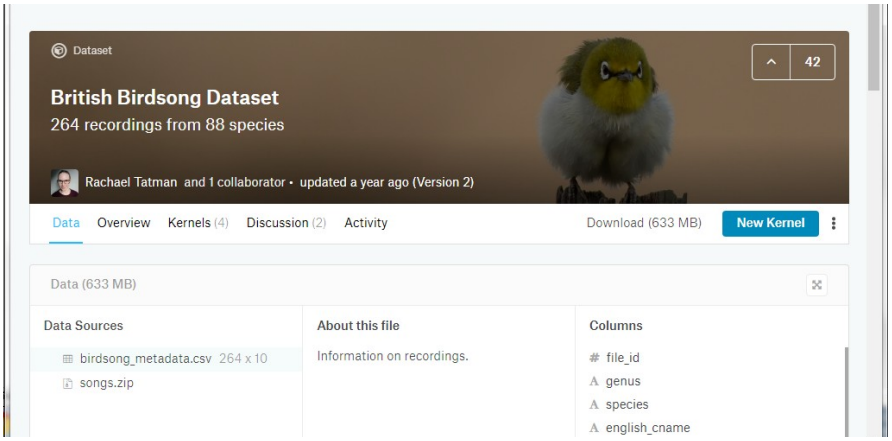
TFM Identificando aves por el canto

8 Edición Master Data Science KSCHOOL

Alberto Becerra Villarejo
Enero 2019

1. Introducción.

La idea del tema de este TFM surge a partir de un dataset publicado en Kaggle, y para el cuál se lanzó el reto de generar un clasificador que fuera capaz de identificar un grupo de aves a partir de su canto.



El dataset en cuestión contenía 264 grabaciones pertenecientes a 88 especies, de las más comunes del Reino Unido, y procedía de la web www.xeno-canto.org, que comparte grabaciones de especies de todo el mundo y las ponen a disposición de todos los usuarios.

A priori, sin tener conocimientos de acústica, y viendo que el número de grabaciones que presenta el dataset es pequeño y que desde Xeno-canto podía acceder a un volumen mayor, opté por generar un nuevo dataset, descargando las grabaciones realizadas en Europa para todas las especies y con una calidad de audio aceptable.

El objetivo se mantiene con respecto al de Kaggle, ver si es posible generar un modelo lo bastante preciso, que permita la identificación de las aves a partir de sus cantos y que pueda ser utilizado con fines de investigación y/o comerciales.

En el mundo de la investigación, y en lo relativo a las aves, uno de los métodos más efectivos y menos costosos para el censo de aves es por el canto, por lo que disponer de una aplicación fiable que permita identificar las aves ahorraría esfuerzo y costes. Del mismo modo, la ornitología como hobby es algo que está en auge, y un complemento ideal a todo el material de campo del ornitólogo, como prismáticos, telescopios, guías, etc., sería disponer de una aplicación que permita avanzar un poco más en la afición, y que les permita identificar las especies de otra forma. Descubrimos que ya existe alguna aplicación en el mercado que lo hace, como Birdgenie.

Como trabajos anteriores de referencia, de lo investigado en la web, del reto de Kaggle no había ningún trabajo publicado, sólo referencias a la conversión del audio en datos, y al uso de Keras.

Si que encontré un trabajo parecido, sobre la identificación de aves de todo el mundo, pero en lugar de a nivel de especie, a nivel de género: https://spark-in.me/post/bird-voice-recognition-five. Para otros propósitos, sobre todo para el reconocimiento de voz, sí que existen bastantes estudios y trabajos relacionados, aunque la forma de enfocarlos es un poco distinta, ya que están focalizados a saber qué dicen más que en identificar a quién lo dice, que es nuestro objetivo.

2. Descripción de los datos de entrada.

Para la obtención del dataset se realiza web scrapping de la web Xeno-canto, utilizando como criterios de filtrado:

- Grabaciones que pertenezcan a aves (no sonido ambiente)
- Grabaciones realizadas en el continente europeo.
- Grabaciones con una calidad de audio aceptable.
- Grabaciones cuya licencia permita su uso para nuestro objetivo.

