Universidad Nacional Autónoma de México

Posgrado en Ciencia e Ingeniería de la Computación

Transferibilidad de representaciones generalistas de imágenes y video a clasificación de avances de películas

TESIS

Que para optar por el grado de:

Doctor en Ciencia e Ingeniería de la Computación

Presenta:

Ricardo Montalvo Lezama

Tutor:

Dr. Gibran Fuentes Pineda

Instituto de Investigaciones en Matemáticas Aplicadas y en Sistemas

Ciudad Universitaria, CD. MX., octubre 2024

PROTESTA UNIVERSITARIA DE INTEGRIDAD Y HONESTIDAD ACADÉMICA Y PROFESIONAL

De conformidad con lo dispuesto en los artículos 87, fracción V, del Estatuto General, 68, primer párrafo, del Reglamento General de Estudios Universitarios y 26, fracción I, y 35 del Reglamento General de Exámenes, me comprometo en todo tiempo a honrar a la institución y a cumplir con los principios establecidos en el Código de Ética de la Universidad Nacional Autónoma de México, especialmente con los de integridad y honestidad académica.

De acuerdo con lo anterior, manifiesto que el trabajo escrito titulado "Transferibilidad de representaciones generalistas de imágenes y video a clasificación de avances de películas", que presenté para obtener el grado de Doctor en Ciencia e Ingeniería de la Computación, es original, de mi autoría y lo realicé con el rigor metodológico exigido por mi Programa de Posgrado, citando las fuentes, ideas, textos, imágenes, gráficos u otro tipo de obras empleadas para su desarrollo.

En consecuencia acepto que la falta de cumplimiento de las disposiciones reglamentarias y normativas de la Universidad, en particular las ya referidas en el Código de Ética, llevará a la nulidad de los actos de carácter académico administrativo del proceso de titulación/graduación.

Atentamente

Ricardo Montalvo Lezama

512013693

Resumen

La trascendencia de la transferencia de conocimiento en el aprendizaje profundo ha impulsado a la comunidad de investigación a estudiar los factores que fomentan la transferibilidad. En el caso de transferencia de representaciones de imágenes, la mayoría de los estudios se han concentrado en la transferibilidad de imágenes generalistas del conjunto ImageNet a diversas tareas de análisis de imágenes, y también algunas de video. Para representaciones de video, el principal foco ha sido la transferibilidad de clips de acciones humanas del conjunto Kinetics a otras tareas de video similares.

En esta tesis llevamos a cabo un estudio de transferibilidad de representaciones aprendidas en ImageNet y Kinetics a una tarea de video no estudiada en la literatura, la tarea de clasificación de géneros en avances de películas. Estudiamos factores que influyen en la transferibilidad, como la representación de entrada de los avances, los conjuntos de preentrenamiento, la arquitecturas preentrenadas, entre otros. En busca de dar solidez al estudio recolectamos Trailers12k, un nuevo conjunto de datos multimodal de alta calidad de avances de películas. Además, proponemos un nuevo método de clasificación llamado DIViTA (*Dual Image and Video Transformer Architecture*) que aprovecha la particular estructura espacio-temporal de los avances para mejorar la transferencia.

Nuestros resultados revelan ciertas direcciones para mejorar el proceso de transferencia para la clasificación de avances de películas. Primero, encontramos factores relacionados con la representación de entrada de los avances que influyen en el desempeño en la clasificación. Además, el uso complementario de ImageNet y Kinetics durante el preentrenamiento produce representaciones que generalizan mejor en comparación con aquellas obtenidas utilizando sólo un conjunto de preentrenamiento. También, la transferencia se ve beneficiada por la representación abreviada de los avances usada por DIViTA, así como de su módulo de agregación espacio-temporal basado en arquitecturas transformadoras. Finalmente, encontramos que es posible obtener modelos eficientes para la clasificación de géneros en avances con resultados competitivos a sus contrapartes con mejor desempeño pero con mucho mayor uso de recursos.

Agradecimientos

Me gustaría que estas líneas sirvieran para expresar mi más profundo y sincero agradecimiento a los seres que me han acompañado para hacer posible el presente trabajo.

Primero, quiero comenzar expresando mi más profundo agradecimiento a mi familia, su amor y dedicación fueron constantes e inquebrantables. En particular, quiero agradecerle a Bere por todo el apoyo y comprensión. Igualmente, gracias a Pasita y Chinito por acompañarme en este viaje.

También quiero expresar un extenso agradecimiento a mi tutor Gibran, su guía ha sido fundamental a lo largo de todo el proyecto, desde las primeras ideas y hasta la última línea de esta tesis. Sus conocimientos, experiencia y paciencia han sido una gran fuente de aprendizaje para mí.

Además, quiero agradecer a Nash, Mony y Beto por su apoyo en diferentes puntos y formas durante este proceso.

Finalmente, este trabajo se realizó con el apoyo de una beca CONACYT y de los proyectos PAPIIT IA104016 e IV100420 de la Dirección General de Asuntos del Personal Académico (DGAPA), UNAM.

Índice general

1 Introducción

El aprendizaje profundo es una familia de algoritmos parte del aprendizaje automatizado que utilizan múltiples capas de procesamiento para aprender representaciones de datos con múltiples niveles de abstracción. Estas capas de procesamiento, también llamadas redes neuronales profundas, están dotadas de la capacidad de descubrir diferentes factores explicativos de variación detrás de los intrincados datos estructurados. Gracias a avances significativos en las arquitecturas de las redes, las estrategias de entrenamiento y la capacidad de cómputo, el aprendizaje profundo ha producido avances en diversas áreas como la visión por computadora, procesamiento del lenguaje natural, procesamiento del habla, procesamiento de imágenes médicas, biología computacional, juegos, entre otros.

A pesar de su extendido éxito en diversas áreas, el aprendizaje profundo continúa enfrentando el desafío significativo de la eficiencia de los datos. En particular, para el entrenamiento de arquitecturas profundas del estado del arte, como las convolucionales (Liu, Mao et al., 2022) y más recientemente las transformadoras (Vaswani et al., 2017), se requieren grandes conjuntos de datos, en el orden de decenas o incluso cientos de millones de muestras, para lograr un buen rendimiento. Sin embargo, recopilar y etiquetar muestras en estas escalas para cada nueva tarea puede ser costoso o incluso prohibitivo. Por ejemplo, en el ámbito médico se requiere de especialistas con años de entrenamiento para realizar el etiquetado, personal para la revisión de protocolos que respeten los derechos de los pacientes, y el tratamiento ético de los datos y seguimiento de la legislación aplicable. Estos factores aumentan considerablemente los tiempos de recolección de datos, los costos del proceso en general, y limitan la cantidad de datos que pueden ser recolectados. Este reto a la disponibilidad de datos dificulta enormemente la adopción del aprendizaje profundo en un espectro más amplio de escenarios.

La transferencia de conocimiento (TL, por las siglas en inglés de *Transfer Learning*), también conocida como aprendizaje por transferencia, es un grupo de técnicas que busca resolver el problema del requerimiento de datos masivo. La TL reúsa conocimiento previamente adquirido en una tarea base para resolver de manera más eficaz una tarea objetivo. De forma general, el

proceso para llevar a cabo TL se puede entender en dos etapas (Jiang et al., 2022). Durante la etapa de preentrenamiento, se entrena una red neuronal profunda aprovechando una tarea base que idealmente cuenta con un conjunto de datos extenso y de gran variabilidad. En la etapa de adaptación, parte de la red neuronal preentrenada se reúsa agregando nuevas capas para adaptarla a la tarea objetivo. Esta red adaptada se entrena con el conjunto de datos objetivo que es considerablemente más pequeño.

La TL ha jugado un papel fundamental para la extensa acogida de los métodos basados en aprendizaje profundo en diversos dominios. En visión computacional (CV, por las siglas en inglés de *Computer Vision*), por ejemplo, una práctica bastante extendida consiste en preentrenar usando el conjunto de datos ImageNet (Deng et al., 2009) y adaptar a una tarea objetivo de un dominio similar. Esto ha permitido obtener modelos del estado del arte para tareas como clasificación de imágenes (Tan & Le, 2019), detección (Redmon et al., 2016) o segmentación de objetos (Long et al., 2015), entre otras.

1.1. Problema de investigación

Los patrones aprendidos en los datos por las redes durante el preentrenamiento se conocen como representaciones (Bengio et al., 2013). La efectividad de la transferencia de conocimiento está fuertemente relacionada con la transferibilidad de estos patrones a otras tareas. Han habido esfuerzos importantes de la comunidad de investigación para estudiar factores que influyen en la transferibilidad y técnicas para mejorarla. En particular, diversos trabajos en la literatura han estudiado la transferencia de representaciones aprendidas en clasificación de imágenes (IC, por las siglas en inglés de *Image Classification*) en ImageNet a otras tareas que involucran imágenes con propiedades similares (Razavian et al., 2014; Yosinski et al., 2014; Zamir et al., 2018). En el caso de preentrenamiento en video, los esfuerzos se han enfocado principalmente en el conjunto de reconocimiento de acciones humanas (HAR, por las siglas en inglés de *Human Action Recognition*) Kinetics (Kay et al., 2017). No obstante, la transferibilidad de ImageNet o Kinetics a tareas objetivo de video más distantes a la tarea de preentrenamiento, sigue siendo un problema abierto en el área. Un ejemplo de este tipo de tareas es la clasificación multietiqueta de géneros de avances de películas (MTGC, por las siglas en inglés de *Movie Trailer Genre Classification*).

MTGC es una tarea de video de naturaleza y contenido muy distintos a HAR, en la que casi no se ha estudiado la transferibilidad. Esta es una tarea difícil porque los géneros pueden no

tener una expresión física específica en un cuadro o una secuencia de cuadros. En consecuencia, los géneros deben inferirse a partir de personajes, escenas, temas, dinámicas, relaciones y otros elementos abstractos. Además, la tarea conlleva una subjetividad natural: distintos observadores humanos pueden asignar géneros diferentes a un mismo avance. A diferencia de HAR, los escenarios de MTGC son más diversos, esto es, la historia no suele presentarse de forma lineal, pueden incluir elementos ficticios (por ejemplo, personajes, paisajes, dispositivos, leyes de la física, etc.) y su duración suele ser mucho mayor. Esto genera importantes disimilitudes entre los contenidos de las imágenes, la estructura del video y la duración en las tareas IC/HAR y MTGC que pueden afectar a la transferibilidad de las representaciones espaciales y espacio-temporales aprendidas por los modelos preentrenados en ImageNet y Kinetics.

Las aproximaciones de los primeros métodos propuestos para resolver MTGC emplearon algoritmos de extracción de características locales de visión computacional en conjunto con modelos de clasificación basados en aprendizaje automatizado (Y.-F. Huang & Wang, 2012; Rasheed et al., 2005; H. Zhou et al., 2010). A lo largo de la última década, se ha vuelto una práctica común el uso de arquitecturas de redes neuronales profundas para resolver esta tarea (Simões et al., 2016; Wehrmann & Barros, 2017). Aunque también se ha explotado TL hasta cierto punto, no se han estudiado explícitamente los factores que afectan la transferibilidad de las representaciones espaciales y espacio-temporales en dicha tarea. Se han recopilado distintos conjuntos de avances de películas para MTGC usando procesos que extraen de forma automatizada títulos y videos de IMDb[1] y YouTube[2]. Estos conjuntos contienen desde unos pocos miles de avances con sólo cuatro géneros hasta decenas de miles de avances y decenas de géneros.

Sin embargo, la automatización del proceso de recolección suele llevar a errores debido a que es común que los videos descargados no correspondan con la película o tengan cantidades significativas de publicidad y/o relleno. Estas inconsistencias deterioran significativamente la calidad de los conjuntos de datos recopilados, lo cual es especialmente relevante para la TL dado que múltiples trabajos han mostrado que la transferibilidad de las representaciones espaciales y espacio-temporales se ve afectada por la calidad de las muestras y del etiquetado tanto del conjunto de datos base como del objetivo (Kataoka et al., 2020; L. Zhang, 2019; W. Zhang et al., 2023). Por lo tanto, es necesario un conjunto con etiquetado y muestras de alta calidad para estudiar la transferibilidad de las representaciones de ImageNet y Kinetics a MTGC.

[1] Internet Movie Database: https://www.imdb.com/
[2] YouTube: https://www.youtube.com/

1.2. Objetivos

El objetivo principal de este trabajo es desarrollar un método que mejore la transferencia de conocimiento de conjuntos de imágenes naturales y videos de acciones humanas a la tarea de clasificación de géneros en avances de películas. Adicionalmente, empleando un conjunto de mayor calidad de muestras y etiquetado, deseamos explorar los factores que influyen la transferibilidad.

Más específicamente, buscamos alcanzar lo siguiente:

- Recolectar y proveer públicamente un conjunto multimodal de alta calidad de avances de películas enfocado a la tarea de clasificación multietiqueta de géneros.
- Proponer un método que mejore la transferencia de representaciones aprendidas en ImageNet o Kinetics a la tarea de MTGC.
- Estudiar factores relacionados a la representación de video que influyen en el desempeño de la tarea de clasificación.
- Comparar la capacidad de transferencia de las representaciones aprendidas usando arquitecturas convolucionales y transformadoras.
- Explorar el compromiso de desempeño y eficiencia que se obtiene al usar arquitecturas profundas ligeras y representaciones de video reducidas.

1.3. Hipótesis

En este proyecto partimos de la hipótesis de que es posible mejorar el proceso de transferencia de conocimiento de las representaciones aprendidas en ImageNet o Kinetics a la tarea de clasificación de géneros en avances de películas. Trabajos previos han mostrado que la similitud entre los conjuntos base y objetivo influye en la transferibilidad (Zamir et al., 2018; W. Zhang et al., 2023). Por lo tanto, consideramos que si durante el proceso de adaptación a MTGC logramos reducir cierto tipo de disimilitudes que surgen entre las representaciones y las estructuras espacio-temporales inherentes a los avances, es posible mejorar la transferibilidad.

1.4. Contribuciones

Las principales contribuciones de este trabajo de investigación se detallan a continuación.

- **Conjunto de datos Trailers12k.** Se provee a la comunidad de investigación de un nuevo conjunto de avances de películas multimodal que incluye, además de los avances, metadatos textuales, sinopsis, carteles y representaciones extraídas con redes profundas preentrenadas.
- **Método DIViTA.** Se propone DIViTA (acrónimo en inglés de *Dual Image and Video Transformer Architecture*), un nuevo método que reduce disimilitudes importantes entre representaciones aprendidas en ImageNet/Kinetics y la tarea clasificación de avances de películas para mejorar la transferibilidad. Las mejoras en la transferencia se miden por medio de un estudio de ablación sencillo y también comparando el desempeño de DIViTA con métodos relacionados en la literatura.
- **Estudio de transferibilidad.** Se presenta un estudio de transferibilidad para explorar y evaluar diversos factores que influyen la transferibilidad, como representaciones de video de entrada, patrones de conectividad para extracción de representaciones visuales, arquitecturas de agregación de representaciones espacio-temporales, entre otros factores.

Este trabajo se presenta en el artículo *Improving Transfer Learning for Movie Trailer Genre Classification using a Dual Image and Video Transformer* (R. Montalvo-Lezama et al., 2023), publicado en la revista *Information Processing & Management*.

1.5. Organización de la tesis

El resto de la tesis está organizado en seis capítulos de la siguiente forma:

- El capítulo 2 describe el panorama general de los avances de películas, incluyendo su historia, tipos, elementos y estructura utilizados para su producción.
- El capítulo 3 hace un repaso sobre los trabajos relacionados en las áreas de transferencia de conocimiento, preentrenamiento en ImageNet y Kinetics, y métodos de clasificación de avances.

- El capítulo 4 introduce Trailers12k, un conjunto de avances de películas recolectado mediante un proceso ideado para obtener muestras de video y etiquetas de alta calidad para este proyecto.
- El capítulo 5 describe DIViTA, un método de clasificación propuesto por este proyecto que reduce disimilitudes entre las representaciones de ImageNet/Kinetics y MTGC para mejorar el proceso de transferencia de conocimiento.
- El capítulo 6 reporta la experimentación realizada para estudiar la efectividad de DIViTA, así como una exploración de diversos factores que influyen en la transferibilidad. También se presentan resultados comparativos de DIViTA con otros métodos similares en la literatura.
- El capítulo 7 presenta las conclusiones y delinea investigación a futuro que puede dar continuidad a este trabajo doctoral.

2 Avances de películas

Las campañas de publicidad cinematográficas son esenciales para el éxito de las películas. Comúnmente comienzan meses antes del estreno de la película, aunque en algunos casos pueden iniciar a la par del proceso de grabación. Para la promoción se usa una combinación bastante diversa de contenidos entre los que se incluyen avances, carteles, bandas sonoras, actos públicos del elenco, mercancía, concursos, juegos, etc. De entre todos estos contenidos, los avances son probablemente el tipo de publicidad más determinante para capturar la atención del público. Por ejemplo, en una encuesta realizada por Jerrick (2013), el 99 % de los participantes aseguraron haber visto un avance antes de acudir a ver la película en salas de cine, mientras que el 96 % consideraron que los avances son el medio de publicidad más efectivo para las películas. En un estudio más reciente llevado a cabo por De Jesus y Shapiro (2020), se encontró que existe una correlación positiva entre el nivel de interacción de los usuarios con avances mostrados en las redes sociales Instagram, Facebook y Twitter, y las ganancias alcanzadas por las películas en taquilla. Conscientes de la importancia de los avances en la promoción exitosa de las películas, los estudios cinematográficos destinan una gran cantidad de recursos a este medio de publicidad. De acuerdo al estudio Theatrical Market Statistics llevado a cabo anualmente por la asociación de cine estadounidense Motion Picture Association (MPAA) (Association et al., 2007), la industria cinematográfica gastó $1.6 millones de dólares en la producción de avances por película en promedio. Además, el costo de producción por avance varió entre los $300,000 y $600,000 dólares, y la industria destinó en total $90 millones de dólares a la producción.

2.1. Historia

El nacimiento de los avances de películas ocurrió durante la época dorada del cine de Hollywood entre 1910 y 1920. Uno de los primeros avances del que se tiene registro fue creado en 1913 por Nils Granlund (Daniel, 2015), administrador de publicidad de la cadena de teatros Marcus Loew. Este avance fue creado para el musical de teatro The Pleasure Seekers que se

Figura 2.1: Imágenes iniciales del avance de la película Casablanca de 1942. Tomadas de Bros. (1942).

presentaba en Broadway. En sus inicios, el proceso de elaboración de los avances era rústico, para su producción se tomaban secciones cortas directamente de las películas o tomas no liberadas previamente al público. Durante este periodo, los avances de películas próximas a estrenarse eran mostrados al final de otras funciones presentadas en salas de cine, una práctica conocida como *trailing* en inglés, y por este motivo se popularizó el término *trailer* (Jason, 2016). Los administradores de salas de cines pronto se dieron cuenta que esta práctica resultaba poco efectiva debido a que los espectadores abandonaban las salas tan pronto como terminaba la película, por lo que posteriormente los avances se comenzaron a mostrar antes de comenzar la función principal como se hace hoy en día.

En 1919 Herman Robbins fundó la National Screen Service (NSS) en Estados Unidos, una compañía privada de servicios para estudios y teatros (Jason, 2016). Durante sus primeros años, los servicios de la NSS estaban limitados a la elaboración de avances de video, sin embargo, posteriormente la compañía incursionó progresivamente en el desarrollo de campañas publicitarias, incluyendo también carteles, fotos y otros materiales publicitarios de la época. Para la mitad de la década de 1940, la mayoría de las casas productoras de cine y medios similares usaban los servicios de la NSS para la promoción de sus películas. El estilo de la NSS definió lo que se conoce como la Era Clásica del Avance (Dornaleteche, 2009), en la que el modelo de video consistía en incorporar varias escenas clave de la película, a menudo acompañadas de textos descriptivos de gran tamaño que puntualizaban la historia, y una banda sonora generalmente extraída de las bibliotecas musicales de los estudios. Uno de los ejemplos más representativos de este estilo quedó plasmado en el avance de la película Casablanca de 1942, del que se muestran algunas capturas en la figura 2.1.

La preponderancia de la NSS en el mercado publicitario continuó creciendo hasta convertirse en un monopolio virtual alrededor de 1960 (Wasko, 2003). A partir de ese momento, el panorama publicitario empezó a transformarse, ya que cineastas como Alfred Hitchcock y Stanley Kubrick comenzaron a elaborar los avances para sus películas ellos mismos. El modelo de promoción

sufrió otro cambio importante después del estreno de la película Tiburón (*Jaws* en inglés) de 1975 del director Steven Spielberg, considerada como el primer éxito de taquilla a nivel mundial. La acogida a escala intencional de esta película ayudó a cimentar el modelo de publicidad vigente hoy en día. Este modelo consiste en diseñar la campaña publicitaria para que los avances sean vistos durante las horas de máxima audiencia. Posteriormente, se busca inundar hasta casi la sobresaturación el mercado con estos avances antes del estreno de la película, con la expectativa de que público potencial sepa que ese fin de semana se estrena esa película y su único plan sea verla.

En este punto, los avances comenzaron a usarse no solo como un medio de promoción hacia los espectadores, sino también como una estrategia de recaudación de fondos. Algunos cineastas empezaron a elaborar avances conceptuales exclusivamente con el fin de persuadir a potenciales inversionistas de financiar nuevas películas. El caso más emblemático de este tipo de películas es *The Evil Dead* de 1981 de Sam Rami (Daniel, 2015), que inició siendo una película conceptual de Rami y su amigo el actor Bruce Campbell cuando ambos eran estudiantes universitarios que terminó por convertirse en una saga de culto dentro de la comunidad cinéfila.

