Reconocimiento de Géneros Musicales Tropicales con Redes Neuronales y Transferencia de Aprendizaje.

Nicolas Abondano, Carlos Salazar
Departamento de Ingeniería de Sistemas y Computación
Universidad de los Andes
Bogotá, Colombia
{nf.abondano, ca.salazara}@uniandes.edu.co

***Abstract*—El reconocimiento de géneros músicales se ha vuelto una tarea bastante interesante y prometedora para el aprendizaje automático pues podría mejorar el desempeño en procesos de automatización de etiquetado de género para una canción. En la actualidad hay múltiples modelos que clasifican con un porcentaje de acierto decente los géneros de multiples canciones. Sin embargo, la mayoría de modelos están limitados a clasificar los géneros más populares a nivel mundial. Se usará el conocimiento de una red con la arquitectura VGG-16 pre-entrenada para clasificación y con ayuda de transferencia de aprendizaje usaremos este conocimiento para reconocer los géneros músicales tropicales más comunes tales como bachata, vallenato, salsa, cumbia y merengue.**

***Index Terms*—música, identificación de género, clasificación, redes neuronales, transferencia de aprendizaje**

I. Introducción

La música es un medio que ha permitido al hombre expresar sus sentimientos y compartirlos con los demás mediante bien sea solo su voz o la de otros o el sonido de uno o más instrumentos o quizá la combinación de todas las anteriores. Con el pasar del tiempo la presencia de algunos instrumentos y la voz humana así como el ritmo en una canción han permitido al hombre encasillarla en uno o más géneros músicales en específico. Los géneros musicales tropicales se caracterizan por ser movidos, por lo general alegres y por ser una combinación del folclore y ritmos del continente Americano con el folclore y ritmos del continente Africano. Clasificar estos géneros para una persona nacida en Centro América o Sur América es una tarea relativamente fácil pues por lo general desde muy temprana edad ha sido expuesto a estos géneros bastante comunes en emisoras de radio y programas de TV. Sin embargo, esta es una de las opciones más lentas para clasificar demasiadas canciones de acuerdo a su género y se ha vuelto necesario automatizar tareas como esta con ayuda de aprendizaje automático, en específico, de redes neuronales. Las redes neuronales, como su nombre lo indica son múltiples neuronas conectadas y ubicadas en diferentes capas que tras un buen entrenamiento y ajuste de hiperparámetros pueden, para el caso de este documento, clasificar con un buen porcentaje de acierto. Una canción es audio, una señal analógica que, para poder ser usada por el modelo necesita ser representada no en su formato digital sino en un formato adecuado, el formato que proponemos usar para representar una canción es el Espectograma de Frecuencias de Mel el cual es la representación visual de las frecuencias en un instante de tiempo en específico de una canción y ha demostrado funcionar en modelos de reconocimiento de voz y de clasificación de géneros musicales. En la actualidad no hay conjuntos de datos con abundante cantidad de géneros tropicales por lo que generaremos un conjunto de datos relativamente pequeño y lo usaremos para entrenar las últimas capas de un modelo previamente entrenado con un conjunto de datos más grande pero con el fin de clasificar imágenes de manera que el conocimiento con el que cuenta este modelo (identificación de ciertos patrones y características) nos sea útil para clasificar géneros musicales tropicales, a esto se le conoce como transferencia de aprendizaje.