Como resultado, se genera un fichero csv, ***Birdsongs_europe_C_20181220213936.csv*****,** con los siguientes campos:

Campo	Descripción
Common	Nombre común en inglés
Scientific	Nombre científico (puede incluir la subespecie)
Length	Duración de la grabación en formato hh:mm
Recordist	Persona que realizó la grabación
Date	Fecha de la grabación
Country	País donde se realiza la grabación
Location	Localización de la grabación
Type	Tipo de canto: llamada, canto, reclamo, etc..
ID	Identificador de la grabación
Class	Calidad del audio. 1.- Muy Alta, 2.- Alta, 0.- No definida
Seconds	Duración de la grabación en segundos
Name	Nombre científico de la especie normalizado

2.1. Dataset

Explorando y analizando el dataset, se realizan ciertos filtrados sobre este (eliminación de especies poco representadas y grabaciones demasiado largas), dando lugar al definitivo y que está guardado como ***Birdsongs_My_Birdsongs_Europe_20181230103204.csv***

Está formado por **17.370 grabaciones**, correspondientes a **103 especies** (*Birdsongs_My_Especies_Europe_20181230103204.csv* contiene la lista)

Campo	Descripción
Common	Nombre común en inglés
Scientific	Nombre científico (puede incluir la subespecie)
Length	Duración de la grabación en formato hh:mm
Recordist	Persona que realizó la grabación
Date	Fecha de la grabación
Country	País donde se realiza la grabación
Location	Localización de la grabación
Type	Tipo de canto: llamada, canto, reclamo, etc..
ID	Identificador de la grabación
Class	Calidad del audio. 1.- Muy Alta, 2.- Alta, 0.- No definida
Seconds	Duración de la grabación en segundos
Name	Nombre científico de la especie normalizado
nType	Tipo de canto normalizado: call, song, isong, icall, others

Incluye un nuevo campo “nType”, que normaliza el tipo de canto en 5 categorías, ya que el campo de tipo de canto es bastante heterogéneo:

Valor	Descripción
Call	Reclamo, llamada
Song	Canto
Isong	Incluye canto en el audio
Icall	Incluye llamada
Others	Otro tipo de canto no recogido en los anteriores

2.2. Audios

A partir del dataset, se descargan los ficheros de audio mediante descarga directa; existe un repositorio en la misma web con las grabaciones. Los audios se caracterizan porque:

- el nombre del fichero es el identificador de la grabación (campo ID).
- están en formato mp3.
- pueden estar grabados en mono o estéreo, y a diferentes frecuencias de muestreo. (sólo utilizaremos un canal para el estudio -mono-)

Están localizados en el repositorio ***/audio.*** Existe un directorio por cada una de las especies presentes en el dataset (en github, debido al volumen de datos que representa, sólo hemos subido 5 especies de ejemplo).