Dornaleteche (2009) además de reconocer la Era Clásica del Avance, identifica otras etapas importantes en la evolución de este medio a lo largo de historia, como son la Era Moderna, la Era de Tránsito y la Era del *Blockbuster*. Las transiciones entre estas eras están marcadas por cambios drásticos en los estilos y técnicas empleados para la producción de avances, el uso de los mismos para fines más allá de la promoción hacia audiencias, así como también en las estrategias y procesos usadas en las campañas publicitarias de las películas.

Hoy en día, los avances han evolucionado a tal punto que constituyen una pequeña subindustria dentro de la industria cinematográfica. Han proliferado empresas especializadas en la creación de este tipo de producción, como Acme Trailer Company, The Ant Farm o Trailer Park, aunque también suelen producirse dentro de los mismos estudios que trabajan en la película. Existen galardones especializados, como el Golden Trailer Awards que se dedica a premiar anualmente los mejores avances. Estos galardones también cuentan con subcategorías como spot de TV, carteles, diseños de títulos y avances de videojuegos. Los estilos y técnicas empleadas en la elaboración de avances también se han enriquecido debido a la diversificación de la industria alrededor del mundo con compañías productoras, como Haddock Films en Argentina, Claudie Ossard Productions en Francia, X Verleih AG en Alemania o China Film Group Corporation en China.

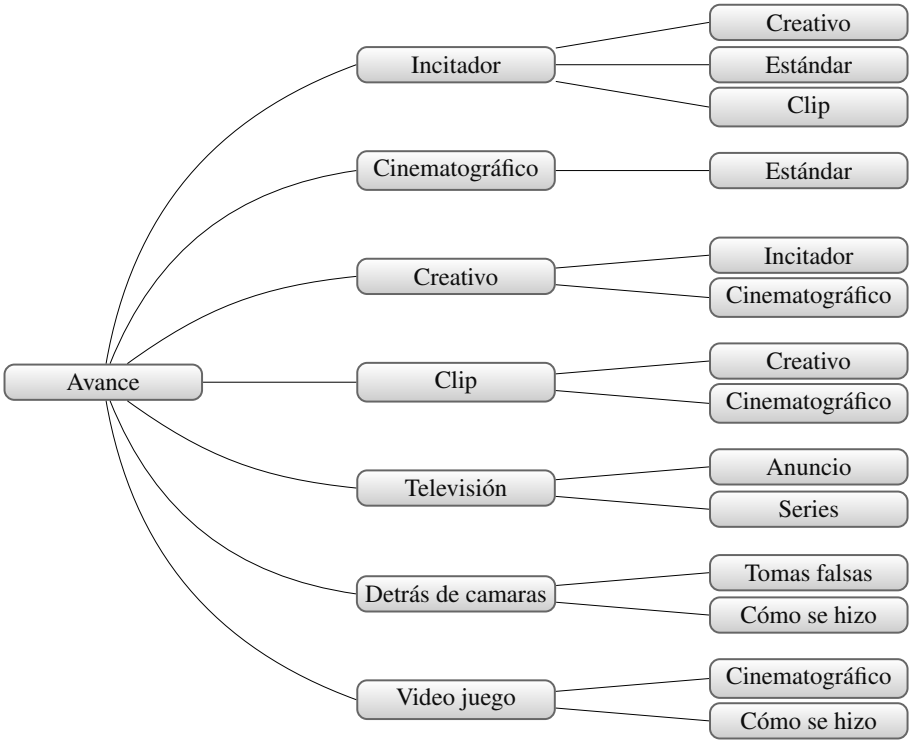


Figura 2.2: Taxonomía del avance como medio de publicidad propuesta por Dornaleteche (2007).

2.2. Tipos

Desde el punto de vista de la mercadotecnia, el principal objetivo de los avances es posicionar un producto efectivamente, en este caso la película, dentro del mercado cinematográfico haciendo uso de material audiovisual. Por estas características se considera al avance como una forma de discurso publicitario (Gil Pons, 2010). Partiendo de esta perspectiva, Dornaleteche (2007) realizó un análisis sobre el papel del avance a lo largo de la campaña publicitaria, partiendo de su objetivo más general como herramienta de la mercadotecnia hasta llegar a ser un medio de publicidad audiovisual. Con base en este análisis, Dornaleteche propone una taxonomía que considera seis tipos de avances, que van apareciendo progresivamente a lo largo de la campaña publicitaria. Cada uno de estos tipos puede contar a su vez con subtipos, como se muestra en la figura 2.2. Durante la campaña publicaría, estos avances se muestran comúnmente en el siguiente orden temporal, dosificando nuevos elementos e información con cada nuevo video para mantener la película presente dentro de la mente de la audiencia potencial.

- Un año antes del estreno. Aparecen primero los avances llamados incitadores (*teasers*). Estos duran menos de un minuto presentando algunos logos, textos o siluetas. Después, aparecen los avances de clip que se conforman por algunos planos o escenas de la película que están disponibles en ese punto de la grabación.
- Seis meses antes del estreno. Se presentan videos promocionales que corresponden más con la noción convencional de avance estándar. Estos están conformados por plano o escenas completas de la película. Comúnmente, se promocionan por redes sociales o plataformas de video bajo demanda. De estos se pueden presentar dos o tres versiones con contenido distinto mostrando progresivamente más detalles de la película.
- Tres meses antes del estreno. Se hace uso principalmente de un único avance, conocido como avance final. Este es una suma de los elementos más atractivos de los avances anteriores. Se muestra en cines y plataformas de video bajo demanda al inicio de otras películas con géneros similares.
- Durante el estreno y después. Se promociona el avance final por todos los medios de comunicación antes mencionados, además de incluir anuncios en otros medios como televisión y eventos deportivos. Es posible que también aparezcan videos con material nuevo, como tomas detrás de cámaras o errores del elenco durante la grabación (*bloopers*).

Si bien la taxonomía de Dornaleteche es útil para entender la evolución de los avances a lo largo de la campaña publicitaria de una película, no aborda las diferencias y objetivos de los avances estándar o finales desde un punto de vista de contenido. En busca de entender los elementos que diferencian a los avances estándar o finales, se han realizado estudios para entender el impacto de diversos elementos de contenido. En un estudio realizado por Johnston et al. (2016), los participantes respondieron que el elenco estrella y equipo de producción, seguidos por la historia y narrativa, son los elementos que tienen una mayor relevancia al recordar la película. Los 10 primeros elementos clave mencionados por los participantes de este estudio se listan en la tabla 2.1.

Antes de ahondar más en la clasificación de avances es relevante conocer el concepto de género cinematográfico. De acuerdo con Bondebjerg (2015), "el género es un concepto utilizado en los estudios y la teoría cinematográficos para describir similitudes entre grupos de películas basadas en aspectos estéticos o sociales, institucionales, culturales y psicológicos". De la misma forma, Bondebjerg indica que las películas clasificadas dentro de un mismo género presentan

Tabla 2.1: Elementos clave memorables en los avances referidos por los participantes del estudio. Solo se muestran los primeros 10 elementos de la tabla original por brevedad. Adaptada de Johnston et al. (2016).

Elementos del avance referenciado	Respuestas
Estrella, actor o elenco (incluyendo una referencia específica al cuerpo del actor)	90
Historia, narración o trama (incluyendo referencias a la adaptación, historia o localización)	71
Elementos estéticos o visuales	64
Banda sonora: música, canción, diálogo o voz en off	46
Divertido, comedia o estrafalario	38
Acción, efectos visuales o espectáculo	36
Reinicio, secuela o franquicia reconocida	24
Personajes	24
Director	21
Género	17

una identidad estilística y temática coherente, lo que las distingue de otras producciones cinematográficas y les otorga características comunes que reflejan su propósito comunicativo único. Por esta razón, Bondebjerg (2015) señala que "un género cinematográfico está constituido por un conjunto de convenciones que influyen tanto en la producción de las películas dentro de ese género como en las expectativas y experiencias de la audiencia". Así mismo, Bondebjerg observa que la industria utiliza los géneros principalmente para crear y promocionar películas, mientras que los críticos y analistas cinematográficos aplican estos mismos géneros en su análisis histórico del cine. Además, la audiencia juega un papel crucial al influir en la selección de las películas a través de sus preferencias y criterios.

En este sentido, Johansen (2013) realiza un análisis considerando los elementos clave y las estrategias de persuasión comúnmente empleadas en los avances para proponer una clasificación. En esta, Johansen identifica que para generar un avance, los productores primero analizan y determinan cuáles son los elementos de la historia de la película que se desean contar. Estos deben ser elementos cruciales para transmitir claramente la historia de la película, pero al mismo tiempo se deben retener ciertos elementos para mantener la expectativa de encontrar sorpresas al ver la película. Una vez definida esta historia, los productores optan por elegir una perspectiva desde la que será contada. Para esto, los productores se apoyan en tres estrategias

de persuasión que van de lo general a lo particular, dando pie a las siguientes categorías.

- **El avance de género**. Es el que presenta una perspectiva más amplia. Este tipo de avance pone énfasis en mostrar elementos asociados a los géneros (acción, comedia, horror, etc.), buscando persuadir al observador que la película incluirá dichos elementos. Por ejemplo, es común encontrar en los avances de acción elementos como peleas, persecuciones, autos, explosiones, etc.

- **El avance narrativo**. La estrategia de este avance consiste en presentar una visión más concisa de la historia; centrándose en los personajes principales, así como sus objetivos y las adversidades que enfrentarán. Estos avances suelen ser de películas con historias familiares para cierta parte del público que pueden provenir de mitos, historietas, libros, películas anteriores, etc.

- **El avance de estrella**. Este tipo de avance apuesta a explotar la popularidad de una persona estrella asociada a la película. Generalmente, se trata de actores o actrices famosos, aunque también puede ser un director, productor o escritor con una amplia trayectoria. La estrategia de persuasión de este avance es simple: si el trabajo anterior de la estrella gustó, verla participar en este avance ayudará a convencer a la audiencia a ver su película más reciente.

2.3. Elementos y estructura

Como sucede con otras obras narrativas, más que guías estrictas que definan la estructura de todos los avances, existen convenciones estructurales en las que los estudios cinematográficos se orientan para la elaboración. En particular, Gil Pons (2010) realiza un paralelismo entre las estructuras normalmente empleadas en los avances y la de las obras literarias, identificando los siguientes cuatro actos narrativos de los que muestra un ejemplo en la figura 2.3.

1. **Introducción**. Comúnmente, en este acto se presenta una escena estableciendo el papel de un personaje principal. También es posible que el avance presente elementos (logotipos, marcas, etc.) de las compañías asociadas (productoras, distribuidoras o directoras). Aquí se busca llamar la atención del público apelando a la popularidad del elenco, personajes, historia o marcas.

(a) Introducción **(b)** Exposición

(c) Argumentación **(d)** Cierre

Figura 2.3: Ejemplos de imágenes en los actos de introducción, exposición, argumentación y cierre del avance de la película El extraño mundo de Jack de 1993. Tomadas de Disney (1993).

2. **Exposición**. Se expone a la audiencia una perspectiva general de la película presentando varios personajes, la historia o los conflictos a los que se enfrentarán. Se trata del acto con mayor duración que tiene como objetivo presentar un relato resumido buscando capturar el interés de la audiencia explotando extractos de la película.

3. **Argumentación**. En este acto aparecen los personajes o actores más conocidos, que pueden no ser los protagonistas, empleando tomas que resalten su participación. También se puede hacer referencia a una obra literaria o hacer uso de cualquier otra información que pueda ayudar a persuadir a la audiencia a ver la película. Además, se pueden usar voz superpuesta o textos en la pantalla para realizar estas referencias.

4. **Cierre**. Hay cuatro pasos que se pueden distinguir en este acto. En el primero se emplean extractos, un momento conocido como clímax, un personaje o una toma nueva para la audiencia. Dependiendo del género de la película, también puede usarse una escena cómica o de terror con la que se busca generar interrogantes en la audiencia. En el segundo paso usualmente se presenta el nombre de la película con un grafismo acorde a la temática abarcando la pantalla completa. El tercer paso muestra la fecha de lanzamiento, que puede ser genérica (próximamente) o tentativa. El cuarto paso opcional es el cierre

en el que aparecen los créditos.

Además de convenciones estructurales, Gil Pons (2010) también identifica tres elementos estéticos que forman parte del proceso de elaboración de avances. El primero es el montaje, que es el proceso que se encarga de armar el video interconectando los personajes, ambientes y objetos. El segundo es la banda sonora. De la misma forma que en la película, se conforma por ruidos, diálogos y música. El último de los elementos estéticos son las imágenes de los personajes. Se eligen tomas de escenas o parte de la narrativa protagonizadas por estos personajes.

3 Clasificación automatizada de avances

La predicción automatizada de géneros en avances de películas ha causado interés en la comunidad de análisis de video en parte debido a que puede permitir asociar géneros a nuevas películas y avances de una misma forma, lo cual es útil para la indización de videos de este tipo de contenido o puede emplearse en el flujo de sistemas de recomendación. Por estos motivos, en la literatura existen diversos trabajos que abordan este problema. Los primeros métodos propuestos (Y.-F. Huang & Wang, 2012; Rasheed et al., 2005; H. Zhou et al., 2010) usaron conjuntamente algoritmos de análisis de imágenes y técnicas de clasificación de texto basadas en bolsas de palabras. Además del video, también se exploró el uso de otras modalidades de información de los avances, como el audio o los metadatos textuales, de forma independiente para la predicción de géneros. La validación del desempeño de estos métodos era bastante limitada, ya que se realizaba sobre conjuntos que apenas contaban con decenas de avances y abarcaban pocos géneros.

Más recientemente, en búsqueda de aprovechar el éxito alcanzado por el aprendizaje profundo en el área de visión computacional, se han propuesto métodos (Simões et al., 2016) basados en redes neuronales del estado del arte ideadas originalmente para tareas de análisis de imágenes. Progresivamente, los métodos (Wehrmann & Barros, 2017) se han vuelto más sofisticados incluyendo nuevas arquitecturas profundas, además de explotar la transferencia de conocimiento de modelos preentrenados principalmente en el conjunto de imágenes ImageNet. En los últimos años, algunos métodos (Bi et al., 2021; Yu et al., 2021) han comenzado a explorar el uso de arquitecturas diseñadas para el procesamiento de video y preentrenadas en el conjunto de video Kinetics para procesar directamente esta modalidad. A diferencia de los trabajos iniciales, los métodos basados en aprendizaje profundo aspiran a aprender representaciones espaciales o espacio-temporales de alto nivel para codificar y explotar la compleja información contenida en los avances.

En este capítulo, primero presentamos el flujo general para llevar a cabo transferencia de conocimiento en redes neuronales profundas. Después, hacemos una revisión de los principales

trabajos sobre trasferencia de conocimiento en el conjunto de imágenes ImageNet y el de videos Kinetics. Finalmente, presentamos los métodos de clasificación de géneros propuestos en la literatura, enfocándonos en aquellos que emplean redes profundas preentrenadas.

3.1. Transferencia de conocimiento

De manera general, la idea de transferencia de conocimiento hace alusión a la capacidad de reutilizar habilidades previamente adquiridas para resolver problemas nuevos. En ciencias de la computación, Bozinovski y Fulgosi (1976) fueron pioneros en introducir el concepto de transferencia de conocimiento, estudiándolo con una red neuronal artificial de cinco capas entrenada para el reconocimiento de caracteres. Esta capacidad también ha sido estudiada por otras disciplinas, como la psicología cognitiva y la pedagogía (Haskell, 2000).

La transferencia de conocimiento toma especial relevancia en el aprendizaje profundo debido a que permite explotar el conocimiento de conjuntos masivos de datos. El proceso para llevarla a cabo puede englobarse en dos etapas: el preentrenamiento y la adaptación (Jiang et al., 2022). El objetivo de la etapa de preentrenamiento es adquirir conocimiento transferible. Para esto, una arquitectura de red neuronal profunda se entrena en una tarea, conocida como tarea fuente o base, que normalmente está equipada con un conjunto de datos a gran escala. Se espera que con este preentrenamiento la red sea capaz de aprender representaciones de los datos que sean generalizables a una familia de tareas similares.

Por otra parte, la meta de la etapa de adaptación es reutilizar el conocimiento aprendido. Para entender esta etapa es conveniente saber que las arquitecturas de redes neuronales se pueden separar en dos partes. La primera parte se conoce como columna y comprende las capas iniciales de la red que tienen como objetivo producir un vector de representación, por ejemplo capas convolucionales o de atención. La segunda parte es la cabeza y se encarga de clasificar el vector de representación. Durante la etapa de adaptación, primero se ensambla una nueva arquitectura adaptada reusando parcial o totalmente las capas preentrenadas de la columna y agregando una nueva cabeza correspondiente a la tarea a resolver. Esta tarea se conoce como tarea objetivo. Finalmente, se entrena parcial o completamente la red adaptada con los datos de la tarea objetivo.

Las técnicas de preentrenamiento han evolucionado de inicialmente ser supervisadas a pasar a técnicas sin supervisión. Hasta hace algunos años, los modelos profundos para visión computacional se preentrenaban de forma supervisada usando una tarea de clasificación.

En el caso de imágenes, ImageNet fue por mucho tiempo el conjunto de preentrenamiento por defecto, mientras que en video un conjunto popular ha sido Kinetics. Sin embargo, esta aproximación no es escalable para explotar conjuntos a mayor tamaño, ya que etiquetarlos puede ser demasiado costoso o infactible. Más recientemente, retomando ideas del área de procesamiento de lenguaje natural se ha comenzado a explorar el uso de tareas sin supervisión para el preentrenamiento. La variante más exitosa de este tipo de métodos se basa en tareas de autosupervisión de tipo contrastivo. Estos métodos buscan aprender representaciones de los datos comparando instancias de tal forma que ejemplos parecidos sean cercanos en el espacio vectorial de representación, mientras que ejemplos diferentes sean distantes en el espacio.

El entrenamiento durante la etapa de adaptación con el conjunto de la tarea objetivo puede llevarse a cabo de dos formas. Se realiza afinado cuando se entrena completamente la red, es decir, tanto las capas y parámetros preentrenados, así como las nuevas capas específicas a la tarea objetivo. En cambio, si las capas y parámetros de la red preentrenada se fijan para que no tengan actualizaciones posteriores, mientras que las capas específicas para la tarea objetivo se entrenan, se dice que la red preentrenada se usa como extractor de características.

3.2. Preentrenamiento en ImageNet y Kinetics

La trasferencia de conocimiento es uno de los pilares en el flujo del aprendizaje profundo que ha permitido a las redes neuronales ser el estado del arte en visión computacional, procesamiento de lenguaje natural, procesamiento de audio, etc. En esta sección hacemos una revisión general de la literatura sobre trabajos que han estudiado o realizado transferencia de conocimiento usando como conjunto fuente ImageNet o Kinetics. En la sección 5.2 realizamos una discusión a mayor profundidad sobre los trabajos que han estudiado los factores que influyen en la transferibilidad, así como su relación en las disimilitudes entre ImageNet/Kinetics y Trailers12k.

En visión computacional, el preentrenamiento en ImageNet ha dado lugar a modelos del estado del arte para una amplia gama de tareas, como la clasificación de imágenes (Kolesnikov et al., 2020; Tan & Le, 2019), la detección de objetos (Girshick et al., 2014; Redmon et al., 2016) y la segmentación de objetos (Girshick et al., 2014; Long et al., 2015). Además de ser ampliamente empleado como conjunto de preentrenamiento para tareas de clasificación generalistas, ImageNet también se ha aplicado como conjunto fuente durante la transferencia hacia conjuntos de datos de grano fino; por ejemplo, aves (Tan & Le, 2019), flores (Kolesnikov et al., 2020), etc. El preentrenamiento en ImageNet incluso se ha aplicado a conjuntos de datos

de dominios específicos, como radiografías de tórax (Ke et al., 2021; Xie & Richmond, 2018), lesiones cutáneas (Lopez et al., 2017), etc. A pesar de que algunas de estas tareas cuentan con conjuntos de imágenes con atributos y formulación de tarea bastante distintos a los de ImageNet, la transferencia se ha vuelto una práctica casi estándar.

Diversos autores se han enfocado en estudiar la transferibilidad de las representaciones aprendidas en ImageNet. Un ejemplo donde se estudió extensamente la transferibilidad es el llevado a cabo por Razavian et al. (2014). En este trabajo se comparó el desempeño de redes neuronales convolucionales preentrenadas usadas como extractores de características con algoritmos hechos "a mano" en varias tareas de análisis de imágenes. En otro trabajo, Yosinski et al. (2014) analizaron el grado de especialización que adquieren las capas en diferentes niveles de redes preentrenadas al transferirlas y comparar el desempeño en tareas objetivo.

De manera similar, Kornblith et al. (2019) investigaron la relación entre el rendimiento en la tarea fuente y el rendimiento en la tarea de objetivo para distintas arquitecturas. Sus resultados indican que existe una fuerte correlación entre la exactitud alcanzada en el preentrenamiento en ImageNet y la exactitud en la tarea objetivo.

Para diversos conjuntos de imágenes, Zamir et al. (2018) propusieron un marco de trabajo de transferibilidad para caracterizar la afinidad entre tareas de visión computacional y generar una taxonomía entre estas. Por otra parte, H.-Y. Zhou et al. (2021) compararon la transferibilidad de arquitecturas convolucionales y transformadoras preentrenadas en ImageNet en varias tareas de análisis de imágenes, encontrando que las transformadoras mostraron una mayor transferibilidad que las convolucionales para la mayoría de las tareas.

