GLOBAL PROJECT

ALUMNO:

GABRIEL FERNANDO REY

PROGRAMA:

POSTGRADO EN INTELIGENCIA ARTIFICIAL Y MACHINE LEARNING

Inteligencia Artificial y Machine Learning (7ª promoción)

Año 2022

NOMBRE DEL PROYECTO:

DETECCION DE PAJAROS A PARTIR DEL AUDIO PARA EL CONTROL POBLACIONAL

ENTREGA:

Marzo 2023

Contenido

RESUMEN 3

INTRODUCCIÓN 4

Problemática 4

Situación actual 5

Planteo 5

Estrategia tecnológica 5

Resultados obtenidos 5

ESTADO DEL ARTE 6

OBJETIVOS 9

Objetivo general 9

Objetivos específicos 9

SOLUCIÓN PLANTEADA 9

Clasificación de audio 9

Conceptos básicos del sonido 10

Transformada de Fourier 11

Espectrogramas 14

Desarrollo del modelo 15

- Los datos 15
- De audio a espectrogramas 17
- Modelo Convolucional 18
- Aplicación 20

EVALUACIÓN y RESULTADOS 20

CONCLUSIONES Y TRABAJOS FUTUROS 22

REFERENCIAS 23

RESUMEN

El impacto que sufre el medio ambiente dado el maltrato proporcionado por el hombre es uno de los desafíos más importantes de la era contemporánea. No solo el hombre padece sus efectos, sino también todo ser vivo sobre la tierra. Poder contribuir para contrarrestar esta situación aun desde un ensayo microscópico y utilizando la inteligencia artificial es la inspiración de este proyecto, teniendo como objetivo disponer de una herramienta que ayude al control poblacional para la conservación de las especies, en particular los pájaros.

En este sentido, se propone el desarrollo de un modelo basado en Deep learning que permita la detección de pájaros a partir del audio de su canto, cuyo propósito se basa en monitorear las especies de una región determinada.

El alcance del proyecto comprende la investigación de las técnicas para lograr el modelo, utilizando redes convolucionales.

El objetivo es la experimentación y su modelado, no desarrollando la aplicación final de usuario que debería contener al modelo en cuestión. Sin embargo, la idea del proyecto comprende que la aplicación final podría estar orientada a lo colaborativo o a la medición en campo. Es decir, una aplicación para dispositivos móviles de uso libre y de contribución colaborativa, en la que el usuario aporta la captura de los datos y la ubicación de la detección en la que se encuentre. La otra aplicación para la medición en campo está orientada a la realización de estudios para el control poblacional que pudieran realizar entidades gubernamentales u organizaciones privadas para la conservación de las especies de pájaros.

Así como existen monitores del sonido ambiental (Sounds of New York City [1] entre otros) utilizando microcontroladores con algoritmos para la clasificación de sonido ambiental, de la misma manera puede establecerse esta tecnología para el caso expuesto de medición en campo.

Los resultados obtenidos son satisfactorios dentro del marco esperado, se demuestra que es posible desarrollar una aplicación para tal fin, aunque para lograr obtener mejores resultados se necesita una mayor inversión de tiempo e investigación para lograr un resultado más fino. La detección de especies está acotada a 11 clases de pájaros del ámbito de la provincia de Buenos Aires, Argentina.

INTRODUCCIÓN

Problemática

Como mencionamos una gran amenaza es el **cambio climático**, las actividades humanas están generando cambios en el clima del planeta. Las aves como indicadoras de la salud de los ambientes nos permiten conocer los impactos del cambio climático de manera rápida, reflejando cómo nuestras actividades impactan negativamente en la biodiversidad.

Dentro de la problemática de la conservación de las especies, podemos encontrar que existen también otras amenazas como el tráfico ilegal de pájaros. Un informe realizado entre el 2020 y 2021 por la asociación avesargentinas.org.ar [2], revela que el tráfico ilegal de aves silvestres en Argentina se ha incrementado en más de un 500% comparado con otro informe realizado entre 2014 y 2015, una cifra alarmante. De ese incremento, el 60% corresponde a la provincia de Buenos Aires, siendo la provincia de mayor porcentaje de tráfico respecto de las demás provincias del interior del país.

Dentro de las especies más traficadas en la provincia de Buenos Aires se encontraron: el jilguero dorado (*Sicalis flaveola*), el cabecitanegra (*Spinus magellanicus*), el cardenal común (*Paroaria coronata*) entre otras especies. Tomaremos estas tres primeras clases de pájaros como parte de nuestro grupo de clases para nuestro proyecto.

Jilguero dorado

Cabecitanegra

Cardenal común

También la especie cardenal amarillo (*Gubermatrix cristata*) será incluida dentro del grupo de clases del proyecto, ya que se encuentra catalogado como en peligro crítico de extinción según la información actualmente publicada en la web por el Ministerio de Ambiente y Desarrollo Sostenible de la República Argentina [3].

El avance de **especies exóticas** por causas humanas a lo largo del planeta está creando una nueva amenaza para la conservación de las especies y los ambientes naturales. Una vez que las especies exóticas se establecen en nuevos ecosistemas, pueden llevar a las especies nativas a la extinción. La extinción es un proceso irreversible y en este caso, la introducción de especies exóticas es considerada como la segunda causa antrópica más importante en la pérdida de diversidad global.

Las especies mencionadas son representativas de ciertas amenazas, incluiremos ocho especies más, totalizando 11 clases para la realización de este proyecto.

Situación actual

En Argentina existe un programa denominado AICA (Áreas Importantes para la Conservación de las Aves), con el fin de identificar y proteger sitios de particular importancia que han sido reconocidos por BirdLife, entre otros. Hoy en día hay más de 10,000 AICAs reconocidas en el mundo, y la manera que cualquier persona interesada puede contribuir es completando un formulario con distintos tipos de datos que no solamente está relacionado a las especies de las aves sino también a ciertos datos que tienen que ver con lo ambiental, como datos de agricultura, desarrollos residenciales, comerciales, mineros, modificaciones del sistema natural, etc.

Este sistema pretende recopilar mucha información, pero al ser basado en el llenado manual de formularios no alienta a la contribución de una masa importante de datos para su posterior análisis.

Planteo

En función de las amenazas planteadas la búsqueda ágil de la información es fundamental para la aplicación de políticas y tomas de decisiones. Por tal motivo un sistema que permita a cualquier usuario identificar a las aves utilizando la tecnología de manera simple induce a una mayor colaboración, independientemente de su nivel de conocimiento sobre las aves. Por lo tanto, con tecnología basada en IA, a través de su canto esta información junto con la geolocalización de su captura pueda ser enviada a un repositorio central, lo que implicaría aportar un mayor caudal de información que permita un monitoreo de las aves más eficiente.

Estrategia tecnológica

Las redes neuronales convolucionales están orientadas a resolver problemas del campo de la visión por computador, obteniendo patrones de las imágenes para poder realizar tareas de clasificación, detección de objetos, etc. Al utilizar como fuente de dato el audio y no una imagen se pretende convertir el audio en imagen para luego poder identificar patrones que permitan la detección del ave con relación a su canto.