II. Revisión bibliográfica

El proyecto publicado en el interspeech 2016 realizado por Deepanway Ghosal y Maheshkumar H. Kolekar [1] consitía en el reconocimiento de géneros musicales haciendo uso de redes neuronales convolucionales basadas en la memoria a largo y corto plazo (CNN LSTM) y un modelo de aprendizaje por transferencia. Los modelos de redes neuronales fueron entrenados usando un conjunto diverso de características espectrales y rítmicas, mientras que el modelo de aprendizaje por transferencia se entrenó en la tarea de etiquetar música.
El proyecto realizado por A.J Krishna [2] se centra principalmente en la implementación del reconocimiento de distintos géneros musicales haciendo uso de redes neuronales de convolución (CNN / ConvNet). Este modelo se entrena utilizando espectogramas Mel de las distintas canciones y etiquetando las mismas. Este proyecto concluyó como en el reconocimiento de géneros musicales las redes convolución tenian un desempeño superior a otros mecanismos especializados en la clasificación de los mismos.
En el estudio realizado por Keunwoo Choi, Gyorgy Fazekas y Mark Sandler [3] presentó un algorimto de etiquetado automático de música basado en contenido que utiliza redes neuronales totalmente convolucionales (FCN). Únicamente se

evaluaron arquitecturas que constan de capas convolucionales 2D y capas de submuestreo. En los experimentos se medía los puntajes AUC-ROC de las arquitecturas con diferentes complejidades y tipos de entrada utilizando el conjunto de datos MagnaTagATune, donde una arquitectura de 4 capas muestra un rendimiento increíble con entrada de espectrograma mel. Además se evaluó el desempeño de las arquitecturas variando el número de capas en un conjunto de datos más grande (Million Song Dataset) y se descubrió que los modelos más profundos superaban a la arquitectura de 4 capas. Los experimentos mostraron que el espectrograma mel es una representación de tiempo-frecuencia efectiva para el etiquetado automático y que los modelos más complejos se benefician de más datos de entrenamiento.

El estudio realizado por Tammina, S [4] estipula como los algoritmos de minería de datos y de aprendizaje automático están diseñados para abordar los problemas de manera aislada. Se emplean para entrenar el modelo en separación en un espacio de características específico y en la misma distribución. Dependiendo del caso de negocio, un modelo se entrena aplicando un algoritmo de aprendizaje automático para una tarea específica. Una suposición generalizada en el campo del aprendizaje automático es que los datos de entrenamiento y los datos de prueba deben tener espacios de características idénticos con la distribución subyacente. Por el contrario, en el mundo real, esta suposición puede no ser válida y, por lo tanto, los modelos deben reconstruirse desde cero si las características y la distribución cambian. Es un proceso arduo recopilar datos de entrenamiento relacionados y reconstruir los modelos. En tales casos, sería deseable la transferencia de conocimientos o la transferencia de aprendizaje desde dominios dispares. El aprendizaje por transferencia es un método para reutilizar un modelo de conocimiento previamente entrenado para otra tarea. El aprendizaje por transferencia se puede utilizar para problemas de clasificación, regresión y agrupación. Este documento utiliza uno de los modelos previamente entrenados - VGG - 16 con Deep Convolutional Neural Network para clasificar imágenes. Esta clasificación sería posible utilizarla con imágenes como lo pueden ser espectogramas mel de audios.

REFERENCIAS

[1] Indian Institute of Technology Patna, India, Deepanway, G., Maheshkumar, H. K. (2018, September). Music Genre Recognition using Deep Neural Networks and Transfer Learning. Interspeech 2018. https://www.isca-speech.org/archive/Interspeech_2018/pdfs/2045.pdf

[2] Krishna Mohana, A. J., Pramod Kumar, P. M., Harivinod, N., Nagaraj, K. (2019). Music Instrument Recognition from Spectrogram Images Using Convolution Neural Network. International Journal of Innovative Technology and Exploring Engineering, 8(9), 1076–1079. https://doi.org/10.35940/ijitee.i7728.078919

[3] Queen Mary University of London, Choi, K., Fazekas, G., & Sandler, M. (2016, June). AUTOMATIC TAGGING USING DEEP CONVOLUTIONAL NEURAL NETWORKS. https://arxiv.org/pdf/1606.00298.pdf

[4] Tammina, S. (2019). Transfer learning using VGG-16 with Deep Convolutional Neural Network for Classifying Images. International Journal of Scientific and Research Publications (IJSRP), 9(10), p9420. https://doi.org/10.29322/ijsrp.9.10.2019.p9420