generarán representaciones de estos títulos y se llevará a cabo un proceso de *clustering*. Se espera que descripciones similares estén cercanas entre sí, mientras que las diferentes se encuentren separadas en el espacio latente. Una vez obtenidos los *clusters* obtenidos a partir de las representaciones de video y texto, se espera una correspondencia. Esto significa que las descripciones que coincidan en el espacio de texto también se agruparán en el mismo *cluster* en el espacio de video. Esta coherencia entre los *clusters* validará la calidad del espacio latente de los videos.

3.2. *Baseline*

Para medir los efectos de entrenar modelos en el contexto del canal, se parte por entrenar o utilizar redes pre-entrenadas en conjuntos de datos públicos para tener tanto un punto de comparación como poder elegir que modelo evaluar y entrenar con los datos del canal. Para esto se entrenan en Kinetics-400 y evaluan en todos los conjuntos de datos públicos (Kinetics-400, UCF101, HMDB51 Y Something-Something V2) las redes SlowFast, TimeSformer, ViViT (FE), Video Swin Transformer y VideoMAE.

La evaluación de los modelos se llevará a cabo utilizando el índice de Rand ajustado (Adjusted Rand Index, ARI) como métrica principal. Esta elección se basa en la necesidad de medir la calidad de las representaciones de videos generadas por cada una de las estructuras evaluadas. La razón fundamental detrás de esta evaluación se debe a que un espacio de representación deficiente puede dar lugar a descripciones inexactas o confusas, lo que a su vez puede comprometer la utilidad de la catalogación de videos en términos de búsqueda y clasificación.

Por otro lado, un espacio de representación bien aprendido permitirá que el modelo capture de manera efectiva las características visuales relevantes de los videos. Esto, a su vez, facilitará la generación de descripciones precisas y relevantes. En consecuencia, el ARI se utiliza para evaluar cuán bien las representaciones de videos de cada modelo se alinean con las etiquetas o clases reales de los videos. Un ARI alto indicaría que las representaciones reflejan de manera coherente las características y relaciones entre los videos, lo que es fundamental para mejorar la catalogación y la recuperación de videos en el contexto del canal televisivo.

En términos de implementación, se utilizaron las implementaciones proporcionadas en los respectivos documentos de investigación para SlowFast, ViViT (FE) y Video Swin Transformer. Para TimeSformer y VideoMAE, se aprovecharon las implementaciones disponibles en la biblioteca Hugging Face. Cabe destacar que todo el código utilizado en este estudio se encuentra disponible en los repositorios de GitHub correspondientes [1] [2] [3] [4] [5].

[1] `https://github.com/facebookresearch/SlowFast`
[2] `https://huggingface.co/docs/transformers/model_doc/timesformer`
[3] `https://github.com/google-research/scenic/tree/main/scenic/projects/vivit`
[4] `https://github.com/SwinTransformer/Video-Swin-Transformer`
[5] `https://huggingface.co/docs/transformers/model_doc/videomae`

3.3. Procesamiento de Datos

Para el procesamiento de datos entregados por el canal se realizaron las siguientes metodologías.

Para el procesamiento de los videos que consistian en capitulos de 4 horas de duración se realizaron los siguientes pasos para recortarlos en los tiempos correspondientes.

1. Leer un archivo json con la metadata y obtener todas sus "entry" correspondientes.
2. Obtener el tiempo de inicio y la duración de la sección "entry" correspondiente por medio de los atributos "start" y "duration".
3. Por medio de la librería moviepy [6] disponible para Python realizar el cortado y guardar las secciones cortadas en la ruta correspondiente.
4. Repetir todos los pasos hasta cortar todos los videos disponibles en sus secciones correspondientes.

Después de haber dividido los videos en segmentos, se obtuvieron un total de 1505 videos, y el siguiente paso crucial implica el procesamiento de los metadatos. Para evaluar y entrenar modelos de manera efectiva, es esencial contar con etiquetas o clases adecuadas. Sin embargo, al analizar el formato de los metadatos presentados en la Figura 3.10, se identificó que no se disponía de atributos directamente aplicables para estas tareas.

Ante esta situación, se optó por aprovechar el atributo "c13.assetTitle" para crear las clases necesarias. Este proceso implicó una serie de pasos específicos que se detallan a continuación.

1. Leer un archivo json con la metadata y obtener todas sus "entry" correspondientes.
2. Obtener el titulo del "entry" correspondiente por medio del atributo "c13.assetTitle".
3. Se utiliza pyspellchecker [7], una herramienta de corrección ortográfica automática disponible en Python, para realizar las correciones necesarias al texto.
4. Se utiliza GoogleTranslator [8], una herramienta de traducción automática disponible en Python, para realizar la traducción del texto en español al ingles (los modelos funcionan mejor en este idioma).
5. Guardamos la traducción obtenida en el "entry" correspondiente mediante un atributo nuevo llamado "c13.assetTitle_translated"
6. Por medio de distintos modelos de NLP (RoBERTa y su variante para oraciones, *sentence* RoBERTa) se obtienen representaciones vectoriales de los textos.

[6] `https://pypi.org/project/moviepy/`
[7] `https://pypi.org/project/pyspellchecker/`
[8] `https://pypi.org/project/deep-translator/`

7. En el caso de obtener los vectores utilizando RoBERTa, se realiza un promedio de los vectores entregados por el modelos, ya que estos se tratan de los vectores por palabra y no por la oración.

8. Guardamos los vectores en el "entry" correspondiente mediante un atributo nuevo llamado "c13.assetTitle_translated_embeddings"

9. Repetir todos los pasos hasta procesar la toda metadata disponibles.

Se desarrolló una segunda metodología para obtener las representaciones vectoriales de los metadatos. Esta técnica involucra el uso de un modelo NER (Named Entity Recognition) para identificar y suprimir las entidades en el texto, como nombres de personas, lugares, entre otros. Este proceso se insertó como un paso adicional entre el paso 5 y el paso 6 del flujo de trabajo.

Para llevar a cabo esta metodología, se utilizó la biblioteca Flair [9] junto con el modelo NER disponible en ella. En este paso, el texto se sometía a un procesamiento en el cual las entidades identificadas se reemplazaban por etiquetas correspondientes. Por ejemplo, un nombre de persona sería reemplazado por la etiqueta "persona". Una vez que se había realizado esta supresión de entidades, el texto resultante se almacenaba en el atributo "c13.assetTitle_translated_norm", mientras que su vector de representación se guardaba en el atributo "c13.assetTitle_translated_norm_embedding".

Obtenidos los vectores de representación de las oraciones o *sentence embeddings*, se utiliza UMAP [10] para la reducción de dimensionalidad de los datos y posteriormente mediante K-Means, una tecnica de clustering, se obtienen 40 clusters de datos similares entre si, estos cluster pasaran a ser las clases correspondientes a cada video. En la Figura 3.11 se observa un ejemplo de un "entry" de los datos ya procesados.


```
{
    "entry2": {"duration": "297.764431097764", "start": "7764.064064
    06406", "c13.thematicDescriptors": [], "c13.description": "
    INFORMA ACCIDENTE DE TRANSITO, VEHICULO VOLCADO ", "c13.
    assetTitle": "MOVIL ANA MARIA SILVA, PROVIDENCIA.", "c13.
    assetTitle_translated": "ANA MARIA SILVA MOBILE, PROVIDENCIA.", "
    c13.assetTitle_translated_embedding": "-0.0024841686, 0.013695076
    , -0.004167279, -0.02855076, 0.0062146364, ... ,0.022005778, 0.01
    8648963, 0.024808254}
}
```


Figura 3.11: Objeto JSON de ejemplo que hace referencia los datos entregados por el canal ya procesados.

[9] https://github.com/flairNLP/flair
[10] https://umap-learn.readthedocs.io/en/latest/

Capítulo 4

Resultados Iniciales y Discusión

En este capítulo se muestran los resultados obtenidos en los experimentos utilizando las arquitecturas planteadas en el capítulo 3, donde se parte por explicar el hardware y el software utilizados para el desarrollo, y luego se discuten los experimentos junto con sus resultados.

4.1. Hardware y Software

Con respecto al servidor utilizado para el entrenamiento y evaluación de los modelos, en un inicio se contaba con 2 GPUs Nvidia TITAN X, con 12Gb de memoria cada una. Más adelante, se actualizaron ambas GPUs por Nvidia TITAN RTX, de 24Gb de memoria cada una, sin embargo por temas de uso, solo se podía utilizar una de esta en su completitud.

En cuanto al software utilizado, se utilizó Python y los repositorios de Github mencionados en el capitulo anterior 3 para la implementación de funciones y modelos.

4.2. Evaluación en Conjuntos Datos Públicos

Para evaluar la calidad de los espacios de representación generados por los modelos previamente discutidos en el capítulo anterior 3, se optó por utilizar las versiones pre-entrenadas de estos modelos en el conjunto de datos Kinetics-400. Luego, se realizó una selección aleatoria de alrededor de 500 videos por conjunto de datos, cada uno compuesto por 10 clases que contenían aproximadamente 50 videos cada una.

Este enfoque permitió llevar a cabo pruebas exhaustivas de los diferentes modelos en dos contextos cruciales. Primero, se evaluó su rendimiento en los conjuntos de datos en los que fueron inicialmente entrenados, lo que ofrece una medida de su capacidad intrínseca. Luego, se examinó cómo se desempeñaban en conjuntos de datos distintos a los de su entrenamiento original, lo que proporciona información valiosa sobre su capacidad de adaptación y la calidad de las representaciones aprendidas.

4.2.1. Análisis Cuantitativo

La Tabla 4.1 muestra la calidad de los espacios de representación aprendidos por los modelos seleccionados, evaluados en cuatro conjuntos de datos públicos diferentes. Estos resultados permiten observar cómo los modelos se desempeñan en distintos contextos y datasets, lo que proporciona una visión de la adaptabilidad de los modelos y las representaciones que han aprendido.

Para realizar la medición, se utilizó la métrica del Indice Rand Ajustado (ARI) descrita en el Capitulo 2 Sección 2.6.1

	Adjusted Rand Index				
Model	**Kinetics 400**	UCF101	HMDB51	SSv2	Canal
TimeSfomer	**0.951**	0.960	0.541	**0.098**	-
VideoMAE	0.911	**1.000**	**0.873**	0.065	-
ViViT (FE)	0.123	0.072	0.048	0.021	-
Video Swin	0.282	0.524	0.346	0.079	-
SlowFast	0.578	0.968	0.577	0.044	-

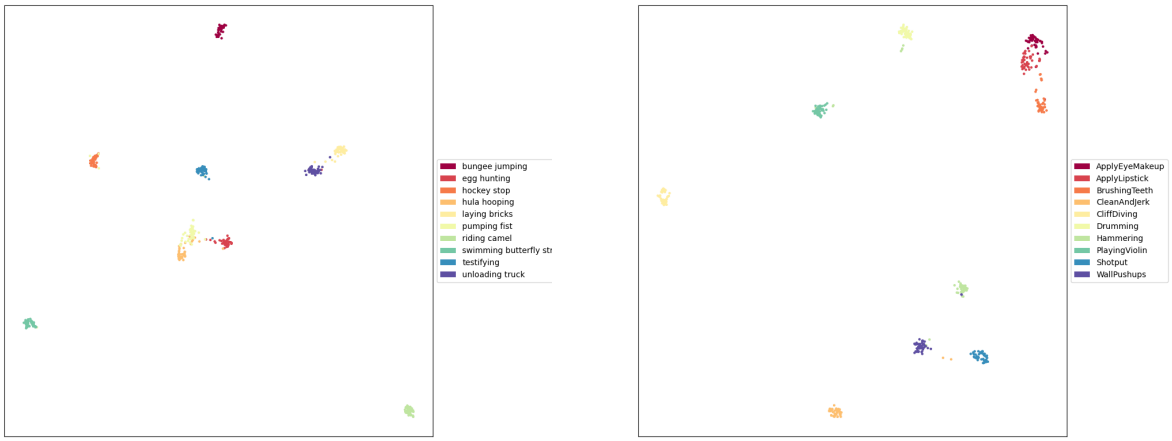
Tabla 4.1: Tabla Adjusted Rand Index para la evaluación de los espacios aprendidos por los modelos en distintos conjuntos de datos públicos.