Resultados obtenidos

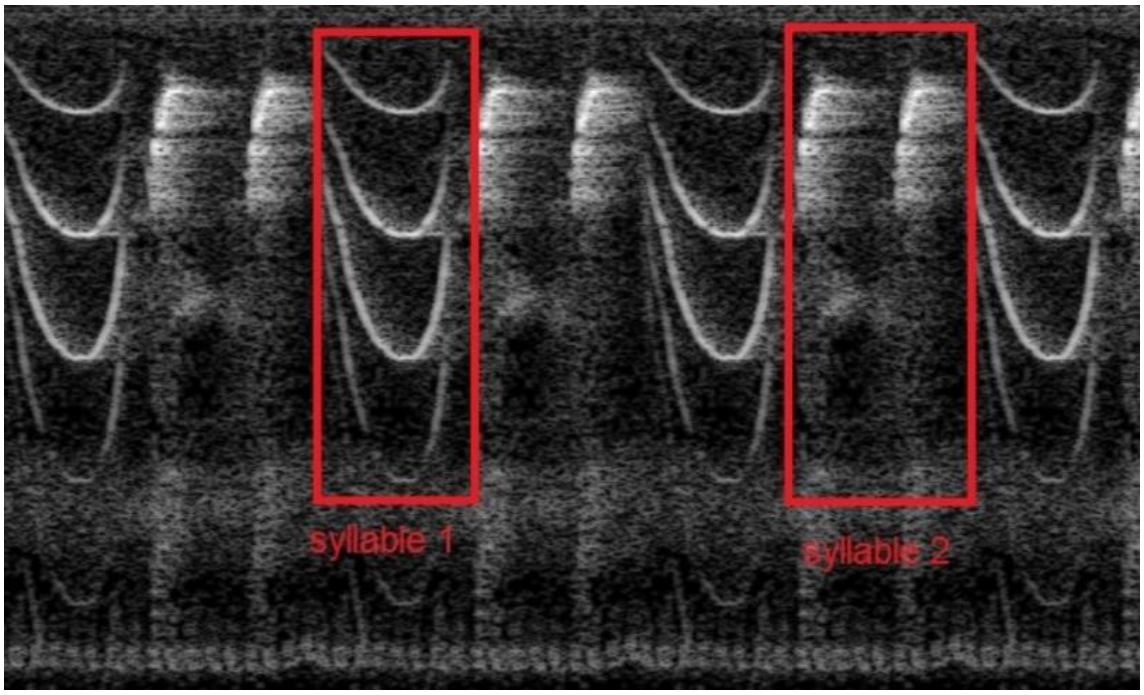
Los resultados obtenidos a nivel de detección de aves demuestran la viabilidad de la implementación de soluciones basadas en esta tecnología.

ESTADO DEL ARTE

Se enumeran algunos artículos investigados, los cuales se han basado en redes neuronales convolucionales y que han sido la base de las técnicas que justifican lo empleado para el desarrollo de este proyecto.

Reconocimiento de llamadas de aves utilizando una red neuronal convolucional profunda, ResNet-50 [4]

Este trabajo desarrollado en 2018 por la Universidad James Cook, Queensland, Australia utilizó un conjunto de datos disponible públicamente que consiste en el audio de llamadas de 46 especies de aves diferentes. Los espectrogramas extraídos de los cantos de los pájaros se totalizaron en 2814, se utilizaron para entrenar en una red neuronal convolucional ResNet-50 el 75% de los datos y el resto para validación, logrando una precisión del 60% al 72% en el reconocimiento del canto del pájaro emisor. Según el documento tomó aproximadamente 100 horas entrenar el modelo en una GPU Nvidia GTX 1070. Una técnica interesante que utilizaron fue el uso de espectrogramas basado en sílabas, las sílabas pueden verse como unidades básicas de reconocimiento.



Redes neuronales convolucionales para la clasificación de sonidos de Búhos.

Este documento describe un sistema automatizado de detección de sonidos de aves desarrollado Alam Ahmad Hidayat et al. de Bina Nusantara University, Jakarta, Indonesia. Específicamente, 7 clases de búhos de Indonesia. Se experimentó con dos modelos, el primero utilizando una red neuronal convolucional simple de cuatro capas convolucionales más dos densas y el segundo utilizando dos redes idénticas al modelo

anterior pero cada una con entradas diferentes. Es decir, para clasificar diferentes especies en función de sus sonidos vocales se extraen dos representaciones comunes de la señal acústica, el espectrograma de Mel con escala logarítmica y el coeficiente cepstral de frecuencia mel (MFCC). El primer modelo utilizó como entrada a los espectrogramas, y el segundo modelo utilizó una entrada con los espectrogramas y la otra entrada con los filtros MFCC. El modelo de doble entrada (figura 2) es el que mejor rendimiento obtuvo en el experimento, alcanzando una precisión media (MAP) del 97,55%. Mientras tanto, el modelo basado solo en espectrogramas obtuvo un MAP del 94,36% y probando solo con los filtros MFCC se obtuvo un MAP de 96,08%

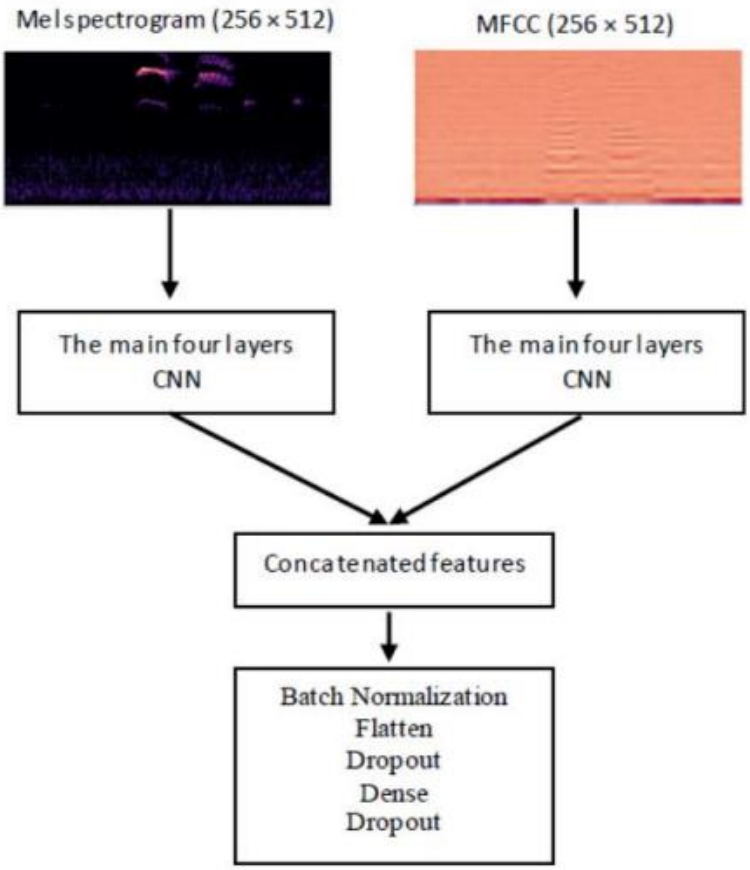


Fig. 2. La red neuronal propuesta que toma dos características acústicas como entrada.

