Instituto Tecnológico y de Estudios Superiores de Monterrey

Campus Querétaro

Momento de Retroalimentación: Módulo 2 Implementación de un modelo de deep learning.

(Portafolio Implementación)

Alumno:

Tomás Pérez Vera – A01028008

TC3007C.501

Inteligencia Artificial Avanzada Para la Ciencia de Datos II

Fecha:

8 de noviembre de 2025

Resumen

El presente trabajo, pretende realizar un análisis comparativo de dos arquitecturas de redes neuronales convolucionales (CNN) aplicadas a la clasificación binaria de audio, con el fin de realizar la distinción entre sonidos de gatos y perros. El conjunto de datos empleado consiste en grabaciones en formato WAV, recortadas a una duración máxima de un segundo y muestreadas a 16 kHz, balanceadas entre ambas clases. El primer modelo utiliza convoluciones en una dimensión sobre las formas de onda crudas, mientras que el segundo emplea representaciones espectrales mediante Coeficientes Cepstrales en la Frecuencia Mel (MFCC) combinados con convoluciones bidimensionales. Los resultados muestran que el modelo basado en MFCC alcanzó una mayor precisión de clasificación y una convergencia más rápida, demostrando la efectividad de las características del dominio frecuencial para representar información acústica relevante en tareas de clasificación de audio.

INTRODUCCIÓN

En los últimos años, el uso de redes neuronales convolucionales (CNN) se ha extendido más allá del procesamiento de imágenes hacia el reconocimiento de audio, gracias a su capacidad para extraer patrones complejos y características jerárquicas. En este contexto, la identificación automática de sonidos de animales representa un caso de estudio para evaluar la efectividad de distintas representaciones del audio.

El presente trabajo evalúa el desempeño de dos enfoques:

1. **CNN 1D**, que procesa directamente las formas de onda en el dominio temporal.
2. **CNN 2D**, que utiliza **MFCC** (Mel-Frequency Cepstral Coefficients) para representar los audios en el dominio frecuencial antes de aplicar convoluciones bidimensionales.

El objetivo es comparar ambas estrategias y determinar cuál proporciona una representación más efectiva para la tarea de clasificación entre sonidos de gatos y perros.

DESCRIPCIÓN DEL CONJUNTO DE DATOS

El conjunto de datos utilizado corresponde a una colección de grabaciones de audio disponible en la plataforma **Kaggle**, particularmente en la competencia de *Audio Cats and Dogs*. Este dataset tiene como propósito evaluar la capacidad de los modelos de aprendizaje profundo para distinguir entre sonidos de gatos y perros.

El conjunto está conformado por archivos en formato WAV con una frecuencia de muestreo uniforme de 16 kHz y duraciones variables. Cada clase presenta un número diferente de muestras:

- Clase gato: 164 archivos WAV, equivalentes a 1,323 segundos de audio.

- Clase perro: 113 archivos WAV, equivalentes a 598 segundos de audio.

Las grabaciones provienen de la base AE-Dataset (Acoustic Events Dataset) y fueron extraídas originalmente de la plataforma FreeSound, una comunidad colaborativa de sonidos abiertos.

Para el presente trabajo, los audios fueron normalizados en amplitud, convertidos a una duración estándar de un segundo y segmentados en tres subconjuntos:

- Entrenamiento (80%)
- Validación (20%)

El objetivo de este preprocesamiento fue garantizar un equilibrio entre clases y homogeneidad en la longitud de las señales, permitiendo un entrenamiento más estable y una comparación justa entre los modelos.

METODOLOGÍA DE TRABAJO

Modelo 1 – CNN 1D

Este modelo utiliza directamente la señal de audio re-muestreada y normalizada como entrada.
La arquitectura consta de capas convolucionales 1D con activaciones ReLU, seguidas de *AdaptiveAvgPool1d* y capas densas.
Su propósito es permitir que la red aprenda los patrones temporales y energéticos de las señales sin una etapa explícita de extracción de características.

