EXTRACCION DE PARTITURAS DE UN AUDIO



NOMBRE: Vera Miranda Luis Gustavo

EXTRACCION DE PARTITURAS DE UN AUDIO

1. Resumen:

A través del uso de modelos de aprendizaje automático y técnicas de procesamiento de señales, se desea desarrollar un software capaz de identificar patrones musicales y realizar tareas de clasificación y transcripción automática de música. Se utilizarán algoritmos de redes neuronales y técnicas de preprocesamiento de datos para mejorar la precisión del modelo. Se intentará crear un software capaz de reconocer la música de forma efectiva y precisa.

Para la primera parte de este proceso se separará el audio original en canales, cada uno de un instrumento diferente para aislar las notas de estos, para este propósito se usará la arquitectura de webe-unet que es una red convolucional mixta que usa capas CNN y LSTM para la segmentación del audio.

Posteriormente se dividirá el audio en frames del tamaño de una semicorchea para poder analizar el audio y esta será transformada a un espectrograma y clasificada por una red CNN para su clasificación y posterior escritura en una partitura

1. Antecedentes:

El análisis de notas musicales ha sido un campo de interés para musicólogos y científicos desde hace siglos. Tradicionalmente, este análisis se realizaba de manera manual, donde expertos en música estudiaban partituras, escuchaban interpretaciones y analizaban las características de las notas, como la frecuencia, la duración y la intensidad. Con el avance de la tecnología, estos métodos han evolucionado, permitiendo un análisis más preciso y eficiente.

La inteligencia artificial (IA) ha revolucionado numerosos campos, y el análisis musical no es una excepción. A partir de la década de 2010, los avances en el aprendizaje automático (machine learning) y el aprendizaje profundo (deep learning) han permitido a los investigadores desarrollar sistemas capaces de analizar grandes volúmenes de datos musicales con una precisión y rapidez sin precedentes. Estos sistemas pueden identificar patrones complejos en las notas musicales que serían difíciles o imposibles de detectar manualmente.

1. Situación problemática:

En la actualidad, la transcripción de música, es decir, la extracción de notas musicales a partir de una grabación de audio, sigue siendo un proceso mayoritariamente manual. Este procedimiento resulta ser laborioso, requiere un oído musical entrenado y es propenso a errores, lo que limita su accesibilidad a músicos aficionados o personas sin formación especializada. Además, la variabilidad de los instrumentos, la calidad de las grabaciones y la presencia de ruido dificultan aún más la precisión en la identificación de las notas.

Los métodos automáticos existentes para la extracción de notas musicales no siempre son lo suficientemente precisos, especialmente cuando se aplican a géneros musicales complejos o en entornos con ruido. Estas herramientas suelen estar diseñadas para funcionar bien solo en condiciones muy controladas, lo que limita su aplicabilidad en la diversidad de situaciones musicales reales.

En un contexto en el que la música digital sigue creciendo y la creación musical asistida por tecnología está en auge, la necesidad de desarrollar herramientas más precisas y eficientes para la extracción automática de notas se ha vuelto evidente.

1. Problema principal

Aun a expertos en sonido les cuesta diferenciar notas el de la música directamente ya sea por el entorno o por lo complicado de la composición de la música por lo cual se requiere de una herramienta que ayude en dicho análisis

1. Objetivo general:

El objetivo de este proyecto es poder obtener partituras a partir de un audio o canción directamente extrayendo estas con inteligencia artificial, para ayudar a los músicos a poder obtener dichas partituras ya sea para practicas u otros objetivos.

1. Objetivos específicos:

1-Extraer un instrumento en particular usando Weve-U-Net el cual es una red U-Net modificada para su uso en el área del sonido

2-Usando otra red neuronal CNN hacer el procesamiento de dicho canal de audio extraído para poder procesar y analizar las notas musicales

3-Escribir la partitura con los resultados del análisis de las notas extraidas en el objetivo 2

1. Fundamento teórico:
2. Ámbito al que se aplica la inteligencia artificial.

* Transcripción musical: Facilita la conversión automática de grabaciones en partituras musicales, reduciendo el esfuerzo manual.
* Análisis musical: Permite el estudio detallado de piezas musicales para identificar patrones, armonías y estructuras melódicas.
* Educación musical: Proporciona herramientas para ayudar a estudiantes a mejorar sus habilidades auditivas y comprensión de la teoría musical.
* Producción musical: Asiste a productores y compositores en la creación, edición y análisis de música de manera más eficiente.