4.2.2. Análisis Cualitativo

A continuación, se presentan de manera visual los resultados obtenidos por cada modelo en forma de gráficos o representaciones visuales. Estos resultados proporcionan una visión clara y comprensible de la calidad de las representaciones aprendidas por los modelos en diferentes contextos y *datasets*, lo que permite una comparación y análisis detallado de su desempeño en cada caso.

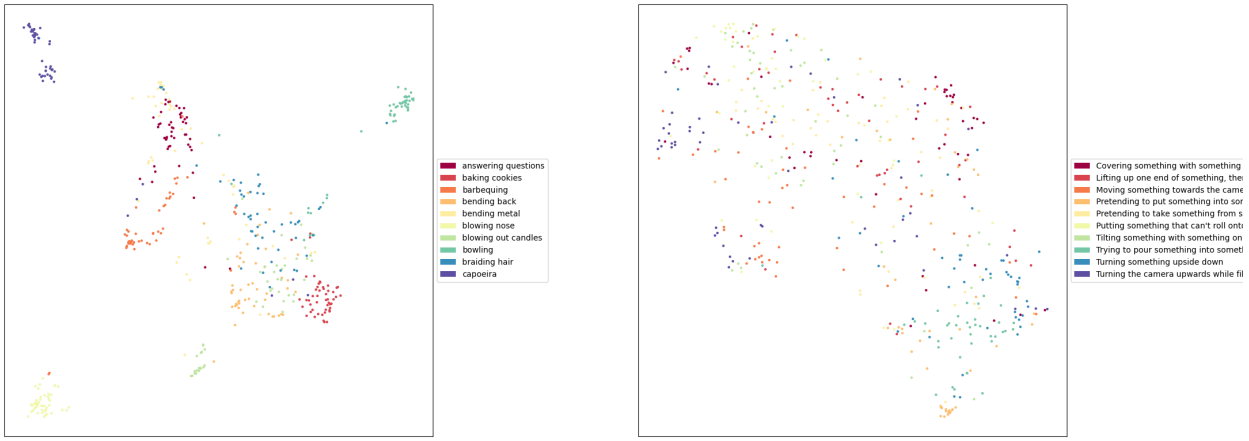
Al realizar estas representaciones visuales, se espera que los espacios latentes de los modelos de representación de videos muestren una clara distinción entre los distintos grupos de clases. Esta distinción se reflejará en la cercanía de los datos pertenecientes a una misma clase o cluster (representados por el color) y su separación de las otras clases o clusters. Cuanto más evidentes sean estas características, mejor será la calidad del espacio aprendido por el modelo en el conjunto de datos evaluado.

- TimeSfomer.



(a) *Dataset* kinetics 400. (b) *Dataset* UCF101.

Figura 4.1: Espacio generado en datasets kinetics 400 y UCF101.

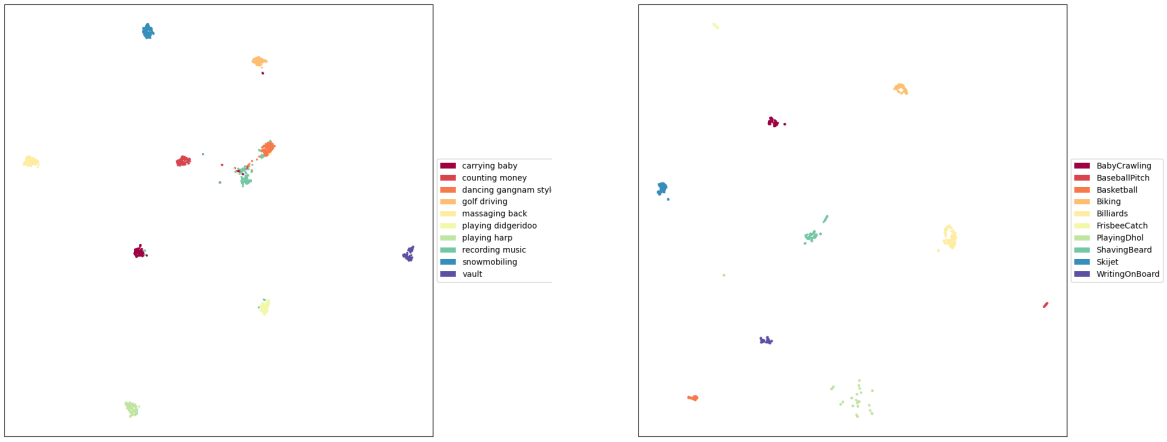


(a) *Dataset* HMDB 51. (b) *Dataset* SSV2.

Figura 4.2: Espacio generado en datasets HMDB 51 y SSV2.

En la Figura 4.1, se observa que con TimeSformer en conjuntos de datos como Kinetics y UCF101, se logra una distinción clara entre los *clusters*, obteniendo resultados sólidos, tal como se esperaba. Por otro lado, al examinar la Figura 4.2 en HMDB51, aunque no se aprecia una distinción tan nítida entre los *clusters* en comparación con los otros dos conjuntos de datos, se puede notar cómo los elementos de una misma clase están mayormente cercanos entre sí. En contraste, en SSV2 se observa una mayor dispersión de los datos en el espacio, sin un orden claro entre ellos.

- VideoMAE.



(a) *Dataset* kinetics 400. (b) *Dataset* UCF101.

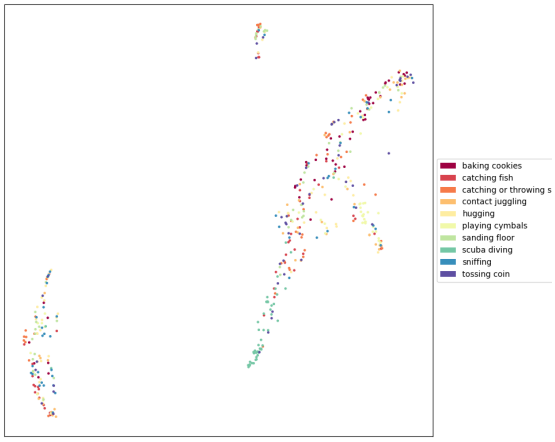
Figura 4.3: Espacio generado en datasets kinetics 400 y UCF101.

(a) *Dataset* HMDB 51. (b) *Dataset* SSV2.

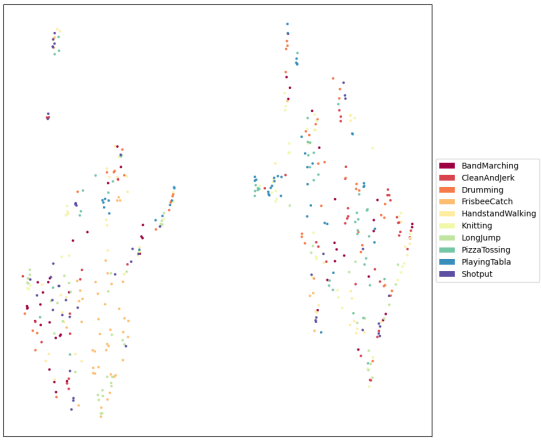
Figura 4.4: Espacio generado en datasets HMDB 51 y SSV2.

En las Figuras 4.3 y 4.4, se observa que con VideoMAE en conjuntos de datos como Kinetics, UCF101 y HMDB51, se logra una distinción clara entre los *clusters*, obteniendo resultados sólidos, tal como se esperaba. Por otro lado, en SSV2 se observa una mayor dispersión de los datos en el espacio, sin un orden claro entre ellos.

- ViViT.

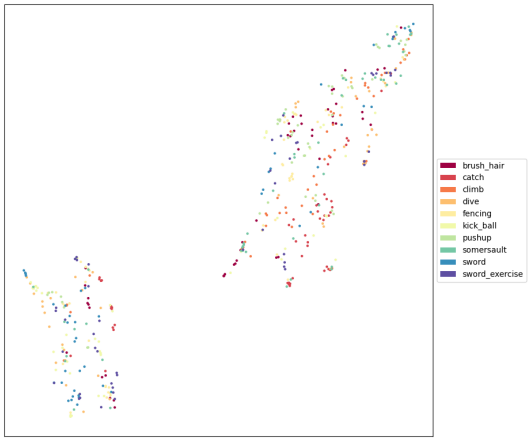


(a) *Dataset* kinetics 400.

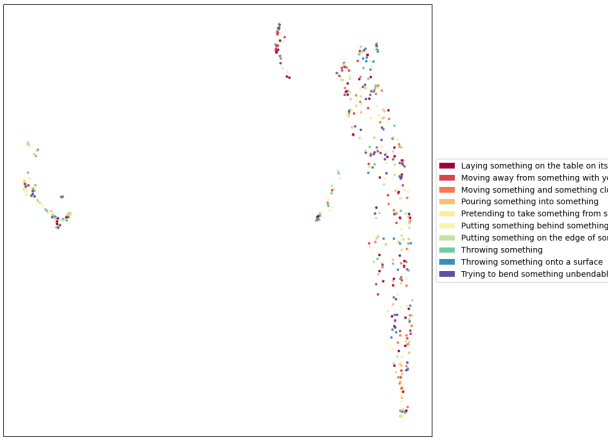


(b) *Dataset* UCF101.

Figura 4.5: Espacio generado en datasets kinetics 400 y UCF101.



(a) *Dataset* HMDB 51.

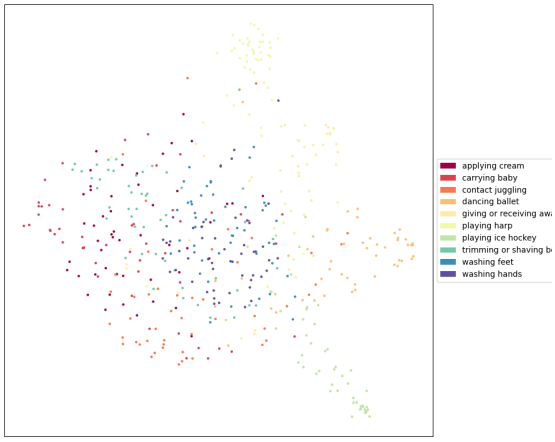


(b) *Dataset* SSV2.

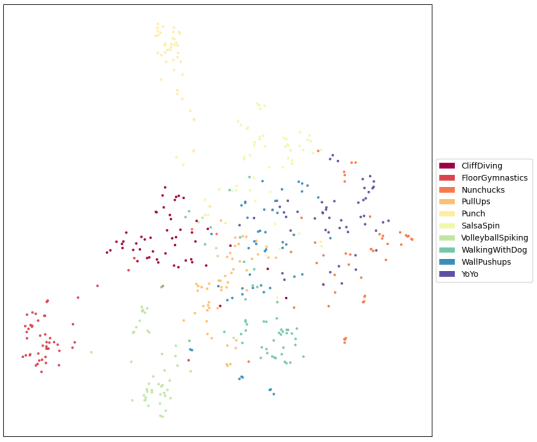
Figura 4.6: Espacio generado en datasets HMDB 51 y SSV2.

En las Figuras 4.5 y 4.6, se observa que con ViViT (FE) como en los conjuntos de datos Kinetics, UCF101, HMDB51 y SSV2, a pesar de haber figuras definidas en el espacio, no existe un orden claro entre los datos evaluados y en ninguno de los conjuntos de datos evaluado.

- Video Swin Transformer.

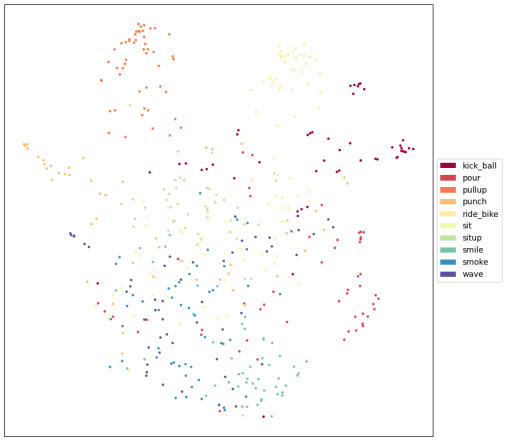


(a) *Dataset* kinetics 400.

