Marco Obregón

Procesamiento y Clasificación de Datos

12 November 2023

Practica 7

Como investigador especializado en procesamiento de señales de audio y aprendizaje profundo, mi estudio se centra en un área específica del reconocimiento de voz: la interpretación precisa de comandos de voz para controles direccionales. Este proyecto busca adaptar la tecnología para ser particularmente útil para personas con discapacidades motoras, permitiéndoles una interacción más efectiva con dispositivos electrónicos a través de comandos de voz. Mi hipótesis es que combinando técnicas avanzadas de procesamiento de señales con modelos de aprendizaje profundo, se puede crear un sistema que interprete con precisión comandos direccionales, contribuyendo significativamente a mejorar la calidad de vida de estas personas.

Objetivos del Estudio:
El objetivo principal es desarrollar un sistema de reconocimiento de voz de alta precisión para interpretar comandos direccionales, mejorando así la interacción de personas con discapacidades motoras con dispositivos electrónicos. Los objetivos secundarios incluyen la investigación y aplicación de técnicas avanzadas de preprocesamiento y extracción de características, y la validación de modelos de aprendizaje profundo que sean más efectivos que los enfoques tradicionales. También busco diseñar e implementar estrategias para mitigar el impacto del ruido externo, aumentando la robustez del sistema en entornos no ideales.

Metodología:
En mi proyecto, comparé dos metodologías distintas para el procesamiento de audios de comandos de voz: un autoencoder simple y un autoencoder convolucional. Cada metodología incorporó diferentes enfoques en varias etapas clave:

- **Preprocesamiento de Audio:**
  - **Autoencoder Simple:** Utilicé re-muestreo, la aplicación de WebRTC VAD, y estandarización mediante padding, asegurando una representación uniforme y eficiente de los datos.
  - **Autoencoder Convolucional:** En esta metodología, realicé la carga y estandarización de archivos de audio y generé espectrogramas Mel, normalizándolos al rango [0, 1].
- **Extracción y Normalización de Características:**
  - **Autoencoder Simple:** Extraje características MFCC, normalizándolas con la media y la desviación estándar del conjunto de datos.
  - **Autoencoder Convolucional:** Utilicé espectrogramas Mel como características, normalizados al rango [0, 1].

- **Limpieza y Preparación de Datos:**
  - **Autoencoder Simple:** Implementé PCA para eliminar outliers y usé WebRTC VAD para descartar audios vacíos, seguido de una división en conjuntos de entrenamiento y validación.
  - **Autoencoder Convolucional:** La metodología se centró solo en la división de los datos en conjuntos de entrenamiento y validación, sin una limpieza de datos tan exhaustiva como en el autoencoder simple.
- **Construcción y Entrenamiento del Modelo:**
  - **Autoencoder Simple:** Construí un modelo de autoencoder simple con capas densas, utilizando generadores de datos y aplicando técnicas como Early Stopping y Model Checkpointing.
  - **Autoencoder Convolucional:** Este modelo incluyó capas de convolución, max pooling, upsampling y sigmoid, siguiendo un enfoque de entrenamiento similar.

Resultados:
Los resultados mostraron que el autoencoder simple superó al convolucional, alcanzando un PSNR más alto. El PSNR indica la eficacia en la reducción de ruido y la fidelidad en la reconstrucción del audio. El autoencoder simple demostró ser más efectivo, probablemente debido a su enfoque más detallado en el preprocesamiento y la extracción de características, así como a una limpieza de datos más profunda que mejoró la calidad del conjunto de entrenamiento.

Conclusión y Trabajo Futuro:
La investigación confirmó que un preprocesamiento y limpieza de datos meticulosos son fundamentales para mejorar la precisión de los autoencoders en el reconocimiento de comandos de voz. En el futuro, planeo realizar pruebas de campo con usuarios reales para adaptar el sistema a situaciones reales y explorar el uso de transfer learning con autoencoders entrenados previamente en conjuntos de datos similares, buscando mejorar aún más la precisión y robustez del sistema en diversos entornos acústicos.