3. Metodología.

3.1. Software

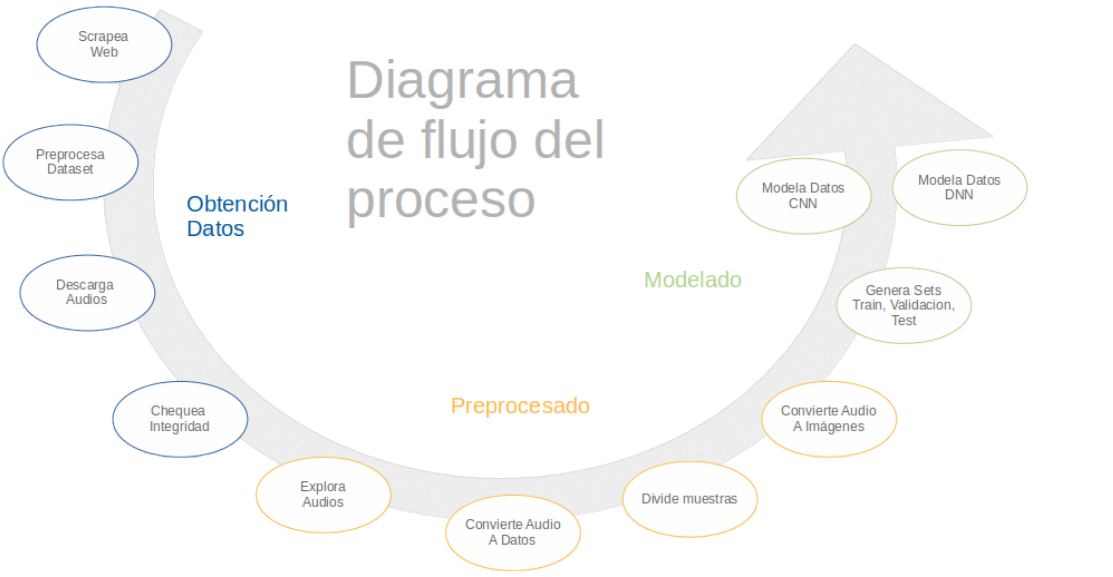
- Anaconda
- Python, versión 3.6.2
- Jupyter Notebook
- Keras sobre Tensorflow
- Scikit-learn
- Sistema operativo Ubuntu 18.04 LTS

3.2. Hardware

- Portátil personal Dell XPS13
- Máquina virtual en Gcloud con GPU y sistema operativo Ubuntu 18.04 LTS.

3.3. Descripción del Proceso

El proceso sigue el flujo normal de trabajo de un problema de ciencia de datos: obtención de los datos, pre-procesado, modelado y evaluación del modelo y resultados, siguiendo un proceso iterativo.



- **Obtención de datos.-** para la obtención del dataset utilizaremos la técnica de Web Scrapping, con las librerías Requests y Beautifulsoup. La descarga de los audios se realiza por descarga directa desde la web a partir del dataset generado.

- **Pre-procesado de datos.-** los ficheros de audio son convertidos a datos utilizando para ello las librerías LibRosa y sciPy. El audio es convertido a uno de estos dos tipos de datos:
    - Espectrogramas: imágenes que nos permiten utilizar algoritmos de redes neuronales convolucionales.
    - Periodogramas: matrices que nos permiten utilizar algoritmos de redes neuronales densas.

  Los audios se pueden tratar completos o dividiendo estos en segmentos más pequeños. Las fases de pre-procesado van generando directorios con la información, de forma que puedan ser reutilizados. A tener en consideración para provisionar el suficiente espacio en las máquinas físicas y/o virtuales.
- **Modelado.-** para el modelado de datos, utilizaremos algoritmos de Deep Learning, y dentro de estos:
    - CNN - Redes Neuronales Convolucionales 2D. Son bastante eficaces para tratamiento de imágenes y las utilizaremos para analizar los espectrogramas
    - DNN - Redes Neuronales Densas. Las utilizaremos para analizar los periodogramas.

3.4. Procesado del Dataset

3.4.1. Dataset

Analizando el dataset inicial generado a partir de la web, se ve que las especies no se encuentran suficientemente balanceadas en cuanto a:

- Número de muestras, el rango oscilaba entre 1 grabación a más de 1000.
- Duración de las grabaciones, algunas llegaban a durar más de 30 minutos.

Al no disponer de la posibilidad de obtener más grabaciones para tener un dataset más grande y mejor balanceado, procedemos a normalizar este, eliminando aquellas especies de las que teníamos un número de grabaciones bajo (inferior a 100) y aquellas grabaciones excesivas en duración (superiores a 60 segundos). El dataset de trabajo quedó en 17.370 grabaciones de 103 especies.

