



CNN y CRNN para Clasificación de Géneros Musicales

TFM Máster Big Data, Data Science & Inteligencia Artificial

Alumno: Miguel Clemente Canedo Rodríguez

Universidad Complutense de Madrid

2023/2024

Tabla de Contenidos

1. Introducción
2. Presentación de los datos
 - Base de datos GTZAN
 - Características relevantes
3. Preprocesamiento
4. Red Convolucional (CNN)
 - Arquitectura de la red
 - Metricas de evaluación
5. Red Convolucional Recursiva (CRNN)
 - Arquitectura de la red
 - Metricas de evaluación
6. Comparación de resultados entre CNN y CRNN
7. Sistema de Recomendacion de Generos Musicales
8. Conclusiones
9. Bibliografía

Introducción

La clasificación automática de géneros musicales se ha convertido en un área de creciente interés dentro del campo de la música y el análisis de señales de audio, especialmente con la proliferación de grandes bibliotecas musicales y plataformas de streaming. Identificar con precisión el género de una canción puede llegar a ser un desafío debido a la naturaleza compleja de la música, que combina elementos rítmicos, melódicos, armónicos y de timbre, sin mencionar los enfoques subjetivos al identificar canciones por uno o varios géneros. A lo largo de los años, se han propuesto diversos enfoques basados en técnicas de procesamiento de señales y machine learning para abordar este problema, destacando el uso del Deep Learning con diferentes aplicaciones de Redes Neuronales (CNN, RNN, CRNN, etc.).

Entre los modelos más populares, las redes neuronales convolucionales (CNN) han demostrado ser muy efectivas para captar patrones locales en señales de audio, mediante los componentes espectrales que caracterizan a diferentes géneros musicales. Sin embargo, las CNN se centran en las relaciones espaciales y pueden no ser suficientes para capturar aquellas dependencias temporales presentes en la música, lo que limita su capacidad para modelar estructuras musicales más complejas. Por otro lado, las redes neuronales recurrentes (RNN), y específicamente las variantes combinadas con capas convolucionales (CRNN), se presentan como una alternativa poderosa, ya que permiten modelar la combinación tanto de esas características espectrales como las dependencias temporales a lo largo de la duración del audio.

El objetivo principal de este trabajo es comparar el rendimiento de las CNN y CRNN en la clasificación de géneros musicales utilizando una base de datos de audios ampliamente utilizada, como es GTZAN. A partir de la extracción de características que combinen dimensiones espectrales y temporales, se propone entrenar y evaluar ambos modelos, explorando sus fortalezas y limitaciones en este contexto. Se espera que la CRNN, al integrar la capacidad de detectar patrones espectrales y temporales, ofrezca un mejor rendimiento comparado con la CNN en términos de precisión y capacidad de generalización.

No obstante, más allá de la identificación de géneros, hoy en día una de las aplicaciones más relevantes en la industria musical es la recomendación de canciones similares, una tarea más compleja que involucra la organización y dimensionamiento de las características más representativas de cada audio en un espacio continuo de manera que las canciones similares queden cercanas entre sí. Para abordar este punto, aprovechando los esfuerzos del entrenamiento de las redes previas, se estudiará la modificación y adaptación de la red CRNN de forma que sirva como un Recomendador de Canciones Similares. Adicional al objetivo principal de este trabajo, se espera con esta propuesta demostrar que las CRNN pueden ser entrenadas para tener la potencia de generar recomendaciones personalizadas basadas en la similitud de canciones.

Presentación de los datos

Base de datos GTZAN

Para el desarrollo de este trabajo de investigación, se usó como insumo la base de datos de **GTZAN**, una de las bases de datos públicas más utilizadas para la tarea de clasificación de géneros musicales, siendo referencia en el campo de análisis de audios. A lo largo de los años, ha sido ampliamente utilizada para investigaciones en el reconocimiento de patrones, aprendizaje automático y sistemas de recomendación de música.

Esta base de datos está compuesta por 1000 clips de audios de 30 segundos de duración, presentados en formato .wav de 22050MHz Mono 16-bit. Estos 1000 audios están organizados en 10 grupos de 100 que representan los géneros musicales, estos son:

- Blues
- Classical
- Country
- Disco
- Hiphop
- Jazz
- Metal
- Pop
- Reggae
- Rock

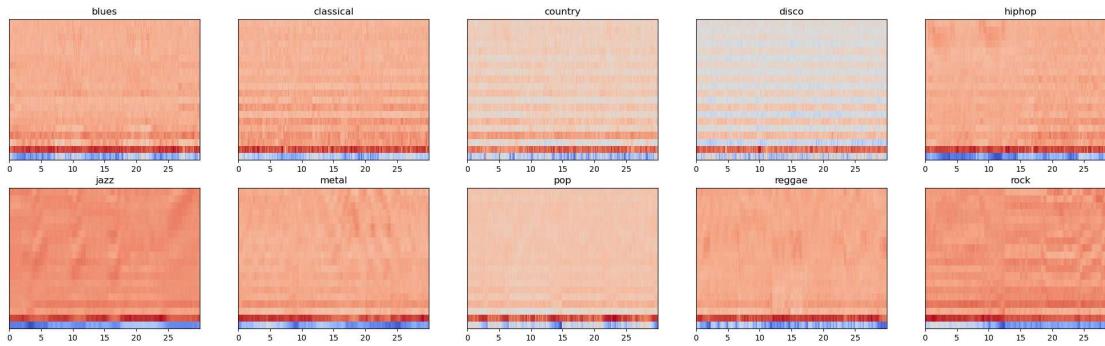
Para efectos de este estudio en específico, debido a que el archivo de audio número 54 del género jazz estaba corrupto, se toma la decisión de descartar todos los audios número 54 de cada género con el fin de mantener el equilibrio entre la cantidad de audios por género. Iniciando así con una base datos de 990 audios (99 por cada género).