ImageNet también se ha utilizado ampliamente para inicializar arquitecturas profundas específicamente diseñadas para tareas de video. Carreira y Zisserman (2017) introdujeron I3D, una de las primeras arquitecturas en esta línea. Para ensamblar I3D, primero se toma una arquitectura de clasificación de imágenes basada en Inception V1 y se inicializa con un preentrenamiento en ImageNet. Después, esta arquitectura se extiende a una de video 3D inflando (copiando) los filtros convolucionales preentrenados a lo largo de la dimensión temporal. La inicialización de I3D con ImageNet supera ampliamente la inicialización aleatoria de parámetros para Kinetics como conjunto de datos objetivo. Otras arquitecturas convolucionales han seguido técnicas similares a I3D para convertir redes de imágenes a video con el fin de mejorar el estado del arte en varios conjuntos de datos de reconocimiento de acciones (Chen et al., 2018; Hara et al., 2018). En los últimos años, el preentrenamiento en ImageNet ha permitido la introducción de arquitecturas transformadoras para tareas de video (Arnab et al., 2021;

Tabla 3.1: Estudios de transferibilidad de representaciones visuales.

Trabajo	Línea de estudio	Conjunto/Tarea	
		Fuente	Objetivo
Razavian et al. (2014)	Características a mano vs. CNNs	ImageNet/IC	Múltiples/IC
Yosinski et al. (2014)	Especialización de capas	ImageNet/IC	ImageNet,Caltech-101/IC
Zamir et al. (2018)	Relación entre tareas	ImageNet/IC	Múltiples/IC
Kornblith et al. (2019)	Desempeño de la tarea fuente vs. objetivo	ImageNet/IC	Múltiples/IC
Kataoka et al. (2020)	Transferibilidad entre conjuntos HAR	Kinetics-700/HAR	Múltiples/HAR
H.-Y. Zhou et al. (2021)	Convolucionales vs. Transformadoras	ImageNet/IC	Múltiples/IC

Bertasius et al., 2021; Liu, Ning et al., 2022).

De manera análoga a los estudios de transferibilidad de ImageNet, diversos trabajos han analizado la transferibilidad de arquitecturas preentrenadas en Kinetics a otros conjuntos de datos populares de reconocimiento de acciones. Para el caso de arquitecturas convolucionales 3D, los estudios existentes (Hara et al., 2018; Tran et al., 2018) han reportado de forma consistente que el preentrenamiento en Kinetics supera a la inicialización aleatoria. Por otra parte, Kataoka et al. (2020) encontraron que el preentrenamiento en conjuntos de datos de video a gran escala ayuda a mejorar el rendimiento de arquitecturas convolucionales 3D. No obstante, los autores puntualizan que la transferibilidad puede beneficiarse más de un conjunto de datos con un etiquetado de mayor calidad, como Kinetics-700, que simplemente de un conjunto de datos de video a mayor escala, como *Moments in Time* (Monfort et al., 2020).

En la literatura también se ha explorado utilizar el preentrenamiento en Kinetics para aplicarlo a otras tareas de reconocimiento de acciones más particulares. Por ejemplo, acciones egocéntricas (Plizzari et al., 2022), reconocimiento de acciones en drones (Choi et al., 2020) o acciones en la oscuridad (Xu et al., 2021). Además, otros trabajos han empleado inicialización de modelos en Kinetics para resolver tareas de video más distantes, como el reconocimiento de lengua de señas (Li et al., 2020) o la toma de decisiones en vehículos autónomos (Singh et al., 2022). La tabla 3.1 lista los principales trabajos de transferibilidad para representaciones de imagen y video.

3.3. Clasificación de géneros en avances de películas

La naturaleza multimodal de la información contenida en los avances y metadatos asociados a las películas ha dado lugar a una amplia variedad de análisis en la literatura desde diferentes perspectivas. Algunos ejemplos de estos trabajos son la recomendación de películas (Deldjoo et al., 2018), predicción de ingresos en taquilla (Ahmad et al., 2020), comprensión de historias completas de películas (Q. Huang et al., 2020), resumen de video (Hu et al., 2022; Kannan et al., 2015), reconocimiento de actores (Shambharkar et al., 2021), clasificación de acuerdo a la edad (Shafaei et al., 2021) o clasificación de géneros basada en cómputo afectivo (Yadav & Vishwakarma, 2020).

Entre estas tareas, el reconocimiento de géneros en avances de películas ha tenido un foco especial en la comunidad de investigación durante las dos últimas décadas. Los trabajos en la literatura pueden agruparse de acuerdo a la aproximación usada por el método propuesto. El elemento central que distingue a la primera oleada de trabajos en abordar este problema es el uso de algoritmos de visión computacional clásicos. En estos trabajos la clasificación de género se planteó como una tarea multiclase, en vez de una formulación multietiqueta más natural. Además, los métodos propuestos se evaluaron en conjuntos de datos pequeños con solo unos cientos de ejemplos. El flujo en general que siguieron estos trabajos (Y.-F. Huang & Wang, 2012; Rasheed et al., 2005; H. Zhou et al., 2010) puede ser descrito de la siguiente forma. En la primera etapa, el video se submuestreaba seleccionando algunos cuadros de forma equidistante o usando un algoritmo para elegir los más relevantes de acuerdo a cierto criterio (por ejemplo, la cantidad movimiento aparente entre cuadros). Después, se extraían características visuales producidas por algoritmos diseñados "a mano" tomados del área de análisis de imágenes. Posteriormente, estas características se usan para calcular las bolsas de características de los cuadros, las cuales son procesadas usando un algoritmo de clasificación de aprendizaje automatizado para predecir el género. Finalmente, se usa alguna estrategia de agregación para determinar la predicción del género global del avance.

Más recientemente, la segunda ola de trabajos ha buscado aplicar técnicas basadas en aprendizaje profundo para proponer métodos para la clasificación de avances. Estos métodos se han evaluado en formulaciones multietiqueta de la tarea, en conjuntos de creciente escala y con mayor número de géneros. Muchos de estos trabajos han abordado la clasificación de avances utilizando únicamente la información visual de los videos y, por lo general, han explotado modelos preentrenados. Simões et al. (2016) presentaron uno de los primeros métodos basados

en redes neuronales profundas donde aplicaron una red convolucional 2D inspirada en la arquitectura VGG inicializada aleatoriamente. Para clasificar, este método primero clasifica de manera independiente cuadros del avance para después utilizar diferentes estrategias de agregación para obtener una predicción de género global del avance. Para evaluar este enfoque, introdujeron LMTD (Simões et al., 2016), un conjunto de datos compuesto por 3,500 avances con cuatro géneros diferentes. En un trabajo posterior, Wehrmann y Barros (2017) propusieron CTT-MMC, un método que aprovecha una arquitectura tipo ResNet preentrenada en los conjuntos ImageNet y Places-365 (B. Zhou et al., 2017). CTT-MMC primero extrae representaciones espaciales a nivel cuadro. Después, estas representaciones son agregadas con un módulo CTT (*Convolution Through Time*), un bloque que emplea convoluciones 1D y produce un vector de representaciones a nivel avance que es clasificado por capas posteriores. Los métodos anteriores se caracterizan por procesar los cuadros de video de forma independiente con redes neuronales profundas y en el segundo caso, aprovechar el preentrenamiento en conjuntos de datos de imágenes. Por estas razones carecen de mecanismos para explotar plenamente las relaciones espacio-temporales codificadas en los cuadros de los avances.

Además de los métodos descritos hasta ahora, existe otra línea de investigación que ha tomado auge reciente dedicada a desarrollar métodos MTGC que exploten fuentes de información más allá de únicamente el video (Behrouzi et al., 2022; Bi et al., 2022; Cascante-Bonilla et al., 2019; Rodríguez Bribiesca et al., 2021). Por ejemplo, Cascante-Bonilla et al. (2019) propusieron un método multimodal que utiliza video, audio, cartel, texto y metadatos. Dada una modalidad secuencial de entrada, este método produce una representación vectorial por cada paso usando un módulo inspirado en fastText (Bojanowski et al., 2017). Este método puede instanciarse para procesar modalidades tanto de forma independiente como conjunta, cuando se instancia para procesar únicamente video se conoce como fastVideo. En este caso, la secuencia de representaciones corresponde a representaciones a nivel cuadro generadas por una red VGG-16 preentrenada en ImageNet, las cuales se agregan usando una capa de atención. Cuando este método usa múltiples modalidades, la representación global del avance se obtiene con un módulo de atención que fusiona las diferentes representaciones a nivel modalidad. En este trabajo el método se evaluó en el conjunto de datos Moviescope (Cascante-Bonilla et al., 2019), conformado por 5,027 ejemplos con avances, carteles y metadatos variados. Rodríguez Bribiesca et al. (2021) extendieron la arquitectura de Cascante-Bonilla et al. (2019) remplazando el mecanismo de agregación basado en fastText por un módulo tipo transformador. Los autores reportan que el uso de mecanismos de atención para la agregación multimodal mejora el rendimiento para

MTGC.

Otro trabajo en esta línea es el realizado por Q. Huang et al. (2020), en el cual introdujeron el conjunto de datos MovieNet. A diferencia de trabajos anteriores, MovieNet es un conjunto para la comprensión holística de películas, por lo que incluye películas completas, subtítulos, avances, carteles, sinopsis, descripciones, metadatos, etc. Además de los datos, con este conjunto se proporcionan varias tareas, como la clasificación de géneros, el reconocimiento de acciones o la clasificación de estilos cinematográficos. Con esto, los creadores de este conjunto buscan fomentar el análisis integral de tareas de películas completas considerando múltiples modalidades. Si bien dentro de este conjunto se agregan avances para las películas incluidas, estos no son el foco del conjunto. Esto queda de manifiesto en que aproximadamente la mitad de las películas en el conjunto tienen más de un avance asociado (diferente versión), y la predicción de géneros, más que ser considerada como la tarea principal, es una más entre varias tareas de diversas modalidades.

Un aspecto que caracteriza a los métodos discutidos hasta el momento es que, a pesar de tener en cuenta la información de múltiples modalidades para llevar a cabo la tarea de MTGC, no explotan las relaciones espacio-temporales codificadas localmente en los cuadros del avance de entrada. Es decir, cada cuadro de video es procesado de forma independiente y la correlación global de esta información se lleva a cabo en etapas finales. En este sentido, trabajos recientes sobre redes convolucionales 3D preentrenadas también se han utilizado para analizar avances buscando una forma más natural de modelar las relaciones espacio-temporales en el video. Por ejemplo, *Video Representation Fusion Network* (VRFN) (Bi et al., 2021) extrae representaciones de clips con una red I3D preentrenada en Kinetics, las cuales se agregan utilizando un módulo basado en LSTM inspirado en CNN-RNN (Wang et al., 2016). Del mismo modo, Yu et al. (2021) propusieron el método *Attention based Spatio-temporal Sequential* (ASTS), que emplea una red BiLSTM seguida de un módulo de atención para clasificar un avance de video a partir de múltiples representaciones de clips extraídas con una red C3D (Tran et al., 2015) preentrenada en el conjunto Sports-1M (Karpathy et al., 2014). Un resumen de los trabajos existentes sobre MTGC utilizando redes neuronales profundas preentrenadas se presenta en la tabla 3.2.

Tabla 3.2: Resumen de los métodos MTGC basados en aprendizaje profundo. Los primeros tres métodos procesan cuadros de manera independiente, los dos últimos usan clips de video.

Trabajo	Preentranmiento/ Columna	Arquitectura para clasificación de géneros	Conjunto de avances	Información procesada
Wehrmann y Barros (2017)	ImageNet/Inception-v3	Conv1D	LMTD-9	Cuadros de video
Cascante-Bonilla et al. (2019)	ImageNet/VGG-16	FastText, Atención	Moviescope	Multimodal
Rodríguez Bribiesca et al. (2021)	ImageNet/CNN	Transformdora, GMU	Moviescope	Multimodal
Yu et al. (2021)	Sports-1M/C3D-like	BiLSTM, Atención	MovieTrailer-14k	Clips de video
Bi et al. (2021)	Kinetics/I3D	C3D-LSTM	LMTD-9	Clips de video

4 Conjunto de datos Trailers12k

En este capítulo describimos a detalle Trailers12k, un conjunto de datos de avances de películas multimodal recolectado durante el desarrollo de este trabajo doctoral. A pesar de existir conjuntos de datos de avances de películas similares en la literatura, estos conjuntos resultaron inadecuados para los fines de este proyecto. Entre las razones están que estos no se encontraron disponibles públicamente durante el desarrollo de este proyecto o porque el procedimiento de recolección empleado para generarlos no garantiza muestras de video con etiquetado de calidad, lo cual es esencial para llevar a cabo el estudio de transferibilidad objetivo de este proyecto. Una comparación de Trailers12k con los conjuntos disponibles se presenta en la tabla 4.1.

En este capítulo primero realizamos una descripción general del conjunto de Trailers12k. Después describimos los pasos seguidos por el procedimiento de recolección para las muestras y verificación del etiquetado en el conjunto. Posteriormente, revisamos los datos y modalidades proporcionados por el conjunto, así como distribuciones y estadísticas relevantes sobre estos datos. Por último, presentamos el protocolo de evaluación que establece tres divisiones del conjunto para la tarea de clasificación multietiqueta de géneros en avances de películas.

4.1. Descripción general

Trailers12k es un conjunto de avances de películas multimodal conformado por 12,000 títulos que se proporciona públicamente a la comunidad de investigación. La figura 4.1 ilustra los principales datos asociados a un título muestra. En este conjunto, los videos fueron obtenidos de YouTube[1], mientras que los carteles y demás datos de IMDb[2]. Si bien Trailers12k es un conjunto de datos ideado para el análisis de video (avances), también considera las modalidades audio

[1] YouTube: https://www.youtube.com/
[2] Internet Movie Database: https://www.imdb.com/

Figura 4.1: Trailers12k es un conjunto de alta calidad de avances de películas conformado por 12,000 títulos. El conjunto provee públicamente datos textuales, URL, representaciones de avances a nivel cuadro y clip, representaciones de cartel y particiones de evaluación para MTGC.

(avances), imagen (carteles), texto (sinopsis, palabras clave, etc.) y otros (elenco, valoración de los usuarios, idiomas, etc.).

Una comparación general de los datos proporcionados por Trailers12k con respecto a otros conjuntos de datos similares presentados en trabajos anteriores se muestra en la tabla 4.1. Trailers12k se distingue por las siguientes cinco fortalezas.

1. Diversidad de datos. Como se muestra en la tabla 4.1, Trailers12k es el conjunto con mayor diversidad de datos asociados en diferentes modalidades y también el segundo con mayor número de muestras.

2. Proceso de recolección. Trailers12k fue generado con un proceso de recolección con el objetivo de obtener un conjunto de alta calidad de muestras y etiquetado. Para esto, durante el proceso de recolección, se realizó una verificación manual de la correspondencia entre el título y su avance, así como también la calidad del vídeo. A su vez, el proceso de recolección considera únicamente avances de películas estrenadas en las últimas dos décadas, buscando homogeneidad en las técnicas de producción cinematográficas usadas

en los avances.

3. Representaciones visuales de modelos profundos. Además de los datos recolectados, con el conjunto también se proveen representaciones espaciales a nivel cuadro de video y espacio-temporales a nivel clip de video computadas con modelos preentrenados en los conjuntos Kinetics e ImageNet-Kinetics, respectivamente. Estas representaciones no se proveen por ningún otro conjunto de avances.

4. Protocolo de evaluación de la tarea MTGC. Junto con los datos y representaciones profundas, se proporcionan tres divisiones, cada una con subconjuntos de entrenamiento, validación y prueba. La meta de estas divisiones es proveer una base para la evaluación sólida de la tarea de clasificación de géneros para futuros trabajos que empleen Trailers12k.

5. Disponibilidad pública garantizada. Otro aspecto importante a resaltar sobre Trailers12k es su disponibilidad, todos los datos del conjunto se encuentran accesibles públicamente en el repositorio de datos abiertos científicos europeo Zenodo[3].

Trailers12k está basando en el conjunto previo Trailers15k (B. Montalvo-Lezama, 2018). A diferencia del conjunto base que solo recolectó las URL de los avances, Trailers12k mejora la calidad del conjunto con el procedimiento de verificación manual, considera múltiples modalidades y provee representaciones computadas con modelos profundos. El apéndice A presenta la hoja de datos del conjunto.

4.2. Procedimiento de recolección

El procedimiento de recolección de Trailers12k fue diseñado principalmente para obtener un conjunto que tenga consistencia entre el título y el avance asociado, lo que se logra con una etapa de verificación manual. Asimismo, se buscó obtener un conjunto con una distribución homogénea en términos de las técnicas de producción cinematográficas usadas en los avances. Para lograr esto, se consideraron aquellos avances con mayor popularidad (puntaje) entre los usuarios y únicamente avances de películas que hayan sido estrenadas en las dos últimas décadas. La segunda restricción se incluyó debido a que los estilos y técnicas de producción cinematográficas evolucionan considerablemente con el tiempo, como se mencionó en el

[3] https://doi.org/10.5281/zenodo.5716409

Tabla 4.1: Comparación de Trailers12k con otros conjuntos de avances de películas. Las columnas marcadas con ✔ indican que los datos se encontraban públicamente disponibles durante la elaboración de esta tabla (marzo del 2023). La columna ImageNet/Kinetics indica que se proveen representaciones extraídas con modelos preentrenados en el conjunto correspondiente. Los conjuntos considerados son H. Zhou et al. (2010), LMTD (Simões et al., 2016), MoviScope (Cascante-Bonilla et al., 2019) y MovieNet (Q. Huang et al., 2020).

Conjunto de datos	Muestras	Verificación manual	Géneros	Avance			Cartel		Argumento	Datos Asoc.
				URL	ImageNet	Kinetics	URL	ImageNet		
Zhou et al.	1,239		4							
LMTD	3,500		9							
MovieScope	5,027		13	✔						
MovieNet	33,000*		28				✔		✔	✔
Trailers12K	12,000	✔	10	✔	✔	✔	✔	✔	✔	✔

*Los autores de MovieNet mencionan 60k avances en su artículo, no obstante, solo 33k de esos videos pertenecen a películas únicas.

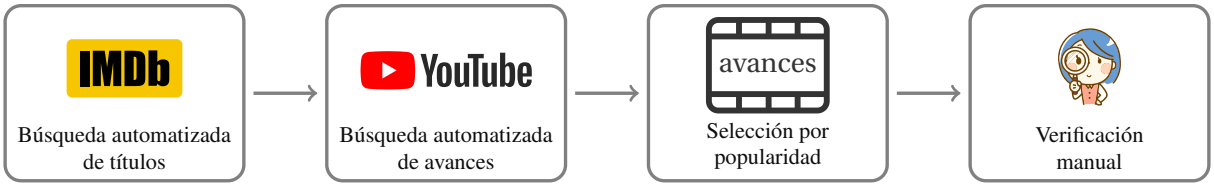


Figura 4.2: Etapas del procedimiento de recolección del conjunto de datos Trailers12k.

capítulo 2. Además, durante el procedimiento de recolección se agregaron una amplia variedad de datos asociados a los títulos (cartel, reparto, duración, palabras clave, etc.) con el fin de enriquecer el conjunto y cubrir diversas modalidades. El procedimiento general es ilustrado en la figura 4.2 y sus etapas se describen a continuación:

1. A partir de una búsqueda automatizada en IMDb, se obtuvieron aquellas películas estrenadas entre los años 2000 y 2019 con el mejor puntaje otorgado por los usuarios.

2. Para cada película, se realizó una búsqueda automatizada en YouTube utilizando el título, añadiendo el año y la palabra “trailer”. De esta búsqueda se descargó el video correspondiente al primer resultado.

3. Se filtraron las películas para conservar únicamente aquellas que tuvieran asignadas al menos uno de los diez géneros más populares en IMDb y que contaran con al menos 500 votos de usuarios.

4. Dado que en una cantidad significativa de casos, el avance resultante no correspondió con el título (pudiendo tratarse de una adaptación o nueva versión, un homónimo, un avance producido por admiradores, etc.), los avances se curaron manualmente. En específico, se verificó manualmente la correspondencia de cada par título-avance y se sustituyeron los avances incorrectos por los mejores disponibles en YouTube. También se sustituyeron avances para cumplir los siguientes criterios de calidad de vídeo: su duración debía estar comprendida entre 60 y 210 segundos, su resolución debía ser de al menos 480p, y debía contener la menor cantidad posible de publicidad y relleno (anuncios, logos o barras de color/cuadros).

En relación con el etiquetado de géneros en IMDb, es relevante conocer que el registro de nuevas películas, junto con sus metadatos, es realizado por usuarios de la plataforma[4]. Comúnmente, estos usuarios son parte del equipo de producción de la película o tienen alguna relación con industria cinematográfica. Además, es habitual que los metadatos sean actualizados con el paso del tiempo. Esto es un detalle a tomar en cuenta, ya que puede que los géneros de las películas recolectadas en Trailers12k sean distintos a los disponibles en IMDb.

4.3. Datos y estadísticas

La tabla 4.2 presenta la lista completa de los datos que conforman el conjunto Trailers12k.

Para las modalidades de video y audio, se proporcionan las URL de los avances. Además, se suministran representaciones espaciales a nivel cuadro de video y espacio-temporales a nivel clip de video computadas con redes convolucionales y transformadoras preentrenadas en los conjuntos Kinetics e ImageNet-Kinetics. En el caso de la modalidad de imagen, se proveen las URL de los carteles y representaciones computadas con una red transformadora preentrenada en ImageNet. Para la modalidad de texto, cada título incluye datos recopilados de IMDb, como argumentos, reparto, puntuación de los usuarios, número de votos de los usuarios, idiomas, sinopsis, clasificaciones por edad, etc.

[4] https://help.imdb.com/article/contribution/titles/genres/GZDRMS6R742JRGAG#

Tabla 4.2: Datos proporcionados por película en Trailers12k.

