Clasificación de Sonidos de la Naturaleza usando el Audio Spectrogram Transformer (AST)

Descripción del Proyecto

Los sonidos en la naturaleza son ricos en información acústica, que se puede clasificar en diversas categorías como aves, agua, viento, entre otros. Este proyecto tiene como objetivo utilizar un modelo que pueda identificar la fuente de estos sonidos a través de un enfoque basado en transformers para transformar audio en representaciones visuales que el modelo puede analizar.

Introducción a los Transformers

Los transformers son modelos revolucionarios en el aprendizaje profundo por su habilidad para capturar relaciones de largo alcance en datos secuenciales, como texto o audio. Originalmente desarrollados para Procesamiento de Lenguaje Natural (NLP), los transformers pueden captar relaciones complejas entre elementos secuenciales. Esto es posible gracias a su arquitectura, basada en mecanismos de auto-atención, que permite que el modelo evalúe todas las relaciones entre los elementos de una secuencia, proporcionando así un contexto completo para cada elemento en una única operación.

El Audio Spectrogram Transformer (AST) adapta esta arquitectura para datos acústicos. AST primero transforma las señales de audio en espectrogramas, representaciones visuales que exhiben cómo cambian las frecuencias en el tiempo. Al aplicar auto-atención a estos espectrogramas, el AST captura patrones acústicos detallados, permitiendo reconocer sonidos de manera robusta en distintas categorías de audio. Este enfoque es especialmente valioso en el análisis de sonidos no estructurados de la naturaleza, donde las características de los sonidos son más impredecibles y contextualmente variables.

Objetivo

Hacer fine tuning y evaluar el modelo preentrenado Audio Spectrogram Transformer para la clasificación de sonidos naturales en categorías como “aves”, “agua”, “viento” y “mamíferos”.

Conjunto de Datos

Se usará un dataset de grabaciones de sonido natural con diversas etiquetas de categoría llamado ESC50 que contiene 2000 grabaciones de audio para clasificación de sonidos de ambiente. Estas grabaciones se transformarán en espectrogramas para que el modelo pueda procesarlas.

Carga de datos

Deben cargar los datos de la siguiente manera:

```
import torch
import torch.nn as nn
import torchaudio
from datasets import load_dataset
# Dataset de entrenamiento
dataset = load_dataset("confit/esc50-demo", "fold1", split="train")
# Dataset de test
val_dataset = load_dataset("confit/esc50-demo", "fold1", split="test")
```

Modelo preentrenado

Para el modelo, deberán usar la biblioteca Transformers de Hugging Face. Pueden encontrar información sobre cómo cargar modelos, las funciones disponibles y más opciones en el siguiente enlace:

https://huggingface.co/docs/transformers/model_doc/audio-spectrogram-transformer

Les recomendamos usar el modelo "MIT/ast-finetuned-audioset-10-10-0.4593".

Actividades por Realizar

1. **Procesamiento de Audio**: Diseñar un pipeline que convierta cada archivo de audio en espectrogramas visuales, normalizando los datos y garantizando consistencia en los tamaños de entrada. Los espectrogramas generados deben estar listos para ser usados como entradas para el modelo AST.

2. **Implementación del Audio Spectrogram Transformer:** Hacer fine tuning de un modelo de AST durante 2 épocas para adaptarlo a la tarea específica.

3. **Evaluación del Modelo:** Realizar predicciones sobre los datos y reportar métricas como precisión, recall y F1-score, además de un análisis detallado de las categorías en las que el modelo presenta mejor o peor rendimiento. La evaluación también incluirá un análisis de la matriz de confusión para observar los patrones de errores y las categorías más confusas para el modelo.