Características relevantes

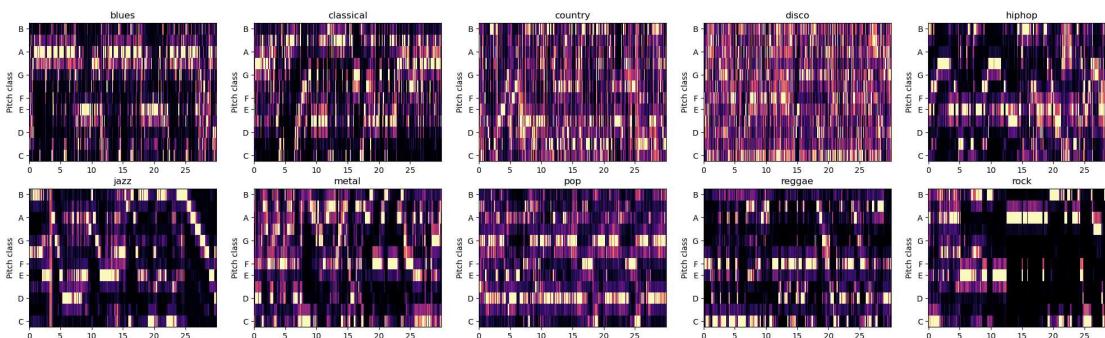
Para este proyecto, buscando capturar diferentes aspectos tanto *espectrales* como *temporales* de los audios, se toman en cuenta **8 características** basadas en su capacidad de captar la información de ambos aspectos. Esta combinación de características espectrales y temporales proporcionará una representación integral de cada audio. Teniendo en cuenta además que estas características son ampliamente utilizadas en tareas de análisis de audio debido a su capacidad de representar los diferentes aspectos de un sonido musical. Estas 8 características son:

- **Mel Frequency Cepstral Coefficients (MFCC)**: son una representación compacta de las frecuencias del espectro de audio que se asemejan a la percepción humana del sonido. Es una de las características cruciales ya que reduce la dimensionalidad de la señal y preserva información relacionada con el timbre. Comúnmente se representa mediante un spectrograma

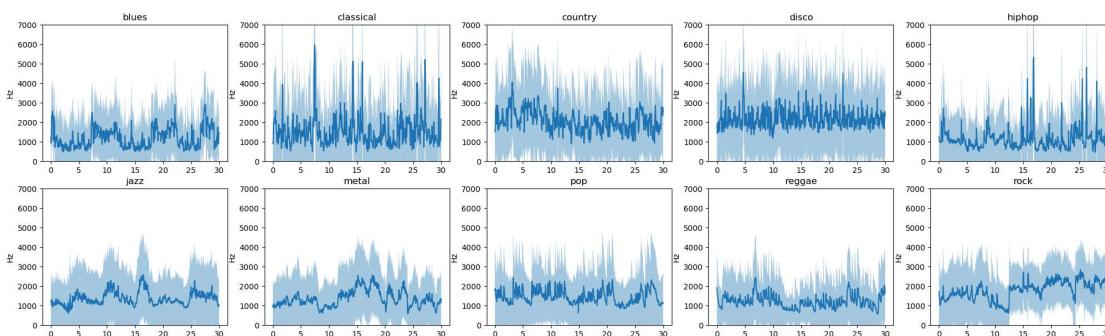
donde el eje X es el tiempo, el eje Y representa los Coeficientes y el Color representa la magnitud de la energía.



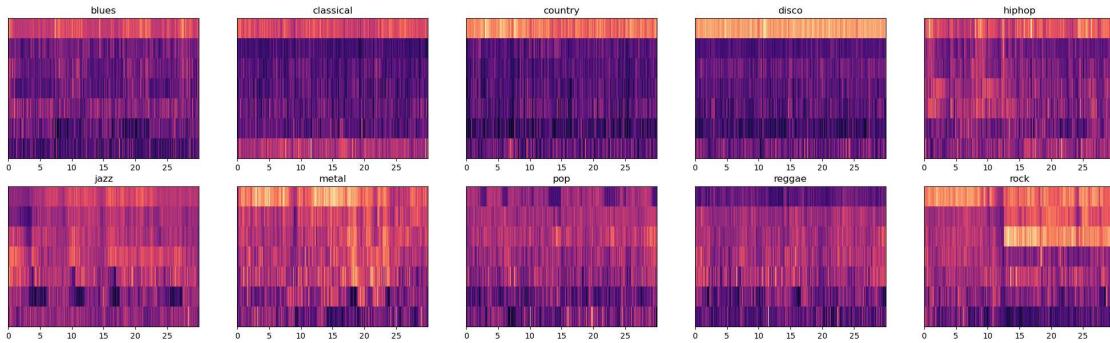
- **Chroma STFT:** representa la distribución de energía en cada una de las 12 clases de tonos musicales. Útiles para capturar información tonal en los audios, clave al identificar géneros como clásica o jazz que presentan tonalidades particulares. Se representa gráficamente mediante un espectrograma donde el eje X representa el tiempo y el eje Y representa las clases de tonos musicales.



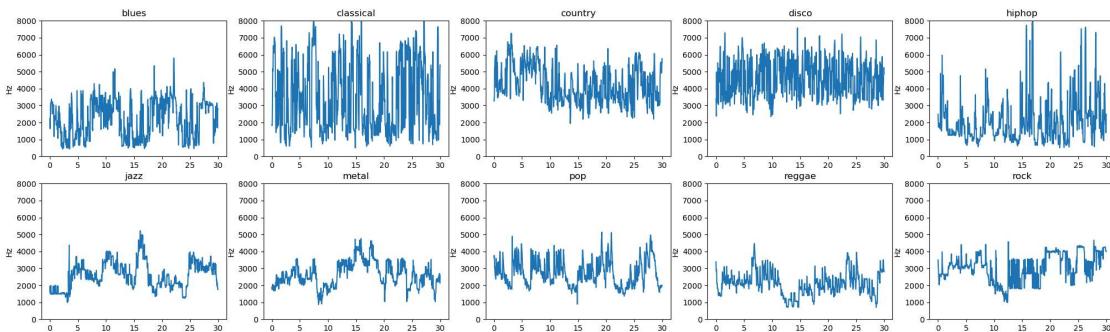
- **Spectral Centroid:** indica la brillantez de una señal, describiendo dónde se concentra la mayor parte de la energía en el espectro de frecuencias. Ayudando a identificar géneros con una energía más alta en frecuencias agudas (como el rock).
- **Spectral Bandwidth:** mide la extensión del espectro de frecuencias de una señal al rededor del centroide, permitiendo diferenciar géneros con texturas sonoras simples (como el reggae) de géneros más densos (como el metal). Tanto el Centroid como el Bandwidth se complementan entre sí para dar más información y por ello se representan juntos en la misma gráfica de líneas, donde el eje X es el tiempo y el eje Y representa los Hz.