Arquitecturas utilizadas para la clasificación y detección de audio

Dentro de las arquitecturas utilizadas que aplican al caso de estudio podemos citar el trabajo de Hinton et al. (2012) [5] que empleó redes neuronales convolucionales en una dimensión para el reconocimiento de voz. Otro estudio de Lee et al. (2009) [6] examinó una red neuronal convolucional para tareas de voz y música. En relación con la música también el trabajo de Van der Oord et al. (2013) [7] se centró en la recomendación automática de música según los gustos del usuario como podemos disfrutar en aplicaciones como Spotify, Apple Music, Amazon Music, Google Play Music.

La evolución natural de estas redes son las redes neuronales convolucionales en dos dimensiones. Su uso es bastante estandarizado, ya que no solo se puede emplear para resolver nuestras tareas de audio, sino que originalmente se desarrollaron para solventar problemas de clasificación de imágenes. La entrada a este tipo de redes tendría que ser una representación en el tiempo y en la frecuencia del audio, que es lo que fue utilizado para este proyecto. Humphrey et al. (2012) [8] emplearon esta arquitectura para el reconocimiento automático de acordes musicales y Schlüter et al. (2015) [9] trabajaron en el problema de detección de voz cantada.

También se ha experimentado con redes neuronales recurrentes y arquitecturas híbridas como por ejemplo que la que se ha utilizado en este proyecto que combina una red neuronal convolucional con una red densamente conectada.

OBJETIVOS

Objetivo general

Desarrollar un modelo basado en inteligencia artificial que permita mediante la detección de audio identificar la clase de pájaro de acuerdo con su canto de entre 11 clases previstas para este proyecto. Para lograrlo se pretende utilizar los conocimientos adquiridos basado en redes neuronales convolucionales.

Objetivos específicos

1. Determinar el grado de complejidad base para lograr el objetivo general
2. Como lograr resultados razonables con una mínima cantidad de muestras
3. Establecer que técnicas son necesarias para adecuar los datos de entrenamiento
4. Investigar que técnicas son necesarias para el desarrollo del objetivo general

SOLUCIÓN PLANTEADA

Clasificación de audio

La clasificación de audio es una tarea de machine learning donde el objetivo es entrenar un modelo para que pueda predecir a que clase pertenece.

Pero en un sentido más amplio, el tratamiento del audio con técnicas de inteligencia artificial nos permite resolver problemas que se relacionan con:

- Detección de audio (Sound event detection)
- Clasificación de audio (Sound Scene Detection)
- Etiquetado de audio (Audio Tagging)

La clasificación automática del audio es un área de investigación en crecimiento con numerosas aplicaciones en el mundo real. La técnica que permite la clasificación de audio se basa en la misma técnica de la clasificación de imágenes, utilizando redes convolucionales. Para esto habrá que representar el audio como una imagen, pero antes repasemos algunos conceptos.

Conceptos básicos del sonido

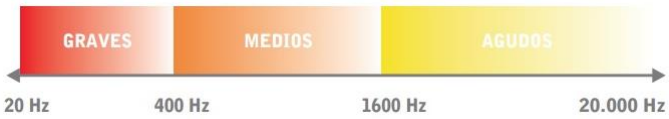
El sonido está producido por pequeñas variaciones de presión en un medio, habitualmente el aire.

Al hablar provocamos un movimiento de las partículas de aire alrededor de nuestra boca. El movimiento de estas partículas, causa pequeñas variaciones sobre el valor de la presión atmosférica, que son detectadas por nuestro oído.

Un sonido se caracteriza por dos propiedades: la **amplitud** y la **frecuencia.**

La **frecuencia** es la cantidad de oscilaciones por unidad de tiempo, normalmente medido en 1 segundo, que realiza una onda en movimiento. La forma de estas oscilaciones ayuda a determinar si los sonidos son más graves o agudos. A mayor frecuencia el tono del sonido es más agudo, mientras que a menor frecuencia el tono del sonido es más grave. La frecuencia se mide en hercios con notación Hz, 1 Hz equivale a un ciclo de compresión y descompresión de onda por segundo.

1000 Hz equivalen a 1khz. El oído humano es capaz de percibir sonidos entre los 20 hz y los 20.000 hz o 20Khz.



La amplitud se refiere a la altura de la onda y hace referencia a la intensidad o el volumen del sonido, amplitud cero equivale al silencio, amplitud media a sonidos leves y amplitud grande a sonidos altos.

La amplitud se mide habitualmente en decibeles con notación dB. La escala auditiva va desde un rango entre 0 dB y 130 dB, por ejemplo, los sonidos por encima de 110 dB producen una sensación dolorosa y la exposición por largos periodos de tiempos a esos niveles puede reducir la capacidad auditiva.

Se debe tener en cuenta que un aumento de 10 dB en el nivel de un sonido equivale a percibir este sonido el doble de intenso.

Está claro que el sonido es una señal analógica que debe digitalizarse para poder ser tratado por ordenador. En la conversión **analógica-digital** se determina la frecuencia de muestreo, las amplitudes de las frecuencias y el tiempo.

Tras esta conversión analógica-digital ya tenemos nuestro audio almacenado en el ordenador y podemos trabajar con él. El siguiente paso para la obtención de una representación tiempo-frecuencia es convertir la señal de audio en el dominio de la frecuencia.

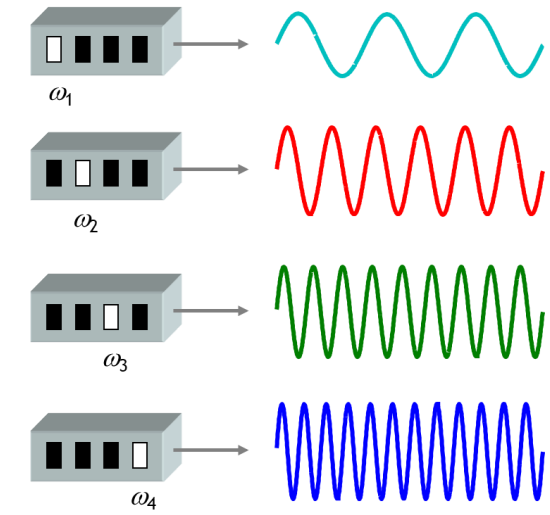
Transformada de Fourier

La transformada de Fourier es una técnica que permite descomponer una señal de audio en las distintas frecuencias que componen esa señal, y también permite el proceso inverso.