Datos	Identificador	Tipo
Video y audio		
Representación espacial del avance	`trailers_i_*`	vector
Representación espacio-temporal del avance	`trailers_k_*`	vector
URL del avance	`yt_url`	cadena
Imagen		
Representación del cartel	`posters_i_swin`	vector
URL del cartel	`cover_url`	cadena
Texto		
Identificador de IMDb	`id`	cadena
Título	`title`	cadena
Año	`year`	entero
Géneros	`genres`	lista de cadenas
Argumentos	`plots`	lista de cadenas
Sinopsis	`synopsis`	cadena
Reparto	`cast`	cadena
Directores	`directors`	lista de cadenas
Escritores	`writers`	lista de cadenas
Compositores	`composers`	lista de cadenas
Productores	`producers`	lista de cadenas
Compañías productoras	`production_companies`	lista de cadenas
Idiomas	`languages`	lista de cadenas
Clasificaciones por edad	`certificates`	lista de cadenas
Duración	`runtime`	entero
Votos	`votes`	entero
Puntuación	`rating`	flotante
Palabras clave	`keywords`	lista de cadenas

Respecto al etiquetado de géneros, en Trailers12k cada par título-avance tiene uno o más géneros asociados. Como se mencionó en el capítulo 2, los géneros están entre los atributos más importantes, ya que indican el contenido de la película y suelen influir en las decisiones de la audiencia. En este sentido, Trailers12k está multietiquetado con los diez géneros más populares de IMDb: drama, suspenso, comedia, acción, terror, crimen, romance, aventura, fantasía y ciencia ficción. Las barras azules en la figura 4.3 muestran el número de ejemplos por género en Trailers12k. Como puede observarse, de forma natural hay un fuerte desbalance entre géneros: el género más frecuente (drama) aparece aproximadamente cuatro veces más que el menos frecuente (ciencia ficción).

Por otra parte, la correlación intragénero influye en otros aspectos de la distribución de

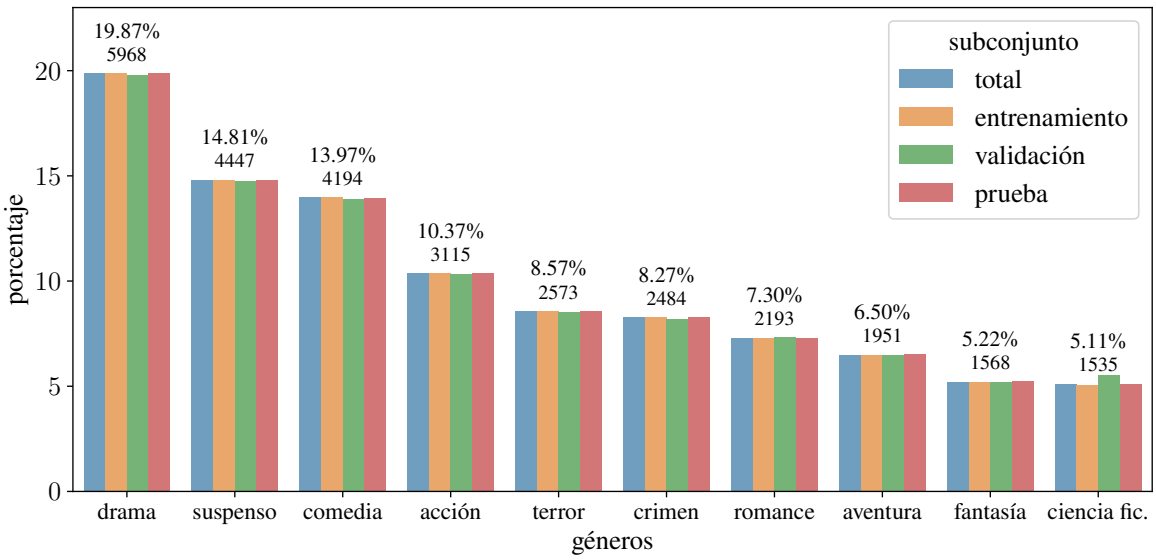
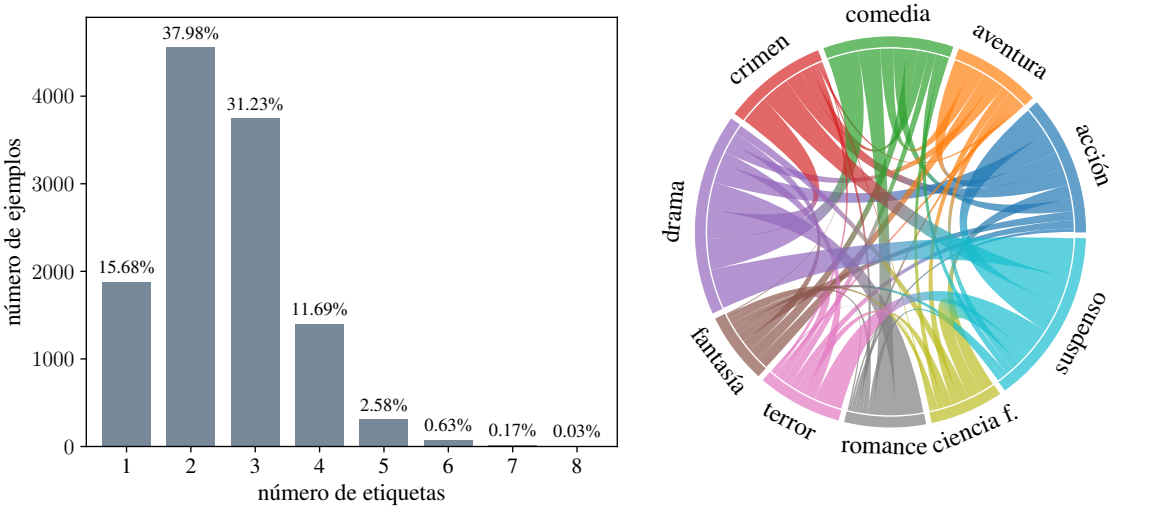


Figura 4.3: Distribución de los géneros en Trailers12k. Gráfica correspondiente a la primera división del conjunto. La distribución de los géneros de forma global (barras azules) se replica aproximadamente en los subconjuntos de entrenamiento (naranjas), validación (verdes) y prueba (rojas). Los porcentajes y números de muestras en la parte superior de las barras de cada género corresponden al conjunto de datos completo (azul).

los géneros. Para arrojar luz sobre esto, la figura 4.4 (a) muestra el histograma del número de etiquetas. Como puede observarse, alrededor del 70 % de los ejemplos tienen 2 o 3 etiquetas. En este sentido, el conjunto de datos tiene una cardinalidad y una densidad (véase Tsoumakas et al. (2010)) de etiquetas de 2.5 y 0.25, respectivamente. Del mismo modo, la figura 4.4 (b) muestra un diagrama de correlación de cuerdas entre géneros. Como puede esperarse, ciertos pares de géneros, como drama-suspenso o comedia-romance, son bastante comunes debido a la afinidad entre estos temas entre sí. De manera opuesta, las películas etiquetadas con pares de géneros poco afines, como crimen-fantasía o aventura-terror, son poco frecuentes.

Sobre la distribución de otros atributos podemos notar lo siguiente. Como se muestra en la figura 4.5 (a), si bien en el conjunto hay 128 países productores, el 85 % de las películas fueron producidas por los 15 primeros países. De un total de 200 lenguas habladas, el 86.1 % de las películas usa una de las 15 lenguas principales, como se aprecia en la figura 4.5 (b). Estados Unidos, Reino Unido y Canadá produjeron el 53.3 % de las películas, lo que explica que el inglés sea el idioma más utilizado, abarcando el 51.1 % de las películas. La figura 4.5 (c) indica una tendencia creciente en los estrenos de películas a lo largo de los años. De acuerdo a la figura 4.5



(a) Histograma del número de etiquetas (b) Diagrama de correlación de cuerdas por género

Figura 4.4: Distribución de géneros: (a) histograma del número de etiquetas y (b) correlación entre pares de géneros.

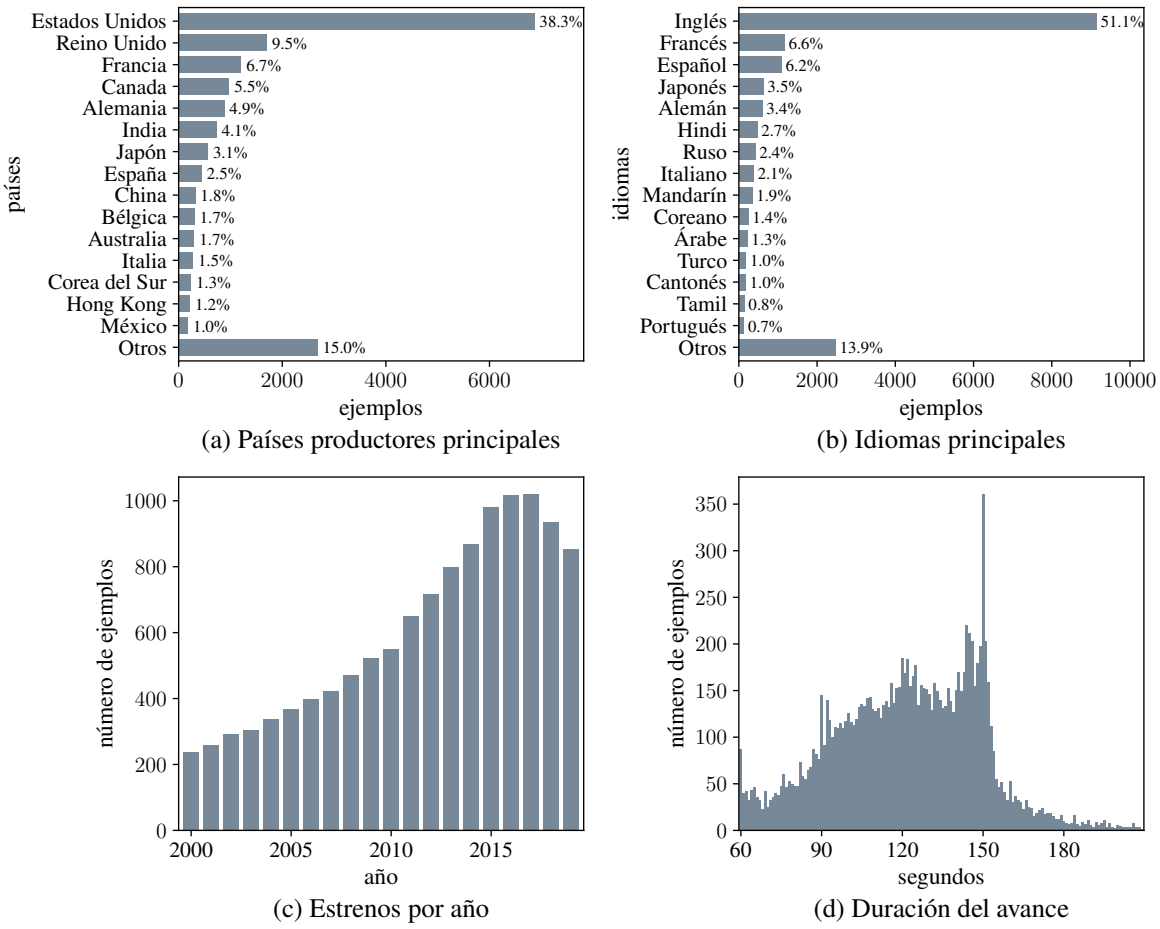
(d), el 87.9 % de los avances tienen una duración menor a los 150 segundos, lo que corresponde a las prácticas habituales para generar avances de la industria (Pepe & Zarzynski, 2016).

Respecto al preprocesamiento de los videos, todos los avances están normalizados a 24 cuadros por segundo. En total, Trailers12k contiene 407.61 horas de video y 35,217,616 cuadros.

Todos los datos de Trailers12k, así como ejemplos de código de uso, se encuentran accesibles en el sitio público del conjunto[5]. Asimismo, los datos están accesibles en el repositorio abierto europeo de datos científicos Zenodo[6]. Este último garantiza que los datos estarán disponibles públicamente por al menos dos décadas[7].

Finalmente, el conjunto de datos Trailers12k se recolectó para fines estrictamente académicos. Por esta razón, el conjunto se distribuye bajo la licencia *Creative Commons Attribution Non Commercial Share Alike 4.0 International*[8], la cual permite el uso libre no comercial del conjunto, y extensiones del mismo siempre que estas se distribuyan bajo la misma licencia. En el apéndice A se pueden encontrar más detalles acerca de la composición, uso y distribución del conjunto.

[5] https://richardtml.github.io/trailers12k
[6] https://doi.org/10.5281/zenodo.5716409
[7] https://about.zenodo.org/principles/
[8] https://creativecommons.org/licenses/by-nc-sa/4.0/legalcode



(a) Países productores principales

(b) Idiomas principales

(c) Estrenos por año

(d) Duración del avance

Figura 4.5: Distribuciones de varios atributos en Trailers12k.

4.4. Protocolo de evaluación

Como se ha mencionado anteriormente, la distribución del conjunto de datos está influida por el desbalance de géneros y la correlación natural intragénero. Esto puede resultar un reto durante la evaluación de modelos predicativos, ya que diversos experimentos pueden emplear diferentes formas de evaluación.

Para mitigar esta dificultad, además de los datos de Trailers12k, se proporcionan tres divisiones (particiones) diferentes de los conjuntos de datos. Esta estrategia está inspirada en siguiendo las formas de evaluación triple usadas en los conjuntos de video de acciones humanas HMDB (Kuehne et al., 2011) y UCF101 (Soomro et al., 2012). Cada división se compone

de tres subconjuntos: entrenamiento (70 %), validación (10 %) y prueba (20 %). Para generar los subconjuntos se utilizó el algoritmo de generación de particiones estratificadas para conjuntos de datos multietiqueta SOIS (Second-order Label Relation) (Szymański & Kajdanowicz, 2017b). SOIS es un algoritmo iterativo que genera particiones en las que los subconjuntos aproximan la distribución global de etiquetas. En particular, se utilizó la implementación de la biblioteca scikit-ml (Szymański & Kajdanowicz, 2017a) para generar tres particiones, la primera de estas se muestra en la figura 4.3. Las otras dos divisiones generadas siguen la distribución global de los géneros aproximadamente del mismo modo. Las divisiones de evaluación se proporcionan en el mismo sitio que los datos del conjunto.

5 Método DIViTA

En este capítulo comenzaremos realizando una revisión de las estructuras narrativas usadas en la producción de avances. Esto nos permitirá comprender las principales disimilitudes entre la información presente de los ejemplos en los conjuntos fuente ImageNet/Kinetics y el conjunto objetivo Trailers12k. Reconocer estas disimilitudes es importante, ya que pueden influir negativamente en el proceso de transferencia de conocimiento hacia la tarea de clasificación de géneros en avances. Finalmente, presentaremos DIViTA, un método de clasificación propuesto por este proyecto que busca atenuar ciertas disimilitudes entre las muestras de los conjuntos base y el objetivo en busca de mejorar el proceso de transferencia de conocimiento.

5.1. Estructura de los avances

Los avances de películas usan una estructura narrativa que busca capturar la atención de la audiencia. Esta estructura hereda elementos de producción de las películas completas, por lo que es útil conocer un poco más las composiciones usadas en el proceso de producción cinematográfico.

Para el desarrollo de la narrativa de una película completa, en el proceso de producción cinematográfico se usa una familia de composiciones jerárquicas complejas de estructuras (Katz, 2019). La unidad más elemental es el cuadro, una imagen fija. Los cuadros se usan principalmente para establecer escenarios. En el siguiente nivel está el plano, una sucesión de cuadros sin corte de cámara. Por lo general, un plano tiene un fondo único poniendo el foco en personajes u objetos que aparecen en la mayoría de los cuadros y pueden realizar algún tipo de dinámica (e.g., dos personas abrazándose). Subiendo en la jerarquía de composición de la película se encuentran las escenas. Estas se emplean para presentar un bloque de narración a través de una secuencia de planos unidos por transiciones con continuidad de lugar, personajes y tiempo. Además de estas estructuras, durante el proceso de producción se emplean composiciones de mayor jerarquía, como las secuencias y tomas.

Para la producción de avances, comúnmente se retoman de la película ciertos planos y escenas especialmente llamativos dado que se trata de videos cortos destinados a capturar la atención de la audiencia (Dornaleteche, 2007). En el avance, es común que los planos y escenas elegidos estén organizados en una secuencia que no suele corresponder con el orden temporal original de la película. Sin embargo, no existen guías definitivas para la producción de avances. De hecho, los estilos utilizados en la producción de avances evolucionan constantemente, y como se mencionó en la sección 2.3, los cambios abruptos en los estilos marcan las eras en la historia de los avances.

Una aproximación más formal para analizar la estructura narrativa de los avances es la teoría de la gramática narrativa visual (GNV) propuesta por Cohn et al. (2012a). La GNV describe un modelo de comprensión secuencial de imágenes que integra herramientas de la lingüística y la experimentación psicológica. Este modelo establece una analogía entre el lenguaje e imágenes secuenciales. La GNV sostiene que una secuencia de imágenes presenta pistas semánticas que corresponden a funciones narrativas organizadas de manera jerárquica para generar un significado discursivo. La GNV considera que los componentes de narrativa usados en las secuencias de imágenes pueden caer en una de las siguientes cuatro categorías básicas.

- Establecedora (*establisher*). Delinea la interacción de diversos elementos de manera pasiva sin presentar acciones concretas.
- Iniciadora (*initial*). Comienza la tensión de la narrativa preparando para presentar la acción principal o como fuente hacia la ruta narrativa principal.
- Cúspide (*peak*). Marca el clímax de la tensión en la narrativa y el punto de máxima estructura de acontecimientos. Comúnmente, presenta una acción terminada o la meta en la ruta argumentativa, también puede presentar una acción interrumpida.
- Liberadora (*release*). Libera la tensión en la interacción, ocurre comúnmente después de la acción final o como colofón.

A nivel global existen paralelismos entre las categorías de la GNV y los actos discursivos (introducción, exposición, argumentación y cierre) de las obras literarias descritos en la sección 2.3. No obstante, la GVN sostiene que a mayor complejidad de la secuencia, la narrativa se compone de una jerarquía de múltiples niveles en la que cada nivel puedes ser descompuesto

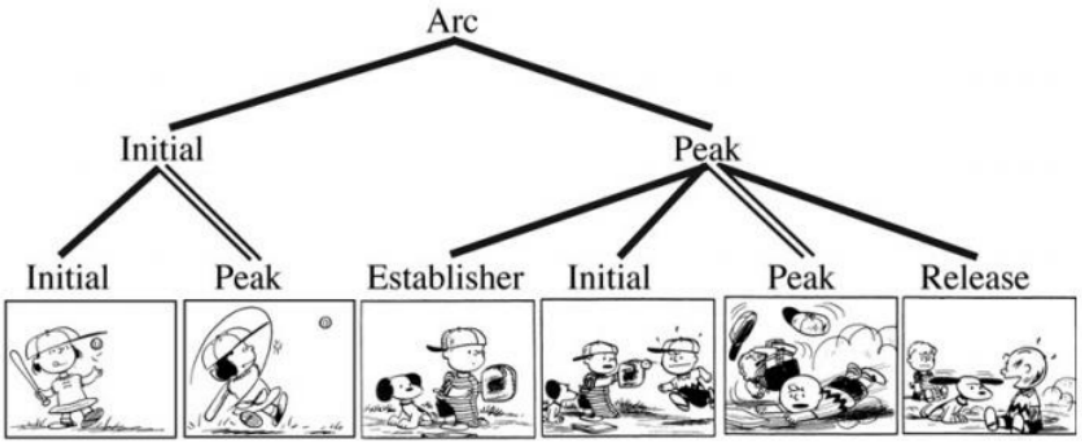


Figura 5.1: Categorías de la estructura jerárquica de una historieta. Imagen tomada de Cohn et al. (2012b).

a su vez en estas categorías. La figura 5.1 ilustra un ejemplo de la estructura jerárquica propuesta por GNV de una secuencia de imágenes.

La GNV ha sido aplicada para el análisis del discurso de historietas (Cohn, 2014) y películas completas (Cohn, 2016). Para el caso de los avances, de Jonge (2019) aplicó la GNV para el análisis de la narrativa encontrando hallazgos particulares en este tipo de contenido. Primero, los avances cuentan comúnmente con solo un nivel discursivo, donde las categorías están construidas principalmente a partir de planos cinematográficos. Segundo, las estrategias narrativas usadas en la producción de los avances han evolucionado a lo largo de la historia de una narrativa que seguía el orden global de la película, a una estructura cada vez más fuera de su orden temporal. De acuerdo con Johnston (2008), estos cambios en las estrategias empleadas en la producción de avances han sido motivados por la búsqueda de incrementar la interactividad del público con los avances, explotando las capacidades de las plataformas de video bajo demanda modernas.

5.2. Disimilitudes entre ImageNet/Kinetics y Trailers12k

Como mencionamos en el capítulo 3, el flujo estándar para llevar a cabo transferencia de conocimiento consta de dos etapas. Primero, en la etapa de preentrenamiento se preentrena un modelo en un conjunto de datos fuente. Después, en la etapa de adaptación se adapta una nueva arquitectura para la tarea objetivo reutilizando parte del modelo preentrenado. Comúnmente, en la adaptación del modelo preentrenado se descartan las capas específicas relacionadas con la

tarea fuente (capas finales) y se remplazan por capas acorde a la tarea de objetivo. Por último, la nueva arquitectura se entrena en el conjunto de datos objetivo.

Diversos autores han encontrado evidencia de que el rendimiento de la transferencia en la tarea objetivo está influenciado por múltiples factores, como el tamaño del conjunto de datos (Cherti & Jitsev, 2022; Kolesnikov et al., 2020; Soekhoe et al., 2016), la variabilidad del dominio (Cherti & Jitsev, 2022), la capacidad de la arquitectura para aprender representaciones generalizables (Tan & Le, 2019), y la similitud entre los conjuntos fuente y objetivo (W. Zhang et al., 2017). Otros trabajos (Razavian et al., 2014; Yosinski et al., 2014; Zamir et al., 2018) han estudiado la transferencia de conocimiento para diferentes tareas de análisis de imágenes y han encontrado consistentemente que una mayor similitud entre las tareas fuente y objetivo resulta en una mejor transferibilidad, produciendo un mayor rendimiento en la tarea objetivo. En general, se produce una transferencia positiva cuando la transferencia beneficia el rendimiento en la tarea objetivo en comparación con la inicialización aleatoria. Por el contrario, si el rendimiento empeora al utilizar transferencia, se habla de transferencia negativa (Rosenstein et al., 2005). Algunas de las posibles causas de transferencia negativa que se han identificado en la literatura son la disimilitud entre los dominios fuente y objetivo, la aplicación de métodos de transferencia ingenuos y la calidad de los conjuntos de datos fuente y objetivo (W. Zhang et al., 2023).