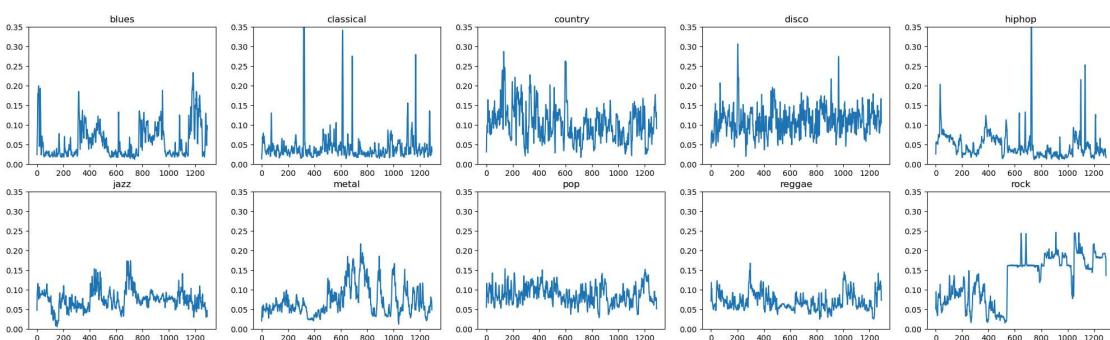
- **Spectral Contrast:** mide la diferencia en la energía entre los picos y los valles del espectro en diferentes bandas de frecuencia. Útil para diferenciar géneros con altos contrastes de volumen y energía, como el metal. Ilustrado mediante spectrogramas donde el eje X es el tiempo y el eje Y representa las bandas de frecuencia.



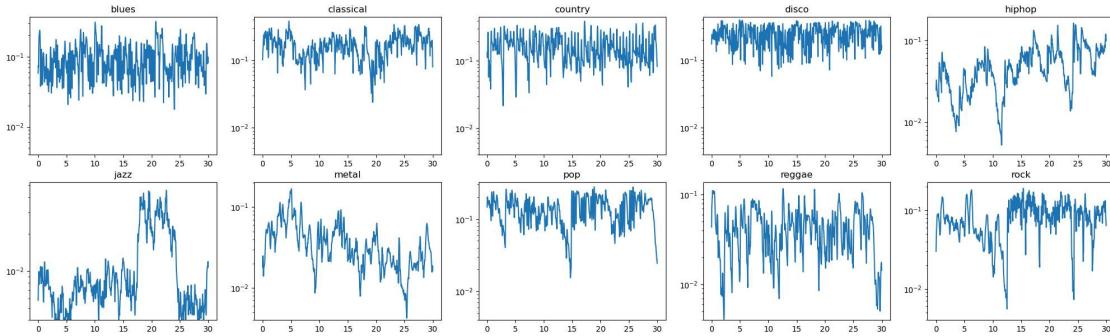
- **Spectral Roll-off:** representa la frecuencia por debajo de la cual se acumula un porcentaje de la energía espectral total (comúnmente 90%). Útil para diferenciar géneros con concentraciones de energía en frecuencias altas (como el rock) o frecuencias bajas (como el jazz o blues). Se ilustra mediante una gráfica de líneas, donde el eje X es el tiempo y el eje Y representa los Hz.



- **Zero-Crossing Rate (ZCR):** mide la cantidad de veces que la señal cruza el eje cero, representando información sobre la rapidez con la que cambian los signos en la señal. Útil para identificar géneros con muchos elementos rítmicos o percusivos, como el hiphop o el rock. Se ilustra mediante una gráfica de líneas, donde el eje X es el tiempo y el eje Y representa el ratio de cruces en cero.



- **Root Mean Square Energy (RMS)**: mide la energía promedio de la señal de audio en cada marco de tiempo, capturando la intensidad o volumen de una canción. Permite diferenciar géneros que varían mucho en dinámica (como el pop) o con dinámicas más suaves y fluctuantes (como el jazz). Representado mediante una gráfica de líneas con escalado logarítmico, donde el eje X es el tiempo y el eje Y el logaritmo del RMS.



En líneas generales, cada género musical presenta un comportamiento diferenciador en cada una de estas 8 características y combinándolas se busca potenciar esta diferenciación a fin de lograr nuestro objetivo de identificarlos mediante los modelos que se definirán en este trabajo. Por ejemplo, el *MFCC* capta variaciones en las frecuencias que son distintivas para géneros vocales como el jazz o hiphop, mientras que el *chroma STFT* destaca en géneros con armonías complejas, como la música clásica. El *spectral centroid* y *bandwidth* revelan la distribución de energía en las frecuencias, ayudando a diferenciar géneros como metal (con un centroid más alto) del reggae (más bajo). El *spectral contrast* y *rolloff* muestran los cambios entre picos y valles en el espectro, diferenciando géneros más dinámicos como el rock del pop. Finalmente, el *zero crossing rate* y el *RMS* ayudan a identificar la naturaleza rítmica o abrupta del audio, siendo útiles para distinguir géneros percusivos como el disco o electrónicos como el hiphop.