Para la clasificación de audio vamos a necesitar la **Transformada corta de Fourier (STFT).**

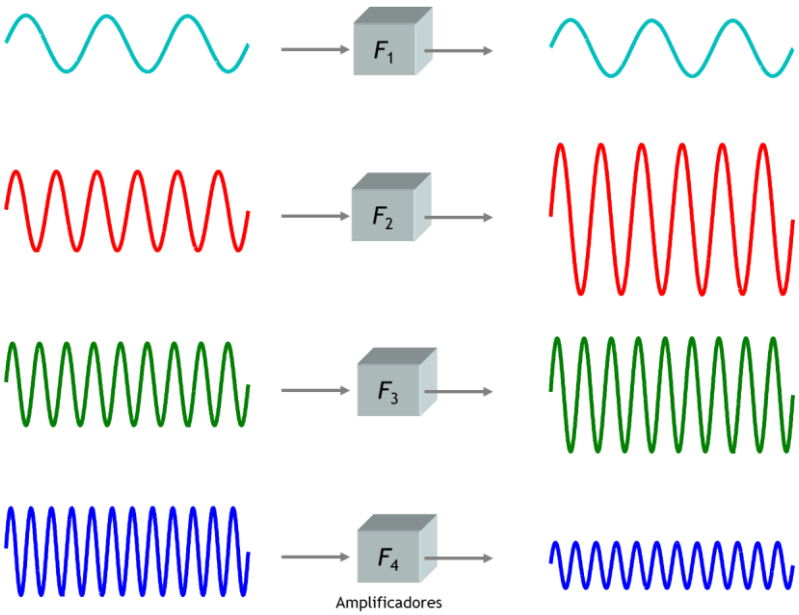
La transformada de Fourier descompone una señal de audio, pero pierde toda la información de tiempo. En comparación, **STFT** divide la señal en ventanas de tiempo y ejecuta una transformada de Fourier en cada ventana, preservando parte de la información de tiempo y devolviendo un tensor 2D en el que puede ejecutar convoluciones estándar.

Para entender la Transformada de Fourier vamos a partir de distintas frecuencias y combinarlas en una señal de audio. Cada señal se origina con la misma amplitud, pero distinta frecuencia.



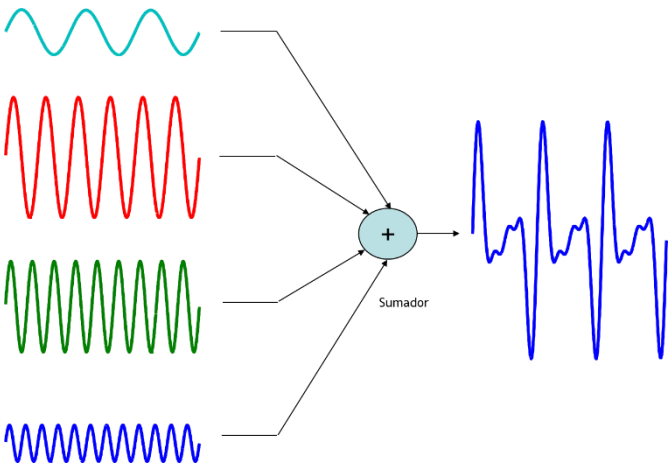
Osciladores sinusoidales

Ahora, vamos a modificar algunas amplitudes.



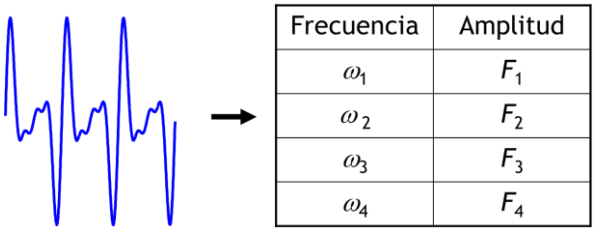
Amplificadores

Y ahora vamos a sumar todas las señales componiendo un audio.



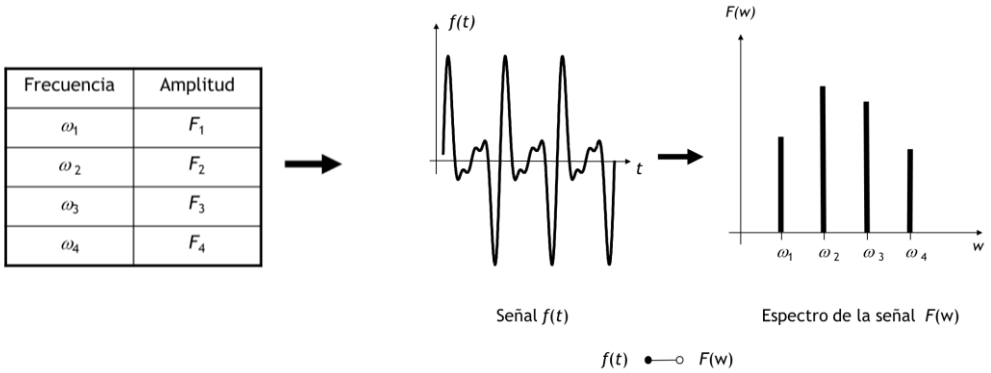
Sumador

Como podemos ver, el audio está compuesto por señales de distinta frecuencia y amplitud.



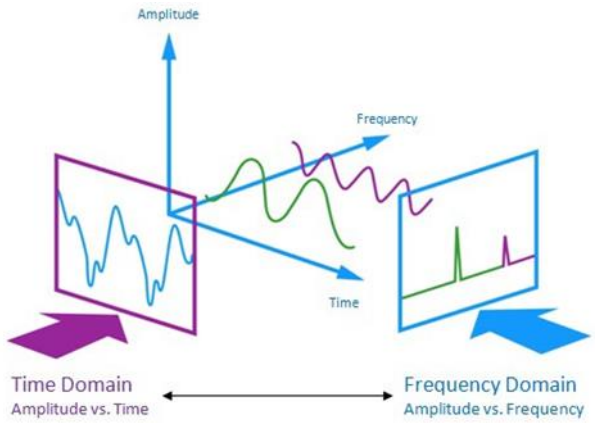
Frecuencia	Amplitud
ω_1	F_1
ω_2	F_2
ω_3	F_3
ω_4	F_4

Podemos entonces graficar este audio en función de la frecuencia y amplitud.



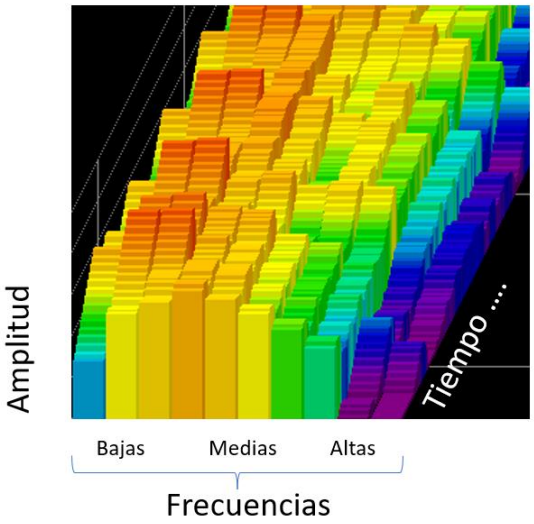
¿Puedo utilizar estos gráficos para entrenar la red?