En este trabajo realizamos una revisión a la etapa de adaptación durante la transferencia de conocimiento y proponemos un procedimiento de adaptación que promueve la transferencia positiva para la clasificación de avances. Para entender este procedimiento, denominado Generación de Fragmentos, primero vamos a analizar las disimilitudes importantes entre las tareas fuente y objetivo. Si bien un avance puede verse como una secuencia de imágenes correlacionadas, su contenido y estructura difieren significativamente de las imágenes de ImageNet y los videos de Kinetics, como se ilustra en la figura 5.2. En particular, nos centramos en las siguientes tres disimilitudes:

(a) **Contenido espacial**. Es habitual que el avance de una película esté conformado por elementos (e.g., personajes, objetos, paisajes, etc.) o acciones ficticias (e.g., dinámicas que violan leyes físicas) que no están presentes en las imágenes naturales contenidas en ImageNet ni en los clips de acciones humanas en Kinetics.

(b) **Estructura del video**. Como se discutió en la sección 5.1, los avances tienen una estructura compuesta por escenas, cada una conformada por planos. Esta estructura

Figura 5.2: Disimilitudes entre ImageNet, Kinetics y Trailers12k. La estructura común de un avance es una composición de planos con contenidos y dinámicas de casi cualquier naturaleza.

difiere significativamente de la estructura de los clips de acciones humanas, que se componen usualmente de unos pocos cuadros enfocados en seres humanos que realizan una acción.

(c) **Duración del video**. Los avances en Trailer12k tienen una duración promedio aproximada de 122 segundos. Esto supera en un orden de magnitud la duración de 10 segundos de los clips de Kinetics-400 (Kay et al., 2017). Esto implica diversos retos para el análisis de video, como la necesidad de modelos con la capacidad de capturar relaciones espacio-temporales más largas. Además, videos más largos obligan a considerar aspectos técnicos como mayores requerimientos computacionales de almacenamiento, memoria y poder de procesamiento.

La etapa de Generación de Fragmentos del método propuesto DIViTA por este proyecto tiene como objetivo mejorar el proceso de transferencia de conocimiento, enfocándose en atenuar las disimilitudes (b) y (c). Esta etapa se describe detalladamente en la sección 5.3.

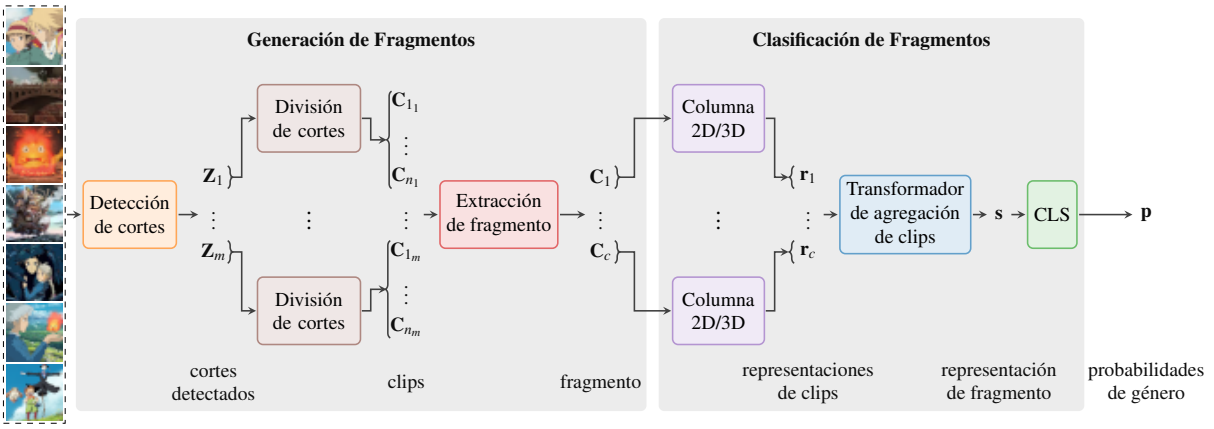


Figura 5.3: Flujo de cómputo de DIViTA durante la fase de entrenamiento.

5.3. Método de clasificación DIViTA

DIViTA es un método de clasificación de avances de películas propuesto en este proyecto. Consta de dos etapas, la Generación de Fragmentos y Clasificación de Fragmentos, como se ilustra en la figura 5.3. A grandes rasgos, la etapa de Generación de Fragmentos extrae del avance de entrada un fragmento de video corto compuesto por una secuencia de clips, donde cada clip se preprocesa para convertirlo en una entrada más adecuada para un modelo preentrenado. La etapa de Clasificación de Fragmentos toma como entrada el fragmento extraído, genera una representación del fragmento agregando representaciones espaciales/espacio-temporales a nivel de clip y clasifica el avance utilizando esta representación. Las instancias específicas de columnas de extracción de representaciones, así como el algoritmo de detección de cortes, se discuten en el capítulo 6.

De manera más formal, el algoritmo 1 presenta el procesamiento de DIViTA durante la fase de inferencia. En la etapa de Generación de Fragmentos (líneas 1 a 9) se toma un avance de entrada $\mathbf{A}$ con l cuadros y se extrae un fragmento $\mathbf{S}$ con c clips, cada uno con f cuadros. En esta etapa se realizan cuatro pasos. En el primer paso (línea 2), un algoritmo de detección de cortes particiona el avance de entrada en una secuencia de m cortes $(\mathbf{Z}_1, \ldots, \mathbf{Z}_m)$. Estos son segmentos cortos de video de longitud variable delimitados por las transiciones detectadas (cuadros negros, fundidos, disoluciones, etc.). Este paso busca aproximar las composiciones de planos utilizadas durante el proceso de producción del avance, tal y como se describe en la sección 5.1. En el segundo paso (línea 4), cada corte $\mathbf{Z}_i$ se divide en segmentos más pequeños denominados clips $\mathbf{Z}'_i = (\mathbf{C}_{1_i}, \ldots, \mathbf{C}_{n_i})$ donde los primeros $n_i - 1$ clips tienen f cuadros. Si el

Algoritmo 1: Flujo de cómputo de DIViTA durante la fase de entrenamiento.

Datos: avance $\mathbf{A}$
Resultado: probabilidades de género $\mathbf{p} = (p_1, \ldots, p_g)$

1 **Etapa de Generación de Fragmentos:**
2 $(\mathbf{Z}_1, \ldots, \mathbf{Z}_m) \leftarrow detectar\_cortes(\mathbf{A})$;
3 **para** $\mathbf{Z}_i \in (\mathbf{Z}_1, \ldots, \mathbf{Z}_m)$ **hacer**
4 $\mathbf{Z}'_i \leftarrow dividir\_corte(\mathbf{Z}_i)$;
5 **si** $|\mathbf{C}_{n_i}| < f$ *donde* $\mathbf{C}_{n_i} \in \mathbf{Z}'_i$ **entonces**
6 $\mathbf{C}_{n_i} \leftarrow rellenar\_clip(\mathbf{C}_{n_i})$;
7 $\mathbf{S} \leftarrow extraer\_fragmento((\mathbf{C}_{1_1}, \ldots, \mathbf{C}_{n_1}, \ldots, \mathbf{C}_{1_m}, \ldots, \mathbf{C}_{n_m}))$;
8 **si** $usando\_columna\_2D()$ **entonces**
9 $\mathbf{S} \leftarrow submuestrear\_clips(\mathbf{S})$;
10 **Etapa de Clasificación de Fragmentos:**
11 **para** $\mathbf{C}_j \in \mathbf{S}$ **hacer**
12 $\mathbf{r}_j \leftarrow computar\_representacion\_de\_clip(\mathbf{C}_j)$;
13 $\mathbf{s} \leftarrow computar\_representacion\_de\_fragmento((\mathbf{r}_1, \ldots, \mathbf{r}_c))$;
14 $\mathbf{p} \leftarrow clasificar\_representacion\_de\_fragmento(\mathbf{s})$;
15 **devolver** $\mathbf{p}$

último clip $\mathbf{C}_{n_i}$ tiene menos de f cuadros, se rellena al final con cuadros negros hasta alcanzar los f cuadros (líneas 5 y 6). En este punto, el avance se ha transformado en una secuencia compuesta por los clips de todos los cortes $\mathbf{T} = (\mathbf{C}_{1_1}, \ldots, \mathbf{C}_{n_1}, \ldots, \mathbf{C}_{1_m}, \ldots, \mathbf{C}_{n_m})$. En el tercer paso (línea 7), se extrae el fragmento seleccionando c clips adyacentes de $\mathbf{T}$ para formar un fragmento $\mathbf{S}$ del avance. Observemos que con estos pasos, $\mathbf{S}$ tiene una alta correlación a dos niveles diferentes. A un nivel inferior entre cuadros debido a que los cuadros de un clip se espera que pertenezcan a un mismo corte detectado. A un nivel superior entre clips porque todos los clips de un fragmento son adyacentes. En el último paso (líneas 8 y 9), que sólo se realiza cuando el método usa una columna 2D (columna de extracción de representaciones de imágenes), cada clip $\mathbf{C}_j \in \mathbf{S}$ se representa seleccionando un único cuadro de los f cuadros totales.

Por otro lado, la etapa de Clasificación de Fragmentos (líneas 10 a 14) es una arquitectura de red neuronal profunda que clasifica el fragmento extraído del avance. Esta consta de tres módulos: una columna 2D/3D para obtener representaciones espacio-temporales de los clips,

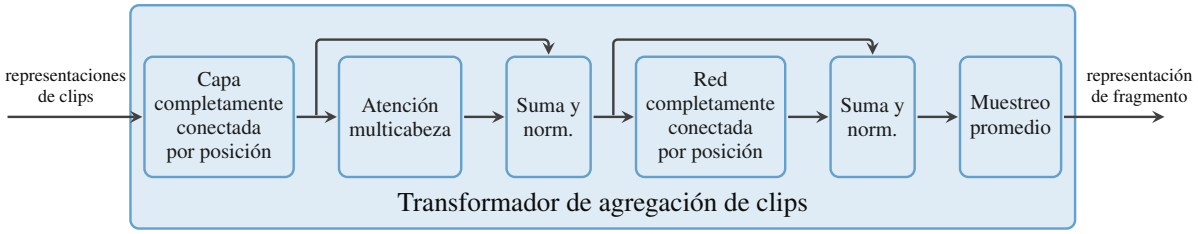


Figura 5.4: El módulo transformador de agregación de clips está basado en la red transformadora original de Vaswani et al. (2017). Al inicio del módulo, se incorpora una capa completamente conectada por posición para reducir el tamaño de los vectores a nivel de clip.

un módulo transformador para agregar información espacio-temporal y una capa lineal (CLS) para la clasificación. En el primer paso (líneas 11 y 12), la columna 2D/3D genera un vector de representación $\mathbf{r}_j \in \mathbb{R}^b$ para cada clip del fragmento $\mathbf{C}_j$, donde b es el tamaño de salida de la columna. Este módulo se construye transfiriendo las capas de extracción de representaciones de una arquitectura de clasificación preentrenada en ImageNet y/o Kinetics. En el caso de una columna 2D para imágenes, $\mathbf{C}_j$ es solo un cuadro, por lo que $\mathbf{r}_j$ codifica información puramente espacial. En cambio, para una columna 3D para video, dado que $\mathbf{C}_j$ es una secuencia de cuadros, $\mathbf{r}_j$ codifica información espacio-temporal. En ambos casos, la salida es una secuencia de representaciones de clips $(\mathbf{r}_1, \ldots, \mathbf{r}_c)$. En el segundo paso (línea 13), esta secuencia es combinada por el módulo de agregación de clips en un único vector $\mathbf{s} \in \mathbb{R}^d$ con información espacio-temporal a nivel de fragmento. La arquitectura del módulo transformador de agregación de clips se ilustra en la figura 5.4.

La arquitectura del módulo transformador consta de un único módulo basado en la red transformadora original introducida por Vaswani et al. (2017), no obstante, incorpora una capa completamente conectada por posición al inicio para reducir cada representación de clip de entrada $\mathbf{r}_j$ a un vector de tamaño $d < b$. La intuición detrás de esta última capa es reducir la explosión de parámetros en las capas siguientes y, por tanto, ayudar a mitigar el sobreajuste. Los cuatro bloques siguientes del módulo transformador de agregación de clips producen una nueva secuencia de representaciones de clips que busca capturar las dependencias entre las representaciones de clips. La capa de submuestreo promedio al final se aplica sobre la dimensión temporal para agregar la secuencia en un único vector de representación $\mathbf{s}$ para todo el fragmento. El último paso (línea 14) consiste de una capa completamente conectada CLS seguida de una función de activación sigmoide que clasifica $\mathbf{s}$, produciendo un vector $\mathbf{p}$ de tamaño g donde cada entrada representa la probabilidad de que el fragmento pertenezca a un

género determinado.

El procesamiento de DIViTA depende de si se encuentra en la fase de entrenamiento o de inferencia. Durante la fase de entrenamiento, se utiliza un único fragmento $\mathbf{S}$ para clasificar el avance de entrada. En el tercer paso de la Generación de Fragmentos (línea 7), se selecciona sólo un fragmento del avance $\mathbf{S}$ compuesto por c clips adyacentes eligiendo una posición inicial de $[1, |T| - c]$ uniformemente al azar. En esta fase, el vector de probabilidad del fragmento $\mathbf{p}$ se considera la clasificación del avance completo. Dado que se pueden generar distintos fragmentos desde distintas posiciones iniciales, esta estrategia de Generación de Fragmentos (línea 7) proporciona un efecto de aumentado de datos implícito en la dimensión temporal y, al mismo tiempo, reduce los requisitos de procesamiento y memoria durante el entrenamiento; lo que es de gran relevancia durante la retropropagación. Durante la fase de inferencia, todos los fragmentos $(\mathbf{S}_1, \ldots, \mathbf{S}_q)$ se utilizan para clasificar el avance de entrada. En el tercer paso de la Generación de Fragmentos, la secuencia de clips del avance $\mathbf{T}$ se divide en una secuencia de fragmentos $(\mathbf{S}_1, \ldots, \mathbf{S}_q)$. La clasificación del avance completo se obtiene promediando por género los vectores de probabilidad $(\mathbf{p}_1, \ldots, \mathbf{p}_q)$ de todos los fragmentos.

6 Experimentos y resultados

En este capítulo presentamos nuestros resultados experimentales evaluando empíricamente la capacidad de DIViTA para llevar a cabo transferencia de conocimiento. Primero, describimos hiperparámetros importantes fijados en la configuración experimental por defecto usada durante el entrenamiento y evaluación. También mencionamos los métodos de referencia como forma de comparación.

Después, presentaremos experimentos para estudiar diversos factores que influyen en el proceso de transferencia. En particular, evaluamos el impacto de la estrategia de generación de clips y número de cuadros por clip en la transferibilidad de las representaciones de ImageNet o Kinetics. También estudiamos diferentes tamaños de fragmentos y estrategias de agregación de fragmentos. Además, analizamos el efecto de usar columnas convolucionales y transformadoras preentrenadas en ImageNet o Kinetics comparando su rendimiento y requisitos computacionales. Finalmente, comparamos el desempeño de DIViTA con métodos pertinentes en la literatura.

6.1. Configuración experimental

Para la evaluación empírica, fijamos ciertos hiperparámetros de entrenamiento y del método, mientras que variamos otros para estudiar su impacto en términos de diferentes métricas de rendimiento. A continuación, detallamos la configuración experimental por defecto. El código para reproducir los principales resultados se encuentra disponible públicamente[1].

6.1.1. Entrenamiento

Como función de pérdida se utilizó la entropía cruzada binaria. Los modelos fueron entrenados durante 100 épocas en lotes de 32 ejemplos. Adoptamos un criterio de paro temprano respecto a la función de pérdida en el conjunto de validación. Utilizamos el optimizador AdamW

[1] https://github.com/richardtml/DIViTA

Tabla 6.1: Hiperparámetros para la configuración de DIViTA con mejor desempeño.

Hiperparámetero	Configuración
Entrenamiento	
Inicializador	Kaiming uniforme
Optimizador	AdamW
Tasa de aprendizaje inicial	1×10^{-4}
Tamaño del lote	32
Épocas de entrenamiento	100
Decaimiento de la tasa de aprendizaje	0.1
Actualización de la tasa de aprendizaje	Reducción en *plateau*
Paciencia para reducción en plateau	20 épocas
Arquitectura	
Clips por fragmento	30
Cuadros por clip	24
Columna 2D	Swin-2D (ImageNet)
Columna 3D	Swin-3D (ImageNet-Kinetics)
Transformador de agregación de clips	4 cabezas de 128

(Loshchilov & Hutter, 2019) con una tasa de aprendizaje inicial de 1×10^{-4} que se reduce en un factor de 10 cada vez que la pérdida de validación se estabiliza durante 20 épocas. También, se congelaron los pesos de la columna preentrenada para reducir los recursos computacionales y el tiempo de entrenamiento. Para todos los experimentos utilizamos un servidor DGX A100.

A menos que se indique lo contrario en una sección en concreto, la configuración por defecto para la etapa de Generación de Fragmentos se fija a 30 clips por fragmento, cada uno de los cuales se compone de 24 cuadros tomados de la salida del detector de cortes. Esta configuración denominada Shot-24 se detalla en la sección 6.2. La etapa de Clasificación de Fragmentos usa una columna Swin-2D (Liu et al., 2021) preentrenada en ImageNet-1K o una columna Swin-3D (Liu, Ning et al., 2022) preentrenada en ImageNet-1K y Kinetics-400. El módulo transformador de agregación de clips consta de 4 cabezas con proyecciones lineales de 128 dimensiones. Los hiperparámetros de la arquitectura y el entrenamiento para la configuración de DIViTA con mejor desempeño se presentan en la tabla 6.1.

6.1.2. Evaluación

Se eligieron cuatro métricas calculadas a partir del área bajo la curva (AUC, por las siglas en inglés de *Area Under the Curve*) de precisión-exhaustividad que se emplean habitualmente en trabajos de clasificación de avances de películas multietiqueta (Cascante-Bonilla et al., 2019; Wehrmann & Barros, 2017).

- Promedio micro μAP. Se calcula una sola AUC utilizando todas las etiquetas globalmente, esto es, se reduce el problema a la evaluación de una sola tarea de clasificación binaria. Esta métrica proporciona información global sobre las predicciones, permitiendo que las clases más frecuentes tengan un mayor impacto en el rendimiento.
- Promedio macro mAP. En esta se calcula una AUC por clase y se promedian los resultados. Esto proporciona información sobre el rendimiento de las clases independientemente de su frecuencia.
- Promedio ponderado wAP. Esta métrica es similar a mAP, pero el promedio se pondera por la frecuencia de la clase. A diferencia de mAP, la métrica wAP tiene en cuenta la frecuencia del género.
- Promedio por muestra sAP. Se calcula una AUC por muestra y se promedian.

Todos los modelos se entrenaron y evaluaron en cada una de las tres divisiones de Trailers12k; para cada métrica se reporta el promedio y la desviación estándar de los tres conjuntos de prueba.

6.1.3. Métodos de referencia

Para estudiar el impacto de los diferentes componentes de DIViTA, se realizó un estudio de ablación sencillo en el que se sustituyeron componentes clave por alternativas más sencillas. También comparamos DIViTA con dos métodos MTGC unimodales que sólo emplean cuadros de video. Estos métodos fueron discutidos en el capítulo 3: CTT-MMC-A y fastVideo. Recordemos que CTT-MMC-A es una de las variantes de la arquitectura CTT propuestas por Wehrmann y Barros (2017), mientras que fastVideo es el método de agregación de representación a nivel cuadro basado en fastText propuesto por Cascante-Bonilla et al. (2019). Para realizar una comparación justa, a diferencia del trabajo original, utilizamos una columna preentrenada

únicamente en ImageNet para CTT-MMC-A y omitimos el preentrenamiento adicional en Places360. Además de estos métodos, comparamos DIViTA con TimeSformer (Bertasius et al., 2021), una arquitectura transformadora para clasificación de video que aprende representaciones espacio-temporales a partir de una secuencia de parches a nivel de cuadro. Reproducimos estas arquitecturas lo más fielmente posible a partir de las descripciones de sus artículos y código públicamente disponible, y las entrenamos y evaluamos en Trailers12k. A diferencia de los métodos de referencia, durante el entrenamiento DIViTA usa una representación abreviada del avance que provee ciertas ventajas, lo cual se discute en la sección 6.4.

6.2. Particionamiento de cortes

Los dos primeros pasos en la etapa de Generación de Fragmentos tienen por objeto reducir las disimilitudes entre imágenes/clips de acciones humanas y los planos en avances de películas. Para esto, primero se segmenta el avance de entrada en cortes de longitud variable y, a continuación, se divide cada corte en clips con un número f fijo de cuadros. Para segmentar los avances en cortes, utilizamos la red de detección de transiciones TransNet V2 (Souček & Lokoč, 2020). Esta es una arquitectura que procesa los videos mediante una pila de bloques convoluciones dilatados denominada SDDCNN. Para entrenar a TransNet V2, se generan muestras seleccionando aleatoriamente dos videos y uniéndolos con transiciones de corte duro o disolución. La figura 6.1 muestra el histograma del número de cuadros por corte para Trailers12k que se obtiene usando TransNet V2.