Modelo 2 – CNN 2D (MFCCs)

El segundo modelo emplea una representación espectral del audio basada en **Coeficientes Cepstrales en las Frecuencias de Mel (MFCC)**.
Los MFCC permiten mapear la energía espectral de la señal al dominio perceptual del oído humano, facilitando la identificación de patrones acústicos relevantes para cada clase.
Cada muestra se convierte en un mapa de características de dimensiones (n_mfcc × tiempo), que posteriormente se procesa con convoluciones 2D, capas de *Batch Normalization* y *Dropout* para evitar el sobreajuste.

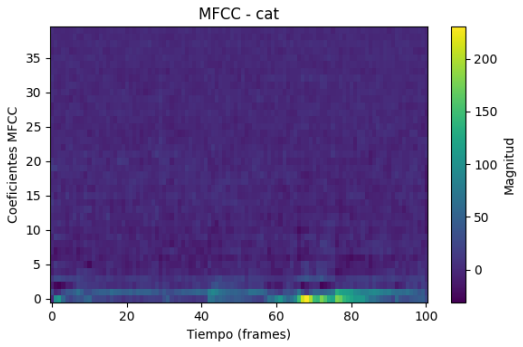

Figura 1. MFFCC de gato

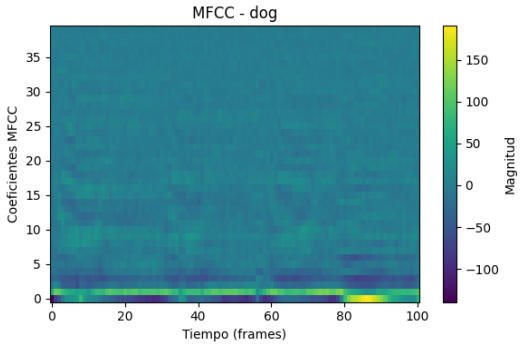

Figura 2. MFFCC de perro

RESULTADOS EXPERIMENTALES

Ambos modelos fueron entrenados bajo las mismas condiciones:

- Optimizador: Adam (lr = 1e-3)
- Función de pérdida: CrossEntropyLoss
- Tamaño de lote: 8
- Número de épocas: 150 para el modelo 2, 150 para el modelo 1.
- Tasa de muestreo: 16 kHz
- Duración máxima: 1 segundo

Tabla 1. Comparativa de Métricas entre Modelos

Métrica	Modelo 1 – CNN 1D	Modelo 2 – MFCC + CNN 2D
Precisión de entrenamiento	0.87	0.92
Precisión de validación	0.76	0.84
Pérdida de validación mínima	0.46–0.50	≈ 0.38–0.40
Épocas totales	300	150
Observación	Ligeramente sobreajustado	Mejor generalización

GRAFICAS OBTENIDAS

Modelo 1 – CNN 1D

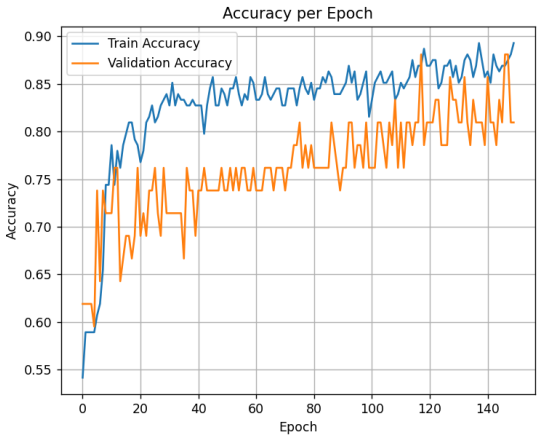

Figura 3. Certeza por época Modelo CNN 1D

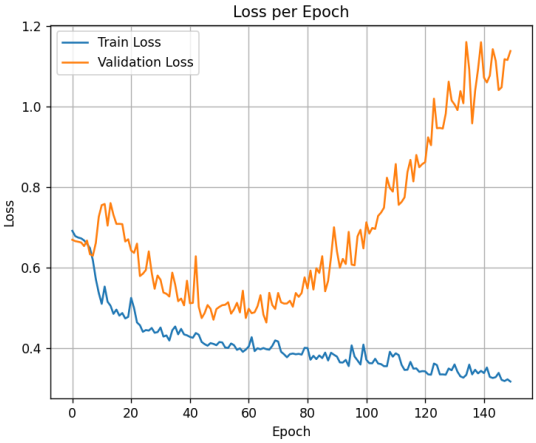

Figura 4. Pérdida por época Modelo CNN 1D

Modelo 2 – CNN 2D (MFCCs)