1. Ámbito de la inteligencia artificial, técnicas, algoritmos, modelos base, entre otros.

Análisis musical

Qué es el sonido

La palabra “sonido” en física, se refiere a la propagación de ondas a través de un fluido**,** estas ondas (para el estudio de la física) pueden ser audibles o no, aunque en general nos referimos al “sonido” como algo que podemos oír. Los sonidos que escuchamos se producen cuando una onda sonora produce una oscilación en la presión del aire y esta oscilación llega a nuestro oído.

Frecuencia

La frecuencia refiere a la cantidad de veces por segundo que oscila una onda. Esta oscilación se mide en Hertz (o “Hercios”), lo cual se abrevia como “Hz”, es así que se dice que una onda que oscila 440 veces por segundo, es una onda de 440 Hz. Si la oscilación es lenta, tendremos sonidos más graves, si es más rápida, percibimos sonidos más agudos.

Tono

Llamamos “tono” a la sensación auditiva que producen los sonidos en nuestro cerebro, los tonos pueden medirse por frecuencias. Para simplificar el estudio y ejecución de la música se les ha dado nombre a determinadas frecuencias y son las que llamamos “notas musicales”.

Armónicos

Un armónico, en música, es un sonido complementario que sale junto con una nota fundamental y que es múltiplo de esa nota fundamental.

La mayoría de sonidos que escuchamos son complejos, no están comprendidos de una sola frecuencia, sino de muchas que viajan simultáneamente. Una nota musical emitida por un instrumento, o por nuestra voz, no se compone sólo de la frecuencia de esa nota musical, sino de varias otras que salen simultáneamente, a las que llamamos “frecuencias armónicas” o, simplemente, “armónicos”.

Para cada sonido tendremos una frecuencia fundamental y frecuencias armónicas, que son siempre múltiplos de la frecuencia fundamental. Por ejemplo, el C2 de un piano tiene una frecuencia de 66 Hz, pero junto con ese C de 66 Hz, van a sonar simultáneamente, aunque con menor intensidad, un C3 de 132 Hz (66×2), un G3 de 198 Hz (66×3), un C4 de 264 Hz (66×4), un E4 de 330Hz (66×5), un G4 de 396 Hz (66×6), y así sucesivamente. Dependiendo de diferentes factores, al tocar una nota pueden sonar mayor o menor cantidad de armónicos y con mayor o menor intensidad.

Tempo

El tempo en la música se refiere a la velocidad o ritmo con el que se ejecuta una pieza musical. Se mide en pulsaciones por minuto (BPM, por sus siglas en inglés), donde un número mayor de BPM indica un ritmo más rápido, y un número menor indica un ritmo más lento.

Por ejemplo:

60 BPM significa que hay 60 pulsaciones (o beats) por minuto, equivalente a un segundo por pulsación.

120 BPM indica un ritmo más rápido, con dos pulsaciones por segundo.

El tempo se suele indicar al inicio de una partitura musical con términos en italiano, como:

Largo: muy lento.

Adagio: lento.

Andante: a un ritmo moderado.

Allegro: rápido.

Presto: muy rápido.

Espectrograma

Un espectrograma es una representación visual de la intensidad de las frecuencias en una señal de audio a lo largo del tiempo. En el contexto del análisis de audio, un espectrograma muestra cómo las frecuencias (o tonos) presentes en una señal de sonido cambian con el tiempo. Los espectrogramas son útiles para entender la estructura y características de un sonido, como la identificación de patrones rítmicos o armónicos, que pueden ser claves para tareas de análisis de audio como reconocimiento de voz, análisis musical, entre otros.