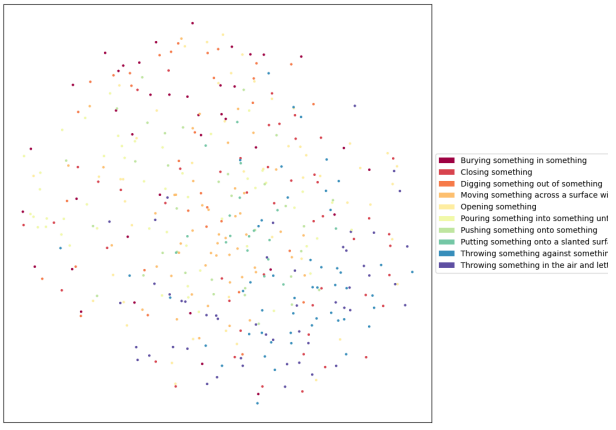


(b) *Dataset* UCF101.

Figura 4.7: Espacio generado en datasets kinetics 400 y UCF101.



(a) *Dataset* HMDB 51.

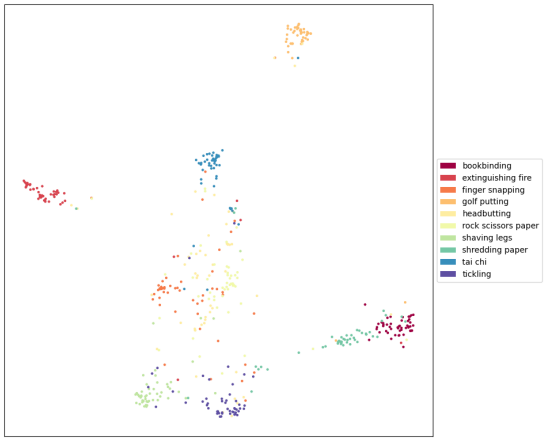


(b) *Dataset* SSV2.

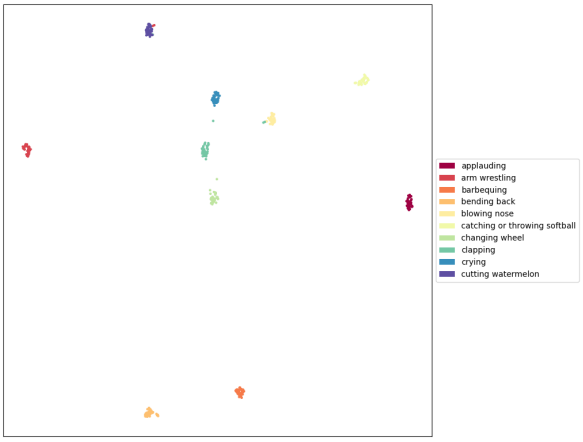
Figura 4.8: Espacio generado en datasets HMDB 51 y SSV2.

En las Figuras 4.7 y 4.8, se observa que con Video Swin Transformer en conjuntos de datos como Kinetics, UCF101, HMDB51 y SSV2, una mayor dispersión de los datos en el espacio a lo esperado, sin un orden claro entre ellos en ninguno de los conjuntos de datos evaluado.

- SlowFast.

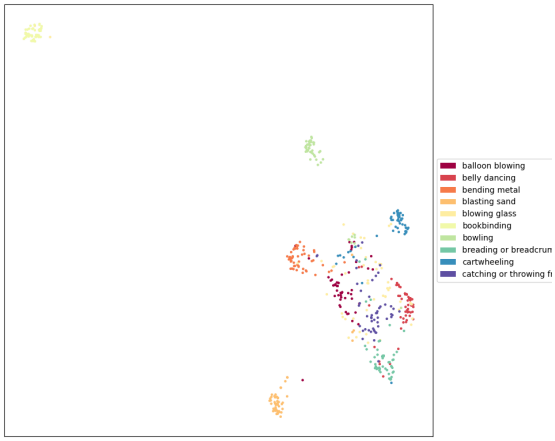


(a) *Dataset* kinetics 400.

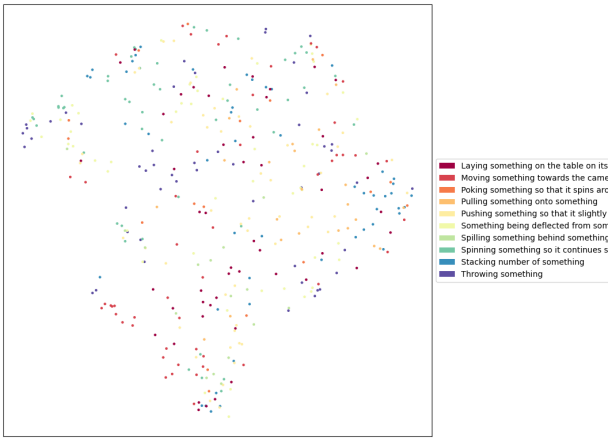


(b) *Dataset* UCF101.

Figura 4.9: Espacio generado en datasets kinetics 400 y UCF101.



(a) *Dataset* HMDB 51.



(b) *Dataset* SSV2.

Figura 4.10: Espacio generado en datasets HMDB 51 y SSV2.

En las Figuras 4.9 y 4.10, se observa que con SlowFast en conjuntos de datos como Kinetics, UCF101 y HMDB51, se logra una distinción clara entre los *clusters*, obteniendo resultados sólidos, tal como se esperaba. Por otro lado, en SSV2 se observa una mayor dispersión de los datos en el espacio, sin un orden claro entre ellos.

4.2.3. Discusión

A partir de los experimentos realizados y observando los resultados obtenidos al evaluar los distintos espacios de representación aprendidos por los modelos seleccionados, se ha determinado que TimesFormer y VideoMae han mostrado un rendimiento superior en las evaluaciones realizadas, obteniendo cada uno las mejores puntuaciones en dos de los cuatros *datasets* utilizados, obteniendo Timesformer 0,951 para Kinetics 400 y un 0,098 para Something Something v2, mientras VideoMae obtuvo 1,000 para UCF101 y 0,873 para HMDB51 como se puede observar en la Tabla 4.1. Estos modelos han demostrado una mayor capacidad para capturar las características visuales y semánticas de los videos como se obersva en las Figuras 4.1, 4.2, 4.3, 4.4, 4.5, 4.6, 4.7, 4.8, 4.9 y 4.10, lo cual los hace ideales para continuar con el desarrollo propuesto. Por lo tanto, se decide continuar con el entrenamiento y evaluación de estos modelos en el contexto del canal televisivo.

Respecto a los resultados obtenidos en el conjunto de datos Something-Something v2, los valores bajos que se observaron se pueden explicar por varias razones. En primer lugar, este conjunto de datos es conocido por su alta complejidad y su naturaleza diferente en comparación con otros conjuntos de datos utilizados. Something-Something v2 se enfoca en capturar diversas interacciones con objetos, mientras que los otros conjuntos de datos se centran principalmente en acciones humanas. Esta diferencia en la naturaleza de los datos hace que la tarea de reconocimiento sea inherentemente más desafiante.

Además, la variabilidad en las interacciones objeto-humano en Something-Something v2 puede llevar a un aumento en la ambigüedad de las clases, lo que dificulta aún más la tarea de clasificación. Además, las acciones relacionadas con objetos pueden tener una gran variabilidad en términos de cómo se realizan, lo que agrega un nivel adicional de dificultad.

Por último, es importante destacar que, si bien los modelos fueron pre-entrenados en el conjunto de datos Kinetics-400, la adaptación a Something-Something v2 puede no ser óptima debido a las diferencias sustanciales entre los conjuntos de datos. Esta discrepancia en la distribución de los datos puede haber contribuido a los resultados más bajos en Something-Something v2.

4.3. Evaluación Metodologías de Procesamiento

Se lleva a cabo una evaluación de las metodologías ideadas para el procesamiento y la creación de clases/*clusters* (RoBERTa + promedio y RoBERTa en su modalidad para oraciones). El propósito principal es analizar estos resultados en profundidad para identificar la mejor metodología a utilizar.

4.3.1. Análisis Cuantitativo

Se lleva a cabo una búsqueda por similitud utilizando los vectores de representación de texto generados por las distintas metodologías evaluadas, de tal manera de observar el

puntaje de similitud obtenido por el Top 5 entregado por la búsqueda. Los resultados de esta búsqueda se presentan en las Figuras 4.11 y 4.12.

```
Consulta: "/data/c13-cut/noviembre2022/TUDIAHDCAP22008-11-2022/entry
   6"
Top 1 "/data/c13-cut/enero2022/TUDIAHDCAP1624-01-2022/entry1"
Similitud: 0.9445844888687134
Top 2 "/data/c13-cut/enero2022/TUDIAHDCAP1624-01-2022/entry1"
Similitud: 0.9445844888687134
Top 3 "/data/c13-cut/enero2022/TUDIAHDCAP1624-01-2022/entry1"
Similitud: 0.9445844888687134
Top 4 "/data/c13-cut/junio2022/TUDIAHDCAP11007-06-2022/entry1"
Similitud: 0.9311782121658325
Top 5 "/data/c13-cut/noviembre2022/TUDIAHDCAP22921-11-2022/entry5"
Similitud: 0.9154930710792542
```

Figura 4.11: Top 5 búsqueda por similitud mediante los vectores obtenidos por RoBERTa y el promedio de los *embeddings* de las palabras.

Al observar los valores de similitud entregados por las búsquedas realizadas en la Figura 4.11 se hace notar como el Top 5 entregado por la búsqueda bajo esta metodología obtiene una similitud muy alta y parecida para todo video/descripción, esto significa que esta metodología no logra diferenciar los datos de buena manera.

```
Consulta: "/data/c13-cut/abril2022/TUDIAHDCAP07-04-2022/entry5"
Top 1 "/data/c13-cut/enero2022/TUDIAHDCAP0610-01-2022/entry10"
Similitud: 0.6309545040130615
Top 2 "/data/c13-cut/junio2022/TUDIAHDCAP12527-06-2022/entry7"
Similitud: 0.5932559370994568
Top 3 "/data/c13-cut/mayo2022/TUDIAHDCAP9413-05-2022/entry0"
Similitud: 0.5606285929679871
Top 4 "/data/c13-cut/enero2022/TUDIAHDCAP1219-01-2022/entry7"
Similitud: 0.5455673336982727
Top 5 "/data/c13-cut/junio2022/TUDIAHDCAP10701-06-2022/entry0"
Similitud: 0.5438978672027588
```

Figura 4.12: Top 5 búsqueda por similitud mediante los vectores obtenidos por la modalidad *sentence* RoBERTa.

Al observar los valores de similitud entregados por las búsquedas realizadas en la Figura 4.11 se hace notar como el Top 5 entregado por la búsqueda bajo esta metodología obtiene una similitud media y distinta para todo video/descripción, esto significa que esta metodología logra diferenciar los datos de buena manera.

4.3.2. Análisis Cualitativo

Se llevaron a cabo búsquedas por similitud que proporcionaron información visual y textual relevante sobre cómo las metodologías procesamiento de texto capturan y representan las similitudes y diferencias entre los contenidos, estos resultados se pueden observar en las Figuras 4.13, 4.14 y 4.15.

- *Query*

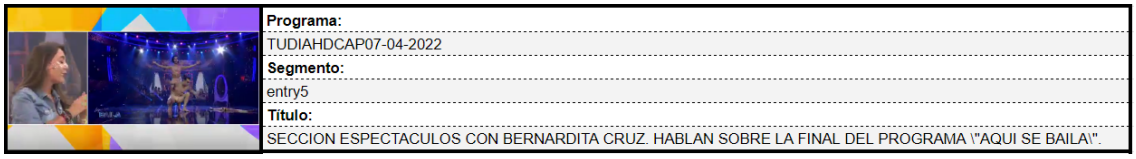


Figura 4.13: Consulta realizada

La Figura 4.13 corresponde a la entrada de la búsqueda por similitud realizada.

- *Sentence* RoBERTa Top 3.