3.4.2. Tipos de cantos

Explorando los espectrogramas vemos que los patrones de los audios varían entre los tipos de cantos; las aves emiten distintos tipos de sonidos dependiendo de la funcionalidad. El patrón que se visualiza para un canto (apareamiento) difiere del que vemos para un reclamo o llamada (alerta, comunicación). Esto a simple vista dificultaría la identificación por parte del clasificador, ya que es como si le

estuviéramos diciendo que dos cosas distintas se identifican con la misma etiqueta. Para analizar este posible efecto, entrenamos el modelo de tres formas distintas:

- Dataset completo
- Dataset filtrando por cantos del tipo “Song” (canto)
- Dataset filtrando por cantos del tipo “Call” (llamada)

Al realizar los filtrados, el número de especies se reduce a 84 y 92 dependiendo del tipo. Verificamos que tras este filtrado el balanceo de las clases no se resiente mucho y eliminamos aquellas especies cuyo número de muestras sea inferior a 10.

3.4.3. Duración (Tamaño de las muestras)

Analizando otros estudios realizados y que tratan también con audios, se ve que la mayoría de ellos, por no decir todos, no tratan la grabación del audio de forma completa, sino que trocean la grabación en muestras más pequeñas. Es verdad que se refieren a estudios relacionados casi todos con el habla humana, cuyos audios no llegan a superar los pocos segundos y que llegan a trocear en ventanas de 100 milisegundos.

Para el TFM opté por asumir tres estrategias en cuanto a tamaño:

- Tratar la grabación completa.
- Generar tramas de dos segundos dejando un salto de otros dos segundos entre cada trama.
- Generar tramas de un segundo dejando un salto de otro segundo entre tramas.

El principal problema aquí es la posibilidad de generar tramas con silencios o con ruidos que en realidad no sirven para identificar a la especie. Además, al realizar los cortes al azar podemos realizar el corte en medio de un patrón y entorpecer al modelo. Si el tamaño del dataset no es suficientemente grande, nos podemos ver bastante influenciados por estas muestras.

Para generar tramas de un segundo entrené el modelo con 10 especies; por motivos de rendimiento y posibilidades de recursos, no tendría la capacidad de entrenar con todas las especies. Contra todo el conjunto de datos, utilizando todas las especies del dataset, el entrenamiento de las épocas daba tiempos medios de proceso de 3 horas por cada una de ellas.

3.5. Notebooks

Relación de notebooks con su lugar en el flujo del proceso.

Proceso	Notebook
Scrapea Web	Birdsongs_01_Obteniendo_Datos_Webscrapping_DataSet.ipynb
Pre-procesa Dataset	Birdsongs_02_Obteniendo_Datos_Seleccionando_Especies.ipynb
Descarga Audios	Birdsongs_03_Obteniendo_Datos_Descargando_Audios.ipynb
Chequea Integridad	Birdsongs_04_Obteniendo_Datos_Verificando_Integridad.ipynb
Explora Audios	Birdsongs_05_Preprocesando_Datos_Explorando_Audios.ipynb
Convierte Audio a Datos	Birdsongs_06_Preprocesando_Datos_Convirtiendo_Audio_a_Datos.ipynb
Divide Muestras	Birdsongs_07_Preprocesando_Datos_Cortando_Datos.ipynb
Convierte Audio a Imágenes	Birdsongs_08_Preprocesando_Datos_Convirtiendo_Datos_a_Imagenes.ipynb
Genera Sets Train, Validación, Test	Birdsongs_09_Modelando_Datos_Determinando_Sets_Train_Validation_Test.ipynb
Modela Datos CNN	Birdsongs_10_Modelando_Datos_Deep_Learning_CNN.ipynb
Modela Datos DNN	Birdsongs_11_Modelando_Datos_Deep_Learning_DNN.ipynb

4. Resumen del resultado.

En el repositorio **/results** se encuentran unos archivos en formato HTML con los notebooks de las ejecuciones de los modelos con los mejores resultados.

4.1. Resultados

Valores de accuracy obtenidos

		Duración de las grabaciones		
Tipo dato	Tipo canto	Completa	2 segundos	1 segundo
Espectrograma	All	45%	47%	**62%**
Espectrograma	Song	49%	48%	**58%**
Espectrograma	Call	36%	42%	**58%**
Periodograma	All	35%	Na	Na
Espectrograma Tratado	Call	Na	Na	53%

4.1.1. Accuracy

Los mejores resultados obtenidos dan un accuracy entre el 58% y 62% sobre test.

