Universidad del Valle de Guatemala Pablo Escobar 20936
NLP , sección 10

Laboratorio 7

1. **¿Qué es Wav2Vec2 y qué problema intenta resolver?** Wav2Vec2 es un modelo avanzado desarrollado por Facebook AI para el reconocimiento automático del habla. Se basa en el aprendizaje auto-supervisado, lo que le permite aprender representaciones de audio a partir de datos sin etiquetar. Tradicionalmente, los sistemas de reconocimiento de voz requieren grandes cantidades de datos transcritos, lo cual es un desafío para muchos idiomas del mundo debido a la falta de datos etiquetados. Wav2Vec2 aborda este problema al utilizar grandes volúmenes de audio no etiquetado durante el preentrenamiento, lo que reduce significativamente la necesidad de datos etiquetados. Este enfoque permite el desarrollo de sistemas de reconocimiento de voz en entornos con recursos limitados de datos etiquetados, facilitando el uso del modelo para una mayor diversidad de idiomas y dialectos.

2. **¿Cómo se diferencia Wav2Vec2 de los métodos tradicionales de reconocimiento de voz?** Wav2Vec2 se diferencia de los métodos tradicionales en varias formas clave. Primero, los métodos convencionales de reconocimiento de voz suelen requerir una gran cantidad de datos de audio etiquetados para entrenar un modelo de manera efectiva. Esto es particularmente complicado para lenguas menos comunes, donde puede no existir una cantidad suficiente de datos etiquetados. En contraste, Wav2Vec2 utiliza un enfoque de aprendizaje auto-supervisado, lo que le permite aprender a partir de grandes cantidades de datos no etiquetados. El modelo se preentrena en datos no etiquetados y luego se ajusta con una pequeña cantidad de datos etiquetados. Además, a diferencia de los enfoques tradicionales que separan los modelos acústicos y de lenguaje, Wav2Vec2 usa una arquitectura Transformer, que permite aprender representaciones de contexto más ricas y completas de los datos de audio.

3. **¿Cuáles son los componentes principales de la arquitectura de Wav2Vec2?** La arquitectura de Wav2Vec2 se compone de varios bloques esenciales:
    a. **Codificador de características**: Es una red convolucional que transforma el audio crudo en una secuencia de representaciones latentes. El audio pasa a través de capas convolucionales que extraen características temporales, proporcionando una representación compacta del audio en distintas etapas temporales.
    b. **Cuantización**: En lugar de trabajar únicamente con representaciones continuas, Wav2Vec2 introduce una etapa de cuantización que transforma estas representaciones en unidades discretas. Este proceso es clave para el aprendizaje auto-supervisado, ya que proporciona las "unidades de referencia" que el modelo debe aprender a identificar durante el preentrenamiento.
    c. **Transformador**: Las representaciones latentes pasan por un modelo Transformer que utiliza mecanismos de atención para generar representaciones contextuales, capturando relaciones a largo plazo en la

secuencia de audio. Esta parte del modelo permite que las representaciones no solo reflejen información local, sino también el contexto global del audio.

d. **Capa de proyección**: Una vez que el modelo ha sido preentrenado, se ajusta para tareas de reconocimiento de voz utilizando una capa de proyección que predice los tokens de salida, como fonemas o caracteres, dependiendo del sistema de etiquetado.

4. **¿Puedes explicar el proceso de entrenamiento de Wav2Vec2? ¿Cómo funciona el aprendizaje auto-supervisado en este contexto?** El entrenamiento de Wav2Vec2 se lleva a cabo en dos fases: preentrenamiento y ajuste fino. En la primera fase, el modelo se preentrena en datos de audio no etiquetados utilizando una técnica de enmascaramiento y aprendizaje contrastivo. Durante el preentrenamiento, partes del audio (representaciones latentes) se enmascaran, y el modelo aprende a predecir estas partes enmascaradas seleccionando la representación correcta de un conjunto de opciones, que incluyen distractores. Este proceso se basa en un objetivo de contraste, donde el modelo debe identificar la representación correcta frente a opciones incorrectas. Para que este proceso sea efectivo, se utiliza la cuantización, que convierte las representaciones continuas en discretas, lo que facilita el aprendizaje. El modelo también utiliza una pérdida de diversidad para asegurarse de que las representaciones discretas se utilicen de manera uniforme. Una vez que el modelo ha aprendido buenas representaciones del audio en la etapa de preentrenamiento, se ajusta con una pequeña cantidad de datos etiquetados para realizar tareas específicas como el reconocimiento de voz.

5. **¿Qué papel juega la cuantización de características de audio en Wav2Vec2?** La cuantización es una parte crucial en Wav2Vec2, ya que convierte las representaciones continuas generadas por el codificador de características en unidades discretas. Este proceso de cuantización se lleva a cabo mediante la selección de una combinación de entradas de múltiples "codebooks", que permiten representar el audio en un conjunto finito de posibles unidades. Este proceso facilita el aprendizaje auto-supervisado porque genera las representaciones que el modelo necesita aprender a predecir durante la tarea contrastiva. En lugar de predecir valores continuos, el modelo aprende a predecir una unidad discreta, lo que hace que el entrenamiento sea más robusto y permite que las representaciones discretas capturen características relevantes de la señal de audio que son útiles para el reconocimiento de voz.

6. **¿Cuáles son las ventajas y desventajas de usar Wav2Vec2 en comparación con otros modelos de reconocimiento de voz?** Las ventajas de Wav2Vec2 incluyen su capacidad para funcionar con una cantidad significativamente menor de datos etiquetados, lo que es beneficioso para lenguas y dialectos con recursos limitados.

Laboratorio 7

El uso de aprendizaje auto-supervisado también permite que el modelo aproveche grandes cantidades de datos no etiquetados, mejorando el rendimiento en tareas de reconocimiento de voz sin depender de un corpus transcrito masivo. Además, el uso de un modelo Transformer permite una captura más eficiente del contexto global en los datos de audio, lo que contribuye a su precisión.

Sin embargo, una de las principales desventajas de Wav2Vec2 es la necesidad de una gran cantidad de datos no etiquetados para el preentrenamiento, lo que requiere recursos computacionales intensivos. Además, aunque el modelo reduce la necesidad de datos etiquetados, el ajuste fino aún requiere algunos datos etiquetados para alcanzar un rendimiento óptimo. Finalmente, el costo computacional del entrenamiento y la inferencia con modelos basados en Transformers puede ser elevado en comparación con métodos tradicionales, lo que puede ser una barrera en ciertos escenarios con limitaciones de hardware.