Deep Learning para el procesamiento de señales de audio

Deep Learning for audio signal processing

Autor: Santiago Ocampo Orrego
IS&C, Universidad Tecnológica de Pereira, Pereira, Colombia
Correo-e: santiago.ocampo1@utp.edu.co

***Resumen*— En este documento abordaremos los principios básicos para el procesamiento y análisis de señales de audio. También hablaremos de algunas de sus aplicaciones en el mundo real y un poco de su importancia en la actualidad. Para lograr nuestro objetivo también hablaremos de Deep Learning y daremos un vistazo a su estructura para el aprendizaje mediante conjuntos de datos.**

***Palabras clave*— aprendizaje, señal, audio, inteligencia artificial, software, redes neuronales.**

***Abstract*— In this document we will cover the basic principles for the processing and analysis of audio signals. We will also talk about some of its applications in the real world and a little about its importance today. To achieve our goal, we will also talk about Deep Learning and take a look at its structure for learning through data sets.**

***Key Word*— learning, signal, audio, artificial intelligence, software, neural networks.**

I. INTRODUCCIÓN

Si bien gran parte de la información sobre el Deep Learning se refiere a la visión por computadora y el procesamiento del lenguaje natural (NLP), el análisis de audio, un campo que incluye el reconocimiento automático de voz (ASR), el procesamiento de señales digitales y la clasificación, el etiquetado y la generación de música, es un subdominio creciente de aplicaciones de aprendizaje profundo. Algunos de los sistemas de aprendizaje automático más populares y extendidos, los asistentes virtuales Alexa, Siri y Google Home, son en gran parte productos construidos sobre modelos que pueden extraer información de señales de audio.

II. DEEP LEARNING

Antes de hablar de lo que es el Deep Learning, debemos tener claro que es el Machine Learning. Más adelante miraremos el porqué de esta aclaración.

El Machine Learning es un subconjunto de la inteligencia artificial asociado con la creación de algoritmos que pueden cambiarse a sí mismos sin intervención humana (aprendizaje no supervisado) para obtener un resultado deseado, alimentándose a sí mismos a través de datos estructurados [1].

Pero ¿por qué es necesario saber que es el Machine Learning para hablar de Deep Learning? Es sencillo. El Deep Learning es un subconjunto del Machine Learning, la forma en la que trabaja es usando redes neuronales, generalmente convolucionales, usando capas con unidades de procesamiento que permiten la extracción y transformación de variables. Cada red dentro de su organización aplica una transformación a su capa de entrada y utiliza esta información para crear un modelo estadístico de salida que itera las veces que sean necesarias para lograr un nivel de aprendizaje y respuesta aceptable. El Deep Learning moderno a menudo implica decenas o incluso cientos de capas sucesivas, y todas aprenden automáticamente a partir de la exposición a los datos de entrenamiento. Mientras que otros enfoques de Machine Learning tienden a centrarse en aprender con solo una o dos capas, por lo que se dice a veces que es superficial [2].

III. DEEP LEARNING EN SEÑALES DE AUDIO

Al igual que todos los demás tipos de datos, el Deep Learning funciona mejor cuando se tiene acceso a grandes conjuntos de datos de entrenamiento. Sin embargo, la diversidad de señales de audio, voz y acústicas, y la falta de grandes conjuntos de datos bien etiquetados, dificulta el acceso a grandes conjuntos de entrenamiento. Al utilizar métodos de aprendizaje profundo en archivos de audio, es posible que se deba desarrollar nuevos conjuntos de datos o ampliar los existentes.

Una vez que tengamos un conjunto de datos inicial, podemos ampliarlo aplicando técnicas de aumento como el cambio de tono, el cambio de hora, el control de volumen y la adición de ruido. El tipo de aumento que vayamos a aplicar depende de las características relevantes para su aplicación de audio, voz o

acústica. Por ejemplo, el cambio de tono (o perturbación del tracto vocal) y el estiramiento del tiempo son técnicas de aumento típicas para el reconocimiento automático de voz. Para el ASR de campo lejano, es común aumentar los datos de entrenamiento mediante reverberación artificial.

Los datos de entrenamiento utilizados en los flujos de trabajo de Deep Learning suelen ser demasiado grandes para caber en la memoria. Acceder a los datos de forma eficiente y realizar tareas comunes de Deep Learning (como dividir un conjunto de datos en conjuntos de trenes, validación y pruebas) puede volverse rápidamente inmanejable.

El preprocesamiento de datos de audio incluye tareas como muestreo de archivos de audio a una frecuencia de muestreo coherente, eliminar regiones de silencio y recortar audio a una duración coherente.

El audio es altamente dimensional y contiene información redundante y a menudo innecesaria. Históricamente, los coeficientes cepstrales en las frecuencias de Mel (MFCC) y las características de bajo nivel, como la velocidad de cruce cero y los descriptores de forma espectral, han sido las características dominantes derivadas de las señales de audio para su uso en sistemas de aprendizaje automático. Los sistemas de aprendizaje automático capacitados en estas características son eficientes desde el punto de vista computacional y normalmente requieren menos datos de entrenamiento.

Los avances en arquitecturas de Deep Learning, un mayor acceso a la potencia informática y un gran conjunto de datos bien etiquetados, han reducido la dependencia de las características diseñadas a mano. Los resultados de hoy en día se logran utilizando espectrogramas de Mel (Mel Spectrogram), espectrogramas lineales o formas de onda de audio sin procesar [3].

Algunas de las aplicaciones que podemos lograr mediante el Deep Learning en señales de audio son:

- Reconocimiento de voz
- Recomendaciones de música en plataformas digitales
- Búsqueda de similitud en archivos de audio
- Procesamiento y síntesis de habla

REFERENCIAS

[1] Parsers. 2019. Deep Learning & Machine Learning: What's The Difference? - Parsers. https://parsers.me/deep-learning-machine-learning-whats-the-difference/#:~:text=The%20main%20difference%20between%20deep,ANN%20(artificial%20neural%20networks)

[2] Chollet, F., 2018. Deep Learning With Python. Shelter Islands: Manning, pp.8-9.

[3] Introduction to Deep Learning for Audio Applications. (s. f.). MathWorks. https://www.mathworks.com/help/audio/gs/intro-to-deep-learning-for-audio-applications.html