Preprocesamiento

Antes de comenzar a modelar y definir las arquitecturas de las redes, primero se deben preparar los audios y para ello se aplicaron los siguientes procedimientos:

- **Data Augmentation**: Teniendo en consideración la limitada cantidad de audios disponibles en la base de datos GTZAN, para este trabajo se aplican técnicas de Data Augmentation con el fin de poder generar nuevos audios a partir de los originales, incrementando de esta manera la variabilidad y diversidad de la base de datos de entrada para las redes, sin necesidad de recolectar nuevos audios.

Especificamente, se optó el fragmentar los audios de 30 segundos en múltiples segmentos más pequeños de 5 segundos, con un solapamiento de 2 segundos entre ellos para asegurar que se use toda la información del audio original y así evitar alguna pérdida de datos importantes en

los cortes de cada segmento. Logrando finalmente con esto que se produzcan 9 audios nuevos de 5 segundos por cada audio original de 30 segundos, y así aumentar el número de muestras totales a 8.910 audios de 5 segundos sin perder el contenido esencial de los audios originales.

Este proceso fue esencialmente importante ya que para la implementación de modelos Deep Learning, en este caso como las CNN y CRNN, se requieren grandes volúmenes de datos para alcanzar buenos desempeños. Adicionalmente, utilizar segmentos más cortos permite mejorar la capacidad de generalización de las redes, evitando el overfitting y optimizando sus rendimientos.

- **Extracción de características:** Para la tarea de clasificación de géneros musicales desde audios requiere convertir la señal de sonido cruda en un conjunto de características numéricas que representen tanto la informaciónpectral como la temporal del audio. Para este trabajo, partiendo de los audios que se obtienen del proceso de data augmentation, se procede a extraer las 8 características seleccionadas y definidas previamente que capturan los elementos más representativos de los audios.

Con el apoyo de *Librosa*, una librería de Python especializada en el análisis de audios, se extraen cada una de estas características de los audios de 5 segundos, que se representan en 216 pasos temporales cada audio. Con esto en cuenta, el dimensionamiento obtenido las características por cada audio de 5 segundos es el ilustrado a continuación:

Característica	Dimensión (por audio de 5 segundos)
MFCC	(20, 216)
Chroma STFT	(12, 216)
Spectral Centroid	(1, 216)
Spectral Bandwidth	(1, 216)
Spectral Contrast	(7, 216)
Spectral Rolloff	(1, 216)
Zero-Crossing Rate	(1, 216)
RMSE	(1, 216)

Luego, tras la extracción, todas las características son concatenadas manteniendo la dimensión temporal de 216 pasos. Dando como resultado una matriz con dimensiones finales de (44, 216) por cada audio, es decir, 44 características para cada uno de los 216 pasos temporales.

- **Normalización de los datos:** Es un paso clave para mejorar la convergencia y la estabilidad del modelo. Por ello, para este trabajo, se emplea la técnica de **normalización Min-Max** para escalar las características extraídas de los audios a un rango estándar, comúnmente entre 0 y 1. Esto es crucial dado que las diferentes características espectrales y temporales extraídas de los audios tienen escalas y unidades muy diferentes, lo que podría dificultar el aprendizaje del modelo si se mantienen en sus valores originales y esta técnica asegurará una representación

homogénea entre las características, evitando el dominio de unas sobre otras y facilitando un entrenamiento más eficiente de las redes neuronales.

Esto se realiza mediante la siguiente fórmula:

$$X_{norm} = \frac{X - X_{min}}{X_{max} - X_{min}}$$

Red Convolucional (CNN)

Arquitectura de la red

La primera red neuronal desarrollada para los objetivos de este proyecto de clasificar géneros musicales es una *Red Neuronal Convolutiva*, también conocida como **CNN**. Este tipo de redes son un tipo de algoritmo de aprendizaje automático que imita, en cierta forma, la manera en que el cerebro humano procesa las imágenes. Aunque su uso común es en el análisis de imágenes, en este caso es aplicada al análisis de los espectrogramas de las canciones, que no son más que visualizaciones del audio en el dominio del tiempo y la frecuencia. Permitiendo así identificar patrones relevantes en las frecuencias de audio que son característicos de cada género musical.

A continuación, se describen brevemente los componentes principales de la arquitectura de la CNN implementada:

1. **Entrada (Input Layer):** compuesta por los espectrogramas de las canciones. En nuestro caso, cada entrada tiene una forma de (44, 216, 1), donde:

- 44 es la cantidad de características extraídas (como MFCC, chroma, centroid, etc.).
- 216 es la cantidad de pasos temporales generados a partir de la segmentación del audio,
- 1 representa que se está trabajando con una imagen en escala de grises.

2. **Convolutional Layers:** son la columna vertebral de la CNN. En este modelo, hemos utilizado 3 *capas convolucionales* con distintos tamaños de filtro y número de canales de salida. Estas capas aplican filtros que recorren el espectrograma buscando patrones de frecuencia y tiempo, como si fueran detectores de características.

3. **Pooling Layers:** luego de cada capa convolucional le sigue una capa de *Max Pooling* de tamaño 2x2, que tienen como propósito reducir la dimensionalidad del espectrograma, lo que disminuye la carga computacional y el riesgo de sobreajuste (overfitting), manteniendo solo las características más importantes.

4. **Flatten Layer:** luego de las capas convolucionales, el tensor resultante es aplanado en un vector unidimensional. Este proceso es necesario para pasar los datos a las capas densas que siguen.

