TRABAJO ACADÉMICO 2

Reconocimiento de formas y apredizaje automático

Autora
Aitana Menárguez Box
Noviembre 2023

Índice

1 Introducción

El objetivo de este trabajo es explorar el entrenamiento, optimización y evaluación de modelos basados en redes neuronales. La tarea que se ha escogido para este trabajo se trata de clasificar diferentes fragmentos de audio según el instrumento predomimante que suene dentro de ellos. A continuación, se muestra el trabajo realizado con los modelos propuestos además de la gestión del *dataset* y las herramientas utilizadas para ello.

2 Datos utilizados

El conjunto de datos utilizado en este trabajo se trata del *set* IRMAS[1]. Dentro de éste se encuentran audios en los cuales predomina un instrumento. Los intrumentos considerados son: cello (cel), clarinete (cla), flauta (flu), guitarra acústica (gac), guitarra eléctrica (gel), órgano (org), piano (pia), saxofón (sax), trompeta (tru), violín (vio) y voz humana cantada (voi). El conjunto de datos se deriva del recopilado por Ferdinand Fuhrmann en su tesis doctoral.
Originalmente, se han dividido los datos en dos partes. Por simplicidad, en este trabajo se experimenta solamente con la primera partición (llamada originalmente *training*), ya que la segunda partición cuenta con audios en los cuales predomina más de un instrumento. Esta primera partición (en adelante será referida como datos sin más) son 6705 archivos de audio en 16 bits en formato estéreo i .wav *sampleados* a 44.1 kHz. Son extractos de 3 segundos de más de 2000 grabaciones diferentes. En la Fig. 1 se puede apreciar la cantidad de datos (audios) de cada instrumento con la que se cuenta. Adicionalmente, algunos de los archivos tienen anotaciones en su nombre referentes a la presencia ([dru]) o no ([nod]) de batería en la grabación, así como el género musical: country-folk ([cou_fol]), clásica ([cla]), pop-rock ([pop-roc]) y latin-soul ([lat-sou]). Por otro lado, datos varían entre música actual y de varias décadas del último siglo, además de en la cualidad de sonido. Cubre una gran variabilidad de tipos de instrumento, intérpretes y articulaciones, además de estilos de producción y grabación.

3 Herramientas

El código de este trabajo está basado en diferentes repositorios de los cuales se han extraído ideas de organización y estructura del código, así como algunos de los modelos utilizados. Principalmente, se ha extraído información de: `https://github.com/OdysseasKr/irmas-cnn` y `https://github.com/claudia-hm/IRMAS_Deep_Learning/tree/master`.
También se mencionan a continuación algunas de las herramientas específicas utilizadas y dónde acceder a su documentación:

- Librosa: `https://librosa.org/`
- Essentia: `https://pypi.org/project/essentia/`

