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

***Resumen*—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, la de otros o el sonido de uno o más instrumentos e incluso 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 Espectrograma 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 espectrogramas de Mel de las distintas canciones y etiquetando las mismas. Este proyecto concluyó que en el reconocimiento de géneros musicales las redes neuronales convolucionales tenían 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 algoritmo 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 espectrogramas de 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 de 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 los espectrogramas de Mel de las canciones.

III. Método propuesto

El método propuesto para resolver este problema de clasificación es una red neuronal con múltiples salidas que hace uso de la función de activación softmax combinado con Transferencia de aprendizaje, La transferencia de aprendizaje consiste en aprovechar el conocimiento obtenido de resolver una o más tareas con la (s) que se cuenta con un gran conjunto de datos, para resolver una tarea similiar con la que se cuenta con un conjunto de datos pequeño pues, entrenar un modelo con un conjunto de datos pequeño puede no llegar a presentar un desempeño tan bueno como con un conjunto de datos más grande [6]. Este método es comúnmente usado en proyectos de redes neuronales dado que, a veces, no se cuenta con un conjunto de datos robusto u otras veces el manejo de grandes conjuntos de datos para entrenar modelos propios es computacionalmente costoso por lo que es necesario recurrir a modelos de quienes cuentan con las capacidades y el acceso a estos conjuntos de datos, entrenan un modelo y lo ponen a disposición de los demás para usarlo en la tarea para la cual fue entrenado o similares. De hecho, en varios de los proyectos revisados se hizo uso de este método con resultados satisfactorios y hay librerías que cuentan con estos modelos pre-entrenados como lo es Tensorflow.

IV. Experimentos

IV-A. Descripción de los datos

El método propuesto no hace uso de archivos de audio mp3 directamente sino que hace uso de espectrogramas de MEL generados en fragmentos de 30 segundos de canciones de los distintos géneros tropicales a clasificar (bachata, vallenato, salsa, cumbia y merengue), las canciones fueron extraídas de vídeos de youtube, separadas en fragmentos de 30 segundos con ayuda de ffmpeg y los espectrogramas fueron generados con las librerías de Python librosa y matplotlib.

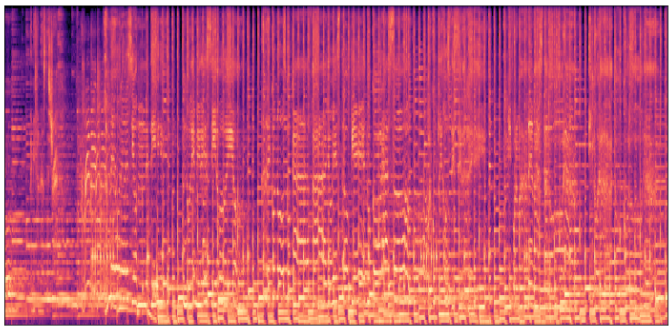

Figura 1. Espectrograma de MEL para un fragmento de una canción de bachata

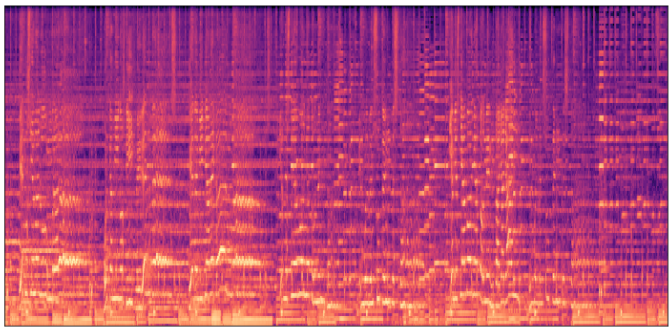

Figura 2. Espectrograma de MEL para un fragmento de una canción de vallenato

Inicialmente se probó con un conjunto de datos que constaba de 1500 datos (300 por género) repartido asignando el 80 % al conjunto de entrenamiento (1200 datos) y 10 % al conjunto de validación y de entrenamiento (150 datos para cada uno). Sin embargo, al probar con canciones que no hacían parte del conjunto de datos sino seleccionadas al azar se observó que el modelo se desempeñaba bien. Sin embargo, este desempeño