Todavía no hay suficiente información, porque además necesitamos representar a nuestro audio con 3 magnitudes, frecuencia, amplitud y tiempo.



Supongamos tener 10 frecuencias y además agreguemos más información que el gráfico anterior representando el transcurso del tiempo a la frecuencia y la amplitud.

Tenemos 3 magnitudes, debemos graficar entonces en 3D.



Aquí, ya estamos aplicando la Transformada corta de Fourier (STFT), estamos graficando frecuencia y amplitud a intervalos de tiempo.

Cada barra representa a una frecuencia en un instante de tiempo dado. Podemos contabilizar 10 frecuencias, desde las bajas (a izquierda) hasta las más altas (las violetas a derecha). A medida que transcurre el tiempo se van encolumnando nuevas barras cuya altura dependerá de la amplitud de la frecuencia en cada instante de tiempo.

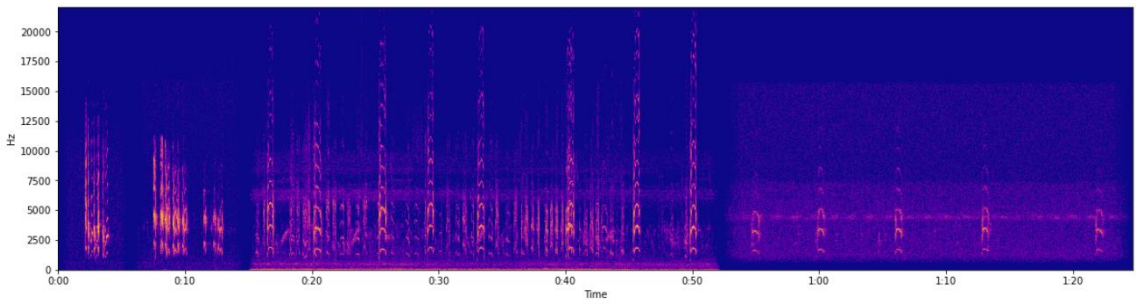
En el primer intervalo de tiempo se pueden observar las primeras 10 barras de frecuencias y sus amplitudes, desde la azul, pasando por las amarillas, naranjas, verdes y violetas.

Si bien agregamos información al gráfico, todavía no es el adecuado como para extraer bien todas las características del audio, podemos ver que hay información de amplitud detrás de las primeras barras que no son interpretables.

Existe otra forma de visualizar la STFT en un gráfico 2D, en donde los ejes representan el tiempo y las frecuencias, y la amplitud se visualiza con los distintos tonos de color, a estos gráficos se los denomina espectrogramas.

Espectrogramas

Los espectrogramas son una técnica útil para visualizar el espectro de frecuencias de un sonido y cómo varían durante un período de tiempo corto.

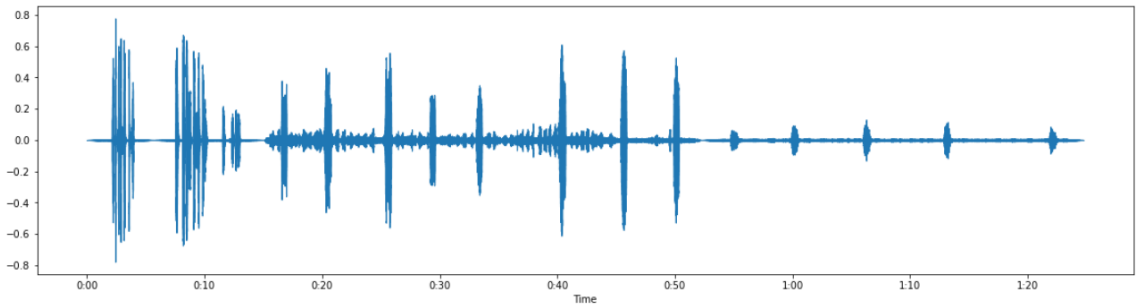


Puede observarse que los tonos que resaltan representan las amplitudes para cada frecuencia, en el gráfico anterior el eje ‘x’ representa el tiempo y en el eje ‘y’ se encuentran las frecuencias que van desde 0 a 20.000hz.

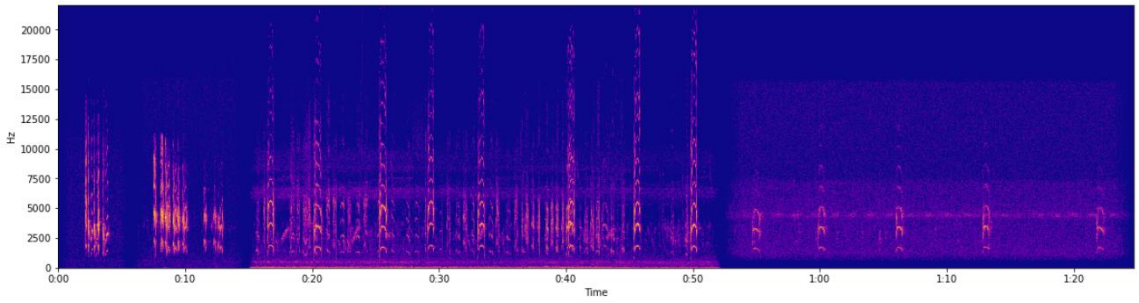
Entonces, podemos tratar a un audio como una imagen y aplicar una red convolucional para entrenarla, en donde tenemos estas similitudes:

- Espectrograma = imagen
- Tiempo, frecuencia = x, y
- Amplitud = valor de un pixel

El siguiente gráfico visualiza la señal del audio original en función del tiempo.



Este otro gráfico visualiza las amplitudes de las frecuencias del audio en función del tiempo



Desarrollo del modelo

Los datos

Una de las primeras cosas a considerar en todo proyecto de IA es saber con qué datos contamos para resolver nuestra tarea, es decir cuál es el volumen disponible, calidad, como los vamos a tratar, etc.

Los modelos basados en la clasificación de audio tienen un gran desafío en verificar que tantos ejemplos se tienen de audio, siendo tal vez más complejo la obtención de éstos en comparación con otros tipos de datos, como estructurados, no estructurados o inclusive imágenes.

El desafío estaba planteado y como recurso se ha conseguido adquirir audios con producción profesional de ornitólogos especializados en vocalizaciones de la fundación Audiornis de Argentina [10]. Aquí, los audios grabados disponibles se ofrecen en formato MP3, cada uno representando al canto de cada ave de la región de Argentina.

En este sentido, solo se obtuvieron muestras de audio en donde por ejemplo en una muestra puede mezclarse el sonido del canto del pájaro con el sonido de llamada de alerta del mismo pájaro. Por supuesto no se ignora que hay varias categorías de los sonidos vocales de los pájaros como por ejemplo la llamada de apareamiento, llamada de vuelo, llamada de alarma, llamada infantil y canto de pájaro, que ha quedo excluido del alcance de este proyecto.

Por lo tanto, nos encontramos con solo una muestra de audio por clase de pájaro y entonces surgió el primer interrogante:

¿Cuántos ejemplos distintos de audio serán necesarios por clase de pájaro para que funcione en un modelo de IA?