[1] Acesible a través de `https://www.upf.edu/web/mtg/irmas`

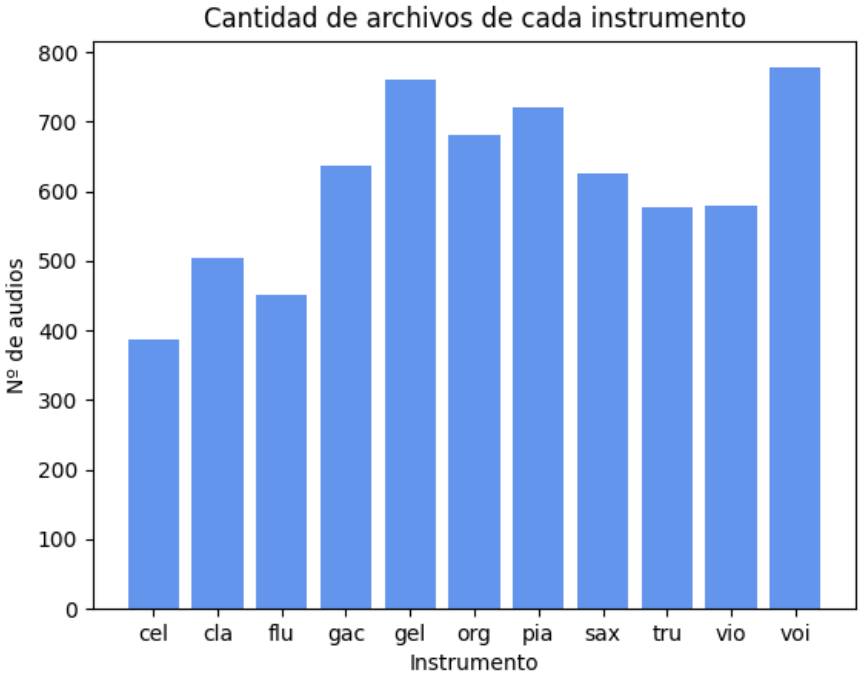


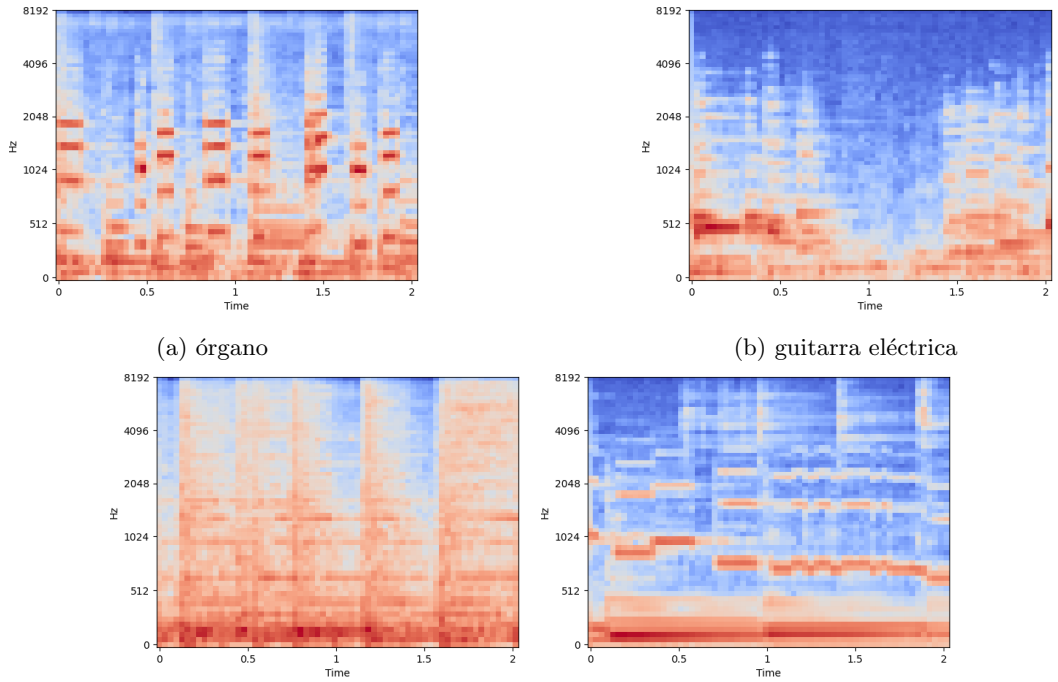
Figura 1: Distribución de la cantidad de archivos de audio en los datos para cada uno de los instrumentos posibles.

4 Preprocesamiento de datos

Para preprocesar los audios musicales se han seguido dos métodos distintos: extracción de espectrograma y extracción del centroide espectral.