no nos daba mucha confianza pues al probar con distintos ritmos de un mismo género el modelo tendía a clasificarlos como un género distinto por lo que se decidió añadir más datos teniendo en cuenta un rango más amplio de ritmos de cada género para tener una mayor certeza en los resultados del modelo. Este nuevo conjunto de datos constaba de los datos del primer conjunto de datos más 1670 datos nuevos para un total de 3170 datos (634 por género) asignando de nuevo el 80 % al conjunto de entrenamiento (2536) y 10 % tanto para el conjunto de prueba como para el conjunto de validación (317 datos para cada uno).

IV-B. Criterios de selección del modelo

Para el ojo humano es difícil distinguir entre un género y otro a partir de los espectrogramas de MEL. Sin embargo, las redes neuronales convolucionales que cuentan con capas de max pooling trabajan bastante bien con este tipo de representación de audios con respecto a las que cuentan con capas de average pooling de acuerdo a Deepanway y Maheshkumar [1]. Según vimos anteriormente en el estudio realizado por Tammina [4] se hizo uso del modelo VGG - 16 y Transferencia de aprendizaje para clasificar imágenes, este modelo constaba de 13 capas convolucionales, 5 capas de max pooling y 3 capas densas [5]. Nosotros de la misma manera iniciamos nuestro modelo haciendo uso del modelo VGG - 16 implementado y pre-entrenado en el conjunto de datos de ImageNet por la librería Tensorflow. Asimismo, se usó como función de pérdida *Categorical Crossentropy*, como optimizador *Adam* y 5 épocas para el entrenamiento pues el uso de más algunas veces el entorno de ejecución no permitía al modelo terminar el entrenamiento por lo que con ayuda del Callback se guardaban los mejores pesos y se cargaban para iniciar un nuevo entrenamiento. Sin embargo, solo se requirió de 1 entrenamiento para alcanzar el mejor desempeño, el uso de más conducía al modelo a sobreajustarse a los datos de entrenamiento y no incrementaba su desempeño sobre el conjunto de validación.

Para aplicar Transferencia de aprendizaje como se evidencia en la arquitectura del modelo, le quitamos las últimas 3 capas (las capas densas) al modelo VGG - 16 y le añadimos dos capas densas de 512 y 64 neuronas respectivamente con activación ReLU y una última capa densa con activación softmax de 5 neuronas para clasificar las imágenes de los espectrogramas de MEL de los distintos géneros trópicales. Asimismo, se optó por congelar o establecer como no entrenables a las capas que no se añadieron al modelo de manera que se pudiera conservar el conocimiento que ya había adquirido el modelo y aprovecharlo para entrenar las nuevas capas. Finalmente, el modelo resultante constaba de 35'982.021 parámetros de los cuales 21'267.333 son entrenables y 14'714,688 no son entrenables (venían en el modelo VGG - 16).

Se intentó variar la cantidad de capas añadidas y de unidades presentes en ellas pero no se obtuvieron diferencias significativas en el desempeño con respecto a tener 3 capas con dicha cantidad de unidades como las presentes en la tabla. Además, el alto consumo de recursos y tiempo no compensaba el ligero cambio en el desempeño.