El número f de cuadros en el clip determina la cantidad de información ingresada a la columna preentrenada, lo que puede influir en la transferibilidad. Esto es de especial importancia en el caso de representaciones aprendidas en Kinetics, ya que son generadas utilizando todos los cuadros del clip. Para estudiar esto, consideramos dos longitudes de clip $f = 24$ y $f = 32$, a las que denominamos configuraciones Shot-24 y Shot-32 respectivamente. Estas longitudes corresponden respectivamente a la moda y la media de la distribución de duraciones de cortes observadas en el histograma de la figura 6.1. Para comprobar las ventajas de la estrategia propuesta para generar clips basados en la partición de cortes, la comparamos con una estrategia de segmentación más sencilla. En esta estrategia alternativa, los clips se generan simplemente tomando secuencias contiguas de 24 y 32 cuadros del avance, que denominamos Seq-24 y Seq-32. La figura 6.2 ilustra un ejemplo comparando las estrategias de generación de clips con las configuraciones Seq-24 y Shot-24.

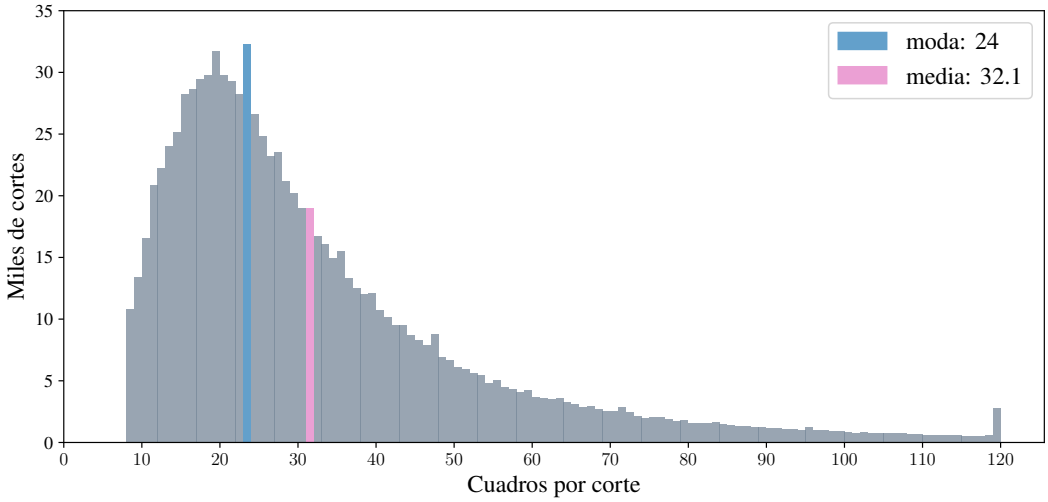


Figura 6.1: Histograma de duraciones de cortes de Trailers12k obtenido con TransNet V2 (Souček & Lokoč, 2020).

La tabla 6.2 reporta resultados en los que se comparan las estrategias de generación de clips Seq-f y Shot-f utilizando columnas preentrenadas en ImageNet e ImageNet-Kinetics. Como puede observarse, la estrategia Shot-f mejora la transferibilidad, especialmente en el caso de ImageNet-Kinetics. Más concretamente, el rendimiento más alto en todas las métricas lo obtiene Shot-24 utilizando una columna ImageNet-Kinetics. Por ejemplo, esta configuración alcanza un μAP promedio de 75.57 %, lo que supone una ganancia de 3.75 y 4.1 con respecto a Seq-24 y Seq-32. Las desviaciones estándar también suelen ser más bajas en las columnas ImageNet-Kinetics con la estrategia Shot-f. Esto podría ser un efecto del detector de cortes que ayuda a generar clips en los que la mayoría de los cuadros están muy correlacionados, reduciendo así el riesgo de que haya cuadros de transición dentro de los clips. Para una columna ImageNet-Kinetics, esto es especialmente importante, ya que consume todos los cuadros de un clip para generar su representación. En las configuraciones que utilizan la columna preentrenada en ImageNet, las ganancias son menores; por ejemplo, el μAP de Shot-24 es alrededor de 1.83 puntos superior al de Seq-24. Esto puede explicarse por el hecho de que una columna 2D toma como entrada solo un cuadro del clip seleccionado utilizando la similitud de su histograma de color al histograma de color promedio del clip, lo que reduce la probabilidad de tomar un cuadro de transición. Para ImageNet-Kinetics, Shot-24 supera ligeramente a Shot-32 en todas las métricas. Esto puede deberse a que como Shot-24 produce clips más cortos, el número de

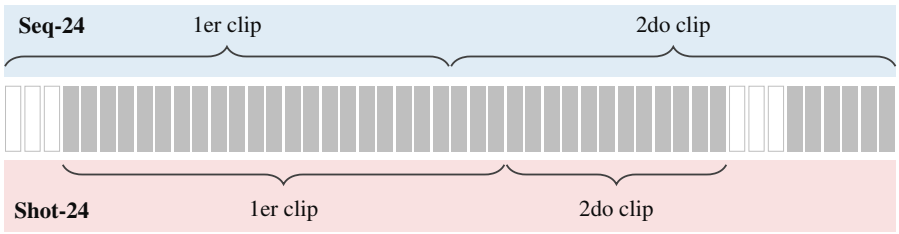


Figura 6.2: Comparación de las estrategias de generación de clips con Seq-24 (azul) y Shot-24 (rosa). En el centro, los cuadros de contenido se representan como rectángulos grises sólidos, mientras que los cuadros de transición son rectángulos vacíos. En Seq-24, los clips están constituidos por cuadros secuenciales, que incluyen tanto cuadros de contenido como de transición. En Shot-24, se espera que los clips estén constituidos sólo por cuadros secuenciales de contenido, excluyendo los cuadros de transición.

muestras de entrenamiento es también un 31 % mayor que Shot-32.

6.3. Frecuencia de cuadros

El movimiento es una pista de información importante para muchas tareas de análisis de video. En los clips de acciones humanas, la principal fuente de movimiento procede de personas realizando una acción y su variabilidad depende del tipo de acción. La frecuencia de cuadros es un factor de interés en el estudio de métodos de reconocimiento de acciones humanas (Varol et al., 2018). En cambio, en avances de películas, los personajes, los objetos, el fondo o los acontecimientos pueden mostrar de forma independiente una gran variabilidad en la cantidad de movimiento, lo que usualmente está asociado con el tipo género (H. Zhou et al., 2010). Para explorar la influencia de este factor en la transferibilidad, reducimos la frecuencia de cuadros de los videos de Trailers12k para aumentar la cantidad de movimiento aparente. El procedimiento de submuestreo simplemente consiste en seleccionar cuadros a intervalos iguales, por ejemplo, para producir un video de 8 cuadros por clip sólo se conserva el primero de cada tres cuadros consecutivos.

Los resultados a diferentes frecuencias de cuadros se reportan en la tabla 6.3. Como podemos observar, el desempeño aumenta a medida que se incrementa la frecuencia, alcanzando el mejor resultado para todas las métricas con la frecuencia original de 24 cuadros. No obstante, el mejor balance entre desempeño y requisitos computacionales se obtiene con frecuencias de cuadro inferiores. Como puede apreciarse en la figura 6.3, se pueden alcanzar resultados competitivos a una fracción del costo computacional. Por ejemplo, 4 cuadros por clip disminuye sólo 3.02

Tabla 6.2: Comparación del desempeño de las estrategias de generación de clips Seq-f y Shot-f en DIViTA.

Estrategia de generación de clips	Métricas ↑			
	μAP	mAP	wAP	sAP
ImageNet				
Seq-24	70.83±1.93	66.39±1.86	70.29±1.03	76.04±1.83
Seq-32	70.13±2.03	66.31±2.12	70.13±2.05	75.95±2.04
Shot-24	72.66±1.37	67.68±1.36	**71.76±1.09**	**77.49±1.18**
Shot-32	**72.90±1.20**	**67.77±1.58**	71.70±1.10	77.45±1.11
ImageNet-Kinetics				
Seq-24	71.82±1.33	66.55±1.24	69.88±1.61	76.01±1.24
Seq-32	71.42±1.09	66.89±1.30	69.93±2.04	75.94±1.72
Shot-24	**75.57±0.66**	**70.48±0.41**	**74.21±0.40**	**80.02±0.47**
Shot-32	75.21±0.43	69.64±0.48	73.32±0.31	79.16±0.29

puntos de μAP, mientras que únicamente se usa $\frac{1}{6}$ de memoria para representar el tensor de entrada.

6.4. Extensión espacio-temporal

DIViTA utiliza el concepto de fragmentos como una manera de simular vagamente las escenas de los avances, y a su vez utilizarlos como representaciones abreviadas de todo el avance. Esto simplifica el proceso de entrenamiento por lotes. Por una parte, introduce implícitamente un mecanismo de aumentado de datos en la dimensión temporal. Por otra, reduce los requisitos de memoria y procesamiento. Sin embargo, la reducción del número de clips por fragmento limita el campo receptivo espacio-temporal del módulo transformador de agregación de clips. Dada la naturaleza supervisada débil del etiquetado de géneros de Trailers12k, esta podría dar lugar a predicciones erróneas a nivel de fragmento. Recordemos que durante el entrenamiento, un fragmento se genera tomando muestras aleatorias de clips contiguos a los que se les asignan todos los géneros del avance completo. Por consiguiente, si los clips del fragmento no contienen información relacionada con uno de los géneros asignados, el proceso de entrenamiento recibe

Tabla 6.3: Incremento gradual del número de cuadros por clip en DIViTA.

Frecuencia de cuadros	Métricas ↑			
	μAP	mAP	wAP	sAP
ImageNet-Kinetics				
4	72.55±0.89	67.50±0.91	71.43±0.86	77.40±0.79
6	72.93±0.60	67.80±0.68	71.73±0.71	77.89±0.50
8	73.04±0.55	67.95±0.60	71.86±0.57	78.24±0.66
12	73.34±0.77	68.07±0.73	71.96±0.39	78.63±0.49
24	**75.57±0.66**	**70.48±0.41**	**74.21±0.40**	**80.02±0.47**

una señal de supervisión engañosa con respecto a ese género. Aumentar el número de clips por fragmento (aproximando progresivamente el avance completo) ayuda a aminorar este problema a costa de reducir las ventajas de una representación abreviada.

En la tabla 6.4 se reportan los resultados de configuraciones aumentando progresivamente el número de clips por fragmento. Desde la perspectiva del módulo transformador de agregación de clips, esto implica campos receptivos crecientes disponibles para el mecanismo de modelado espacio-temporal. De los resultados podemos observar que el mejor rendimiento se obtiene con fragmentos de entre 30 y 40 clips, superando incluso a configuraciones con un mayor número de clips. La pérdida de rendimiento con fragmentos más grandes puede deberse a la reducción del espacio de muestreo de fragmentos durante el proceso de selección aleatoria. Esto a su vez, decrementa la variabilidad en el mecanismo de aumentado de datos durante la generación de fragmentos. Es interesante notar que usando 10 clips se obtiene el mejor balance entre desempeño y requisitos computacionales. Con 10 clips sólo se decae 0.64 μAP puntos por debajo de la configuración con 30 clips, sin embargo, únicamente se utiliza $\frac{1}{3}$ de la memoria para el tensor de entrada.

6.5. Modelado espacio-temporal

El módulo transformador de agregación de clips genera una representación espacio-temporal condensada a nivel fragmento. Desde el punto de vista de la tarea de MTGC, este módulo aspira vagamente a captar las relaciones a nivel escena del avance. Para evaluar la capacidad

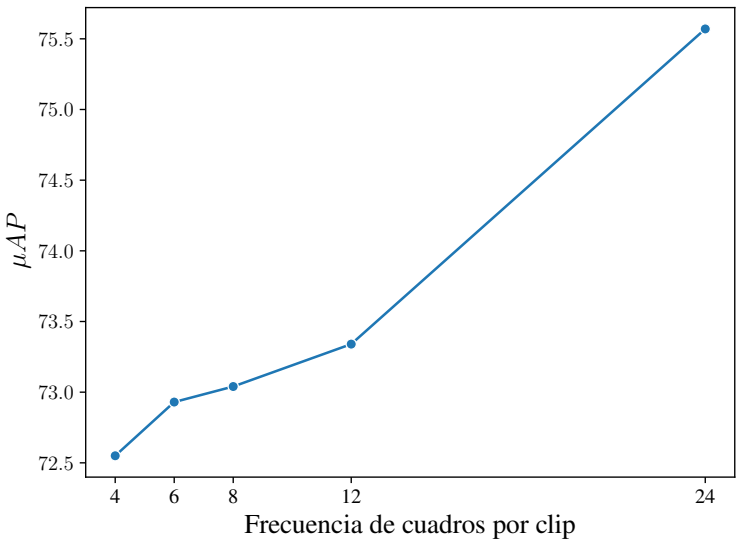


Figura 6.3: Frecuencia de cuadros por clip vs. desempeño en μAP. La memoria requerida durante el procesamiento es proporcional al número de cuadros.

Tabla 6.4: Incremento gradual del campo receptivo espacio-temporal del módulo transformador de agregación de clips.

Clips por fragmento	Métricas ↑			
	μAP	mAP	wAP	sAP
ImageNet-Kinetics				
5	72.86±0.69	68.78±0.56	72.90±0.92	77.16±0.98
10	74.93±0.77	70.25±0.67	74.14±1.01	79.12±0.38
15	75.17±0.69	70.36±0.52	74.02±0.68	79.44±0.45
20	75.39±0.65	70.42±0.48	**74.25±0.59**	79.69±0.51
30	**75.57±0.66**	70.48±0.41	74.21±0.40	80.02±0.47
40	75.53±0.68	**70.71±0.42**	74.17±0.39	**80.06±0.45**
50	75.46±0.75	70.24±0.40	74.02±0.42	79.97±0.45
60	74.87±0.73	70.17±0.53	74.02±0.45	79.99±0.50

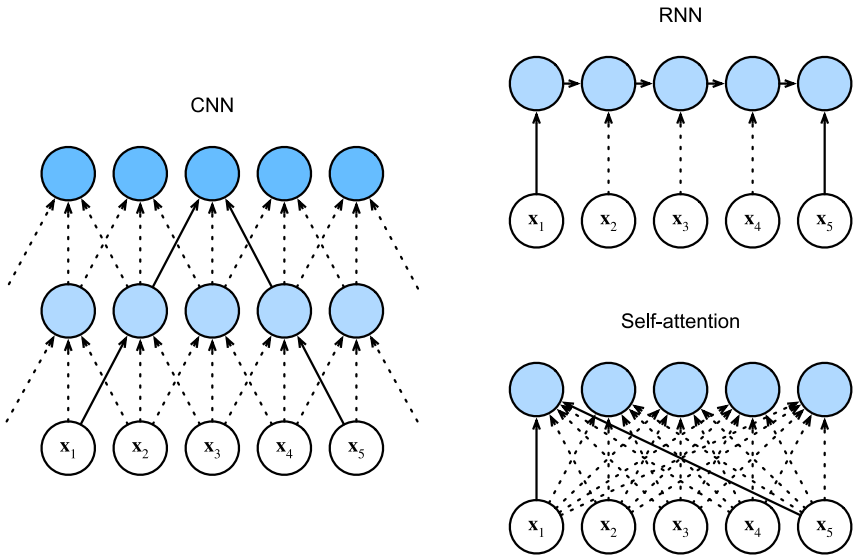


Figura 6.4: Comparación del modelado de un conjunto de elementos usando mecanismos convolucionales, recurrentes y de autoatención. Tomada de (A. Zhang et al., 2021).

de este módulo para modelar relaciones espacio-temporales entre clips, lo comparamos con alternativas convolucionales y recurrentes.

La principal ventaja de una arquitectura transformadora es su mecanismo de autoatención. Este permite capturar las relaciones entre cualquier par de clips, independientemente de su posición relativa dentro del fragmento. En cambio, un enfoque convolucional sólo puede encontrar relaciones entre clips dentro de su campo receptivo determinado por el tamaño del filtro. Del mismo modo, un módulo recurrente estándar se limita a modelar relaciones secuencialmente. Las diferencias entre estos mecanismos de modelado de conjuntos se ilustran en la figura 6.4.

Para comparar estos mecanismos, llevamos a cabo una búsqueda de hiperparámetros para encontrar las mejores configuraciones. Para la experimentación reportada, utilizamos las configuraciones que tuvieron aproximadamente el mismo número de parámetros. En el caso de la versión recurrente, usamos una celda GRU con 115 unidades ocultas. Por otro lado, la versión convolucional utiliza una capa Conv1D con 128 filtros de tamaño 3. Esta versión es similar a la estrategia de agregación de la arquitectura CTT-MMC-A (Wehrmann & Barros, 2017), pero con un número de filtros y tamaños de filtro distintos. Como se mencionó anteriormente, la versión transformadora consta de 4 cabezas con proyecciones lineales de 128 dimensiones.

La tabla 6.5 reporta los resultados correspondientes a las tres versiones del módulo de

Tabla 6.5: Comparación de las versiones recurrente, convolucional y transformadora del módulo de agregación de clips.

Modelado espacio-temporal	Métricas ↑			
	μAP	mAP	wAP	sAP
ImageNet				
Recurrente	71.43±2.14	65.24±1.49	69.93±1.59	76.22±2.88
Convolucional	72.13±2.01	66.88±1.44	70.89±1.31	77.03±2.78
Transformador	**72.66±1.37**	**67.68±1.36**	**71.76±1.09**	**77.49±1.18**
ImageNet-Kinetics				
Recurrente	74.21±0.97	68.28±1.03	72.67±0.69	78.95±0.66
Convolucional	74.40±0.98	69.19±1.12	73.04±1.02	79.30±1.03
Transformador	**75.57±0.66**	**70.48±0.41**	**74.21±0.40**	**80.02±0.47**

agregación de clips. En todas las métricas, el módulo transformador supera a las versiones convolucional y recurrente tanto para ImageNet como para ImageNet-Kinetics. Esto puede explicarse por la capacidad del mecanismo de autoatención para capturar mejor el tipo de relaciones espacio-temporales entre las tomas de los avances. Como se describió en la sección 5.1, es habitual que en los avances dos tomas correlacionadas aparezcan en posiciones arbitrarias.

6.6. Transferibilidad de ImageNet y Kinetics

Utilizamos DIViTA para estudiar el grado de transferibilidad de ImageNet y Kinetics a Trailers12k. Nos enfocamos en tres factores: el conjunto de datos de preentrenamiento, la arquitectura de la columna preentrenada y los requisitos computacionales de la columna. En nuestros experimentos, consideramos las arquitecturas convolucionales ligeras ShuffleNet-2D (Ma et al., 2018) y ShuffleNet-3D (Köpüklü et al., 2019); las convolucionales pesadas ResNet (He et al., 2016) y R(2+1)D (Tran et al., 2018); y las transformadoras para visión Swin-2D (Liu et al., 2021) y Swin-3D (Liu, Ning et al., 2022). La tabla 6.6 reporta los resultados de estos experimentos.

Tabla 6.6: Comparación del desempeño de distintas arquitecturas de columna y conjuntos de preentrenamiento (P) ImageNet (I) y Kinetics (K). Los porcentajes de número de parámetros y FLOPS están calculados respecto al mejor resultado en general (Swin-3D).

Columna	P		Métricas ↑				Paráms. ↓		FLOPS ↓	
	I	K	μAP	mAP	wAP	sAP	(M)	%	(G)	%
Conv ligera										
ShuffleNet-2D	✔		71.14±0.68	66.01±0.46	70.17±0.44	75.80±0.85	1.7	6.09	4.3	0.27
ShuffleNet-3D		✔	63.43±1.54	58.18±1.50	63.59±1.46	69.49±1.58	1.7	6.09	8.5	0.53
ShuffleNet-F	✔	✔	72.11±0.56	67.08±0.37	71.42±0.41	76.66±0.73	3.3	11.82	12.9	0.81
Conv pesada										
ResNet	✔		71.42±0.59	66.63±0.36	70.64±0.34	76.41±0.38	23.9	85.66	122.7	7.71
R(2+1)D		✔	70.88±1.37	64.99±1.37	69.88±1.30	75.15±1.08	31.7	113.62	1823.4	114.67
ResNet-F	✔	✔	73.28±0.66	68.03±0.63	72.14±0.72	77.76±0.44	55.6	199.28	1946.1	122.39
Transformadora										
Swin-2D	✔		72.66±1.37	67.68±1.36	71.76±1.09	77.49±1.18	27.9	100.00	114.0	7.16
Swin-3D	✔	✔	**75.57±0.66**	**70.48±0.41**	**74.21±0.40**	**80.02±0.47**	**27.9**	**100.00**	**1590.0**	**100.00**

6.6.1. ImageNet vs. Kinetics

Las columnas convolucionales 2D preentrenadas en ImageNet superan a las columnas 3D que únicamente fueron preentrenadas en Kinetics. Esto concuerda con la intuición de que puede ser más fácil predecir géneros a partir de elementos espaciales (personajes, objetos, ambientes, etc.) que a partir de información dinámica (acciones, movimiento, etc.). Sin embargo, la diferencia de rendimiento entre ResNet y R(2+1)D es de sólo 0.54 μAP puntos. Desde el punto de vista del preentrenamiento, esto sugiere que los clips de Kinetics (que incluyen información espacio-temporal, pero menos diversa espacialmente) proporcionan una fuente de conocimiento distinta pero competitiva con las imágenes de ImageNet (con mayor diversidad de información espacial, pero puramente estáticas).