Ejes de un espectrograma:

Eje X (horizontal): Representa el tiempo.

Eje Y (vertical): Representa la frecuencia.

Colores o Intensidad: Representan la magnitud o amplitud de las frecuencias, donde los colores más brillantes o cálidos indican mayor intensidad.

Tipos de espectrogramas de audio:

Espectrograma de magnitud o lineal:

Representa la magnitud de las frecuencias en una escala lineal.

Se obtiene aplicando la Transformada de Fourier de corta duración (STFT) a la señal de audio.

Las frecuencias se muestran en una escala lineal, lo que significa que las distancias entre las frecuencias son equidistantes.

Espectrograma en escala logarítmica (log-spectrograma):

Las frecuencias se representan en una escala logarítmica.

Es útil para analizar señales de audio donde las frecuencias más bajas tienen mayor importancia perceptual, como en la música o la voz humana.

A menudo, la intensidad o magnitud también se puede mostrar en escala logarítmica, para mejorar la visualización de sonidos débiles.

Espectrograma de potencia:

Similar al espectrograma de magnitud, pero en lugar de la magnitud, representa la potencia de las frecuencias, lo que es el cuadrado de la magnitud.

Resalta más claramente las diferencias en las intensidades de las distintas frecuencias.

Mel-spectrograma:

Utiliza la escala Mel para representar las frecuencias, la cual se basa en cómo el oído humano percibe el sonido. En la escala Mel, las frecuencias más bajas tienen más resolución.

Es comúnmente utilizado en procesamiento de voz y en tareas de análisis de audio con redes neuronales, debido a su relación más directa con la percepción humana.

Notas musicales por su duración

Las notas musicales se clasifican según su duración, es decir, el tiempo que deben sonar en una pieza musical. Los tipos principales son:

Redonda : Dura 4 tiempos.

Blanca : Dura 2 tiempos, la mitad de una redonda.

Negra : Dura 1 tiempo, la mitad de una blanca.

Corchea : Dura 1/2 tiempo, la mitad de una negra.

Semicorchea : Dura 1/4 de tiempo, la mitad de una corchea.

Delimitación de muestras y muestras

Escalas

Las escalas que se usaran será referente a la escala cromática que conta de 12 notas musicales y se delimitara las notas a sus 3 principales segmentos que son la escala mayor ,escala normal y escala menor dando asi un total de 36 notas a ser clasificadas

Muestreo

Tiempo de cada muestra

El tiempo de cada muestra o frame se hará con respecto al tempo y la duración de la nota mas rápida que en nuestro caso será da corchea a un tempo de 120 bps que seria una canción rápida esto para no saltarse notas por error, el avance se hará no un salto de 50 milisegundos

Tamaño de la muestra

El tamaño de la muestra esta definida en la resolución del audio y su formato( Mono - Estereo) , para el proyecto se opto por transformar el audio a mono esto porque lo que oímos seria la suma de ambos audio para una maquina, y la resolución se refiere a la calidad del audio que es la cantidad de muestras por segundo que tiene un audio , para este proyecto se uso audios con 44.1 kherz de calidad de audio lo que significa que se tendrán 44100 datos por segundo de audio.

Espectrograma

Para el Proyecto se decidio hacer un espectrograma normal que aunque consume mayor

Cantidad de datos mejorara la respuesta pues las muestras son muy cortas después de usar la función espectrogram de pytorch se obtuvo un espectrograma :