Cuadro I
ARQUITECTURA DEL MODELO

Capa (Tipo)	Tamaño de salida	# Parámetros
Entrada_1 (Entrada)	(300,300,3)	0
Bloque1_conv1 (Conv2D)	(300,300,64)	1792
Bloque1_conv2 (Conv2D)	(300,300,64)	36928
Bloque1_pool (MaxPooling2D)	(150,150,64)	0
Bloque2_conv1 (Conv2D)	(150,150,128)	73856
Bloque2_conv2 (Conv2D)	(150,150,128)	147584
Bloque2_pool (MaxPooling2D)	(75,75,128)	0
Bloque3_conv1 (Conv2D)	(75,75,256)	295168
Bloque3_conv2 (Conv2D)	(75,75,256)	590080
Bloque3_conv3 (Conv2D)	(75,75,256)	590080
Bloque3_pool (MaxPooling2D)	(37,37,256)	0
Bloque4_conv1 (Conv2D)	(37,37,512)	1180160
Bloque4_conv2 (Conv2D)	(37,37,512)	2359808
Bloque4_conv3 (Conv2D)	(37,37,512)	2359808
Bloque4_pool (MaxPooling2D)	(18,18,512)	0
Bloque5_conv1 (Conv2D)	(18,18,512)	2359808
Bloque5_conv2 (Conv2D)	(18,18,512)	2359808
Bloque5_conv3 (Conv2D)	(18,18,512)	2359808
Bloque5_pool (MaxPooling2D)	(9,9,512)	0
Plana (Flatten)	(41472)	0
Densa (Densa)	(512)	21234176
Densa_1 (Densa)	(64)	32832
Densa_2 (Densa)	(5)	325

Adicionalmente para el modelo se implementó una función callback para el entrenamiento, la cuál, al terminar una época, guardaba los pesos en caso de que generaran mayor precisión y menor costo de forma que el modelo fuera más preciso. Estos pesos se reutilizaban al entrenar el modelo nuevamente en otra época.

IV-C. Evaluación de la solución

Las métricas de precisión y pérdida del modelo durante y después del entrenamiento se muestran a continuación. El eje x representa el número de la época, la época número 0 correponde a la primera y la número 4 a la última, cada entero es una época. El eje y representa el porcentaje de precisión en la figura 3, y el porcentaje de pérdida en la figura 4

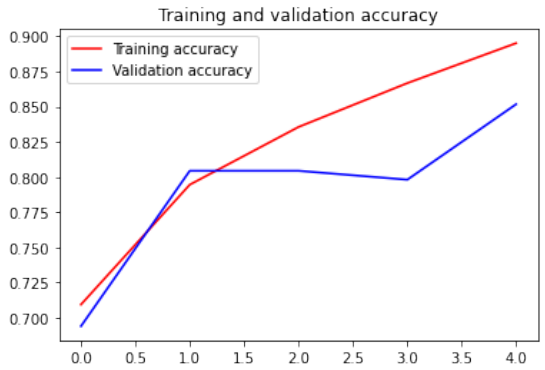

Figura 3. Acierto del modelo

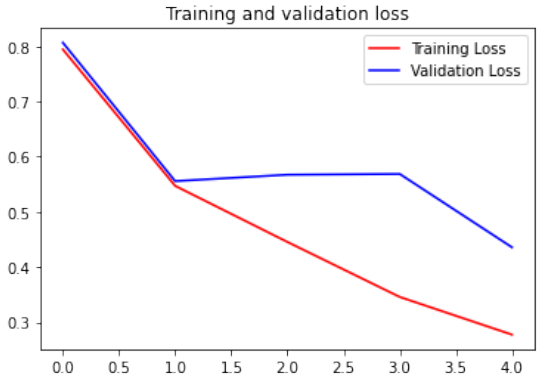

Figura 4. Pérdida del modelo

Finalmente al evaluar con el conjunto de prueba evidenciamos una precisión del 87 % y una pérdida del 38 %. Se puede reconocer que la precisión del modelo es bastante buena, aunque la pérdida no es tan baja como se esperaría. Esto comparando con los otros trabajos mostrados en la parte IV-D

```
Test loss: 0.3815688490867615, Test acc: 0.8769716024398804
```