Adicionalmente, investigamos si ambos conjuntos de datos de preentrenamiento pueden ser complementarios utilizando dos enfoques diferentes. El primer enfoque, denominado fusión (denotado con el sufijo F), utiliza un procesamiento inspirado en la arquitectura de doble flujo propuesta por Simonyan y Zisserman (2014). Esta arquitectura cuenta con dos flujos de

procesamiento idénticos, uno para ImageNet y otro para Kinetics, que producen dos vectores de predicción que se promedian en la etapa final (fusión tardía). Observamos que con este enfoque se obtienen mejoras en los resultados con respecto a un único preentrenamiento en 0.97 μAP puntos para ShuffleNet-F y 1.86 para ResNet-F. En el segundo enfoque se preentrena la misma columna en ambos conjuntos. Para esto, una arquitectura Swin-2D se preentrena primero en ImageNet, después sus capas de encajes lineales se inflan a lo largo de la dimensión temporal para formar una Swin-3D, que posteriormente se entrena en Kinetics (Liu, Ning et al., 2022). Esta aproximación con preentrenamiento doble supera al preentrenamiento simple de ImageNet (Swin-2D) en 2.5 μAP puntos.

6.6.2. Convolucionales vs. transformadoras

Comparamos las arquitecturas convolucionales ResNet y ShuffleNet con las arquitecturas transformadoras Swin. Notamos que la arquitectura transformadora Swin-3D supera a las arquitecturas convolucionales tanto en ImageNet como en Kinetics, con desviaciones estándar similares a las de las arquitecturas convolucionales. Cabe destacar que Swin-3D alcanza los mejores resultados utilizando un enfoque de preentrenamiento estándar para arquitecturas de video transformadoras, en contraste con el enfoque de fusión de dos flujos. Por otro lado, aunque Swin-2D tiene un desempeño superior al de las convolucionales 2D y convolucional con fusión, su desviación estándar también es mayor.

6.6.3. Requisitos computacionales

Analizamos el compromiso entre desempeño y requisitos computacionales, como se muestra en la figura 6.5. Observamos que la arquitectura convolucional ligera con fusión (ShuffleNet-F) tiene un rendimiento 3.46 μAP inferior al de la mejor configuración (Swin-3D). No obstante, el número de parámetros es un orden de magnitud menor y el número de FLOPS es dos órdenes de magnitud menor para ShuffleNet-F. Podemos notar que aunque el costo computacional de ShuffleNet-3D es bajo, también su desempeño es considerablemente más bajo en comparación con cualquier otro modelo. También, ResNet-F es el más costoso entre todos los modelos y su rendimiento es similar al de Swin-2D y ShuffleNet-F. Cabe señalar que el número de parámetros y FLOPS de este modelo proviene principalmente del flujo de video de R(2+1)D.

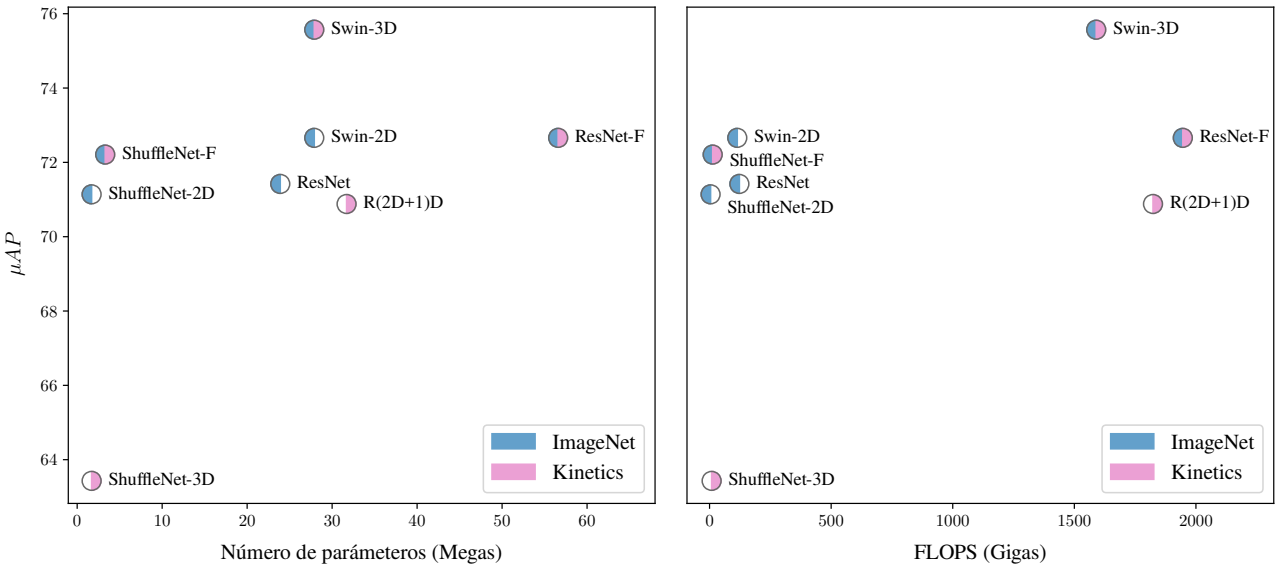


Figura 6.5: Comparación de diferentes configuraciones de DIViTA en términos de número de parámetros (izquierda) y FLOPs (derecha) vs. desempeño en μAP. Los modelos preentrenados sólo con ImageNet se representan como burbujas mitad azul, mientras que los preentrenados sólo con Kinetics se representan como burbujas mitad rosa. Los modelos que tienen preentrenamiento tanto en ImageNet como en Kinetics se representan como burbujas mitad azul y mitad rosa.

6.7. Comparación con métodos de referencia

Para evaluar el rendimiento global de DIViTA, consideramos las configuraciones con el mejor desempeño en μAP y las comparamos con CTT-MMC-A (Wehrmann & Barros, 2017), fastVideo (Cascante-Bonilla et al., 2019) y TimeSformer (Bertasius et al., 2021). La tabla 6.7 reporta el rendimiento de estos métodos y de las configuraciones Swin-2D y Swin-3D de DIViTA. Podemos observar que las puntuaciones de ambas configuraciones de DIViTA son más altas para todas las métricas que los métodos de referencia. Estos resultados sugieren que la combinación de la etapa de Generación de Fragmentos, el módulo transformador de agregación de clips y la columna transformadora Swin-3D puede ser eficaz para mejorar el desempeño al realizar transferencia de conocimiento de ImageNet y/o Kinetics a Trailers12k.

Tabla 6.7: Comparación de DIViTA con los métodos de referencia y su preentrenamiento (P) en ImageNet (I) y Kinetics (K).

Método	P		Métrica ↑			
	I	K	μAP	mAP	wAP	sAP
CTT-MMC-A (Wehrmann & Barros, 2017)	✔		69.27±2.87	65.37±1.61	68.93±2.09	75.09±3.01
fastVideo (Cascante-Bonilla et al., 2019)	✔		68.21±0.73	61.19±0.53	65.86±0.57	74.68±0.68
TimeSformer (Bertasius et al., 2021)	✔		64.98±1.16	59.00±1.07	63.26±0.92	70.77±0.94
DIViTA Swin-2D (nuestro)	✔		72.66±1.37	67.68±1.36	71.76±1.09	77.49±1.18
DIViTA Swin-3D (nuestro)	✔	✔	**75.57±0.66**	**70.48±0.41**	**74.21±0.40**	**80.02±0.47**

7 Conclusiones

En este proyecto buscamos estudiar factores que influyen en el proceso de transferencia de conocimiento de conjuntos de imágenes generalistas y videos de acciones humanas a la tarea de clasificación de avances de películas. En específico, realizamos un estudio de transferibilidad por medio de experimentación donde analizamos diversos factores que influencian la transferencia de representaciones aprendidas en los conjuntos ImageNet y Kinetics a Trailers12k. Este último es un nuevo conjunto de avances de películas recopilado para este trabajo.

7.1. Conclusiones generales

Para llevar a cabo el estudio de transferibilidad propusimos DIViTA, un nuevo método de clasificación de avances que busca explotar su estructura espacio-temporal intrínseca resultante de su proceso de producción. En pro de dar mayor solidez a este estudio, recopilamos el conjunto Trailers12k por medio de un proceso de recolección pensado para obtener un conjunto de avances de alta calidad en muestras y etiquetado.

Tomando como base Trailers 12k y DIViTA, llevamos a cabo el estudio de transferibilidad por medio de experimentación empírica para determinar la influencia que tienen diversos elementos que forman parte del proceso de transferencia de conocimiento. Por una parte, exploramos aspectos relacionados con la de representación de entrada del avance de video. Asimismo, estudiamos el impacto de los conjuntos de preentrenamiento y arquitecturas de columnas convolucionales y transformadoras preentrenadas en estos conjuntos. Además, analizamos el efecto de ciertos componentes internos de DIViTA por medio de ablación.

Las contribuciones de este proyecto se pueden agrupar en las siguientes direcciones. Primero, por medio del estudio se pudieron identificar diversos factores relacionados con la representación de entrada del avance que promueven una transferencia de conocimiento positiva teniendo como resultado un incremento en el desempeño en la clasificación géneros de avances en Trailers12k. Además, nuestros resultados indican que aunque el preentrenamiento en ImageNet proporciona

desempeños ligeramente superiores a los de Kinetics, explotar el preentrenamiento en ambos conjuntos es complementario y mejora los resultados en MTGC. También, encontramos que las columnas transformadoras proporcionan un mayor desempeño que sus contrapartes convolucionales.

Por otra parte, por medio de ablación pudimos verificar ventajas de componentes del nuevo método propuesto DIViTA. Un componente es la etapa de Generación de Fragmentos, que segmenta el avance de entrada, explotando su estructura temporal en busca de mejorar el proceso de transferencia. En comparación con una alternativa más simple, la Generación de Fragmentos mejora los resultados de clasificación. Otro componente es el módulo transformador de agregación de fragmentos que supera a las estrategias recurrentes y convolucionales. Esto puede deberse a que el mecanismo de atención es capaz de correlacionar cualquier par de elementos en una secuencias de clips, lo que es una característica intrínseca a la estructura de los avances.

La última contribución del proyecto es el conjunto multimodal Trailers12k, que incluye URL, metadatos, representaciones profundas de videos, carteles y un protocolo de evaluación; se hacen disponibles públicamente a la comunidad de investigación. En comparación con otros conjuntos, Trailers12k es el único que cuenta con un proceso de verificación manual, además de incluir una amplia variedad de metadatos y representaciones profundas computadas con arquitecturas de video.

Las contribuciones de este proyecto se presentan en el artículo *Improving Transfer Learning for Movie Trailer Genre Classification using a Dual Image and Video Transformer* (R. Montalvo-Lezama et al., 2023), publicado en la revista *Information Processing & Management*.

7.2. Trabajo a futuro

Si bien este trabajo de investigación analizó parte del proceso de transferencia a la tarea de clasificación de avances, al mismo tiempo los resultados indican ciertas líneas de investigación promisorias.

- **Explorar modelos de fundación.** Los resultados de la experimentación realizada indican que el preentrenamiento en ambos conjuntos, ImageNet y Kinetics, mejora el desempeño en la tarea objetivo de clasificación de avances. Aumentar la diversidad de datos y tareas de preentrenamiento de la columna de extracción de representaciones de avances es una

línea de investigación. En este sentido, sería interesante explorar modelos de fundación, ya que son el resultado de preentrenar arquitecturas en múltiples conjuntos de variadas modalidades y diversas tareas, esto en busca de modelos con una base de conocimiento más amplia con la expectativa de que generalicen mejor a otras tareas.

- **Investigar los recursos mínimos para MTGC.** A través de la exploración de variaciones en la representación de entrada del avance y el uso de arquitecturas ligeras obtuvimos modelos que requieren significativamente menos recursos computacionales pero al mismo tiempo manteniendo un desempeño competitivo con los mejores modelos en general. Una línea interesante para continuar este trabajo consiste en investigar cuáles son los factores y sus configuraciones con los mínimos requisitos computacionales que ofrezcan un desempeño competitivo.

- **Proponer una segmentación más eficiente.** El uso de la red TransNet V2 para detectar las transiciones y segmentar en cortes el avance de entrada introduce un costo computacional considerable en DIViTA. Puede ser relevante proponer mecanismos alternativos más simples y eficientes, que estén posiblemente integrados en DIViTA y sean parte de un entrenamiento punto a punto.

- **Explotar información multimodal.** El foco de este trabajo fue estudiar la transferencia de representaciones de ImageNet y Kinetics. Sin embargo, no hay ningún impedimento para extender DIViTA para explotar otras modalidades de los avances, como audio o texto, en busca de mejorar los resultados. De forma alternativa, DIViTA puede ser integrada dentro de métodos multimodales, como Moviescope, simplemente agregándola como un flujo dentro de estos métodos.

- **Estudio de otras tareas de análisis de video.** Como se mencionó al inicio de esta tesis, en MTGC las etiquetas pueden contener subjetividad o representar temáticas que no están presentes físicamente en un cuadro o escena. Como trabajo a futuro sería interesante estudiar la efectividad de métodos automatizados para la resolución de tareas de video con características similares.

Apéndice A: Hoja de datos de Trailers12k

Este apéndice contiene la hoja de datos (Gebru et al., 2021) de Trailers12k.

Hoja de datos del conjunto Trailers12k

Motivación

A ¿Para qué propósito fue creado el conjunto de datos?

El conjunto Trailers12k[1] fue creado para estudiar la transferibilidad de representaciones visuales de imágenes y video aprendidas con redes neuronales. Los detalles sobre el proyecto de investigación que motivó la creación de Trailers12k pueden encontrarse en el artículo *Improving Transfer Learning for Movie Trailer Genre Classification using a Dual Image and Video Transformer* (R. Montalvo-Lezama et al., 2023).

B ¿Quién creó el conjunto de datos?

Este conjunto fue creado por Ricardo Montalvo Lezama, Berenice Montalvo Lezama y Gibran Fuentes Pineda.

C ¿Qué patrocinios tuvo?

Durante parte del proceso de recolección, Ricardo Montalvo Lezama recibió una beca doctoral del Consejo Nacional de Ciencia y Tecnología (CONACYT) de México. Además, este trabajo se realizó con el apoyo del proyecto PAPIIT IA104016 de la Dirección General de Asuntos del Personal Académico (DGAPA) de la UNAM.

D ¿Otros comentarios?

No.

Composición

A ¿Cómo son las instancias que componen el conjunto?

El conjunto está compuesto por registros de películas, donde cada instancia contiene información general de la película obtenida de IMDb, una URL de un avance en YouTube y representaciones vectoriales del cartel y el avance.

B ¿Cuántas instancias son en total?

El conjunto está conformado por 12,000 instancias en total.

[1] https://richardtml.github.io/trailers12k

Tabla 1: Datos proporcionados por instancia en Trailers12k.

Datos	Identificador	Tipo
Video y audio		
Representación espacial del avance	`trailers_i_*`	vector
Representación espacio-temporal del avance	`trailers_k_*`	vector
URL del avance	`yt_url`	cadena
Imagen		
Representación del cartel	`posters_i_swin`	vector
URL del cartel	`cover_url`	cadena
Texto		
Identificador de IMDb	`id`	cadena
Título	`title`	cadena
Año	`year`	entero
Géneros	`genres`	lista de cadenas
Argumentos	`plots`	lista de cadenas
Sinopsis	`synopsis`	cadena
Reparto	`cast`	cadena
Directores	`directors`	lista de cadenas
Escritores	`writers`	lista de cadenas
Compositores	`composers`	lista de cadenas
Productores	`producers`	lista de cadenas
Compañías productoras	`production_companies`	lista de cadenas
Idiomas	`languages`	lista de cadenas
Clasificaciones por edad	`certificates`	lista de cadenas
Duración	`runtime`	entero
Votos	`votes`	entero
Puntuación	`rating`	flotante
Palabras clave	`keywords`	lista de cadenas

C ¿El conjunto contiene todas las instancias o es una muestra de un conjunto más grande?

El conjunto contiene todas las instancias. No obstante, Trailers12k está basado en un conjunto previo llamado Trailers15k[2] recolectado por Berenice Montalvo Lezama. Aproximadamente el 75 % de ejemplos de Trailers12k aparecen en Trailers15k.

D ¿Qué datos componen cada instancia?

La lista completa de atributos por instancia se presenta en la tabla 1.

E ¿Hay una etiqueta asociada a cada instancia?

[2] https://turing.iimas.unam.mx/~bereml/project/trailers/

Cada instancia está multietiquetada en 10 posibles géneros.

F ¿Hay alguna información faltante en las instancias?
No.

G ¿Hay relaciones explícitas entre las instancias individuales?
No de forma explícita, sin embargo, existen relaciones implícitas como secuelas de películas o actores que pueden ser parte del reparto en múltiples películas.

H ¿Hay alguna partición recomendada de los datos?
Sí, el conjunto cuenta con tres particiones estratificadas, cada una con subconjuntos de entrenamiento, validación y prueba pensadas para evaluación de la tarea de clasificación de géneros.

I ¿Hay errores, fuentes de ruido, o redundancias en los datos?
Los datos fueron obtenidos de IMDb y estos son de autoría de los usuarios de dicha plataforma, por lo que es posible que haya errores en algunas instancias.

J ¿El conjunto contiene datos que puedan considerarse confidenciales?
No.

K ¿El conjunto contiene datos que puedan considerarse como ofensivos, insultantes, amenazantes o puedan causar ansiedad?
Sí, es posible que algunos videos contenga escenas con estas características o algunas reseñas contengan lenguaje ofensivo.

L ¿El conjunto identifica subpoblaciones?
No.

M ¿Es posible identificar individuos de forma directa o indirecta?
Sí, los nombres del reparto y equipo están incluidos en el conjunto, no obstante, esta información era pública desde antes de que se recopilara el conjunto.

N ¿El conjunto contiene información que pueda ser considerada como sensible?
Sí, los ejemplos pueden contener información relacionada con etnias, creencias religiosas, opiniones políticas, desnudos, etc.

Ñ ¿Otros comentarios?
No.

Proceso de recolección

A ¿Cómo se adquirieron los datos asociados con cada instancia?
Los datos fueron adquiridos de forma automatizada de las plataformas de IMDb y YouTube.

B ¿Qué mecanismos o procedimientos se usaron para recolectar los datos?
El procedimiento de recolección constó de las siguientes etapas.

a) A partir de una búsqueda automatizada en IMDb, se obtuvieron aquellas películas estrenadas entre los años 2000 y 2019 con el mejor puntaje otorgado por los usuarios.

b) Para cada película, se realizó una búsqueda automatizada en YouTube utilizando el título, añadiendo el año y la palabra "trailer". De esta búsqueda se descargó el video correspondiente al primer resultado.

c) Se filtraron las películas para conservar únicamente aquellas que tuvieran asignadas al menos uno de los diez géneros más populares en IMDb y que contaran con al menos 500 votos de usuarios.

d) Dado que en una cantidad significativa de casos, el avance resultante no correspondió con el título (pudiendo tratarse de una adaptación o nueva versión, un homónimo, un avance producido por admiradores, etc.), los avances se curaron manualmente. En específico, se verificó manualmente la correspondencia de cada par título-avance y se sustituyeron los avances incorrectos por los mejores disponibles en YouTube. También se sustituyeron avances para cumplir los siguientes criterios de calidad de vídeo: su duración debía estar comprendida entre 60 y 210 segundos, su resolución debía ser de al menos 480p, y debía contener la menor cantidad posible de publicidad y relleno (anuncios, logos o barras de color/cuadros).

C ¿Quién estuvo involucrado en el proceso de recolección?
Ricardo Montalvo Lezama y Berenice Montalvo Lezama.

D ¿En qué ventana de tiempo fueron recolectados los datos?
Los datos fueron recolectados entre 2019 y 2021.

E ¿Se llevó a cabo algún proceso de revisión ética?
No.

F ¿Los datos fueron recolectados de forma directa u obtenidos por medio de terceras partes u otras fuentes?
Como se mencionó anteriormente, se recolectó de IMDb y YouTube por medio de un proceso automatizado.

G ¿Se notificó a los individuos de la recolección?
No, los datos fueron recolectados de fuentes públicas.

H ¿Los individuos consintieron la recolección y uso de sus datos?
No (ver la pregunta anterior).

I ¿Otros comentarios?
No.

Procesamiento/limpieza/etiquetado

A ¿Se usó algún proceso de preprocesamiento, limpieza o etiquetado?
En el proceso de recolección descrito se detallan los pasos de preprocesamiento, limpieza y etiquetado. Por otra parte, el proceso de cómputo de representaciones vectoriales se describe a detalle en el artículo *Improving Transfer Learning for Movie Trailer Genre Classification using a Dual Image and Video Transformer* (R. Montalvo-Lezama et al., 2023).

B ¿Se guardaron los datos previos a cualquier preprocesamiento?
No.

C ¿El software usado para el preprocesamiento se encuentra disponible?
No.

D ¿Otros comentarios?
No.

Usos

A ¿Se ha usado el conjunto de datos?
El conjunto ha sido usado en los siguientes artículos.

- *Improving Transfer Learning for Movie Trailer Genre Classification using a Dual Image and Video Transformer* (R. Montalvo-Lezama et al., 2023).
- *Emotion Assessment of YouTube Videos using Color Theory* (Cakmak et al., 2024).

B ¿Hay algún repositorio de enlaces a artículos o sistemas que usen el conjunto?
No.

C ¿Para qué otras tareas se puede usar el conjunto?
El conjunto puede usarse para tareas de clasificación o predicción de atributos como las palabras clave o votos. También puede usarse para la generación de sinopsis, carteles o avances.

D ¿Hay algo sobre la composición del conjunto o algún aspecto sobre la recolección o preprocesamiento que pueda impactar futuros usos?
Sí, en un futuro los videos de los avances pueden ser removidos de YouTube.

E ¿Hay tareas para las que el conjunto no debe ser usado?
El conjunto no puede ser usado con fines comerciales.

F ¿Otros comentarios?
No.

Distribución

A ¿El conjunto de datos será distribuido por terceros en representación de la entidad en la que fue creado?
Sí, el conjunto está públicamente disponible.