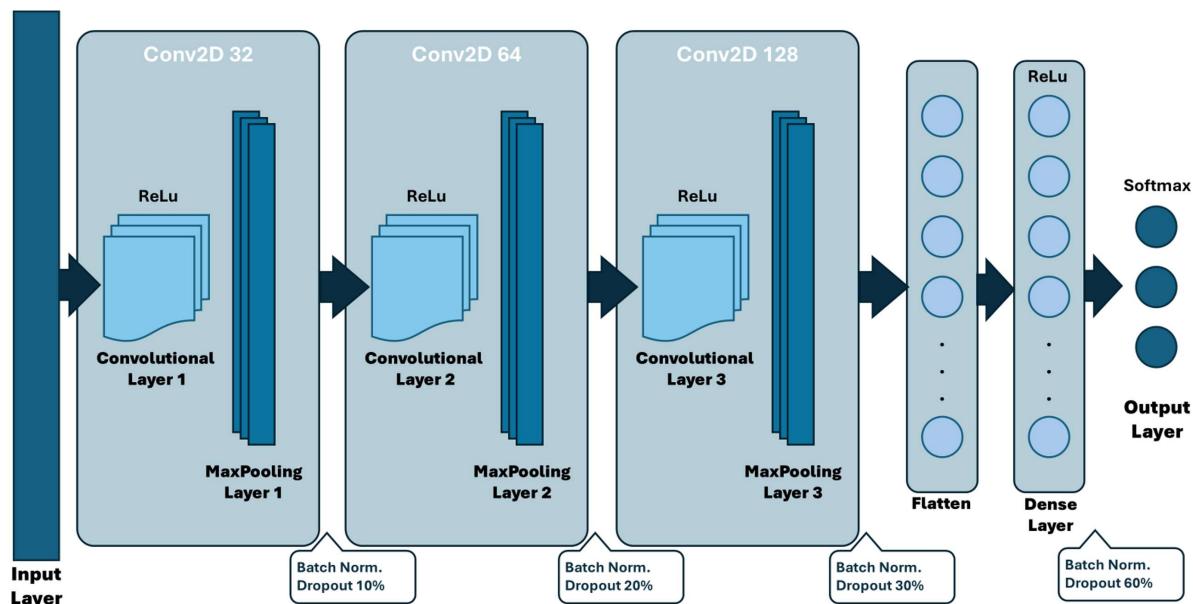
5. **Dense Layer:** una vez que los datos han sido aplanados, entran en 1 *capa densa* de 128 neuronas y activación *ReLU, la cual combina las características aprendidas por las capas

convolucionales para entender mejor las relaciones entre las diferentes frecuencias y temporalidades del audio.

6. **Output Layer:** siendo una capa densa con 10 neuronas (correspondientes a los 10 géneros musicales de la base de datos GTZAN) y una activación *Softmax*, que convierte las predicciones de la red en probabilidades. De esta forma, el modelo genera una probabilidad para cada género, eligiendo aquel con la mayor probabilidad como la clasificación final.

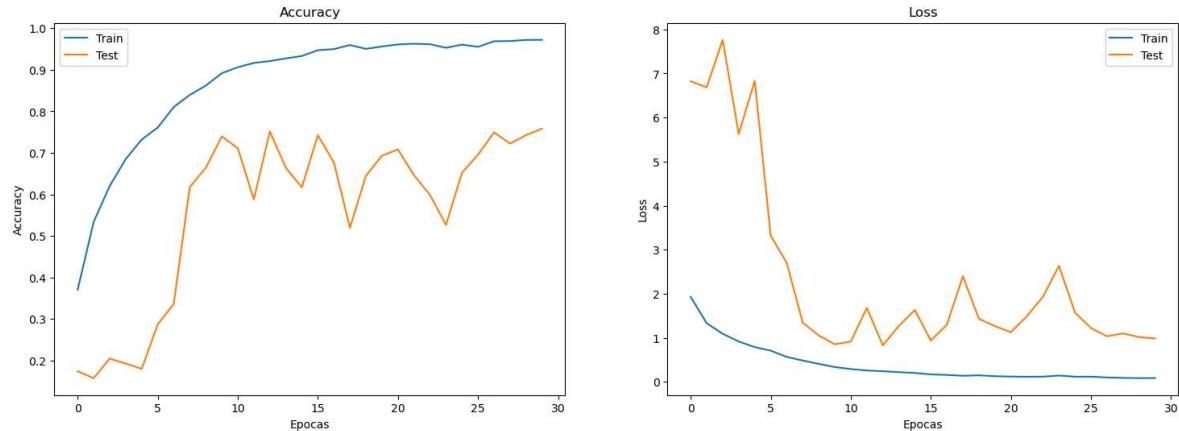
Adicionalmente, a fin de evitar le overfitting de la red, se incorporan *Dropout Layers* y *Batch Normalizations*.

Por ultimo, para el entrenamiento de la red y por ser un problema de clasificación multiclas, se emplea como función de perdida la *categorical cross-entropy* y como optimizador se hace uso de *Adam*.

