Arnoldo Oliva. Clasificación de sentimiento a partir de audio y CV

Script Poster:

INTRODUCCIÓN:

Hay mucha curiosidad hoy en día en tratar de saber que emoción es la que impera en una opinión, o frases dichas, en la primera se suele usar análisis de texto pero en esta ocasión se utilizarán técnicas de análisis de audio y Computer Vision para tratar de clasificar la emoción imperante en diversos audios de personas pronunciando frases en diferentes tonalidades, ya sea Feliz o Triste. Cabe aclarar que en el presente proyecto no se va a enfocar en el contenido de la oración para predecir dicho sentimiento, sino que se enfocara en las tonalidades de como la persona pronuncio la frase.

El objetivo principal de este proyecto es, Mediante el uso de técnicas de análisis de audio y Machine Learning, predecir la emoción preponderante de dicho audio. Y como segundos objetivos son, el procesamiento de los datos adecuados para que nos permitan obtener aprendizajes interesantes a través de las características de los audios.

METODOLOGÍA:

La base de datos se encuentra en Kaggle. Esta base fue descargada en formato de diversos folders con archivos .wav o .mp3 de no más de 15 segundos, y cada folder contiene más de 2100 records. Dichos registros fueron parseados y convertidos a arrays al mismo tiempo de obtener su sample rate ("muestras de audio por segundo" o "hertz" (Hz)), con las librerías de librosa y torchaudio de Python.

Un paso importante fue separar el proceso en muestras tratadas por remoción de ruido de fondo con un cutoff frequency (frecuencia a la cual se establece un límite para el filtrado de una señal) de 300 (escogido arbitrariamente) vs sin remoción de ruido, para comparar si las muestras tratadas tienen una mejora en el accuracy de predicción. A manera de comparación se plotearon algunas muestras mostrando sus formas de onda originales vs con tratamiento.

Como segundo paso se procedió a realizar análisis descriptivos (y comparativos) de las diversas ondas, para la representación visual solo se trabajaron con pequeñas muestras randomizadas y para características generales, como por ejemplo, duraciones de los audios, amplitud de los sonidos, y “energies” (cuantifican la "fuerza" o "amplitud" general de una señal). Un aprendizaje que se pudo obtener es que los audios de Sad duraron un poco más que los de Happy, e igual, en los gráficos de fuerza de la señal, los audios de Happy fueron mucho más relevantes que los de Sad. (puede ser que lo que se pronuncia se hace con más “enjundia”).

Esto ocurrió en ambos tratamientos, con la excepción de que en el caso de sin ruido disminuyeron un poco los valores en las gráficas.

Debido a que no existe manera visual sencilla para representar todas las muestras de manera visual (serían 4000 gráficos individuales) se exploraron gráficos tanto espectogramas y MFCC en una submuestra de 30 audios. El espectograma es una representación visual de las diferentes variaciones de la frecuencia y la intensidad del sonido a lo largo del tiempo/duración del audio. Decidí no mostrar un comparativo de espectogramas de las dos categorías de ruido vs sin ruido debido a que son casi impercetibles las diferencias entre ambas.

También se exploraron en estas submuestras gráficos de MFCC, o Coeficientes Cepstrales en la Frecuencia Mel, que en síntesis es una extracción de características.

Modelo:

Debido a que es más intuitivo el espectograma, se eligió este gráfico para ser el input del modelo, el cuál fue una Red Neuronal Convolucional. Se plotearon los gráficos en imagenes usando Matplotlib para los dos tipos de audio (Feliz y Triste), para los dos tratamientos: con ruido vs sin ruido de fondo.

Se utilizó la librería CV2 para parsear dichas imágenes, que fueron sampleadas aleatoriamiente, y acotarlas a arrays de 224 x 244 pixeles.Se trabajo con escalas a colores por lo tanto se usaron los tres canales de RGB. Para la parte de la convolución, se uso la red de TransferLearning de TensorflowKeras VGG16, con 13 capas convolucionales y 5 de pooling, con kernels de 3 x 3, resultando en 512 neuronas.

Después estos arrays fueron mandados a la segunda red neuronal Dense también de Keras, La primera capa consto de 2048 neuronas. Se trabaja con lotes de 32 inputs (batch size). Todas las capas, a excepción de la última, contaban con función de activación “relu” (para forzar que los valores sean positivos, o de lo contrario, cero). De esta capa a la segunda (de 1024 neuronas) se implementa “Dropout”, el cual es una técnica de regularización lo que ayuda a evitar el sobreajuste (overfitting). La tercera capa consta de 512 neuronas, la cuarta de 64, la quinta de 10 y la última de 2, debido a que solo son dos categorías, Happy siendo 1, Sad siendo 0, se pudo representar con la función de activación “sigmoid”, para obtener la predicción dicotómica. Se recorrio el dataset en 10 epochs.

RESULTADOS Y CONCLUSIONES:

Habiendo entrenado los modelos bajo la misma partición aleatoria, el resultado agregado del accuracy de ambos fueron 78.78% y 52.36% para los datos sin ruido y con ruido, respectivamente.

Se pueden observar varias cosas en las matrices de confusión. Uno, que en el modelo con ruido el modelo confundió las diversas muestras a pesar de las diferencias con la etiqueta de “Happy”, (casi la totalidad de las predicciones las catalogo de esa manera). Esto no ocurrió en el caso del modelo sin ruido, donde, a pesar de tener menor precisión para la categoría de Happy, se tiene una clasificación más balanceada de las dos categorías, (el 70% de los valores de Sad los clasifico correctamente, y para Happy fue del 88%. Esto ultimadamente mejoro el accuracy del modelo.

Cabe aclarar que ambos modelos tenían un training accuracy iniciando alrededor del 78% para terminar en 88%, pero la diferencia real entre ambos fue cuando se trabajo con los datos de testing, indicándonos que el modelo sin ruido fue más apto para generalizar los resultados del aprendizaje del training de la red, esto pudiendo ser debido a que los modelos sin ruido pudieran ser más limpios y por lo tanto más aptos para generalizar.

El accuracy y lo ya mencionado nos muestra que el hecho de remover el ruido de fondo si incidió en una mejora en los modelos. Aunque cabe aclarar que dependía del random seed colocado, por lo tanto se debería de hacer una exploración de ese modelo en cada posible combinación de seed aleatorio para hacerlo determinístico.

A futuro:

Como trabajo futuro se podrían explorer las opciones de trabajar el mismo modelo pero usando esta vez los gráficos MFCC, explorar con otro cutoff rate en el ruido de fondo, tal vez implementar más categorías de emociones e inclusive implementar técnicas de análisis de sentimiento en las frases; modificar la red neuronal o usar otro modelo convolucional; y debido a que no dio tiempo suficiente se podría realizar una comparativa de todos los modelos posibles bajo los diferentes seeds para sacar una conclusion deterministica, entre otros.