Sin duda la respuesta es la mayor cantidad posible. Pero no hay tal disponibilidad de recursos, por lo tanto, se experimentó simulando tenerlos a partir de una única muestra, siendo la única forma posible encontrada hasta el momento en la que se pudo avanzar.

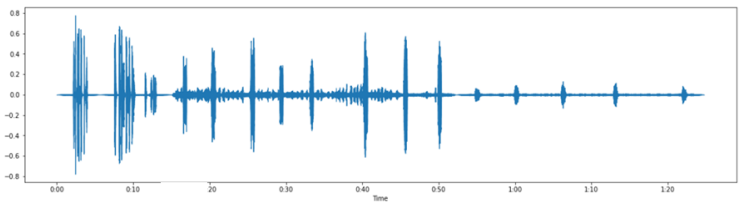
Podríamos decir que se trata de algo similar a la técnica de aumentación de datos, técnica que se utiliza para generar datos sintéticos a partir de datos reales y así disponer de mayor cantidad de ejemplos para entrenamiento.

En este caso, no se generaron datos sintéticos, sino que se ha utilizado una segmentación de cada audio original para obtener múltiples ejemplos de menor duración.

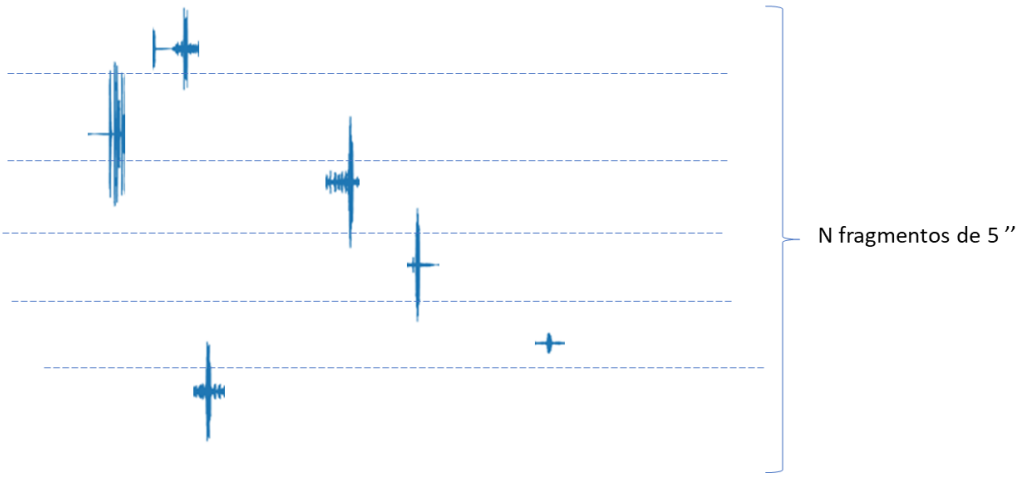
En primer lugar, fue necesario convertir de MP3 a WAV.

En segundo lugar, se realizó el proceso de fragmentación del audio. Para esto se estableció que la duración de cada fragmento sea de 5 segundos, basado en la experimentación y en resultados de investigaciones anteriores, teniendo la idea de que este tiempo es adecuado, pero no el único para la detección de patrones.

Ejemplo de segmentación de un audio original:



Audio original 1’ 20 ’’



De esta forma se estableció que para cada clase se generaran 1000 fragmentos de audio de longitud fija como muestras para entrenar al modelo. Los fragmentos de audio podrían haber resultado como divisiones consecutivas, pero aquí nos enfrentábamos a diferentes problemas:

- Todos los audios son de distinta duración
- La duración mínima de los audios es de 33 segundos
- La duración máxima de los audios es de 151 segundos.

Al dividir en partes iguales no se hubiera logrado la misma cantidad de muestras por audio, por lo que se utilizó la extracción de muestras solapadas y así obtener la misma cantidad de muestras para cada audio, lo que permite hacer un mejor uso de un conjunto de datos limitado.

Pero como se expuso, al modelo lo vamos a entrenar con imágenes, cuya transformación de audio a imagen se explica a continuación.

De audio a espectrogramas

Las 11 clases de pájaros elegidas para este proyecto se pueden conocer mediante el nombre del archivo de audio con extensión WAV que forma parte del material de este proyecto:

BENTEVEO_COMÚN.wav,

CABECITANEGRA_COMÚN.wav,

CALANDRIA_GRANDE.wav,

CARDENAL_AMARILLO.wav,

CARDENAL_COMUN.wav,

COTORRA.wav,

GOLONDRINA_DOMÉSTICA.wav,

GORRIÓN.wav,

HORNERO.wav,

JILGUERO_DORADO.wav,

ZORZAL_COLORADO.wav

En los primeros ensayos los audios no se clasificaban bien. Al activar el micrófono del computador esperando el sonido de algún pájaro, el ruido ambiente previo generaba clasificaciones erróneas dado que el modelo no podía identificar un patrón que desconocía. Por tal motivo fue necesario incorporar al entrenamiento una nueva clase "ruido" y con esto se resolvió en parte el problema. Por lo tanto, se generó un archivo wav de audio con ruido.

Por lo tanto, se utilizaron 12 clases incluyendo: RUIDO.WAV

Como se mencionó, cada audio fue particionado en 1000 muestras de 5 segundos y convertidos a espectrogramas. El comienzo desde donde comenzar a tomar cada muestra es aleatorio y se tiene en cuenta que cada segmento sea estrictamente de 5 segundos.

Ejemplo de espectrogramas del canto del Benteveo común.

Cada uno de estos espectrogramas es una imagen de 128x128 píxeles. Podemos observar gráficamente como existe similitud de patrones visuales en cada una de las imágenes. Esto es lo que será interpretado por la red convolucional que aprenderá a identificar estos patrones y saber de qué pájaro proviene el audio.

Podemos observar que dentro de los 5 segundos la información relevante aparece en distintos sectores e inclusive puede que contengan un silencio que ocurre entre dos cantos de un mismo pájaro sin embargo la etiqueta es la misma, a estos datos de los denomina débilmente etiquetados.

Una vez logrado obtener todos los espectrogramas de todas las muestras se preparó la arquitectura del modelo de red convolucional.

Modelo Convolucional

Se utilizó una red neuronal convolucional, utilizando Tensorflow en Python, sin recurrir a ninguna de las redes convolucionales conocidas como por ejemplo MobileNet [11], etc., al contrario, se buscó probar una arquitectura más simple.

De las 12000 imágenes generadas entre todas las clases, se utilizó el 80% para entrenamiento y el 20% para validación.