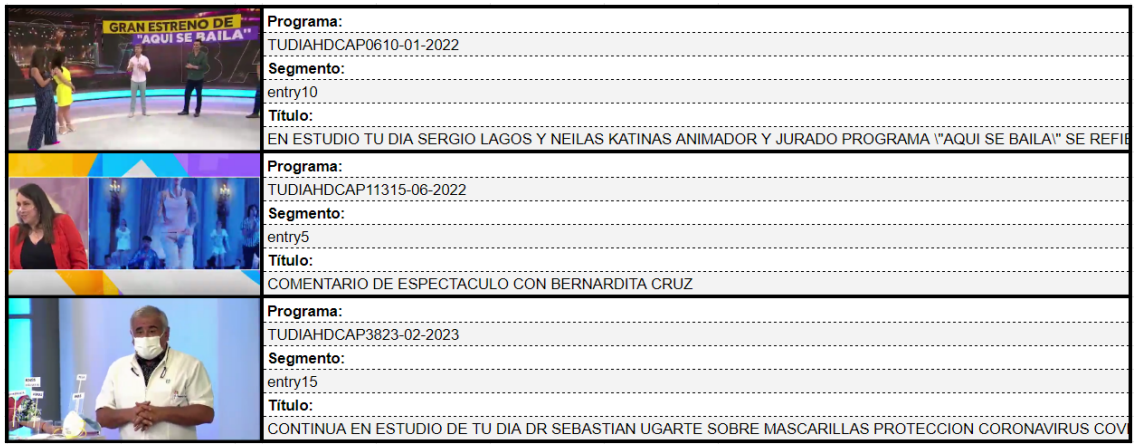


Figura 4.14: Top 3 búsqueda por similitud de coseno.

Al observar la Figura 4.14 se hace notar como las descripciones del Top 3 entregado por la búsqueda bajo esta metodología tienen correspondencia con la búsqueda realizada, siendo el resultado obtenido en el tercer puesto aquel que pierde coherencia con la *query* realizada.

- *Sentence* RoBERTa + NER Top 3.

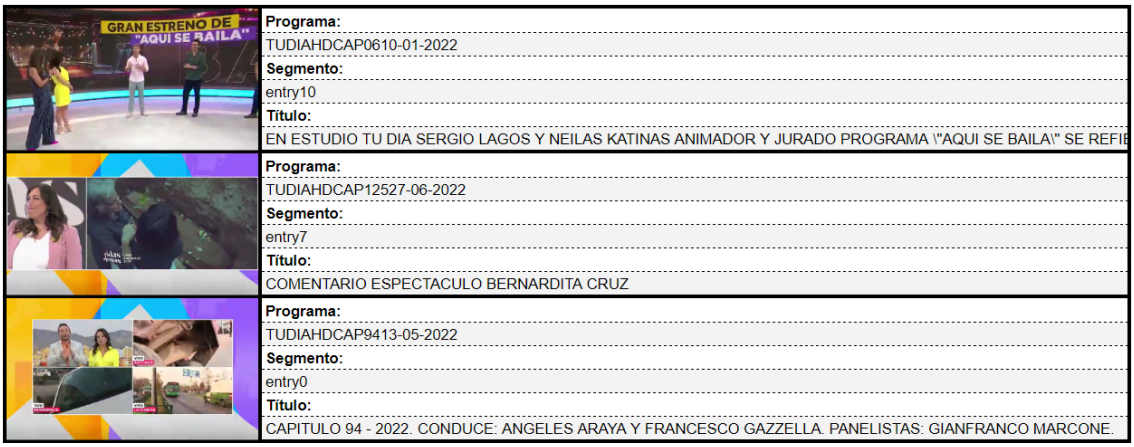


Figura 4.15: Top 3 búsqueda por similitud de coseno.

Al observar la Figura 4.15 se hace notar como las descripciones del Top 3 entregado por la búsqueda bajo esta metodología tienen correspondencia con la búsqueda realizada, siendo el resultado obtenido en el tercer puesto aquel que pierde coherencia con la *query* realizada al igual que en la metodología anterior.

4.3.3. Discusión

Al observar los valores de similitud entregados por las búsquedas realizadas en las Figuras 4.11 y 4.12, se decide abandonar la modalidad de RoBERTa y el promedio de los *embeddings* de las palabras, y continuar con la modalidad de *sentence* RoBERTa. Esto se debe a que los resultados obtenidos por la primera metodología no distingue bien entre los datos como se puede inferir al entregar esta similitudes muy altas y parecidas estos a diferencia de la segunda metodología.

También, al examinar las Figuras 4.13, 4.14 y 4.15 no se ha observado una mejora significativa al utilizar la metodología con un modelo de reconocimiento de entidades nombradas (NER). Por lo tanto, se ha decidido continuar trabajando sin esta propuesta en el corto plazo. Sin embargo, se deja abierta la posibilidad de realizar evaluaciones más exhaustivas en los datos con esta metodología en el futuro, lo que podría influir en una elección diferente.

4.4. Evaluación en Contexto del Canal

Se realizó una evaluación exhaustiva de la calidad de los espacios de representación generados por los modelos seleccionados, TimeSformer y VideoMAE, en el contexto de C13. Para llevar a cabo esta evaluación, se aplicaron ciertas técnicas de preprocesamiento a los videos de entrada. Estas técnicas incluyeron la normalización de los videos, donde se muestrearon los 16 cuadros centrales de cada video y se ajustó el tamaño de cada video a una resolución de 256x256 píxeles. Esto permitió estandarizar las entradas y garantizar que los modelos trabajaran con datos coherentes y comparables.

4.4.1. Análisis Cuantitativo

Para realizar el análisis cuantitativo de los resultados, se empleó el índice Rand, al igual que en la Sección 4.2. El índice Rand es una métrica ampliamente utilizada en evaluación de *clustering* que permite medir la similitud entre dos particiones de datos. En este caso, se comparó la partición obtenida a partir de los espacios vectoriales generados por los modelos de videos con la partición de clases obtenida a partir del procesamiento de los metadatos.

	Adjusted Rand Index	
Model	Metadatos	Metadatos + NER
TimesFormer	0.020	0.014
VideoMAE	0.024	0.017

Tras obtener los resultados de la evaluación en el contexto de C13, se procedió con el entrenamiento de los modelos seleccionados específicamente para el canal. Siguiendo el procedimiento detallado al inicio de esta sección (Sección 4.4), se normalizaron los videos de entrada antes de iniciar el entrenamiento. A continuación, se presenta una representación gráfica del proceso de entrenamiento en las Figuras 4.16 y 4.17, que abarca cinco épocas y utiliza la función de pérdida de entropía cruzada (*cross-entropy loss*) como métrica de evaluación. Esta representación gráfica muestra la evolución del rendimiento de los modelos a medida que se ajustan a los datos específicos del canal.

- TimeSfomer.

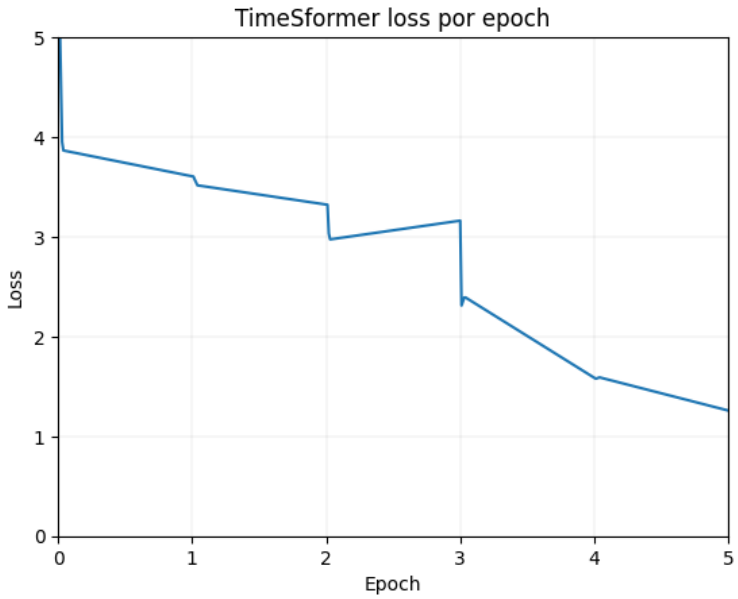


Figura 4.16: Cross Entropy loss durante las 5 primeras épocas del entrenamiento.

- VideoMAE.

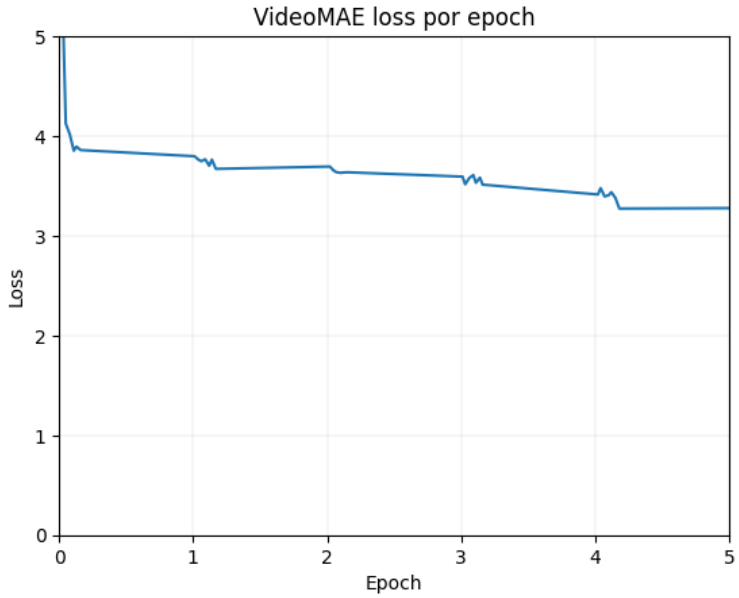


Figura 4.17: Cross Entropy loss durante las 5 primeras épocas del entrenamiento.

Una vez concluido el proceso de entrenamiento durante un total de 20 épocas, se procedió a la evaluación de los modelos mediante el cálculo del Índice Ajustado de Rand (ARI), tal como se detalla en la Tabla 4.2.

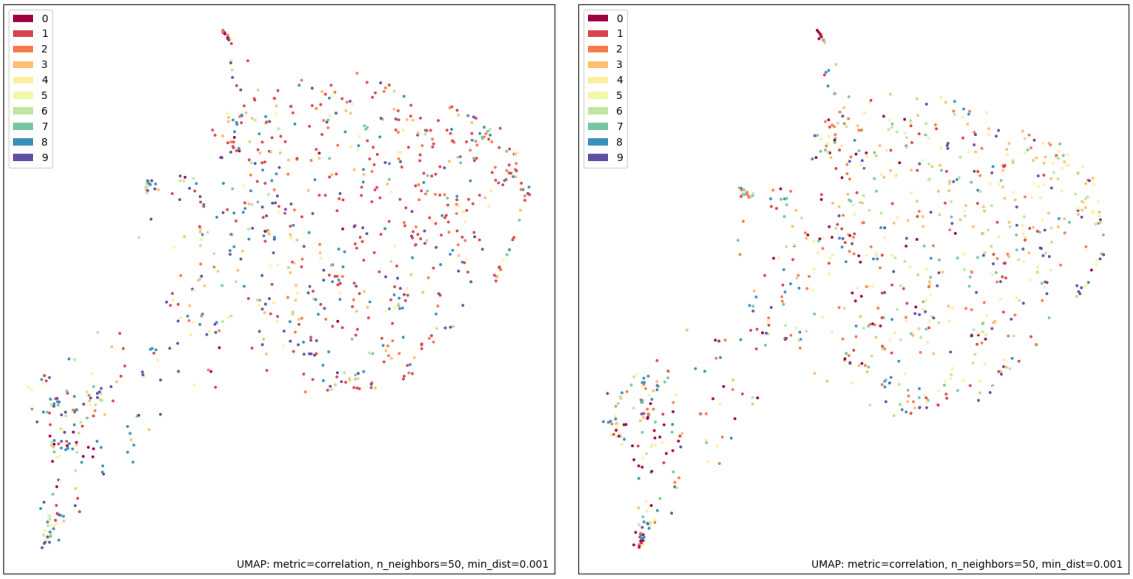
Model	Preentrenamiento	Entrenamiento en canal	Adjusted Rand Index
TimesFormer	K400	no	0.020
VideoMAE	K400	no	0.024
TimesFormer	K400	si	0.032
VideoMAE	K400	si	0.007

Tabla 4.2: Tabla Adjusted Rand Index para la evaluación de los espacios aprendidos en el contexto del canal con y sin entrenamiento en este.