Figura 5. Prueba del modelo

IV-D. Comparación con otros métodos

Otros métodos usados para clasificación de géneros musicales son las redes neuronales recurrentes que demuestran tener un buen desempeño incluso en conjuntos de datos relativamente pequeños y sin hacer uso de transfer learning tal y como demuestra Irvin [7] que entrenó con 1200 espectrogramas diversos modelos entre los cuales destacaron 2: una red neuronal *vanilla* con *LSTM* con un acierto del 79 % sobre el conjunto de prueba y una red neuronal con *Attention* con un acierto de 75.8 % sobre el conjunto de prueba. Tanto *LSTM* como *Attention* son mecanismos usados por redes neuronales recurrentes para incrementar su desempeño.Otro método empleado es Adaboost el cual consiste en varios predictores en secuencia de manera que cada predictor aprenda de los errores de su predecesor, este método fue empleado por Bergstra [8] en 3 conjuntos de datos pequeños aplicando una técnica basada en extraer distintas características acústicas de cada segmento de audio mediante funciones matemáticas y pasarlas como entrada al modelo obteniendo un acierto del 82.5 % en el conjunto de prueba. Nuestro modelo tras entrenarlo con ayuda del callback logró un 87 % de acierto sobre el conjunto de prueba con una pérdida relativamente baja, un desempeño mejor que los mencionados previamente y que puede ser superado con el uso de las arquitecturas propuestas por Irvin [7] o por la extracción de características usada por Bergstra [8] sobre el conjunto de datos creado.

V. Conclusiones, posibles mejoras y trabajo a futuro

En este documento, aplicamos Transfer Learning a redes neuronales convolucionales, específicamente el modelo VGG-16 para clasificar géneros tropicales. Este modelo clasificó de manera correcta al 89 % de los datos del conjunto de entrenamiento, al 85 % de los datos del conjunto de validación y al 87 % de los datos del conjunto de prueba con una pérdida relativamente baja. Creemos que tanto el acierto como la pérdida pueden disminuir con un conjunto de datos más grande que incluya a los subgéneros de los 5 géneros tropicales propuestos y con el uso de más épocas y regularización.
Al igual que como Irvin [7] considera sería interesante a futuro tratar de transformar los pesos a espectrogramas y de espectrogramas a archivos de audio de manera que se pueda escuchar la información que el modelo aprendió de cada género y tiene en cuenta para clasificar. También consideramos que sería interesante identificar el género de una canción completa con base en espectrogramas de Mel generados para cada segmento de 30 segundos de la canción, esto se puede hacer de dos formas, una es eligiendo el género más común entre todos los fragmentos (votación) y otra es sumar todas las probabilidades obtenidas para cada género por los fragmentos de la canción y seleccionar la que tenga una mayor probabilidad pues puede darse el caso en el que dos canciones tengan la misma cantidad de votos pero distintas probabilidades totales, esto último podría ser el factor de desempate y habría que evaluar el desempeño de ambos métodos elegidos para clasificar el género de una canción.

Finalmente, validamos que el uso de redes neuronales convolucionales y espectrogramas de Mel para clasificar géneros musicales garantiza buenos resultados con ayuda de Transfer Learning con conjuntos de datos pequeños.

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

[5] Simonyan, Karen, and Andrew Zisserman. «Very Deep Convolutional Networks for Large-Scale Image Recognition». arXiv:1409.1556 [cs], abril de 2015. arXiv.org, http://arxiv.org/abs/1409.1556.

[6] L. Torrey & J. Shavlik (2009). Transfer Learning. In E. Soria, J. Martin, R. Magdalena, M. Martinez & A. Serrano, editor, Handbook of Research on Machine Learning Applications. IGI Global.

[7] J. Irvin, E, Chartock & N. Hollander. Recurrent Neural Networks with Attention for Genre Classification. http://cs229.stanford.edu/proj2016/report/IrvinChartockHollander-RecurrentNeuralNetworkswithAttentionforGenreClassification-report.pdf

[8] J. Bergstra, N. Casagrande, D. Erhan, D. Eck & B. Kegl. Aggregate features and AdaBoost for music classification. Machine Learning, Vol. 65, No. 2-3. 2006.

VI. Créditos

Nicolás Abondano - Realizó la investigación en la literatura, evaluación del desempeño, realización parcial del conjunto de datos, construcción de la presentación y mejora del modelo.
Carlos Salazar - Realizó la investigación en la literatura, creación del conjunto de datos, construcción de la arquitectura del modelo y funciones auxiliares.