4.1 Espectrograma

El espectrograma (o sonograma) se trata de una representación de señales a partir de su contenido frecuencial. Se trata de una gráfica de tres dimensiones que muestra la variación de la frecuencia de la señal en función de la intensidad y del tiempo. Esta representación se puede interpretar como una proyección en dos dimensiones de una sucesión de transformadads de Fourier de tramas consecutivas, donde la energia y la frecuencia van variando a lo largo del tiempo. En Fig. 2 se muestran varios ejemplos de representaciones por espectrograma del conjunto de datos.



(a) órgano

(b) guitarra eléctrica

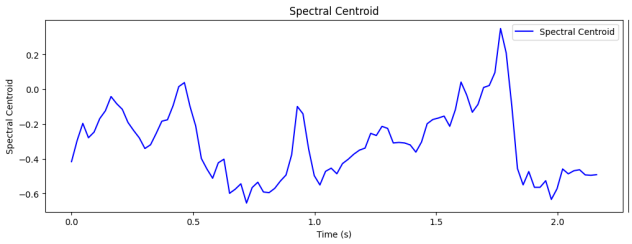
(c) guitarra acústica

(d) flauta

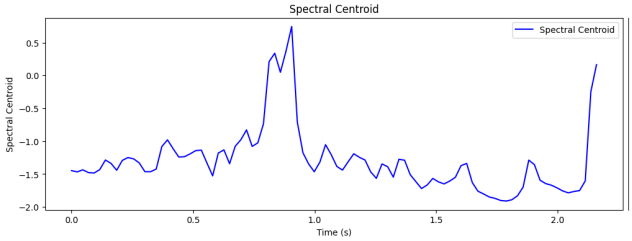
Figura 2: Ejemplos de representación por espectrograma de algunas de las muestras de datos

4.2 Centroide espectral

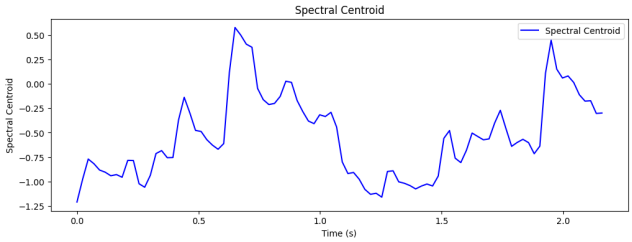
El centroide espectral o *spectral centroid* es una medida utilizada en procesamiento de señales digitales para caracterizar un espectro. Indica dónde se sitúa el centro de la masa de éste. Perceptualmente, tiene conexión directa con cómo de brillante se percibe auditivamente un sonido. En la Fig. 3 se pueden ver ejemplos de esta representación para el set de datos del trabajo.



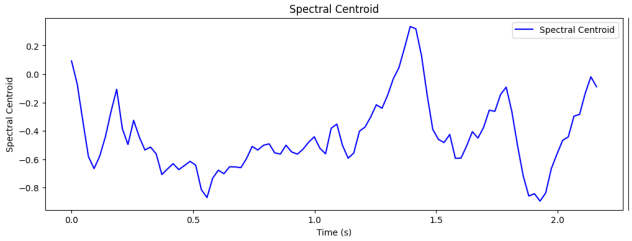
(a) órgano



(b) guitarra acústica



(c) voz humana cantada



(d) clarinete

Figura 3: Ejemplos de representación por centroide espectral de algunas de las muestras de datos

5 Arquitecturas

Las arquitecturas que se han utilizado en el trabajo han sido cuetro diferentes, dos por cada tipo de procesamiento de datos. A continuación se muestra cuáles han sido:

- **Red VGG**: se trata de una red neuronal convolucional propuesta en la Universidad de Oxford, originalmente planteada para el reconocimiento visual a gran escala de ImageNet. Se ha utilizado para el procesamiento por espectrograma.
- **Red convolucional estándar**: se trata de un modelo simple de una red neuronal convolucional. Se ha utilizado para el procesamiento por espectrograma.
- **Red sencilla**: se ha utilizado para el procesamiento por centroide espectral.
- **VGG versionada**: se trata de una red VGG pero con convoluciones en 1D, ya que se ha utilziado para el procesamiento por centroide espectral, en el cual los datos vienen dados en vectores unidimensionales.