4.4.2. Análisis Cualitativo

Además del análisis cuantitativo basado en el índice Rand, también se puede realizar un análisis cualitativo para examinar visualmente la distribución y separación de los datos en los espacios vectoriales. Esto proporciona una comprensión más completa y detallada de la calidad y la capacidad de representación de los modelos de videos.

- TimeSfomer.



(a) Metadatos. (b) Metadatos + NER.

Figura 4.18: Espacio generado en el contexto del canal (metadatos y metadatos + NER)

En la Figura 4.18, con TimeSfomer y bajo conjunto de datos del canal independientemente de la metodología utilizada, se observa una mayor dispersión de los datos en el espacio a lo esperado, sin un orden claro entre ellos en ninguno de los conjuntos de datos evaluado. Estos resultados corresponden con los obtenidos cuantitativamente.

- VideoMAE.

(a) Metadatos. (b) Metadatos + NER.

Figura 4.19: Espacio generado en el contexto del canal (metadatos y metadatos + NER

En la Figura 4.19, con VideoMAE y bajo conjunto de datos del canal independientemente de la metodología utilizada, se observa una mayor dispersión de los datos en el

espacio a lo esperado, sin un orden claro entre ellos en ninguno de los conjuntos de datos evaluado. Estos resultados corresponden con los obtenidos cuantitativamente.

4.4.3. Discusión

Al analizar los resultados tanto cuantitativos como cualitativos presentados en la Tabla 4.2 y las Figuras 4.18 y 4.19, se evidencia un hallazgo significativo. A pesar de que el modelo TimeSformer con un 3.2 % logra obtener métricas de desempeño ligeramente superiores en comparación con VideoMAE que obtiene un 0.7 %, es importante destacar que estos valores son considerablemente más bajos de lo esperado. Este fenómeno sugiere un desafío sustancial en la tarea de catalogación de videos en el contexto del canal. Las cifras obtenidas reflejan la complejidad y la diversidad del contenido de video, así como la necesidad de profundizar en el procesamiento realizado a la metada brindada por el canal.

Capítulo 5

Análisis, Mejoras y Ajustes Metodológicos

5.1. Análisis Procesamiento

Después de haber evaluado los modelos y constatado una mejora menor a la esperada en su rendimiento, se centró la atención en la búsqueda de soluciones. Se llevaron a cabo tres enfoques adicionales con el objetivo de obtener clases de datos más precisas y, en consecuencia, evaluar si la mejora en la calidad de los *clusters* conlleva un mejor rendimiento de los modelos.

Estos enfoques se basaron en la combinación de dos enfoques de procesamiento de datos distintos y el uso de dos técnicas de agrupación (*clustering*) diferentes. Por un lado, se aplicó el procesamiento de la metadata, siguiendo el mismo procedimiento empleado en el Hito 2. Por otro lado, se implementó el procesamiento de la metadata con la eliminación de las entidades, llevada a cabo mediante un modelo NER (FLAIR). Estos enfoques se presentaron como alternativas clave para abordar la problemática y mejorar la precisión en la clasificación de los datos.

Es fundamental señalar que, en contraposición al enfoque empleado en Capítulos y Secciones anteriores, en esta ocasión se optó por la Eliminación de las entidades en lugar de reemplazarlas. Además, se aplicaron las técnicas de *clustering* K-Means y HDBSCAN en cada uno de los procesamientos para la obtención de clases.

Para realizar una distinción llamaremos a cada enfoque evaluado para la obtención de las clases de la siguiente forma:

- **Metodología 1**: Es el método que se lleva utilizando desde el inicio, se corrige la ortografía de la metadata, se traduce al ingles, se obtienen los *embeddings* correspondientes mediante *Sentence* RoBERTa y se realizan los *clusters* para la obtención de las clases por medio de K-Means.
- **Metodología 2**: Igual que la metodología 1 pero utilizando HDBSCAN para el *clustering* en lugar de K-Means.

- **Metodología 3**: Igual que la metodología 1 pero antes de la obtención de los *embeddings* se suprime las entidades en la metadata por medio del mismo modelo NER descrito en Capítulos y Secciones anteriores.
- **Metodología 4**: Igual que la metodología 3 pero utilizando HDBSCAN para el *clustering* en lugar de K-Means.

5.2. Resultados

Las Figuras 5.1, 5.2, 5.3 y 5.4 ilustran la calidad de los agrupamientos obtenidos mediante cada una de estas metodologías, brindando una visión completa de los resultados en función de las técnicas de *clustering* utilizadas y las distintas estrategias de procesamiento de datos.

- Metodología 1: K-Means.

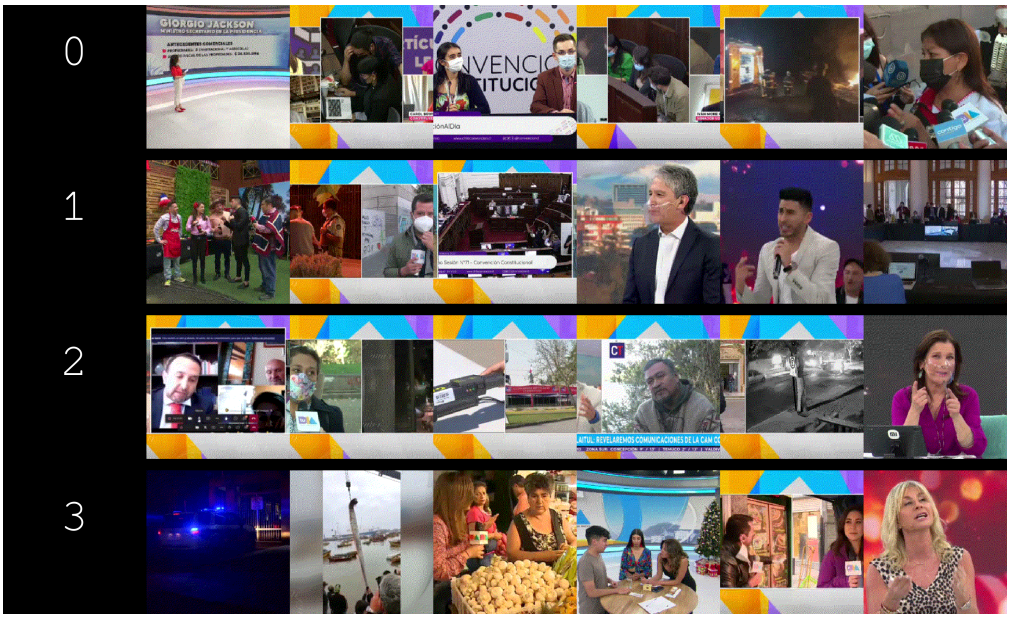


Figura 5.1: Muestreo de los *clusters* obtenidos mediante K-Means.

- Metodología 2: DBSCAN.

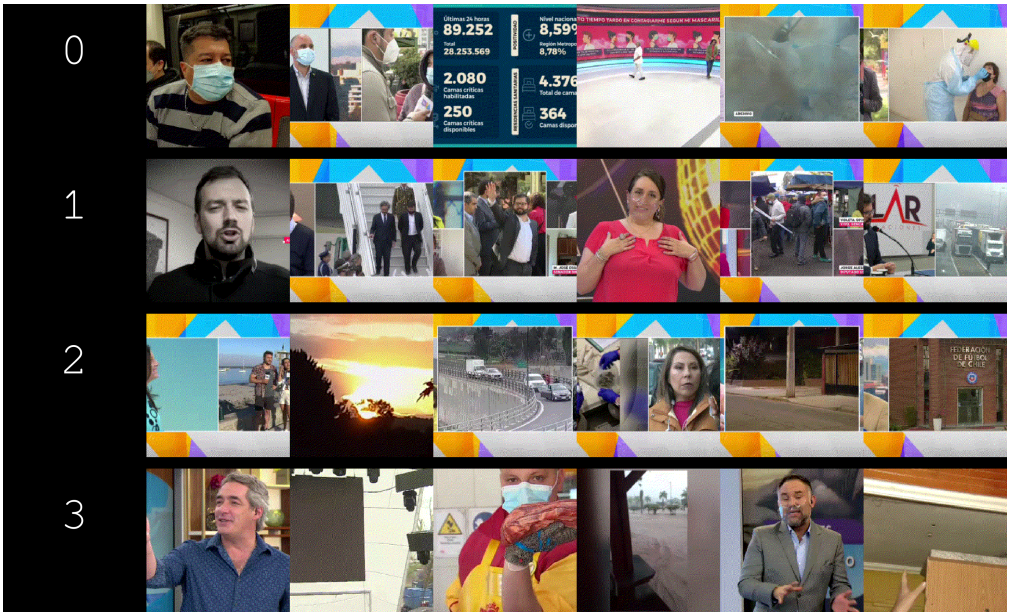


Figura 5.2: Muestreo de los *clusters* obtenidos mediante DBSCAN.

- Metodología 3: Eliminación de entidades (NER) y K-Means.

Figura 5.3: Muestreo de los *clusters* obtenidos mediante K-Means + NER.

- Metodología 4: Eliminación de entidades (NER) y DBSCAN.

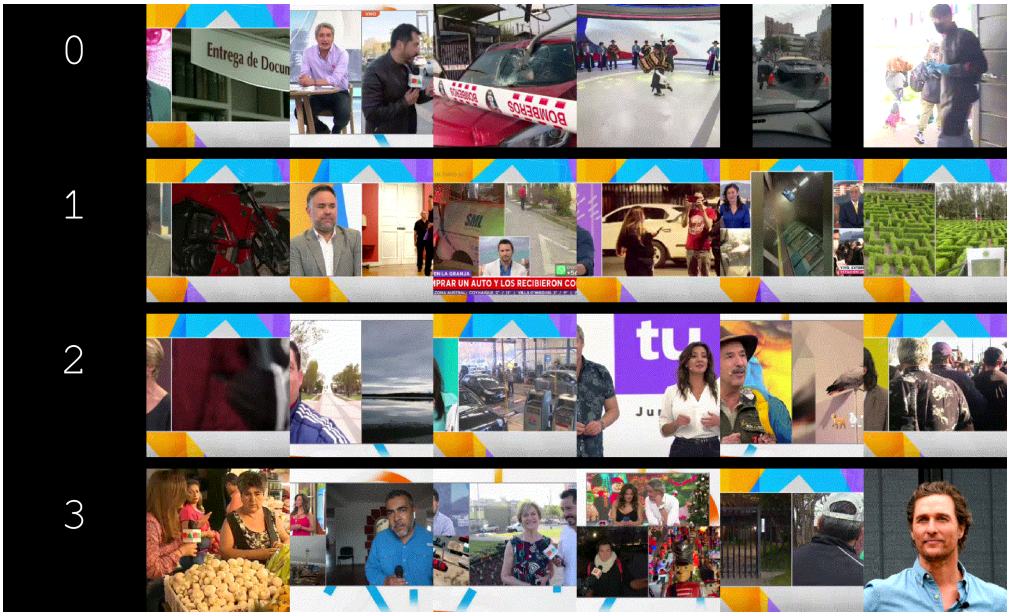


Figura 5.4: Muestreo de los *clusters* obtenidos mediante DBSCAN + NER.

La visualización de los video perteneciente a cada *cluster* proporcionan información visual relevante sobre la similitud de la calidad de los datos a utilizar, esto permite observar como en realidad visualmente un *cluster* tiene una gran diferencia no solo entre *clusters*, sino que también dentro de un mismo *cluster* independiente de la metodología utilizada.

Para un mayor estudio a profundidad, a continuación en las Figuras 5.5, 5.6, 5.7 y 5.8 se incluyen en orden los títulos asociados a cada segmento evaluado en las Figuras 5.1, 5.2, 5.3 y 5.4 en el orden correspondiente (izquierda a derecha y solamente pertenecientes a la clase 0).

