Clasificación de Maullidos y Ladridos Utilizando Características MFCC y Redes Neuronales Profundas

Álvaro Salgado López

I. Introducción

La clasificación de sonidos es un campo de creciente interés en el reconocimiento de patrones y aprendizaje automático. Su aplicación abarca desde el reconocimiento de voz hasta la detección de sonidos en entornos de vigilancia y monitoreo de la fauna. En particular, la clasificación de sonidos de animales domésticos, como maullidos y ladridos, puede proporcionar información valiosa en estudios de comportamiento animal, bienestar de mascotas y detección de emergencias.

La motivación de este estudio surge de una tarea previa en la que se analizaron espectrogramas de sonidos de maullidos y ladridos, observando que presentaban patrones diferenciables en el espectro de frecuencia. Esto sugirió que podría ser viable entrenar un modelo de aprendizaje profundo para clasificar estos sonidos automáticamente.

Inicialmente, se consideró la utilización de transformadas wavelets para la extracción de características. Sin embargo, se encontró que este método era computacionalmente costoso y requería una mayor capacidad de procesamiento. Debido a esto, se optó por emplear coeficientes de frecuencia cepstral en Mel (MFCC), ya que estos han demostrado ser altamente efectivos en tareas de reconocimiento de voz y son computacionalmente más eficientes.

Otro problema inicial fue la variabilidad en la duración de los archivos de audio, lo que generaba vectores de características de diferentes longitudes. Para abordar esta situación, se implementó un procedimiento de normalización donde todos los vectores se ajustaban a una longitud fija mediante el relleno con ceros. Este paso permitió que los datos fueran compatibles con la arquitectura de la red neuronal utilizada.

El objetivo de este estudio es desarrollar un modelo basado en redes neuronales profundas que pueda clasificar de manera automática maullidos y ladridos a partir de sus representaciones MFCC. Se empleó una arquitectura convolucional debido a su capacidad para extraer patrones espaciales relevantes en representaciones espectro-temporales.

II. Metodología

A. Adquisición y Preprocesamiento de Datos

Para este estudio, se recopilaron archivos de audio de maullidos y ladridos en formato WAV. Los datos se obtuvieron directamente de Kaggle, para garantizar una variedad en la calidad y características de los sonidos.

El procesamiento de los archivos de audio incluyó los siguientes pasos:

- Carga del audio: Se utilizó la librería Librosa para cargar los archivos de audio con una frecuencia de muestreo estándar de 22,050 Hz.
- Extracción de MFCC: Se calcularon 13 coeficientes MFCC por ventana de tiempo, que representan la envolvente espectral del sonido.
- Normalización de la duración: Como los audios tenían diferentes longitudes, se truncaron o rellenaron con ceros hasta un máximo de 500 cuadros de tiempo.
- Etiquetado de los datos: Cada audio se asignó a una de las dos clases: maullido o ladrido.

Se dividió el conjunto de datos en 80% para entrenamiento y 20

B. Arquitectura del Modelo

Para la clasificación, se construyó una red neuronal convolucional con la siguiente arquitectura:

- Capa convolucional 2D: 32 filtros con tamaño de kernel 3x3 y activación ReLU.
- Capa de max pooling: Reducción de dimensionalidad con un filtro de 2x2.
- Segunda capa convolucional: 64 filtros con tamaño de kernel 3x3 y activación ReLU.
- Segunda capa de max pooling: Reducción de dimensionalidad adicional con filtro 2x2.
- Capa completamente conectada: 128 neuronas con activación ReLU para capturar patrones complejos.
- Capa de salida: Activación softmax con dos neuronas para la clasificación binaria.

El modelo fue compilado utilizando el optimizador Adam, el cual es ampliamente utilizado por su eficiencia en el ajuste de los pesos durante el entrenamiento. Como función de pérdida, se empleó la entropía cruzada categórica, adecuada para problemas de clasificación multiclase y binaria, permitiendo optimizar la separación entre categorías. Finalmente, la métrica de evaluación seleccionada fue la precisión, ya que proporciona una medida clara del porcentaje de predicciones correctas realizadas por el modelo.

C. Entrenamiento y Validación

El modelo fue entrenado con los siguientes hiperparámetros:

- Número de épocas: 10

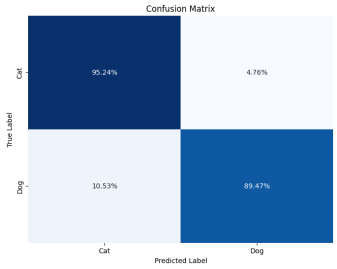

Fig. 1. Matriz de confusión

- Tamaño de lote: 16

Durante el entrenamiento, se monitorizó la precisión y la pérdida en el conjunto de validación para evaluar el desempeño del modelo y evitar sobreajuste.

Se realizaron pruebas adicionales ajustando el número de filtros y la cantidad de capas para optimizar el rendimiento del modelo sin aumentar excesivamente el tiempo de entrenamiento.

III. Resultados

El modelo fue evaluado en un conjunto de prueba utilizando la precisión como métrica principal. Se obtuvo una precisión del 93% en la clasificación de maullidos y ladridos. La matriz de confusión obtenida se observa en la Fig 1

Esto indica que el modelo clasificó correctamente 20 maullidos y 17 ladridos, con solo 3 errores en total. Se observa que la mayor parte de los errores provienen de la clasificación errónea de maullidos como ladridos y viceversa, lo que sugiere cierta similitud en algunos patrones espectrales de ambos sonidos.

El rendimiento del modelo también fue evaluado a través de la pérdida en la función de entropía cruzada y la precisión en el conjunto de validación. Durante el entrenamiento, la pérdida disminuyó de manera constante mientras que la precisión aumentó, lo que indica que el modelo fue capaz de aprender patrones relevantes en los datos.

IV. Conclusión

Se demostró la viabilidad del uso de MFCC y redes neuronales convolucionales para la clasificación de sonidos de animales domésticos. La obtención de una precisión del 93% indica que la arquitectura utilizada es efectiva para la diferenciación de maullidos y ladridos. Sin embargo, algunos errores en la clasificación sugieren la posibilidad de mejorar el modelo mediante el uso de un conjunto de datos más amplio y variado, lo que permitiría una mejor generalización.

Para futuras investigaciones, se podrían explorar técnicas avanzadas como el uso de modelos de aprendizaje profundo preentrenados, la incorporación de transformadas wavelet para enriquecer las características extraídas o la implementación de estrategias de data augmentation que permitan aumentar la diversidad de los datos de entrenamiento.