B ¿Cómo se distribuye el conjunto?
El conjunto se encuentra disponible a través de dos fuentes.

- Sitio oficial: https://richardtml.github.io/trailers12k
- Zenodo: https://doi.org/10.5281/zenodo.5716409

C ¿Cuándo estará disponible el conjunto?
El conjunto fue publicado en 2023.

D ¿El conjunto es distribuido bajo derechos de autor, licencia o algún otro término de uso?
Los datos recolectados pertenecen a sus autores. El conjunto se distribuye bajo la licencia *Creative Commons Attribution Non Commercial Share Alike 4.0 International*[3]. Finalmente, los autores de Trailers12k solicitan citar el artículo original en caso de usar el conjunto.

E ¿Terceras partes han impuesto restricciones de propiedad intelectual u otro tipo en los datos asociados a las instancias?
Sí, como se mencionó anteriormente, los datos recolectados pertenecen a los autores originales.

F ¿Aplican controles de exportación u otras restricciones regulatorias al conjunto o instancias?
No

G ¿Otros comentarios?
No.

Mantenimiento

A ¿Quién dará soporte al conjunto?
Ricardo Montalvo Lezama y Berenice Montalvo Lezama.

B ¿Cómo se puede contactar al propietario?
Los correos de contacto son:

- Ricardo Montalvo Lezama: ricardoml@turing.iimas.unam.mx
- Berenice Montalvo Lezama: bereml@turing.iimas.unam.mx

[3] https://creativecommons.org/licenses/by-nc-sa/4.0/legalcode

C ¿Hay alguna fe de erratas?

No, en caso de haberla en un futuro será publicada en el sitio del conjunto.

D ¿Habrá actualizaciones al conjunto?

En caso de haber actualizaciones, estas serán publicadas en el sitio del conjunto.

E ¿Si el conjunto está relacionado con personas, hay límites aplicables en la retención de los datos asociados en las instancias?

No.

F ¿Versiones anteriores del conjunto continuarán contando con soporte?

El conjunto está actualizado, versiones anteriores serán mantenidas en el sitio.

G ¿Si otros desean extender/aumentar/contribuir al conjunto, hay un mecanismo para hacerlo?

Otros pueden realizar extensiones, pero deben contactar a los autores para incorporar estos cambios, además deben consultar la licencia de uso.

H ¿Otros comentarios?

No.

Bibliografía

Ahmad, I. S., Bakar, A. A., & Yaakub, M. R. (2020). Movie Revenue Prediction Based on Purchase Intention Mining Using YouTube Trailer Reviews. *Information Processing & Management*, *57*(5), 102278.

Arnab, A., Dehghani, M., Heigold, G., Sun, C., Lučić, M., & Schmid, C. (2021). ViViT: A Video Vision Transformer. *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 6836-6846.

Association, M. P., et al. (2007). Theatrical Market Statistics.

Behrouzi, T., Toosi, R., & Akhaee, M. A. (2022). Multimodal movie genre classification using recurrent neural network. *Multimedia Tools and Applications*, 1-22.

Bengio, Y., Courville, A., & Vincent, P. (2013). Representation learning: A review and new perspectives. *IEEE transactions on pattern analysis and machine intelligence*, *35*(8), 1798-1828.

Bertasius, G., Wang, H., & Torresani, L. (2021). Is Space-Time Attention All You Need for Video Understanding? *International Conference on Machine Learning*.

Bi, T., Jarnikov, D., & Lukkien, J. (2022). Shot-Based Hybrid Fusion for Movie Genre Classification. *International Conference on Image Analysis and Processing*, 257-269.

Bi, T., Jarnikov, D., & Lukkien, J. (2021). Video representation fusion network for multi-label movie genre classification. *Proceedings of the International Conference on Pattern Recognition*, 9386-9391.

Bojanowski, P., Grave, E., Joulin, A., & Mikolov, T. (2017). Enriching word vectors with subword information. *Transactions of the association for computational linguistics*, *5*, 135-146.

Bondebjerg, I. (2015, marzo). Film: Genres and Genre Theory.

Bozinovski, S., & Fulgosi, A. (1976). The influence of pattern similarity and transfer learning upon training of a base perceptron b2. *Proceedings of Symposium Informatica*, *3*, 121-126.

Bros., W. (1942). Casablanca - Original Theatrical Trailer [[Accessed 13-11-2023]].

Cakmak, M. C., Shaik, M., & Agarwal, N. (2024). Emotion Assessment of YouTube Videos using Color Theory. *Proceedings of the 2024 9th International Conference on Multimedia and Image Processing*, 6-14.

Carreira, J., & Zisserman, A. (2017). Quo Vadis, Action Recognition? A New Model and the Kinetics Dataset. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 4724-4733.

Cascante-Bonilla, P., Sitaraman, K., Luo, M., & Ordonez, V. (2019). Moviescope: Large-scale Analysis of Movies using Multiple Modalities. *arXiv Preprint*, *1908.03180*.

Chen, Y., Kalantidis, Y., Li, J., Yan, S., & Feng, J. (2018). Multi-Fiber Networks for Video Recognition. *European Conference on Computer Vision*, 352-367.

Cherti, M., & Jitsev, J. (2022). Effect of Pre-Training Scale on Intra- and Inter-Domain Full and Few-Shot Transfer Learning for Natural and Medical X-Ray Chest Images. *Proceedings of the International Joint Conference on Neural Networks*, 1-9.

Choi, J., Sharma, G., Schulter, S., & Huang, J.-B. (2020). Shuffle and Attend: Video Domain Adaptation. *Proceedings of the European Conference on Computer Vision*, 678-695.

Cohn, N. (2014). You're a good structure, Charlie Brown: The distribution of narrative categories in comic strips. *Cognitive Science*, *38*(7), 1317-1359.

Cohn, N. (2016). From Visual Narrative Grammar to Filmic Narrative Grammar: The narrative structure of static and moving images. *Film text analysis: New perspectives on the analysis of filmic meaning*, 94-117.

Cohn, N., Paczynski, M., Jackendoff, R., Holcomb, P. J., & Kuperberg, G. R. (2012a). (Pea) nuts and bolts of visual narrative: Structure and meaning in sequential image comprehension. *Cognitive psychology*, *65*(1), 1-38.

Cohn, N., Paczynski, M., Jackendoff, R., Holcomb, P. J., & Kuperberg, G. R. (2012b). (Pea)nuts and bolts of visual narrative: Structure and meaning in sequential image comprehension. *Cognitive Psychology*, *65*(1), 1-38.

Daniel, D. (2015). A Brief History of Film Trailers, or: Turns Out This Post Is Not About Peter Orner [[Accessed 14-11-2023]].

De Jesus, K., & Shapiro, M. (2020). Social Media Engagement and Film Box Office.

de Jonge, M. (2019). *The evolution of movie trailers* [Tesis de maestría, Tilburg University].

Deldjoo, Y., Constantin, M. G., Schedl, M., Ionescu, B., & Cremonesi, P. (2018). MMTF-14K: A Multifaceted Movie Trailer Feature Dataset for Recommendation and Retrieval. *Proceedings of the ACM Multimedia Systems Conference*, 450-455.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., & Fei-Fei, L. (2009). ImageNet: A large-scale hierarchical image database. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 248-255.

Disney. (1993). The Nightmare Before Christmas (1993) Official Trailer [[Accessed 13-12-2023]].

Dornaleteche, J. (2007). Definición y naturaleza del trailer cinematográfico: Definition and nature of movie trailers. *Pensar la publicidad: revista internacional de investigaciones publicitarias*, *1*(2), 99-117.

Dornaleteche, J. (2009). El trailer cinematográfico: historia de un género publicitario en EE. UU./Movie trailers: the history of an advertising genre in USA. *Pensar la publicidad*, *3*(1), 163.

Gebru, T., Morgenstern, J., Vecchione, B., Vaughan, J. W., Wallach, H., Iii, H. D., & Crawford, K. (2021). Datasheets for datasets. *Communications of the ACM*, *64*(12), 86-92.

Gil Pons, E. (2010). La narrativa del tráiler cinematográfico. *Congreso Euro-Iberoamericano de Alfabetización Mediática y Culturas Digitales (2010),*

Girshick, R., Donahue, J., Darrell, T., & Malik, J. (2014). Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 580-587.

Hara, K., Kataoka, H., & Satoh, Y. (2018). Can Spatiotemporal 3D CNNs Retrace the History of 2D CNNs and ImageNet? *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 6546-6555.

Haskell, R. E. (2000). *Transfer of learning: Cognition and instruction*. Elsevier.

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

Hu, Y., Jin, L., & Jiang, X. (2022). A GCN-Based Framework for Generating Trailers. *Proceedings of the International Conference on Computing and Artificial Intelligence*, 610-617.

Huang, Q., Xiong, Y., Rao, A., Wang, J., & Lin, D. (2020). MovieNet: A Holistic Dataset for Movie Understanding. *Proceedings of the European Conference on Computer Vision*, 709-727.

Huang, Y.-F., & Wang, S.-H. (2012). Movie Genre Classification Using SVM with Audio and Video Features. *Proceedings of the International Conference on Active Media Technology*, 1-10.

Jason, S. (2016). Why Are Movie Previews Called "Trailers"? [[Accessed 14-11-2023]].

Jerrick, D. (2013). The Effectiveness of Film Trailers: Evidence from the College Student Market. *UW-L Journal of Undergraduate Research*, *16*, 1-13.

Jiang, J., Shu, Y., Wang, J., & Long, M. (2022). Transferability in deep learning: A survey. *arXiv preprint arXiv:2201.05867*.

Johansen, S. J. (2013). Coming Attractions: An Essay on Movie Trailers & Preliminary Statements. *Legal Comm. & Rhetoric: JALWD*, *10*, 41.

Johnston, K. M. (2008). 'The Coolest Way to Watch Movie Trailers in the World' Trailers in the Digital Age. *Convergence*, *14*(2), 145-160.

Johnston, K. M., Vollans, E., & Greene, F. L. (2016). Watching the trailer: Researching the film trailer audience. *Participations*, *13*(2), 56-85.

Kannan, R., Ghinea, G., & Swaminathan, S. (2015). What do you wish to see? A summarization system for movies based on user preferences. *Information Processing & Management*, *51*(3), 286-305.

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

Kataoka, H., Wakamiya, T., Hara, K., & Satoh, Y. (2020). Would Mega-scale Datasets Further Enhance Spatiotemporal 3D CNNs? *arXiv Preprint*, *2004.04968*.

Katz, S. D. (2019). *Film Directing: Shot by Shot - 25th Anniversary Edition: Visualizing from Concept to Screen*. Michael Wiese Productions.

Kay, W., Carreira, J., Simonyan, K., Zhang, B., Hillier, C., Vijayanarasimhan, S., Viola, F., Green, T., Back, T., Natsev, P., et al. (2017). The Kinetics Human Action Video Dataset. *arXiv Preprint*, *1705.06950*.

Ke, A., Ellsworth, W., Banerjee, O., Ng, A. Y., & Rajpurkar, P. (2021). CheXtransfer: Performance and Parameter Efficiency of ImageNet Models for Chest X-Ray Interpretation. *Proceedings of the Conference on Health, Inference, and Learning*, 116-124.

Kolesnikov, A., Beyer, L., Zhai, X., Puigcerver, J., Yung, J., Gelly, S., & Houlsby, N. (2020). Big Transfer (BiT): General Visual Representation Learning. *Proceedings of the European Conference on Computer Vision*, 491-507.

Köpüklü, O., Kose, N., Gunduz, A., & Rigoll, G. (2019). Resource Efficient 3D Convolutional Neural Networks. *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshop*, 1910-1919.

Kornblith, S., Shlens, J., & Le, Q. V. (2019). Do Better ImageNet Models Transfer Better? *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2661-2671.

Kuehne, H., Jhuang, H., Garrote, E., Poggio, T., & Serre, T. (2011). HMDB: A large video database for human motion recognition. *Proceedings of the International Conference on Computer Vision*, 2556-2563.

Li, D., Rodriguez, C., Yu, X., & Li, H. (2020). Word-level Deep Sign Language Recognition from Video: A New Large-scale Dataset and Methods Comparison. *Proceedings of the IEEE Winter Conference on Applications of Computer Vision*, 1459-1469.

Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., & Guo, B. (2021). Swin Transformer: Hierarchical Vision Transformer using Shifted Windows. *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 10012-10022.

Liu, Z., Ning, J., Cao, Y., Wei, Y., Zhang, Z., Lin, S., & Hu, H. (2022). Video Swin Transformer. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3202-3211.

Liu, Z., Mao, H., Wu, C.-Y., Feichtenhofer, C., Darrell, T., & Xie, S. (2022). A convnet for the 2020s. *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 11976-11986.

Long, J., Shelhamer, E., & Darrell, T. (2015). Fully Convolutional Networks for Semantic Segmentation. *IEEE Conference on Computer Vision and Pattern Recognition*, 3431-3440.

Lopez, A. R., Giro-i-Nieto, X., Burdick, J., & Marques, O. (2017). Skin lesion classification from dermoscopic images using deep learning techniques. *Proceedings of the IASTED International Conference on Biomedical Engineering*, 49-54.

Loshchilov, I., & Hutter, F. (2019). Decoupled Weight Decay Regularization. *International Conference on Learning Representations*.

Ma, N., Zhang, X., Zheng, H.-T., & Sun, J. (2018, enero). ShuffleNet V2: Practical Guidelines for Efficient CNN Architecture. En *Proceedings of the European Conference on Computer Vision* (pp. 122-138).

Monfort, M., Andonian, A., Zhou, B., Ramakrishnan, K., Bargal, S. A., Yan, T., Brown, L., Fan, Q., Gutfreund, D., Vondrick, C., & Oliva, A. (2020). Moments in Time Dataset: One Million Videos for Event Understanding. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *42*(2), 502-508.

Montalvo-Lezama, B. (2018). *Clasificación multi-etiqueta de videos cortos usando unidades recurrentes reguladas*.

Montalvo-Lezama, R., Montalvo-Lezama, B., & Fuentes-Pineda, G. (2023). Improving Transfer Learning for Movie Trailer Genre Classification using a Dual Image and Video Transformer. *Information Processing & Management*, *60*(3), 103343.

Pepe, P. J., & Zarzynski, J. W. (2016). *Documentary Filmmaking for Archaeologists*. Routledge.

Plizzari, C., Planamente, M., Goletto, G., Cannici, M., Gusso, E., Matteucci, M., & Caputo, B. (2022). E2 (GO) MOTION: Motion Augmented Event Stream for Egocentric Action Recognition. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 19935-19947.

Rasheed, Z., Sheikh, Y., & Shah, M. (2005). On the Use of Computable Features for Film Classification. *Proceedings of the IEEE Transactions on Circuits and Systems for Video Technology*, *15*(1), 52-64.

Razavian, A. S., Azizpour, H., Sullivan, J., & Carlsson, S. (2014). CNN Features off-the-shelf: an Astounding Baseline for Recognition. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 806-813.

Redmon, J., Divvala, S., Girshick, R., & Farhadi, A. (2016). You Only Look Once: Unified, Real-Time Object Detection. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 779-788.

Rodríguez Bribiesca, I., López Monroy, A. P., & Montes-y-Gómez, M. (2021). Multimodal Weighted Fusion of Transformers for Movie Genre Classification. *Proceedings of the Workshop on Multimodal Artificial Intelligence*, 1-5.

Rosenstein, M. T., Marx, Z., Kaelbling, L. P., & Dietterich, T. G. (2005). To Transfer or Not To Transfer. *Proceedings of the Neural Information Processing Systems Workshop on Inductive Transfer: 10 Years Later*.

Shafaei, M., Smailis, C., Kakadiaris, I., & Solorio, T. (2021). A Case Study of Deep Learning-Based Multi-Modal Methods for Labeling the Presence of Questionable Content in Movie Trailers. *Proceedings of the International Conference on Recent Advances in Natural Language Processing*, 1297-1307.

Shambharkar, P. G., Mehrotra, G., Thakur, K. S., Thakare, K., & Doja, M. N. (2021). Multi-Class Classification of Actors in Movie Trailers. *Proceedings of the International Conference on Paradigms of Computing, Communication and Data Sciences*, 953-965.

Simões, G., Wehrmann, J., Barros, R., & Ruiz, D. (2016). Movie genre classification with Convolutional Neural Networks. *Proceedings of the International Joint Conference on Neural Networks*, 259-266.

Simonyan, K., & Zisserman, A. (2014). Two-Stream Convolutional Networks for Action Recognition in Videos. *Proceedings of the International Conference on Neural Information Processing Systems*, 568-576.

Singh, G., Akrigg, S., Di Maio, M., Fontana, V., Javanmard alitappeh, R., Khan, S., Saha, S., Jeddisaravi, K., Yousefi, F., Culley, J., Nicholson, T., Omokeowa, J., Grazioso, S., Bradley, A., Di Gironimo, G., & Cuzzolin, F. (2022). ROAD: The ROad event Awareness Dataset for Autonomous Driving. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1-1.

Soekhoe, D., van der Putten, P., & Plaat, A. (2016). On the Impact of Data Set Size in Transfer Learning Using Deep Neural Networks. *Proceedings of the Advances in Intelligent Data Analysis XV*, 50-60.

Soomro, K., Zamir, A. R., & Shah, M. (2012). UCF101: A Dataset of 101 Human Actions Classes From Videos in The Wild. *arXiv Preprint*, *1212.0402*.

Souček, T., & Lokoč, J. (2020). TransNet V2: An effective deep network architecture for fast shot transition detection. *arXiv Preprint*, *2008.04838*.

Szymański, P., & Kajdanowicz, T. (2017a). A scikit-based Python environment for performing multi-label classification. *ArXiv e-prints*.

Szymański, P., & Kajdanowicz, T. (2017b). A Network Perspective on Stratification of Multi-Label Data. *Proceedings of the International Workshop on Learning with Imbalanced Domains: Theory and Applications*, *74*, 22-35.

Tan, M., & Le, Q. (2019). EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks. *Proceedings of the International Conference on Machine Learning*, 6105-6114.

Tran, D., Bourdev, L., Fergus, R., Torresani, L., & Paluri, M. (2015). Learning spatiotemporal features with 3d convolutional networks. *Proceedings of the IEEE International Conference on Computer Vision*, 4489-4497.

Tran, D., Wang, H., Torresani, L., Ray, J., LeCun, Y., & Paluri, M. (2018). A Closer Look at Spatiotemporal Convolutions for Action Recognition. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 6450-6459.

Tsoumakas, G., Katakis, I., & Vlahavas, I. (2010). Mining Multi-label Data. En *Data Mining and Knowledge Discovery Handbook* (pp. 667-685). Springer US.

Varol, G., Laptev, I., & Schmid, C. (2018). Long-Term Temporal Convolutions for Action Recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *40*(6), 1510-1517.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is All You Need. *Proceedings of the International Conference on Neural Information Processing Systems*, 6000-6010.

Wang, J., Yang, Y., Mao, J., Huang, Z., Huang, C., & Xu, W. (2016). Cnn-rnn: A unified framework for multi-label image classification. *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, 2285-2294.

Wasko, J. (2003). How hollywood works. *How Hollywood Works*, 1-248.

Wehrmann, J., & Barros, R. (2017). Movie genre classification: A multi-label approach based on convolutions through time. *Applied Soft Computing*, *61*, 973-982.

Xie, Y., & Richmond, D. (2018). Pre-training on Grayscale ImageNet Improves Medical Image Classification. *Proceedings of the European Conference on Computer Vision Workshops*, 476-484.

Xu, Y., Yang, J., Cao, H., Mao, K., Yin, J., & See, S. (2021). ARID: A New Dataset for Recognizing Action in the Dark. *Proceedings of the International Workshop on Deep Learning for Human Activity Recognition*, 70-84.

Yadav, A., & Vishwakarma, D. K. (2020). A unified framework of deep networks for genre classification using movie trailer. *Applied Soft Computing*, *96*, 106624.

Yosinski, J., Clune, J., Bengio, Y., & Lipson, H. (2014). How transferable are features in deep neural networks? *Proceedings of the Advances in Neural Information Processing Systems*, *27*, 3320-3328.

Yu, Y., Lu, Z., Li, Y., & Liu, D. (2021). ASTS: attention based spatio-temporal sequential framework for movie trailer genre classification. *Multimedia Tools and Applications*, *80*(7), 9749-9764.

Zamir, A. R., Sax, A., Shen, W. B., Guibas, L. J., Malik, J., & Savarese, S. (2018). Taskonomy: Disentangling Task Transfer Learning. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3712-3722.

Zhang, A., Lipton, Z. C., Li, M., & Smola, A. J. (2021). Dive into deep learning. *arXiv preprint arXiv:2106.11342*.

Zhang, L. (2019). Transfer Adaptation Learning: A Decade Survey. *IEEE Transactions on Neural Networks and Learning Systems*, *PP*.

Zhang, W., Fang, Y., & Ma, Z. (2017). The effect of task similarity on deep transfer learning. *Neural Information Processing: 24th International Conference, ICONIP 2017, Guangzhou, China, November 14-18, 2017, Proceedings, Part II 24*, 256-265.

Zhang, W., Deng, L., Zhang, L., & Wu, D. (2023). A Survey on Negative Transfer. *IEEE/CAA Journal of Automatica Sinica*, *10*(2), 305-329.

Zhou, B., Lapedriza, A., Khosla, A., Oliva, A., & Torralba, A. (2017). Places: A 10 million image database for scene recognition. *IEEE transactions on pattern analysis and machine intelligence*, *40*(6), 1452-1464.

Zhou, H.-Y., Lu, C., Yang, S., & Yu, Y. (2021). ConvNets vs. Transformers: Whose visual representations are more transferable? *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, 2230-2238.

Zhou, H., Hermans, T., Karandikar, A., & Rehg, J. (2010). Movie Genre Classification via Scene Categorization. *Proceedings of the ACM International Conference on Multimedia*, 747-750.