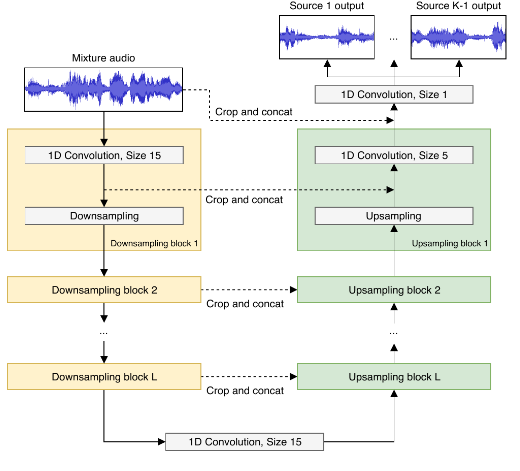


1. COMPONENTE IA:

Weve U-Net

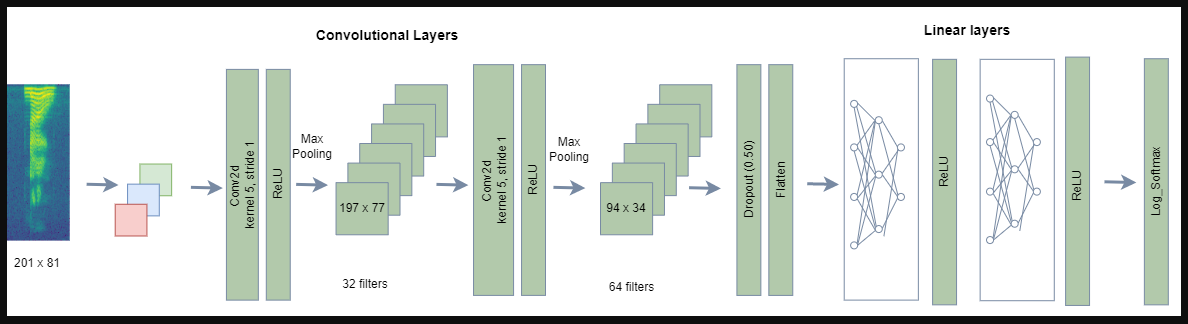
Wave-U-Net es una red neuronal convolucional aplicable a tareas de separación de fuentes de audio, que trabaja directamente sobre la forma de onda de audio sin procesar.

Wave-U-Net es una adaptación de la arquitectura U-Net al dominio temporal unidimensional para realizar la separación de fuentes de audio de extremo a extremo. A través de una serie de bloques de submuestreo y sobremuestreo, que implican convoluciones 1D combinadas con un proceso de submuestreo y sobremuestreo, se calculan las características en múltiples escalas/niveles de abstracción y resolución temporal, y se combinan para hacer una predicción.



Red CNN

Se uso una red CNN para analizar los espectrogramas de los frames de las notas para que este aprendiese a diferenciarlos, este fue entrenado con 50 epochs y un learning rate de 0.001, su acuraci fue del 92.01% y el dataset estaba formado por 450000 espectrogramas



Adquisición de datos:

Para la red webe-Unet se recurrió a un dataset de una universidad estadounidense, este dataset se llama musdb18 y tiene varias canciones en formato mp3 para dividir sus pistas en frecuencias y asi entrenar la red con estas pistas por instrumentos.

Para la CNN se uso un dataset llamado tyniSOL la cual es un dataset de notas tocadas y clasificadas por la universidad de Stanford en su departamento de música, estas tienen una duración arbitraria y también no están bien recortadas , es por ello que se requirió un preprocesamiento el cual fue:

1. Primero se debe convertir los audios a formato wav y a 44.1khz de calidad en formato mono
2. Se debe dividir el audio en frames de 125 milisegundos con un step\_hop de 50 milisegundos
3. Este audio de debe convertir en un espectrograma el cual podrá ser usado por la red

Métricas:

El primer modelo trabaja con su propio código para hacer fine tuning y te entrega el modelo pre-entrenado y tu solo le pasas los datos pero no te devuelve un resultado que precisión,

El segundo modelo tubo un 92% de aciertos.

Técnicas de depuración:

Se uso las técnicas para evitar el sobre ajuste y la visualización de resultados de pytorch para poder corregir posibles fallas

1. DESCRIPCION DEL TRABAJO REALIZADO:
2. Arquitecturas frameworks y componentes:

Se uso pytorch para los modelos de inteligencia artificial, se uso pydub y librosa para el preprocesamiento de los datos.

1. Hardware:

Se uso una computadora (laptop) con un Intel i5 de de la generación 13 con un tarjeta de video rtx3050

1. Herramienta:

Se usaron