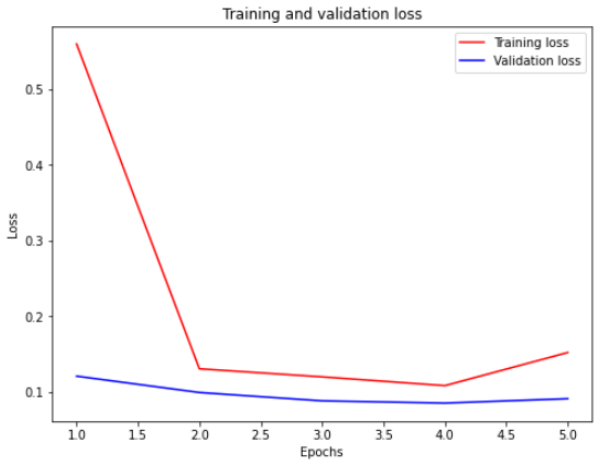
La red está compuesta por cuatro capas convolucionales y dos capas completamente conectadas, utilizando una softmax para obtener la distribución de probabilidades a la salida de la red.

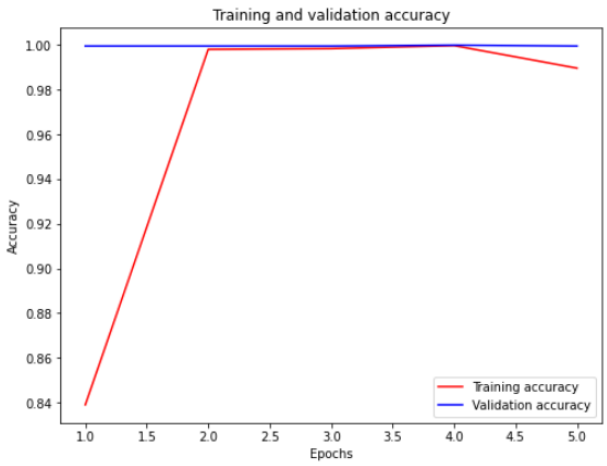
```
model = Sequential([
  layers.Rescaling(1./255, input_shape=(img_height, img_width, 3)),
  layers.Conv2D(16, 3, padding='same', activation='relu'),
  layers.BatchNormalization(),
  layers.MaxPooling2D(),

  layers.Conv2D(32, 3, padding='same', activation='relu'),
  layers.MaxPooling2D(),

  layers.Conv2D(64, 3, padding='same', activation='relu'),
  layers.MaxPooling2D(),

  layers.Conv2D(128, 3, padding='same', activation='relu'),
  layers.MaxPooling2D(),

  layers.Flatten(),
  layers.Dense(256, activation='relu'),
  layers.Dropout(0.5),

  layers.Dense(256, activation='relu'),
  layers.Dropout(0.5),

  layers.Dense(num_classes, activation='softmax')
])
```

El entrenamiento se estableció en 5 epochs, resultando satisfactorio el bajo overfitting y buen rendimiento de accuracy.

```
Epoch 1/5
600/600 [==============================] - 234s 388ms/step - loss: 0.5596 - accuracy: 0.8390 - val_loss: 0.1207 - val_accuracy: 0.9996
Epoch 2/5
600/600 [==============================] - 249s 416ms/step - loss: 0.1305 - accuracy: 0.9981 - val_loss: 0.0991 - val_accuracy: 0.9996
Epoch 3/5
600/600 [==============================] - 236s 393ms/step - loss: 0.1198 - accuracy: 0.9984 - val_loss: 0.0881 - val_accuracy: 0.9996
Epoch 4/5
600/600 [==============================] - 239s 398ms/step - loss: 0.1082 - accuracy: 0.9998 - val_loss: 0.0851 - val_accuracy: 1.0000
Epoch 5/5
600/600 [==============================] - 238s 396ms/step - loss: 0.1519 - accuracy: 0.9897 - val_loss: 0.0909 - val_accuracy: 0.9996
```

Se realizó una comprobación del rendimiento y obtuvimos los siguientes gráficos:





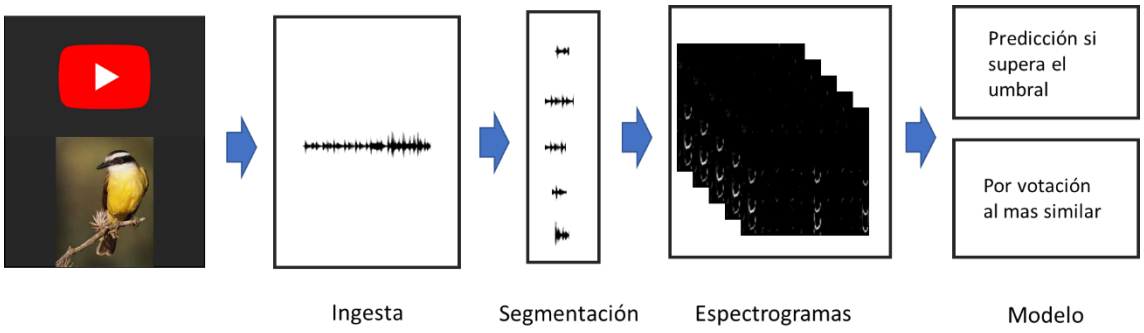
Aplicación

Para poder tener un ambiente controlado para este proyecto y simplificar la demostración de funcionamiento del modelo, en lugar de utilizar el micrófono del computador para la ingesta de audio, se ha utilizado una función que permite elegir una *url* de youtube con el canto de algún pájaro que se encuentre dentro de las aves comprendidas para este modelo.

El audio seleccionado es procesado de manera similar a la fase de preentrenamiento, se obtendrán en este caso 10 muestras de 5 segundos cada una. Cada muestra será un dato de entrada para obtener la predicción y luego habrá una función que devuelva una predicción unificada.

Esto proceso se debe a que como el proceso de segmentación comienza de manera aleatoria puede haber segmentos de audio de silencio, que son lo que por azar quedaron entre medio de dos cantos del mismo pájaro y no sea útil para la detección. Por esta razón se determinó obtener 10 muestras. El número de muestras podría incrementarse lo que ayudaría a clasificar mejor, pero con mayor costo computacional.

La función llamada “predicciones” entonces toma como parámetro la url del video elegido y extrae el audio, obtiene las 10 muestras, las convierte en espectrogramas y las somete a predicción del modelo. El umbral indicado implica que tan seguro quiero que sea el modelo antes de clasificar. Por ejemplo, si el umbral es 0.9 sólo las predicciones que superen esta probabilidad serán clasificadas.



Además, en caso de que no se lograra superar el umbral de certidumbre se añadió la función de votación al más similar. Esto implica que cada una de las 10 muestras ha obtenido una probabilidad de clase, y en caso de mayoría se sugerirá esta clasificación aun estando por debajo del umbral.

EVALUACIÓN y RESULTADOS

Una vez que se realizó el entrenamiento y sobre la base de resultados aceptables se efectuaron otros entrenamientos para su evaluación. Como criterio de evaluación en primer lugar se ha observado la validación de aciertos y la validación de pérdida variando la duración de los espectrogramas.

				model.evaluate()	
	Espectrograma (segundos)	Cantidad de Espectrogramas	Epochs	Val_Loss	Val_accuracy
Modelo_1	5	1000	5	0,085569	1
Modelo_2	10	1000	5	0,080023	1

La variación del largo del espectrograma no ha significado cambios significativos en los modelos y en consecuencia se ha optado por el valor de 5 segundos para mejorar el costo computacional.