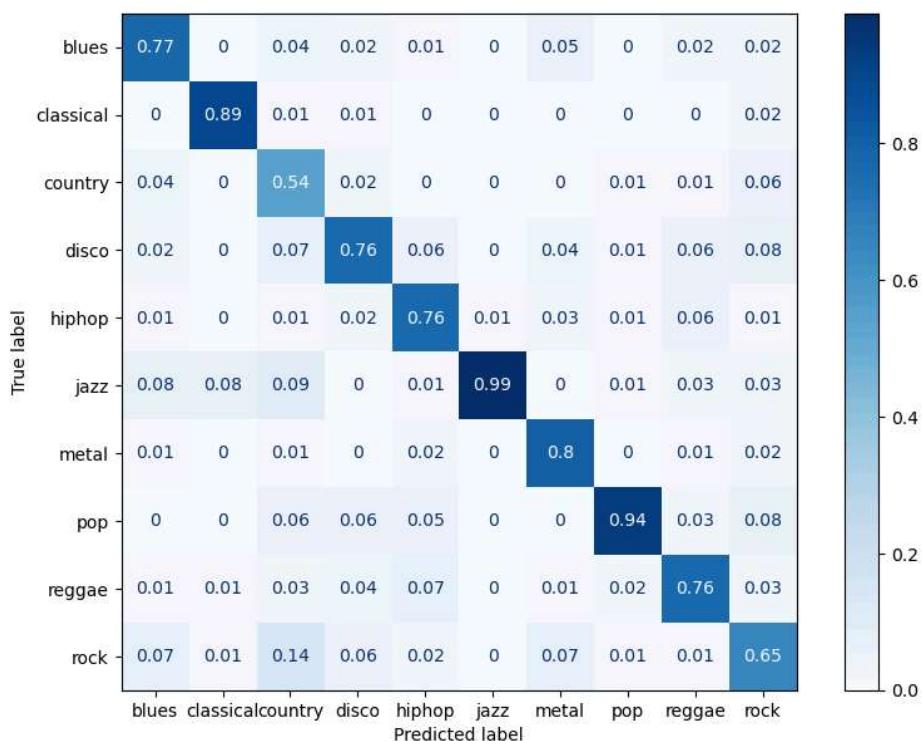


Metricas de evaluación

Evaluando el modelo CNN, se obtuvo un **accuracy de 0.9440 para el conjunto train**, indicando que el modelo es capaz de aprender y ajustarse a los datos de forma eficiente. Sin embargo, **en el conjunto test, el accuracy disminuye a 0.7594**, lo que sugiere un posible *overfitting*, es decir, que está capturando patrones específicos del conjunto train que no se generalizan bien a nuevos datos. Esta diferencia se hace evidente al analizar los gráficos de *accuracy y loss a lo largo de las épocas de entrenamiento, donde se observa que, mientras el accuracy del conjunto train sigue mejorando, el de validación comienza a estabilizarse y fluctuar, indicando que el modelo alcanzó un punto en el que ya no mejora su capacidad de generalización.



Siguiendo con la evaluación del modelo, se incluye una *matriz de confusión* que ilustra el rendimiento del modelo en la clasificación de los diferentes géneros musicales. En esta matriz, se observa que el modelo tiene un buen desempeño en la clasificación de géneros como *pop* y *jazz*, que poseen un accuracy superior a **0.9**. No obstante, existe una notable confusión entre géneros como *rock* y *country*, con accuracy de **0.65** y **0.54** respectivamente.



Red Convolucional Recursiva (CRNN)

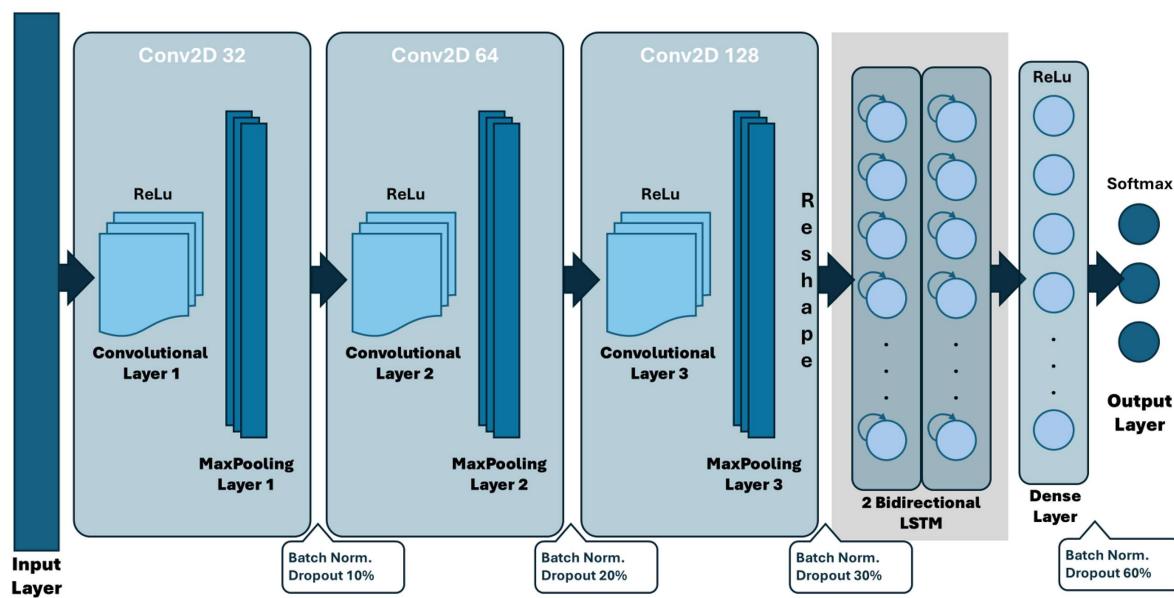
Arquitectura de la red

Ahora, como segunda red neuronal se implementa una *Red Neuronal Convolucional Recurrente*, también conocida como **CRNN**, las cuales combinan la capacidad de las *redes convolucionales* (CNN) para capturar patrones espaciales con las ventajas de las *redes recurrentes* (RNN) para modelar secuencias temporales. Este tipo de arquitectura es especialmente adecuada para el objetivo de este trabajo de clasificación de géneros musicales, donde es crucial analizar tanto las características

espectrales como la evolución temporal del audio. La capacidad de las CRNN para captar patrones en el tiempo permite profundizar el análisis de los cambios rítmicos y melódicos en las canciones, lo que puede mejorar la precisión de la clasificación en comparación con las CNN tradicionales.

La CRNN propuesta en este trabajo comparte la arquitectura descrita previamente de CNN implementada, aprovechando las capas convolucionales para extraer características locales relevantes de los audios. Sin embargo, a diferencia de la CNN, no se utiliza una capa de *Flatten* después de las convoluciones. En su lugar, se aplica una **Reshape Layer** que transforma las salidas de las capas convolucionales en secuencias temporales. Esto permite preservar la estructura temporal del audio, esencial para capturar la evolución de los patrones musicales a lo largo del tiempo. Posteriormente, se integran **2 Bidirectional LSTM Layers** (Long Short-Term Memory) que permiten procesar las secuencias temporales en ambas direcciones, mejorando así la capacidad del modelo para aprender dependencias tanto a corto como a largo plazo dentro del audio.