- Metodología 1

```
Clase 0

1* RESENTACION PERIODISTICAS PATRIMONIO NUEVAS AUTORIDADES -
   GABIENTE  GABRIEL BORIC
2* TEMA: CONVENCION CONSTITUCIONAL. FRANCESCO GAZZELLA SE REFIERE A
   NORMAS APROBADAS.
3* INFORME, REPORTE DE LA CONVENCION CONSTITUCIONAL "CONVENCION AL
   DIA"
4* TEMA PLENO CONVENCION CONSTITUCIONAL DECIDE PONER FIN AL SENADO Y
    CREACION DE CAMARA DE LAS REGIONES
5* TEMA - TOMA VIP DE CONSTITUCION  -  Y NUEVA CONSTITUCION - EN EL
   ESTUDIO MAURICIO DAZA EX CONVENCIONAL - JORGE CORREA SUTIL
   ABOGADO CONSTITUCIONALISTA  DEBATE PLEBISCITO
6* EL EL ESTUDIO GABRIEL ALEGRIA INFORMA SOBRE ELECCION DE NUEVA
   DIRECTIVA DE LA CONVENCION CONSTITUCIONAL
```

Figura 5.5: Objeto JSON con los títulos asociados a los videos de la clase 0 utilizando la metodología 1.

- Metodología 2

```
Clase 0

1* EN ESTUDIO DOCTOR SEBASTIAN UGARTE FIN DE LAS MASCARILLAS AUN EN
   PANDEMIA USO O NO EN TRANSPORTE PUBLICO
2* COVID 19 - AUMENTO DE CASOS - NUEVA OLA DE CONTAGIOS - EN EL
   ESTUDIO DR. CRISTIAN GARCIA JEFE DEPTO. EPIDEMIOLOGIA MINSAL
3* INFORME DIARIO CASOS COVID - 19  CONTACTO CON MANUEL PALOMINOS
   AUMENTO DE CASOS - EVOLUCION PANDEMIA
4* EN ESTUDIO DE TU DIA DR SEBASTIAN UGARE CIFRAS CONTAGIOS COVID 19
    CORONAVIRUS A LA BAJA " ES UNA BUENA SEÑAL" PANTALLA VIRTUAL
5* MOVIL MANUEL PALOMINOS SUBEN 4.069 CASOS NUEVOS COVID 19
   CORONAVIRUS PANTALLA DIVIDIDA ALZA EN CONTAGIOS
6* PANEL SE REFIERE A AUMENTO EN CASOS NUEVOS DE COVID
```

Figura 5.6: Objeto JSON con los títulos asociados a los videos de la clase 0 utilizando la metodología 2.

- Metodología 3

```
Clase 0

1* CHORIPAN MAS LARGO EN CALERA DE TANGO PERIODISTA ANA MARIA SILVA
2* CONCEPCION DAYANE SALAZAR PARQUE BICENTENARIO FONDAS
3* CALERA DE TANGO GRUPO FOLKLORICO PAREJAS PIE DE CUECA
4* CONTINUA MOVIL GABRIEL ALEGRIA DESDE VIÑA DEL MAR EN REMATE DEL
   HOTEL O'HIGGINS
5* TEMA: PLEBISCITO DE SALIDA. INVITADOS: HERALDO MUÑOZ Y MARIO
   DESBORDES.
6* FRANCESCO GAZZELLA TEMA ISAPRES REGISTRAN PERDIDAS HISTORICAS
```

Figura 5.7: Objeto JSON con los títulos asociados a los videos de la clase 0 utilizando la metodología 3.

- Metodología 4

```
Clase 0

1* EN ESTUDIO DOCTOR SEBASTIAN UGARTE FIN DE LAS MASCARILLAS AUN EN
   PANDEMIA USO O NO EN TRANSPORTE PUBLICO
2* COVID 19 - AUMENTO DE CASOS - NUEVA OLA DE CONTAGIOS - EN EL
   ESTUDIO DR. CRISTIAN GARCIA JEFE DEPTO. EPIDEMIOLOGIA MINSAL
3* INFORME DIARIO CASOS COVID - 19  CONTACTO CON MANUEL PALOMINOS
   AUMENTO DE CASOS - EVOLUCION PANDEMIA
4* EN ESTUDIO DE TU DIA DR SEBASTIAN UGARE CIFRAS CONTAGIOS COVID 19
    CORONAVIRUS A LA BAJA " ES UNA BUENA SEÑAL" PANTALLA VIRTUAL
5* MOVIL MANUEL PALOMINOS SUBEN 4.069 CASOS NUEVOS COVID 19
   CORONAVIRUS PANTALLA DIVIDIDA ALZA EN CONTAGIOS
6* PANEL SE REFIERE A AUMENTO EN CASOS NUEVOS DE COVID
```

Figura 5.8: Objeto JSON con los títulos asociados a los videos de la clase 0 utilizando la metodología 4.

Además, en la Figura 5.9, se presenta un histograma que muestra la distribución de la duración de cada instancia en el conjunto de datos del canal. Este gráfico destaca la alta variabilidad en las duraciones de los videos, lo que indica una alta entropía en los datos. Es importante considerar que durante el entrenamiento del modelo, solo se utilizan los 16 cuadros o *frames* centrales de cada video, lo que podría no estar capturando adecuadamente la información presente en videos más largos o más cortos.

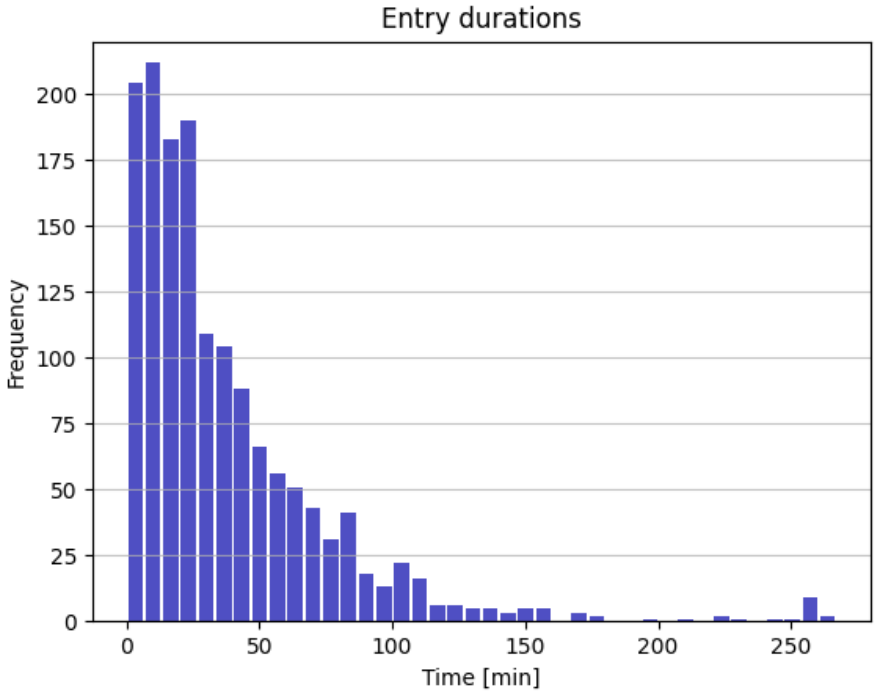


Figura 5.9: Duración de las instancias en el conjunto de datos del canal

5.3. Mejoras y Ajustes Metodológicos

El análisis previo del proceso de procesamiento de datos ha destacado la necesidad de adoptar un enfoque más refinado en la creación de *clusters* debido a la alta entropía de los datos. Como respuesta a este desafío, se han tomado tres decisiones significativas. En primer lugar, se ha optado por utilizar solo videos con una duración inferior a 30 minutos, con el objetivo de reducir la entropía de los datos y mejorar su manejo. En segundo lugar, se ha decidido utilizar las descripciones de los videos en lugar de los títulos de los "entry" para la creación de clases, ya que las descripciones brindan información más rica y relevante. En tercer lugar, se han definido manualmente 10 clases que son pertinentes para el contexto del canal.

En relación a la segunda decisión, se ha observado que algunas descripciones de la metadata del canal contienen marcas de tiempo dentro de ellas, como se observa en la Figura 5.10. Aprovechando esta información, se ha realizado un proceso de segmentación adicional en los videos, lo que ha permitido obtener más fragmentos de datos con una duración menor, contribuyendo así a la reducción de la entropía. Como resultado de estas acciones, se ha obtenido un conjunto de datos compuesto por 3,132 videos, cada uno con su correspondiente metadata.

```
{"entry0": {"duration": "14668.6353019686", "start": "0", "c13.
   thematicDescriptors": ["FIESTAS PATRIAS", "TIEMPO (METEOROLOGIA)"
   , "DELINCUENCIA", "ANTOFAGASTA"], "c13.description": "TU DIA
   CONDUCE  ANGELES ARAYA - MIRNA SCHINDLER -  FRANCESCO GAZZELLA \n
   \n\n\n00:02 CONTACTO CON JUAN PABLO BENITO DESDE TALCA CON MANTA
   DE HUSO - BAILA CUECA - PAYA \n\n00:05 CONTACTO CON ANA MARIA
   SILVA DESDE BARRIO BELLAVISTA  CONVERSAN SOBRE POSIBLES  PRECIOS
   DE EMPANADAS, ANTICUCHOS ETC PARA FIESTA DEL 18\n\n00:11 CONTACTO
    CON RODRIGO PEREZ  DESDE SANTIAGO CENTRO  LOCAL VENTA DE
   SOPAIPILLAS - LA PICA DE LA TIA\n\n00:21 GIANFRANCO MARCONE  -
   INFORME  Y PRONOSTICO DEL TIEMPO - MEDICION PRECIPITACIONES\n\n00
   :37 POLEMICA .... \n\n03:18 ADELANTO PROGRAMA  TIEMPOS DE BARRIO
   - LORETO ARAVENA \n\n03:25 ANA MARIA SILVA Y GABRIEL ALEGRIA
   ENTREGAN DATOS DE PRECIOS DE JUGUETES POR EL DIA DEL NINO\n\n03:3
   7 ADELANTO PROGRAMA DE TU A TU CON MARTIN CARCAMO - LUIS JARA \n\
   n03:40 AFIFE DOCMAN \n\n- KATHERINE LARRAGUIBEL - TRIGLICERIDOS
   ALTOS -  VITAMINA B KOMPLEX - FLORADIX\n- MAQUILLAJE TERAPEUTICO
   CAROLA DONELLO - VITILIGO \n- FASHION PARK - MODA KIDS \n\n\n ",
   "c13.assetTitle": "TU DIA CONDUCE  ANGELES ARAYA - MIRNA
   SCHINDLER -  FRANCESCO GAZZELLA"
}
```


Figura 5.10: Ejemplo objeto JSON con las descripciones asociadas a los videos, con marcas de tiempo en ellas.

Para la tercera decisión, se describen los pasos que se han seguido en la creación y asignación de las clases necesarias para el entrenamiento.

1. Se definen 10 posibles clases para los datos.
2. Se obtiene el *top* 700 para cada clase mediante una búsqueda por similitud con las descripciones de los datos.
3. Se eliminan los datos que se repitan en mas de una clase excepto en la clase con la puntuación mayor.
4. Se analizan las clases obtenidas
5. Si el análisis resulta negativo, se vuelven a realizar los pasos.

Bajo la metodología anteriormente descrita, finalmente se obtienen las siguientes clases.

1. Policial, Crímenes y Delincuencia. Con 453 datos.
2. Clima, Tiempo, Temporal y Medio ambiente. Con 250 datos.
3. Finanzas, Negocios y Economía. Con 276 datos.
4. Entretenimiento, Celebraciones, Fiestas y Concursos. Con 286 datos.

5. Programa de Entrevistas y Conversaciones. Con 281 datos.
6. Salud, Bienestar, Belleza y Cuidado. Con 360 datos.
7. Viajes, Turismo y Vacaciones. Con 220 datos.
8. País, Estado, Política y Leyes. Con 371 datos.
9. Educación, Ciencia y Tecnología. Con 193 datos.
10. Accidentes de transito y automóviles. Con 442 datos.

En la Figura 5.11 se puede ver como las clases generadas por estas 10 etiquetas son visualmente mas parecidas que en casos anteriores.

Figura 5.11: Muestreo de las clases obtenidas

Capítulo 6

Resultados Finales y Discusión

En este capítulo se muestran los resultados obtenidos en los experimentos utilizando las mejoras planteadas en el capítulo anterior 5.