6 Experimentación y resultados

6.1 Pasos previos

Antes de realizar el trabajo con los modelos, se han instalado las librerías necesarias y se han preprocesado los datos. Las funciones para el preprocesamiento pueden consultarse en el *notebook* asociado a esta memoria. Por otro lado, se ha montado un directorio en Drive[2] para poder acceder de forma sencilla a los datos de IRMAS, ya que no hay niguna librería que los proporcione para poder gestionarlos en Python.

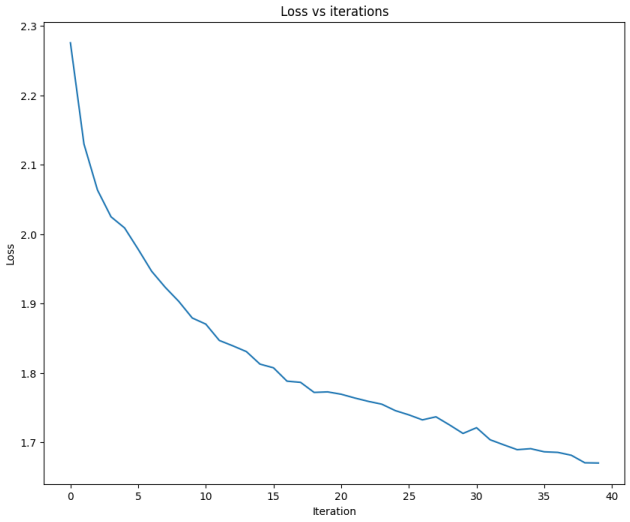
6.2 Entrenamiento y evaluación

El entrenamiento se ha realizado con el 80% de los datos, mientras que el *test* se ha hecho con el 20% restante. En total, hay 5364 muestras de entrenamiento y 1341 de prueba.
La evolución de la pérdida en el entrenamiento de cada modelo para su set de datos correspondiente se muestra en las figuras Fig. 4 y Fig. 5. Tras intentar clasificar los datos de prueba a partir de los modelos anteriores, se obtienen los resultados de precisión mostrados en Tab. 1.

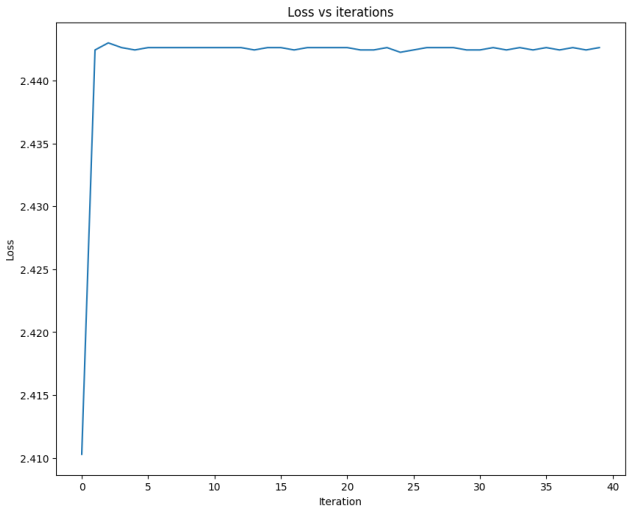
	Precisión
VGG	58.17%
CNN	10.67%
Modelo simple	24.61%
VGG-2	24.09%

Tabla 1: Resultados de precisión evaluados a partir de los datos de IRMAS para los modelos de VGG, CNN, Modelo simple y VGG2.