Comparando esta accuracy, respecto a lo que obtendríamos por azar, un 1% si consideramos 100 especies de media, o por la clase mayoritaria, cuyo porcentaje no supera nunca el 15%, los resultados obtenidos serían aceptables.

Realicé una prueba consistente en filtrar manualmente las muestras del set de train y validación del dataset de 10 especies con tramas de un segundo, eliminando todas aquellas imágenes que no mostraban nada, mostraban ruido o eran imágenes distorsionadas. El accuracy en validación se disparó hasta el 80%, pero luego no funcionaba muy bien en test (En la tabla, el resultado correspondiente a “Espectrograma Tratado”).

Dado los resultados obtenidos, no he llegado a analizar valores de “Precision” y “Recall”, pero serían importantes en el caso de realizar estudios sobre especies que estén en peligro de extinción y/o protegidas, y cuyas acciones a tomar en base a los censos puedan comprometer o no su viabilidad como especie.

4.1.2. Transformación de los datos

De las dos estrategias de tratamiento de los audios, generando espectrogramas y periodogramas, la opción de los periodogramas es la que da peores resultados, sobre un 35% de accuracy como máximo. Buscando información por internet no encontré ninguna referencia a estudios basados en los periodogramas.

Un periodograma trata la información de la señal en el dominio de la frecuencia. Enfrenta las frecuencias respecto a la potencia de la señal. Esto hace que se trate de forma global toda la grabación, por lo que puede incluir ruidos u otros sonidos que aparezcan en la grabación que no pertenezcan al ave. El modelo podría confundir un sonido de un ave con una grabación donde las frecuencias y potencias sean parecidas, no sería muy fiable.

Respecto a los espectrogramas, estos arrojan mejores resultados, hasta un 62%, lo que está avalado también por la bibliografía al respecto sobre su uso para generar clasificadores en base a grabaciones de audio de otro tipo. Se ven patrones claros, como las notas musicales del canto.

4.1.3. Tipos de cantos

Por los resultados obtenidos respecto a los tipos de cantos, existen variaciones entre separar el dataset según tipo y realizarlo de forma completa, pero tampoco pueden ser concluyentes; parece en primera instancia que daría igual tratar el dataset de forma completa que filtrando, pero creo que esto estaría más influenciado por el tamaño del dataset; habría que analizarlo con un número mayor de muestras.

4.1.4. Duración (tamaño de las muestras)

Parece que hay una tendencia positiva en cuanto a tratar las grabaciones dividiendo estas en tramos más pequeños. Se obtienen mejoras de hasta un 15% en algunos casos. Pero también podría estar afectado por el tamaño de las muestras, a medida que dividimos hacemos más grande el dataset y puede que esto sea lo que esté produciendo estas mejoras.

5. Conclusiones

Los resultados obtenidos no son del todo satisfactorios en cuanto a tener un modelo lo suficientemente preciso para implantarlo en una aplicación de uso comercial o con fines de investigación.

Aun así, los resultados son bastante positivos. Teniendo un dataset mucho más grande y profundizando más en el tratamiento del audio, de forma que fuera posible filtrar aquellas grabaciones que tengan mucho ruido, silencios, etc. se podrían conseguir resultados aceptables. Probé a realizar clustering de las grabaciones para alguna especie, para ver si era capaz de agrupar estas y que posteriormente pudiera descartar aquellas muestras que no correspondieran, pero el problema era luego saber con qué cluster quedarse.

Viendo que ya existe alguna aplicación comercial que lo hace, y que el tratamiento de audio para el habla está cada vez más avanzado, seguramente se resuelva pronto.

6. Agradecimientos.

Agradecimientos a:

- A todos los profesores que han dado clases en el Master.
- A los compañeros, por el buen rollo demostrado.
- A la web de Xeno-Canto, por compartir las grabaciones.