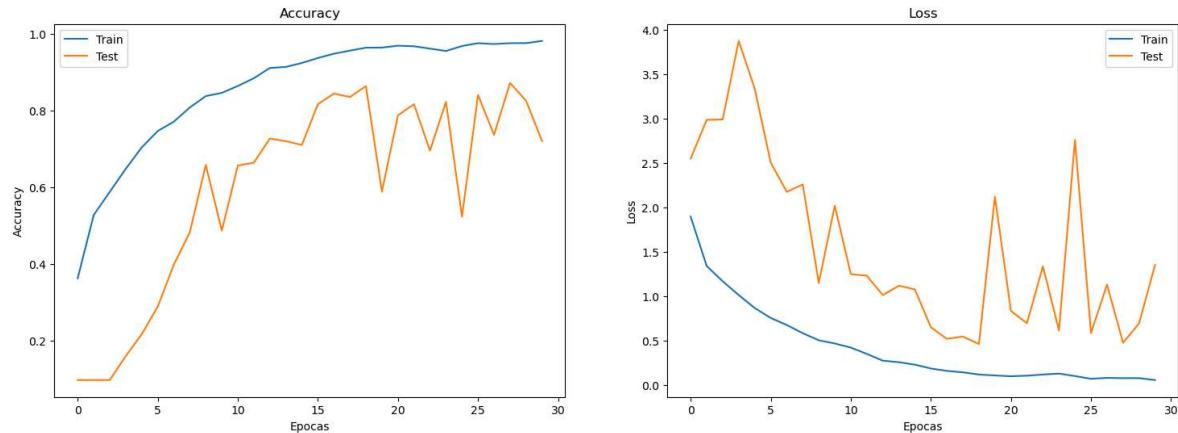
Este enfoque es clave para identificar géneros musicales que dependen de variaciones rítmicas o melódicas complejas a lo largo del tiempo, como el jazz o el rock progresivo. Al combinar estas capas convolucionales y recurrentes, esta CRNN ofrece una solución robusta para la clasificación de géneros musicales, buscando superar las limitaciones de las arquitecturas CNN tradicionales al incorporar una dimensión temporal en el análisis.



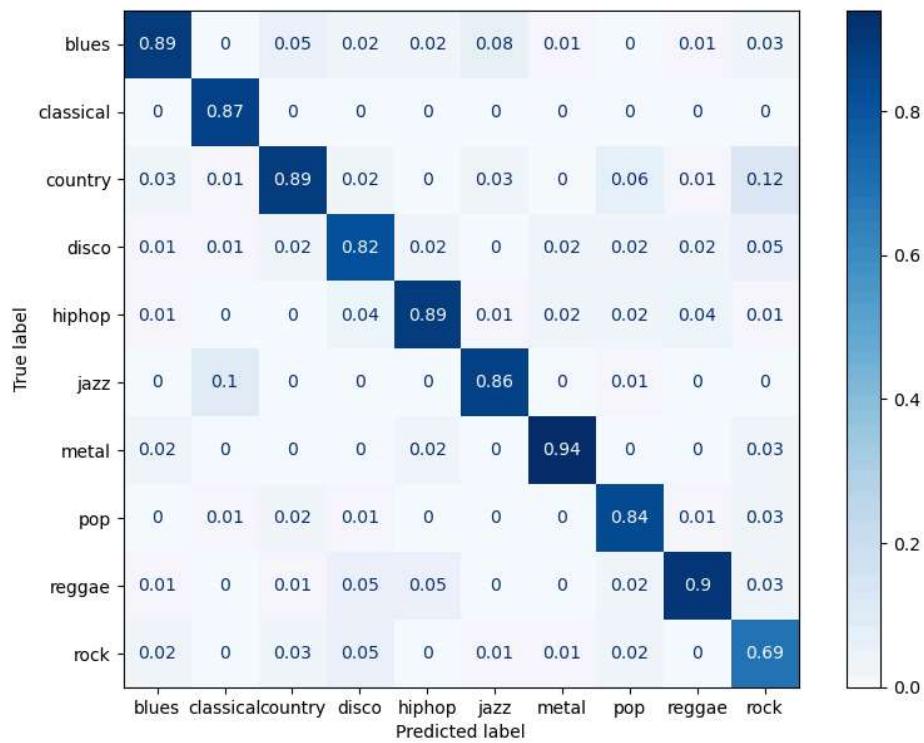
Métricas de evaluación

Ahora evaluando la CRNN, se obtuvo un **accuracy de 0.9573 para el conjunto train**, indicando que este modelo también es capaz de aprender y ajustarse a los datos de forma eficiente. Sin embargo, **en el conjunto test, el accuracy disminuye a 0.8559**, mejorando a la red CNN previa, pero se sigue evidenciando un ligero *overfitting*. Esta diferencia se sigue evidenciando al analizar los gráficos de ***accuracy y loss** a lo largo de las épocas del entrenamiento, donde se observa inicialmente un comportamiento similar a las épocas de entrenamiento de la CNN donde mientras el accuracy del

conjunto train mejora, el de validación comienza a estabilizarse y fluctuar, indicando también que el modelo alcanzó un punto en el que ya no mejora su capacidad de generalización.



Siguiendo con la evaluación de la red CRNN, observamos la *matriz de confusión* entre los diferentes géneros musicales. En ella, se observa que esta red mantiene un buen desempeño en la clasificación de géneros que se clasificaban bien en la CNN como *pop* y *jazz*, manteniendo un accuracy superior a **0.9**. Y adicionalmente, presenta una notable mejora en géneros como *country*, llegando a un accuracy de **0.89**. Sin embargo, el modelo sigue presentando la confusión en el género *rock*, con el mismo accuracy que la red anterior de **0.69**



Comparación de resultados entre CNN y CRNN

Al comparar los resultados entre la red CNN y la CRNN, se aprecia una mejora significativa en el accuracy de la CRNN para el conjunto test. La CNN alcanzó un accuracy de 75.94%, mientras que la CRNN, al integrar tanto las características espaciales como las temporales de los audios, logró un

accuracy del 85.59%, significando que la red CRNN clasifica los géneros musicales de un audio con una precisión de casi 10pp. por encima de la red CNN. Confirmando así que la arquitectura CRNN, al incluir capas recurrentes (LSTM) capaces de modelar secuencias temporales, mejora la capacidad del modelo para clasificar géneros musicales que presentan variaciones rítmicas y melódicas a lo largo del tiempo, que no son fácilmente capturadas por una CNN convencional.