6.1. Evaluación en Contexto del Canal

La evaluación se lleva a cabo mediante el entrenamiento de TimeSformer, el cual ha demostrado obtener mejores resultados, utilizando las nuevas clases generadas a través de las metodologías propuestas en el Capítulo 5.

6.1.1. Análisis Cuantitativo

Una vez concluido el entrenamiento durante un total de 20 épocas, se procedió a la evaluación del modelo mediante el cálculo del Índice Ajustado de Rand (ARI), tal como se detalla en la Tabla 6.1.

Model	Preentrenamiento	Entrenamiento canal	Adjusted Rand Index
TimesFormer	K400	no	0.020
VideoMAE	K400	no	0.024
TimesFormer	K400	si	0.032
VideoMAE	K400	si	0.007
Timesformer	K400	si (nueva metodología)	**0.195**

Tabla 6.1: Tabla Adjusted Rand Index para la evaluación de los espacios aprendidos en el contexto del canal con, sin entrenamiento en este y, antes y luego de la nueva metodología.

En la Tabla 6.2, hemos incluido el valor de precisión (*accuracy*) del modelo en el contexto específico del canal. Esta métrica nos brinda una visión general de cómo se desempeña el modelo en la tarea de clasificación de videos dentro del ámbito del canal. Además, para una

evaluación más detallada, en la Tabla 6.3 presentamos el *accuracy* calculado para cada clase individual. Esto nos permite comprender mejor cómo el modelo se comporta en la clasificación de videos en las diversas categorías temáticas definidas manualmente.

Model	Preentrenamiento	Entrenamiento en contexto del canal	Accuracy
TimesFormer	K400	si (nueva metodología)	0.202

Tabla 6.2: Tabla Accuraccy para la evaluación de los espacios aprendidos en el contexto del canal luego de la nueva metodología.

Clase	Accuracy
Policial, Crímenes y Delincuencia	0.220
Clima, Tiempo, Temporal y Medio ambiente	0.222
Finanzas, Negocios y Economía	0.229
Entretenimiento, Celebraciones, Fiestas y Concursos	0.259
Programa de Entrevistas y Conversaciones	0.164
Salud, Bienestar, Belleza y Cuidado	0.215
Viajes, Turismo y Vacaciones	0.215
Pais, Estado, Politica y Leyes	0.199
Educación, Ciencia y Tecnología	0.199
Accidentes de transito y automoviles	0.238

Tabla 6.3: Tabla Accuraccy para la evaluación de los espacios aprendidos en el contexto del canal por clase luego de la nueva metodología.

6.1.2. Análisis Cualitativo

Además del análisis cuantitativo basado en el índice Rand ajustado, también se realiza un análisis cualitativo para examinar visualmente la distribución y separación de los datos en los espacios vectoriales. Esto proporciona una comprensión más completa y detallada de la calidad y la capacidad de representación de los modelos de videos.

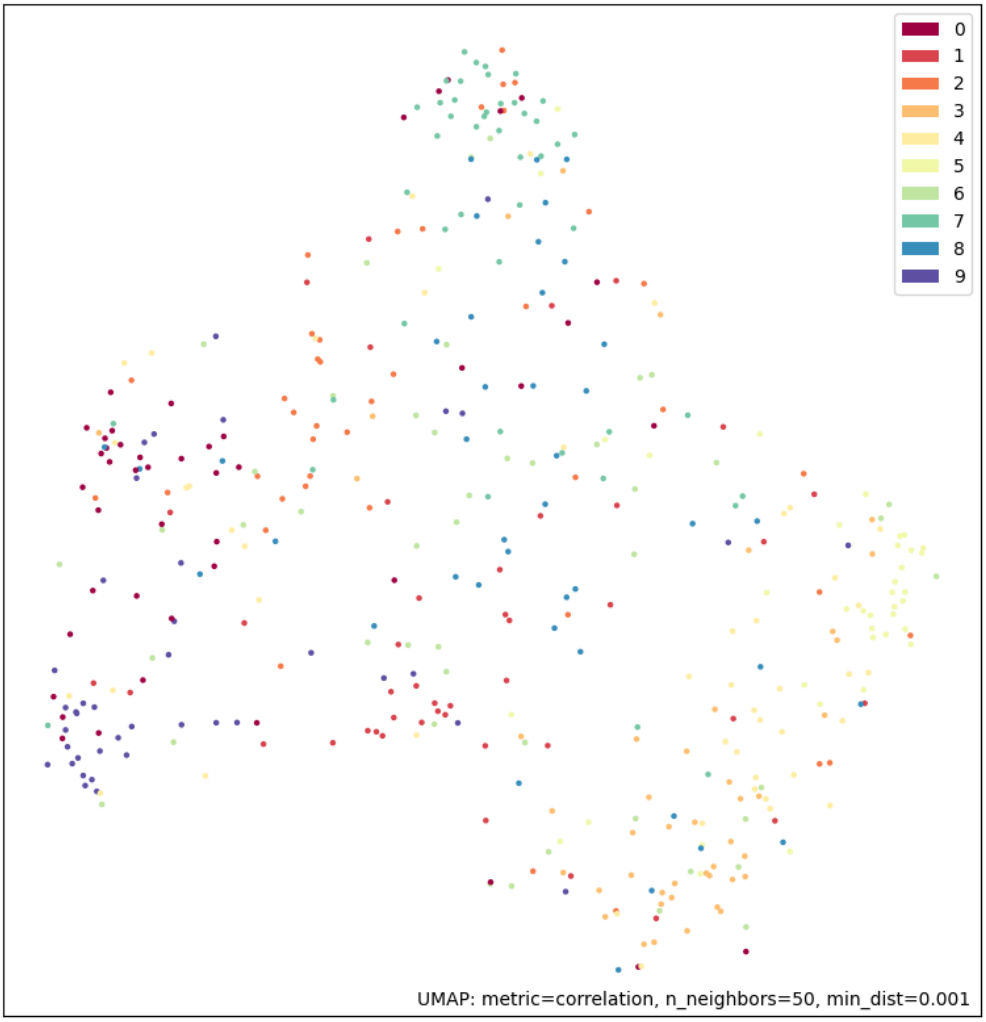


Figura 6.1: Espacio generado en el contexto del canal, TimeSformer

En la Figura 6.3 se lleva a cabo un proceso de búsqueda por similitud de coseno por video, comúnmente conocido como video retrieval. Este proceso se realiza con el propósito de evaluar la efectividad de las representaciones generadas por los modelos. El enfoque de video retrieval permite recuperar videos similares en función de la similitud de sus vectores de representación. Esto es esencial para aplicaciones que requieren espacios de representación de alta calidad, ya que videos similares deberían quedar más cerca entre sí en el espacio (lo que se traduce en una menor similitud de coseno) que videos que son distintos entre sí y, por lo tanto, deberían estar más alejados en el espacio (lo que se traduce en una mayor similitud de coseno).

Figura 6.2: *Query* o consulta realizada por medio de búsqueda por similitud en los videos

Figura 6.3: *Top 5* obtenido con la consulta realizada por medio de búsqueda por similitud en los videos

A continuación, en la Figura 6.4 se añaden las clases reales, predichas, descripción y grado de similitud de los datos asociados a la consulta realizada.

```
Consulta: (cap: TUDIAHDCAP25222-12-2022/entry7)
Descripcion: FRANCESCO GAZZELLA  SOBRE LA REALIZACION DE ANO NUEVO EN EL MAR - FUEGOS ARTIFICIALES
- CONTACTO TELEFONICO CON DIPUTADO ANDRES CELIS - SE REFIERE A REALIZACION DE ESPECTACULO DE FUEGOS ARTIFICIALES VALPARAISO
Clase Original: 3
Clase Predicha: 5

Top 1: (cap: TUDIAHDCAP25729-12-2022/entry1)
Descripcion: CONTACTO CON GABRIEL ALEGRIA DESDE ISLA NEGRA
 - RESTAURANTE BACO -  CONVERSAN CON ELIANA DUENA DEL RESTAURANT - CUENTA DE SUS COMIENZOS EN ISLA NEGRA
Clase Original: 6
Clase Predicha: 8
Similitud: 0.4474399983882904

Top 2: (cap: TUDIAHDCAP22921-11-2022/entry0_3)
Descripcion:  CONTACTO CON RODRIGO PEREZ DESDE PAINE - ACTUALIZA SOBRE PARO CAMIONEROS - CONGESTION
Clase Original: 9
Clase Predicha: 5
Similitud: 0.40614956617355347

Top 3: (cap: TUDIAHDCAP20924-10-2022/entry2_0)
Descripcion: INVITADO: ANTONIO BRIONES, COORDINADOR DE PROYECTO. TEMA: PREMIO INTERNACIONAL WORLD BEST SCHOOL PRIZES PARA ESCUELA BASICA EMILIA LASCAR.
Clase Original: 3
Clase Predicha: 5
Similitud: 0.3874279260635376

Top 4: (cap: TUDIAHDCAP20924-10-2022/entry0_6)
Descripcion:  HABLAN SOBRE ASADO. INGRESAN PANELISTAS: LIBARDO BUITRAGO, FRANCESCO GAZZELLA Y GIANFRANCO MARCONE. COMENTAN LA CARNE. ENTREGAN PREMIOS SEGUN CATEGORIAS DE LA FIESTA:
- "LA QUE LO DIO TODO" ES PARA CARLA, INGRESA LE PONEN BANDA. VIDEOS DE CARLA BAILANDO.
Clase Original: 3
Clase Predicha: 5
Similitud: 0.3781149983406067

Top 5: (cap: TUDIAHD20821-10-2022/entry0_0)
Descripcion: TU DIA CONDUCE JOSE LUIS REPENNING, PRISCILLA VARGAS SE VEN CAMAROGRAFOS PERSONAL TECNICO "CHIQUI" DIA DE EVALUACION, "COCA" EN CAMARA RESPODEN SOBRE COMO EVALUAN A LOS ANIMADORES, ROBERTO CAMAROGRAFO, "MANE". SE INCORPORA FRANCESCO GAZZELLA. REPENNING COMIENDO DULCE DEBE PONERSE A DIETA LE RECOMIENDAN, SOBRENOMBRE.
Clase Original: 4
Clase Predicha: 5
Similitud: 0.3698650002479553
```


Figura 6.4: Datos asociados a la búsqueda por similitud ilustrada en las Figuras 6.2 y 6.3.

6.1.3. Discusión

Al analizar tanto los resultados cuantitativos como cualitativos presentados en la Tabla 6.1, la Figura 6.1 y la búsqueda por similitud ilustrada en las Figuras 6.2 y 6.3, se concluye que gran parte de los resultados anteriores se debían a la calidad y entropía de las clases generadas. Después de las mejoras implementadas, TimeSformer logra un aumento del 19.5 %, obteniendo los mejores resultados en este trabajo. TimeSformer también logra un accuracy general de 20.2 %, como se evidencia en la Figura 6.2. Además, al observar el accuracy por clase en la Figura 6.3, se nota una distribución más equitativa de rendimiento entre las clases, lo que indica que se ha logrado reducir significativamente el desbalanceo de clases y mejorar la capacidad del modelo para generalizar a través de todas las categorías de videos. En resumen, las mejoras implementadas no solo han aumentado el rendimiento global del modelo, sino que también han contribuido a una representación más equilibrada y precisa de las diferentes clases de videos en el conjunto de datos del canal.

La búsqueda por similitud realizada ofrece una visión interesante de cómo los modelos están representando y relacionando los videos en función de sus vectores de representación. Al observar las Figuras 6.3, 6.2 y 6.4, es evidente que los modelos están logrando agrupar videos visualmente similares en una región cercana del espacio de representación, lo que es un resultado prometedor.

Sin embargo, también es evidente que el hecho de que todos los videos en el "top 5"pertenezcan a la misma clase predicha, pero no necesariamente a la misma clase real, sugiere que los modelos aún tienen dificultades para realizar predicciones precisas en algunos casos. Esto podría deberse a varias razones.

Capítulo 7

Conclusión

En este estudio, hemos abordado el desafío de catalogación de videos en el contexto de un canal de televisión, un proceso que es crucial para organizar y acceder al vasto contenido audiovisual disponible en la actualidad. A lo largo de esta investigación, hemos explorado diversas metodologías y enfoques, lo que nos ha permitido obtener valiosos *insights* sobre cómo mejorar la precisión y eficacia de los modelos de aprendizaje automático en esta tarea específica.