Por otra parte, la evaluación del acierto del modelo con audios no vistos se realizó con la extracción de audio de videos de pájaros tomados al azar de youtube, repitiendo el proceso predictivo 10 veces con el mismo audio.

Nota: Cada repetición toma 15 muestras aleatorias del mismo audio para obtener la predicción.

url	Etiqueta	Aciertos	No aciertos
https://www.youtube.com/watch?v=D_OvrN0ekao&t=228s	Zorzal colorado	10	0
https://www.youtube.com/watch?v=bRPgMuPJeY4&t=331s	Benteveo comun	9	1
https://www.youtube.com/watch?v=R2rW363ydE8&t=29s	Cabecita negra	10	0
https://www.youtube.com/watch?v=GRNA_GFeZIs	Calandria Grande	10	0
https://www.youtube.com/watch?v=f0S6ANb7W1Q&t=1325s	Cardenal amarillo	7	3
https://www.youtube.com/watch?v=cO0bsV8vePg	Cardenal comun	10	0
https://www.youtube.com/watch?v=oibxurOe1zE	Cotorra	10	0
https://www.youtube.com/watch?v=zsvapbpmwNY&t=110s	Golondrina domestica	2	8
https://www.youtube.com/watch?v=TPuG-FOa1yw&t=971s	Gorrion	5	5
https://www.youtube.com/watch?v=6hZtOsnc7SE	Hornero	3	7
https://www.youtube.com/watch?v=kviBd6K-eF4&t=438s	Jilguero dorado	10	0

En general los resultados han sido favorables, obteniendo aciertos para 8 de 11 clases, y solo 3 no favorables.

Para los casos no favorables existen varios factores para tener en cuenta:

1. El audio original no es lo suficientemente representativo del canto del pájaro, esto implica que tal vez ciertos llamados o variaciones simplemente no se encuentran en el audio original.

2. El audio original es de escasa duración, en comparación con otros que duplican o triplican a los de menor duración y por ende menor variabilidad de datos.

3. El audio original contiene ruido o silencio de forma proporcional al contenido del canto, esto no permite un patrón distintivo entre el canto vs ruido/silencios.

4. La muestra seleccionada tiene ruido, la fuente de audio de youtube muchas veces tiene ruido o está mal grabada, el audio natural de los pájaros no contiene este tipo de contaminación, de todas formas, no hay un buen sistema desarrollado para combatir el ruido en este modelo.

5. Las clases de menor acierto suelen confundir repetidamente con otra clase de pájaro, existen similitudes en algunas especies en su canto que permite la confusión de patrones, siendo este otro de los aspectos a mejorar del modelo.

CONCLUSIONES Y TRABAJOS FUTUROS

Uno de los desafíos planteados fue el de poder encontrar un clasificador que funcione razonablemente considerando solo una "única muestra" de audio por tipo de pájaro. Por supuesto esto es demasiado utópico, pero los resultados no han sido malos considerando que se utilizado solo una muestra de audio por clase de pájaro, esto implica poca variabilidad.

Por supuesto, incrementando el número de muestras de distintos registros de un mismo pájaro implicaría un mejor rendimiento del modelo. A su vez se ha detectado que hay pájaros que tienen cierta similitud en el canto o en la llamada que realizan los pájaros y estos son temas que se deberían pulir para mejorar al modelo.

El tratamiento de silencios también es otro tema para tener en cuenta, descartando aquellos segmentos de audio que contengan silencio o ruido.

Las técnicas de aumentación de datos no fueron utilizadas, pero en la literatura existen técnicas comunes como time-shift, pitch-shift y time-stretch.

Pero la idea fundamental en este trabajo es exponer con algunas simples técnicas la posibilidad de clasificación de audio a partir de la conversión a imágenes o espectrogramas utilizando redes neuronales convolucionales.

La maquetación de este trabajo se ha establecido en Google Colab a fin de simplificar el proyecto, obteniendo los audios de prueba de videos de pájaros cantando de la web, pero la aplicación final debería captar audio en vivo con un micrófono interno o externo de un computador o incluso de un microcontrolador, lo cual se ha probado, pero no se incluye dentro del alcance de este trabajo. Aquí existen otros desafíos más ligados a cuestiones técnicas del sonido y su captura, tratamiento de ruidos, normalización de volumen, etc. que es fundamental para lograr el resultado, pero que va en carriles paralelos al Deep learning.

Otra cuestión por trabajar a futuro es la detección multi-etiqueta, es decir, poder discernir y clasificar todos los cantos de pájaros que se estén escuchando a la vez o que los contengan en el audio de entrada.

REFERENCIAS

[1] “Sound of new york (sonyc) homepage.” http://wp.nyu.edu/sonyc

[2] Editor. (2021). Caer en las redes. https://www.avesargentinas.org.ar/noticia/caer-en-las-redes.

[3] Ministerio de Ambiente y Desarrollo Sostenible. (2023) Lista de especies seleccionadas. https://www.argentina.gob.ar/ambiente/biodiversidad/extincion-cero/especies

[4] Reconocimiento de llamadas de aves utilizando una red neuronal convolucional profunda, ResNet-50 (2018) https://www.researchgate.net/profile/Dmitry-Konovalov-2/publication/328418948_Bird_Call_Recognition_using_Deep_Convolutional_Neural_Network_ResNet-50/links/5bcd62dc458515f7d9d02755/Bird-Call-Recognition-using-Deep-Convolutional-Neural-Network-ResNet-50.pdf

[5] G. Hinton et al., “Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups,” IEEE Signal Process. Mag., vol. 29, no. 6, pp. 82–97, 2012, doi: 10.1109/MSP.2012.2205597

[6] H. Lee, Y. Largman, P. Pham, and A. Y. Ng, “Unsupervised feature learning for audio classification using convolutional deep belief networks,” 2009, doi: 10.1109/SCET.2012.6342000.

[7] A. van den Oord, S. Dieleman, and B. Schrauwen, “Deep content-based music recommendation,” 2013.

[8] E. J. Humphrey and J. P. Bello, “Rethinking automatic chord recognition with convolutional neural networks,” in Proceedings - 2012 11th International Conference on Machine Learning and Applications, ICMLA 2012, 2012, vol. 2, pp. 357–362, doi: 10.1109/ICMLA.2012.220.

[9] J. Schlüter and T. Grill, “Exploring data augmentation for improved singing voice detection with neural networks,” in Proceedings of the 16th International Society for Music Information Retrieval Conference, ISMIR 2015, 2015, pp. 121–126.

[10] Audioronis (2022) https://audiornis.org/

[11] A. G. Howard et al., “Mobilenets: Efficient convolutional neural networks for mobile visión applications,” arXiv preprint arXiv:1704.04861, 2017.

[12] Convolutional Neural Networks for Scops Owl Sound Classification (2021) https://www.sciencedirect.com/science/article/pii/S1877050920324492