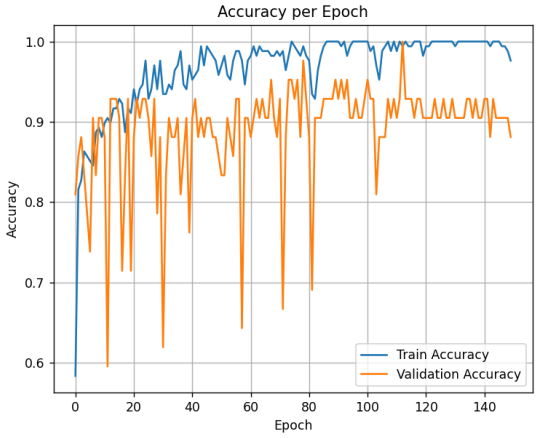

Figura 5. Certeza por época Modelo CNN 2D

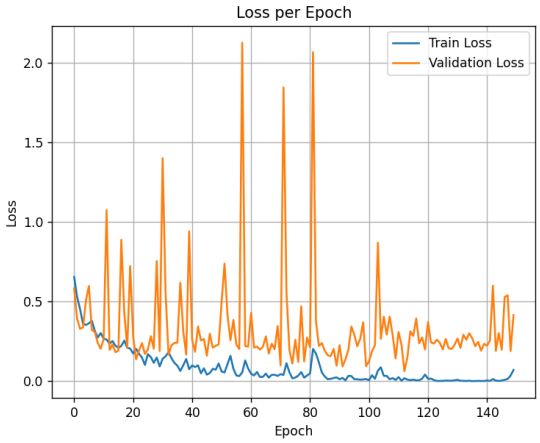

Figura 6. Pérdida por época Modelo CNN 2D

ANÁLISIS DE RESULTADOS

El modelo CNN 1D, al operar sobre la forma de onda cruda, debe aprender simultáneamente tanto las propiedades espectrales como las temporales de la señal, lo que incrementa su complejidad y demanda una mayor cantidad de datos para generalizar correctamente.

Por el contrario, el modelo CNN 2D aprovecha la representación MFCC, que compacta la información espectral más relevante en un espacio más discriminativo, reduciendo la carga de aprendizaje y mejorando la robustez ante variaciones de ruido o volumen.

Los resultados empíricos muestran que el segundo enfoque logra una mayor precisión de validación (≈84%), con menor pérdida y una brecha más reducida entre entrenamiento y validación, lo que indica una mejor generalización.

En consecuencia, se concluye que la utilización de MFCCs como representación intermedia mejora significativamente el desempeño del modelo y su estabilidad durante el entrenamiento.

CONCLUSIÓN

En conclusión, el modelo basado en MFCC + CNN 2D demostró un desempeño superior al modelo CNN 1D, destacando en precisión, estabilidad y capacidad de generalización. La utilización de características perceptuales mediante los coeficientes MFCC permitió simplificar el proceso de aprendizaje y disminuir la tendencia al sobreajuste, al proporcionar una representación más compacta y relevante del contenido acústico. Por ello, este enfoque resulta especialmente recomendable para tareas de clasificación de audio con conjuntos de datos limitados o con alta variabilidad temporal. Además, se desarrolló un script de predicción por terminal que posibilita cargar archivos .wav y obtener la probabilidad correspondiente a cada clase (“Cat” o “Dog”), lo que facilita la evaluación práctica y la validación del modelo entrenado.