[2]Este directorio es accesible desde el siguiente enlace:
`https://drive.google.com/drive/folders/1zU5f25zaeQb3lKK-uuqc3TOnWQFtD9ub?usp=sharing`

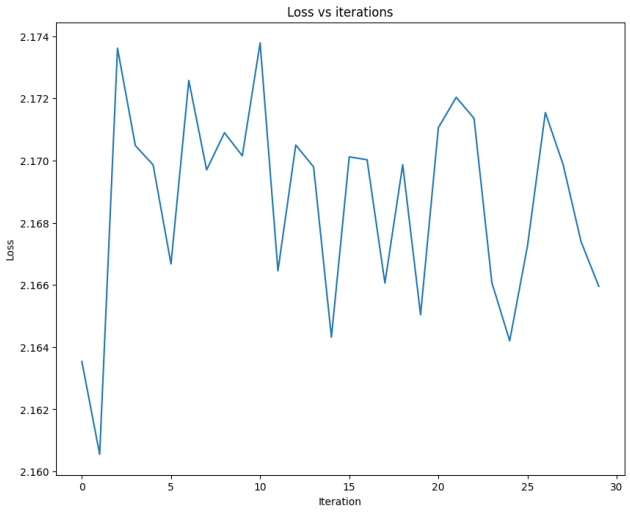


(a) Modelo VGG

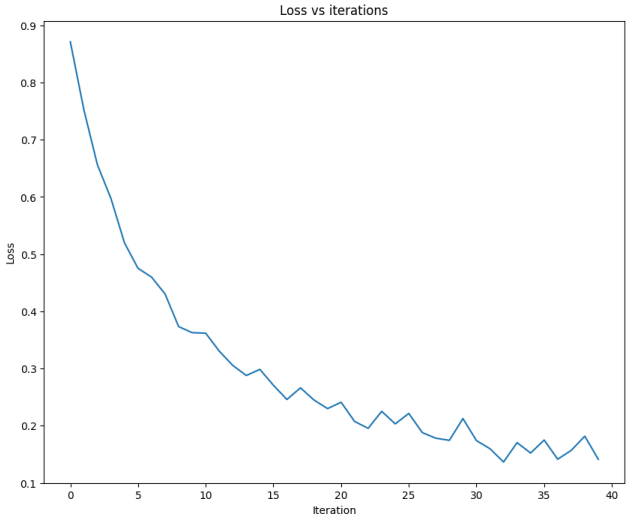


(b) Modelo CNN

Figura 4: Evolución de la *loss* en el entrenamiento para la representación de espectrograma.



(a) Modelo simple



(b) Modelo VGG-2

Figura 5: Evolución de la *loss* en el entrenamiento para la representación de centroide espectral.

7 Conclusiones

De los resultados obtenidos se observa que el modelo que mejor funciona es el VGG para el preproceso de datos con espectrograma. El peor resultado obtenido es el de la CNN. Además, los resultados para el preproceso a partir de centroide espectral son realmente similares. Eso último puede ser debido a que, con el objetivo de distinguir qué instrumento suena en cada audio, no sea suficiente con extraer el centroide espectralm sino que se necesite más información sobre cada pista. Es decir, en este caso no ha sido suficientemente buena la representación del instrumento.
Por otro lado, se puede apreciar que los resultados de precisión son bastante bajos, o al menos si se comparan con los resultados de otras tareas de clasificación resueltas a partir de redes neuronales. Esto puede ser debido a la poca cantidad de datos que han sido utilizados para entrenar los modelos.

En conclusión, como trabajo futuro, se podrían representaciones de los audios diferentes (como el *roll-off* espectral, el *zero-crossing rate* o el ancho de banda espectral), con tal de buscar cuál de ellas (o de sus combinaciones) permiten extraer los datos representativos de cada tipo de instrumento. Además, también sería interesante investigar nuevos tipos de arquitecturas específicas que, no han podido ser probadas por falta de capacidad de tiempo, de cómputo y, sobre todo, de datos. Por esto último cabe destacar la necesidad de más tipos de datos, ya que gran parte del tiempo destinado a este trabajo se ha basado en encontrar cómo procesarlos.