Sin embargo, a pesar de esta mejora general en la CRNN, el género *rock* sigue presentando confusión con otros géneros, como el *country*, lo que se refleja en un **accuracy individual de 69%* para dicho género. Esto puede deberse a la similitud entre las características espectrales y temporales de ambos géneros, que comparten estructuras rítmicas y patrones melódicos similares. Esta confusión sugiere que, aunque la CRNN capta mejor los patrones temporales, los géneros con características musicales cercanas aún son difíciles de distinguir.

Sistema de Recomendación de Géneros Musicales

La red CRNN implementada en este trabajo, originalmente fue diseñada para la clasificación de géneros musicales, demostrando su eficacia al combinar la extracción de características espectrales con la captura de la evolución temporal de los audios. Aprovechando esta capacidad de la CRNN para identificar patrones complejos presentes en la música, se propone una modificación clave en su arquitectura de forma de poder aprovechar con la tarea de *recomendación de canciones similares*.

Para lograr este nuevo modelo de *recomendación de canciones similares*, se modifica la red CRNN de la siguiente manera:

1. **Eliminación de la capa Softmax:** partiendo de la red CRNN entrenada previamente, la cual fue construida como un clasificador multclases, se modifica su arquitectura eliminando la capa Softmax del final del modelo con el fin de cumplir ahora con la función de *generación de embeddings*, siendo esto lo es más adecuado para sistemas de recomendación.
2. **Generación de Embeddings:** Con la modificación, el modelo generará para cada audio de entrada un vector de salida que representará dicho audio en un espacio de características que captura las propiedades musicales más importantes identificadas por la red. Este vector de salida toma el nombre de **embeddings**, los cuales no son más que representaciones numéricas de baja dimensión que capturan las relaciones de las características de los audios.

Ahora mediante la CRNN, para cada audio de la base de datos inicial, se transforma en su respectivo embedding que contiene una representación compacta de todas sus características musicales, lo que permitirá comparar diferentes canciones en un espacio común.

3. **Calculo de Similitud:** Luego de obtener el universo de embeddings, para lograr identificar aquellos audios similares se aplica la **Similitud del Coseno**. Esta es una métrica utilizada para medir cuán cercanos son dos vectores, para el caso de este trabajo, indicará qué tan similares son dos canciones en términos de sus características musicales.

Finalmente de esta forma, mediante la aplicación de la CRNN modificada y la *similitud del coseno* entre embeddings, se **pudiese recomendar canciones similares a los usuarios** en función de sus proximidades en este espacio, lo que permite una personalización y descubrimiento de música más afinada a los gustos musicales.

Conclusiones

En conclusión, la implementación y comparación de las arquitecturas *CNN* y *CRNN* para la clasificación de géneros musicales ha demostrado la superioridad de la CRNN en la captura de las complejas relaciones temporales y espectrales presentes en los audios. Mientras que la CNN logró un *accuracy* de 75.94% en el conjunto de prueba, la CRNN, al integrar capas recurrentes, alcanzó un *accuracy* de 85.59%, mostrando una mejor capacidad de casi 10pp. para distinguir géneros musicales con variaciones rítmicas y melódicas a lo largo del tiempo. Sin embargo, se observó que se mantiene una ligera confusión en géneros cercanos en ambas redes, como *rock* y *country*, lo que sugiere la necesidad de ajustes adicionales o un mayor enriquecimiento de las características utilizadas. Esta diferencia en el rendimiento sugiere que, si bien la CRNN es más eficaz para capturar la dinámica temporal de los géneros musicales, podrían ser necesarias estrategias adicionales, como *data augmentation* más específico o técnicas de *regularización*, para reducir esta confusión y mejorar el rendimiento en géneros con características similares. Además, técnicas como el **fine-tuning de hiperparámetros* o la inclusión de más características acústicas podrían ayudar a mejorar la clasificación específica del rock.

Es importante tener en cuenta que la base de datos *GTZAN* es limitada en cuanto al tamaño de la muestra, lo que puede afectar la generalización de los modelos de deep learning. Si bien en este trabajo se aplicó *data augmentation* sobre los audios para mitigar este problema, las redes neuronales aún requieren de un mayor volumen de datos para mejorar su rendimiento y les permita evitar la confusión entre géneros como *rock* y *country*.

Por último, la modificación de la CRNN para convertirla en un *recomendador de canciones similares* resalta la flexibilidad y aplicabilidad de este enfoque. A través del uso de *embeddings* y la *similitud del coseno*, fue posible desarrollar un sistema que no solo clasifica géneros, sino que también recomienda canciones en función de características musicales compartidas. Los resultados obtenidos muestran el potencial de las CRNN no solo en tareas de clasificación, sino también en aplicaciones más avanzadas como la recomendación personalizada.

Bibliografía

- **Repositorio GitHub que respalda este informe**
(https://github.com/migcanedo/TFM_MusicGenreClasification)
- Base de Datos GTZAN (<https://www.kaggle.com/datasets/andradaolteanu/gtzan-dataset-music-genre-classification>)
- Documentación de Librosa (<https://librosa.org/doc/main/index.html>)
- Documentación de TensorFlow (https://www.tensorflow.org/api_docs/python/tf)
- Musical Genre Clasification of Audio Signals
(<https://www.cs.cmu.edu/~gtzan/work/pubs/tsap02gtzan.pdf>)
- Music Genre Classification Using a Divide & Conquer CRNN
(<https://towardsdatascience.com/music-genre-classification-using-a-divide-conquer-crnn-2ff1cf49859f>)
- Music Genre Classification: A Review of Deep-Learning and Traditional Machine-Learning Approaches (<https://ieeexplore.ieee.org/document/9422487>)
- Convolutional Recurrent Neural Networks for Music Classification
(<https://qmro.qmul.ac.uk/xmlui/bitstream/handle/123456789/30083/Choi%20Convolutional%20recu-sequence=1&isAllowed=y>)

