UANL
UNIVERSIDAD AUTÓNOMA DE NUEVO LEÓN

FCFM
FACULTAD DE CIENCIAS FÍSICO MATEMÁTICAS

Autoencoders y GANs en proyecto de procesamiento de sonido
Procesamiento y Clasificación de Datos

Dr. Mayra Berrones Reyes

Lic. Hugo Arnoldo Oliva Castillo

Primeramente, se describirá brevemente que son los Generative Adversarial Networks (GAN) y los Autoencoders, ambos modelos de machine learning que sirven para reconstruir/generar datos sintéticos que podrían servir para procesar datos faltantes en sonido o corregir audios con mucho ruido de fondo. Así mismo, se hablará de sus aplicaciones en el presente trabajo (procesamiento de audios para clasificar la emoción del sujeto que habla, ya sea feliz o triste)

Generative Adversarial Networks (GAN):

Los Generative Adversarial Networks (GAN) son un tipo de modelo de aprendizaje profundo que consiste en dos redes neuronales: un generador y un discriminador. El generador toma ruido aleatorio como entrada y genera datos sintéticos, en este caso, audio. El discriminador evalúa si un fragmento de audio es real o generado. Estas dos redes compiten en un proceso de entrenamiento: el generador busca mejorar su capacidad para crear audio realista, mientras que el discriminador busca volverse más efectivo en la detección de datos falsos. Los GAN son útiles en trabajos de audio para la generación de música, la creación de efectos de sonido auténticos y la síntesis de audio realista.

Los GAN como se acaba de describir, sirve para crear contenido de audio muy realista a partir de ciertas muestras. Esto es útil cuando un audio es de muy pobre calidad/hay mucho ruido de fondo adicional y no es muy audible/entendible las frases que se están diciendo; o también cuando se quiere generar un audio a partir de la voz de alguien, pudiéndose usar como “fake news”. No hay tanta necesidad factible para usar dicha técnica en el presente proyecto ya que los audios están demasiado entendibles lo que las personas están diciendo (en inglés), y no se desea trabajar con el contenido de la oración en sí, sino de como lo dice, en que tono, amplitud de onda, para así saber que emoción es la que influye más (no es lo mismo como dice una oración alguien que está feliz, a alguien que esta triste, el primero podría decirlo en voz más alta, el segundo en tono más bajo, y las ondas se verían impactadas por ello) en como dice el audio (se puede representar en espectograma).

Pero aterrizando dicha técnica en nuestro proyecto, para trabajos futuros se podría utilizar el GAN replicando frases que ciertas personas dijeron en las muestras para así simular la manera en que ciertas personas en específico dirían una frase en particular (por lo tanto, datos “fake”, pero simulando muy convincentemente el como una persona en determinada emoción diría la frase en la vida real). Resumiendo, se podría utilizar para fines de simulación de las emociones de las personas en caso de que dichas personas no estén

presentes a la hora de pronunciar las frases. De igual manera, podría servir en caso de que ciertas muestras contengan mucho ruido adicional de fondo y no se percibe con una calidad suficientemente la pronunciación de la frase de la persona, por lo que se podría simular el cómo lo dijo.

Autoencoders:

Los autoencoders son modelos de aprendizaje profundo diseñados para reducir la dimensionalidad de los datos y extraer características significativas. En el contexto del audio, un autoencoder toma una señal de audio y la transforma en una representación de dimensionalidad reducida, a través de un codificador. Luego, el decodificador reconstruye la señal de audio a partir de esa representación reducida. Esto puede ser útil en aplicaciones de audio como la compresión, donde se busca reducir el tamaño de las señales de audio, o en la eliminación de ruido, donde se intenta mejorar la calidad del audio al eliminar componentes no deseados. Además, los autoencoders también pueden utilizarse para generar nuevas señales de audio a partir de características latentes aprendidas.

Los autoencoders podrían ser más útiles en nuestro proyecto debido a que hay audios que duran un poco más de 8 segundos y como son muchos audios en términos de procesamiento tal vez valdría la pena comprimir los archivos sin perder muchas características relevantes de los audios. Asimismo, se podría utilizar al mismo tiempo para eliminación del ruido de fondo de los audios donde las personas pronuncian las frases (en nuestro caso no es tan necesario debido a que los audios fueron construidos para elaborar muestras especificar de personas pronunciando oraciones, por lo que no hay mucho ruido de fondo); si es que hubiera alguna muestra en la que contiene mucho ruido de fondo que no permita una audición correcta.

Fuente: material del curso.