Uno de los hallazgos clave de este estudio ha sido la importancia de la calidad de los datos y la estructuración de las clases en la catalogación de videos. Inicialmente, observamos resultados subóptimos, lo que nos llevó a analizar en profundidad los factores que contribuían a esta situación. La alta entropía en los datos y la falta de clases homogéneas se destacaron como problemas fundamentales. Para abordar esto, implementamos estrategias que incluyeron la selección de videos de menor duración, la explotación de marcas de tiempo en las descripciones y la creación manual de clases relevantes. Estas decisiones condujeron a una mejora significativa en el rendimiento de los modelos.

En particular, el modelo TimeSformer demostró ser altamente efectivo en la catalogación de videos en este contexto, superando al modelo VideoMAE con un aumento del 19.5 % en la métrica ARI. Este resultado subraya la importancia de elegir una arquitectura de modelo adecuada y cómo la calidad de los datos de entrenamiento puede tener un impacto sustancial en el rendimiento final.

Este estudio no solo proporciona una base efectiva para la catalogación de videos en el contexto del canal de televisión, sino que también establece una base sólida para investigaciones futuras. La capacidad de mejorar la precisión y eficacia de los modelos mediante la optimización de datos y clases es una lección valiosa que puede aplicarse en otros escenarios de procesamiento de videos. Además, la elección de arquitecturas de modelos adecuadas, como TimeSformer, abre nuevas oportunidades para abordar desafíos en el campo de la catalogación de contenido audiovisual.

Con esto, se cumple con los objetivos principales del trabajo de sentar bases sólidas para un trabajo a futuro.

7.1. Limitaciones y Desafíos

A lo largo de este estudio, nos encontramos con varias limitaciones y desafíos que influyeron en la dirección y los resultados de la investigación. Estas limitaciones arrojaron luz sobre las complejidades inherentes a la catalogación de videos en el contexto de un canal de televisión y proporcionaron valiosas lecciones para futuros proyectos de investigación en esta área.

La decisión de enfocarnos inicialmente en la información visual de los videos se fundamenta en la potencialidad de los modelos de visión por computadora basados en redes neuronales para extraer patrones visuales significativos. Si bien el audio también es una fuente rica de información, la elección inicial de centrarnos en lo visual fue estratégica. Buscamos establecer un sólido modelo visual como punto de partida para luego incorporar información de audio. Esta progresión permite una comprensión más profunda y un refinamiento del modelo, mejorando su capacidad para capturar la complejidad de los datos audiovisuales.

La selección de conjuntos de datos públicos se basó en la variedad y amplitud de datos disponibles. Estos conjuntos proporcionan una diversidad de escenarios, lo que permite a los modelos aprender un espacio de representaciones latentes con mayor generalización. Utilizar datos públicos al inicio del estudio nos permitió aprovechar el conocimiento preexistente y los conjuntos de datos de calidad para comenzar con un modelo visual sólido desde el principio.

Uno de los desafíos más significativos fue el costo computacional asociado con el entrenamiento de redes neuronales para videos. La capacidad de las GPU disponibles limitó el tamaño del lote (*batch size*) que podíamos usar para el entrenamiento. Esto resultó en un proceso de entrenamiento más prolongado y menos eficiente, ya que se requerían más épocas para alcanzar resultados satisfactorios. Inicialmente, el tiempo de entrenamiento para cada modelo se estimó en una semana, pero después de una serie de optimizaciones y ajustes, pudimos reducirlo a aproximadamente 26 horas para 20 épocas. Aunque logramos mejorar la eficiencia, esta limitación sigue siendo un factor que debe abordarse en futuras investigaciones, ya que podría restringir la capacidad de explorar modelos más grandes y complejos.

Otra limitación importante estuvo relacionada con la calidad de los datos. La alta entropía de los videos y la falta de clases homogéneas en los conjuntos de datos iniciales influyeron significativamente en los resultados. Si bien implementamos estrategias efectivas para abordar esta limitación, como la selección de videos de menor duración y la creación manual de clases, aún es un desafío en curso. La calidad de los datos es esencial en la catalogación de videos, y futuras investigaciones deben centrarse en la recopilación y curación de datos de alta calidad para mejorar aún más el rendimiento de los modelos.

Finalmente, otro desafío clave fue la elección de la arquitectura del modelo. Aunque obtuvimos resultados prometedores con TimeSformer, este estudio solo rasca la superficie de las posibilidades. Explorar y evaluar una gama más amplia de arquitecturas de modelos podría conducir a mejoras adicionales en el rendimiento y proporcionar una comprensión más profunda de cuáles son las más adecuadas para esta tarea específica.

En resumen, este estudio enfrentó desafíos significativos relacionados con el costo computacional, la calidad de los datos y la selección del modelo. Aunque superamos muchos de

estos obstáculos y logramos mejoras notables en el rendimiento, queda trabajo por hacer para abordar estas limitaciones de manera más completa y para avanzar en la catalogación de videos en el contexto de los medios de comunicación.

7.2. Trabajo a Futuro

Las siguientes son algunas de las áreas y enfoques que deberían explorarse en trabajos futuros para mejorar aún más la catalogación de videos en este entorno.

- **Procesamiento de Metadata Mejorado**: Aunque se han realizado avances significativos en la creación de clases y etiquetas a partir de la metadata, existe un amplio espacio para mejorar el procesamiento de la metadata. Esto podría incluir el desarrollo de modelos de procesamiento de lenguaje natural (NLP) específicamente diseñados para extraer información relevante de la metadata de los videos, lo que podría conducir a una representación de clase más precisa y a un mejor rendimiento general.
- **Exploración de Nuevas Arquitecturas de Modelos**: El campo de la visión por computadora y el procesamiento de videos sigue evolucionando rápidamente, con nuevos modelos y arquitecturas que surgen constantemente. Futuros estudios deberían explorar y evaluar estas nuevas arquitecturas para determinar si pueden superar a los modelos existentes. Esto podría incluir modelos basados en transformers aún más avanzados, así como enfoques híbridos que combinen características de diferentes arquitecturas.
- **Mejoras en el Procesamiento de Datos**: La calidad de los datos sigue siendo un desafío importante en la catalogación de videos. Futuras investigaciones podrían centrarse en la recopilación y curación de datos de video de alta calidad. Además, se podrían explorar técnicas de aumento de datos y estrategias de selección de ejemplos más efectivas para abordar la alta entropía de los datos.
- **Optimización de Recursos de Hardware**: Continuar optimizando el uso de recursos de hardware, como GPU, podría acelerar significativamente el proceso de entrenamiento de modelos y permitir la exploración de arquitecturas más grandes y complejas.
- **Evaluación de Modelos Multimodales**: Dado que los videos contienen tanto información visual como de audio, futuros estudios podrían centrarse en la evaluación de modelos multimodales que puedan aprovechar ambas fuentes de información para una catalogación más precisa y enriquecida.

En conclusión, el campo de la catalogación de videos en el contexto de los medios de comunicación es un campo en evolución constante, y las oportunidades de investigación y desarrollo son abundantes. Futuros estudios pueden aprovechar estos enfoques para mejorar aún más la catalogación de videos y su utilidad en la gestión y distribución de contenido audiovisual en la industria mediática.

Bibliografía

[1] Meta AI. Action recognition. https://paperswithcode.com/task/action-recognition-in-videos.

[2] Meta AI. Machine translation. https://paperswithcode.com/task/machine-translation.

[3] Meta AI. Named entity recognition (ner). https://paperswithcode.com/task/named-entity-recognition-ner.

[4] Meta AI. Sentence embeddings. https://paperswithcode.com/task/sentence-embeddings.

[5] Meta AI. Spelling correction. https://paperswithcode.com/task/spelling-correction.

[6] Meta AI. Video captioning. https://paperswithcode.com/task/video-captioning.

[7] Meta AI. Video object detection. https://paperswithcode.com/task/video-object-detection.

[8] Meta AI. Video object tracking. https://paperswithcode.com/task/video-object-tracking.

[9] Alan Akbik, Tanja Bergmann, Duncan Blythe, Kashif Rasul, Stefan Schweter, and Roland Vollgraf. FLAIR: An easy-to-use framework for state-of-the-art NLP. In *NAACL 2019, 2019 Annual Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, pages 54–59, 2019.

[10] Alan Akbik, Duncan Blythe, and Roland Vollgraf. Contextual string embeddings for sequence labeling. In *COLING 2018, 27th International Conference on Computational Linguistics*, pages 1638–1649, 2018.

[11] Jose Mariano Alvarez. El perceptrón como neurona artificial. http://blog.josemarianoalvarez.com/2018/06/10/el-perceptron-como-neurona-artificial/.

[12] Anurag Arnab, Mostafa Dehghani, Georg Heigold, Chen Sun, Mario Lučić, and Cordelia Schmid. Vivit: A video vision transformer. In *International Conference on Computer Vision (ICCV)*, 2021.

[13] Gedas Bertasius, Heng Wang, and Lorenzo Torresani. Is space-time attention all you need for video understanding? In *Proceedings of the International Conference on Machine Learning (ICML)*, July 2021.

[14] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding, 2019.

[15] C. Feichtenhofer, H. Fan, J. Malik, and K. He. Slowfast networks for video recognition. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 6201–6210, Los Alamitos, CA, USA, nov 2019. IEEE Computer Society.

[16] R. Goyal, S. Kahou, V. Michalski, J. Materzynska, S. Westphal, H. Kim, V. Haenel, I. Fruend, P. Yianilos, M. Mueller-Freitag, F. Hoppe, C. Thurau, I. Bax, and R. Memisevic. The "something something" video database for learning and evaluating visual common sense. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 5843–5851, Los Alamitos, CA, USA, oct 2017. IEEE Computer Society.

[17] Jie Huang, Wengang Zhou, Houqiang Li, and Weiping Li. Sign language recognition using 3d convolutional neural networks. In *2015 IEEE International Conference on Multimedia and Expo (ICME)*, pages 1–6, 2015.

[18] Will Kay, J. Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, Mustafa Suleyman, and Andrew Zisserman. The kinetics human action video dataset. *CoRR*, 05 2017.

[19] H. Kuehne, H. Jhuang, E. Garrote, T. Poggio, and T. Serre. HMDB: a large video database for human motion recognition. In *Proceedings of the International Conference on Computer Vision (ICCV)*, 2011.

[20] Scikit learn Developers. Comparing different clustering algorithms on toy datasets. `https://scikit-learn.org/stable/auto_examples/cluster/plot_cluster_comparison.html`.

[21] Mason Liu and Menglong Zhu. Mobile video object detection with temporally-aware feature maps. *CoRR*, abs/1711.06368, 2017.

[22] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach, 2019.

[23] Ze Liu, Jia Ning, Yue Cao, Yixuan Wei, Zheng Zhang, Stephen Lin, and Han Hu. Video swin transformer. *arXiv preprint arXiv:2106.13230*, 2021.

[24] MonkeyLearn. Named entity recognition: Concept, tools and tutorial. `https://monkeylearn.com/blog/named-entity-recognition/`.

[25] Khurram Soomro, Amir Zamir, and Mubarak Shah. Ucf101: A dataset of 101 human actions classes from videos in the wild. *CoRR*, 12 2012.

[26] TIBCO. ¿qué es una red neuronal? `https://www.tibco.com/es/reference-center/what-is-a-neural-network`.

[27] Zhan Tong, Yibing Song, Jue Wang, and Limin Wang. Videomae: Masked autoencoders are data-efficient learners for self-supervised video pre-training. In *Advances in Neural Information Processing Systems*, 2022.

[28] Manuel Traub, Sebastian Otte, Tobias Menge, Matthias Karlbauer, Jannik Thümmel, and Martin V. Butz. Learning what and where: Disentangling location and identity tracking without supervision, 2023.

[29] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. *CoRR*, abs/1706.03762, 2017.

[30] Yi Zhu, Xinyu Li, Chunhui Liu, Mohammadreza Zolfaghari, Yuanjun Xiong, Chongruo Wu, Zhi Zhang, Joseph Tighe, R. Manmatha, and Mu Li. A comprehensive study of deep video action recognition. *CoRR*, abs/2012.06567, 2020.