UNCUYO
UNIVERSIDAD NACIONAL DE CUYO

FACULTAD DE INGENIERÍA

Inteligencia Artificial II

Anteproyecto

Demix: Análisis comparativo entre U-Nets y Vision Transformers para la separación de fuentes de audio

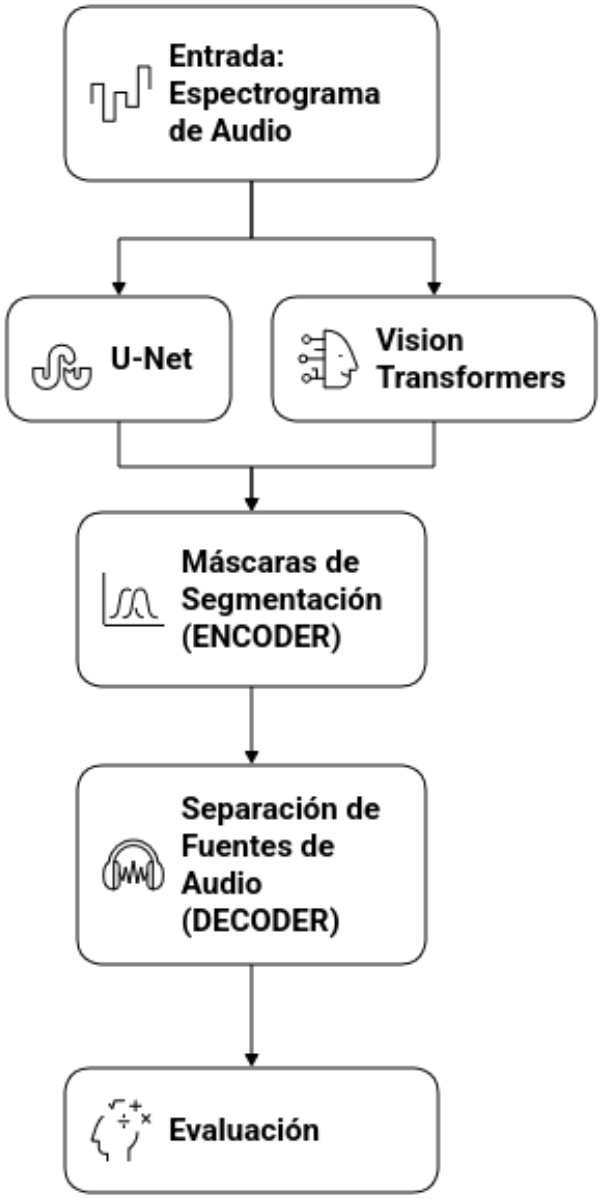
Gabriel Lopez Romero

November 11, 2025

Resumen

El presente proyecto, Demix, propone comparar la efectividad de las U-Nets y los *Vision Transformers* (ViT) para la tarea de separación de fuentes de audio (ASS). Las U-Nets han sido, históricamente, el estándar *de facto* en ASS, debido a que su tarea original (segmentación de imágenes médicas) es análoga al objetivo de este problema: la generación de máscaras que permitan segmentar el espectrograma de una pista de audio en sus fuentes constitutivas (p. ej., voz, batería y bajo).
El objetivo principal de este proyecto es desafiar esta dominancia utilizando *Vision Transformers* Dosovitskiy et al. (2020). La hipótesis de este trabajo es que la capacidad de los *Transformers* para capturar contexto global mediante mecanismos de autoatención solventará la principal desventaja de las convoluciones: su naturaleza inherentemente local. Esta limitación dificulta que los modelos convolucionales relacionen patrones temporales distantes, como un motivo rítmico al inicio de una canción con su repetición al final.
En ambos modelos, la entrada será la representación tiempo-frecuencia de una canción, es decir, el espectrograma, y la salida serán cuatro máscaras de segmentación (una por cada fuente: voz, batería, bajo y "otros"). Estas máscaras serán utilizadas para generar estimaciones de las señales aisladas de cada una de dichas fuentes. En consecuencia, el problema se enmarca en el aprendizaje supervisado y puede abordarse como una tarea de regresión. Para el entrenamiento se utilizará el *dataset* MUSDB18 Rafii et al. (2017) y la evaluación se realizará mediante la métrica estándar *Signal-to-Distortion Ratio* (SDR).

Diagrama de Cajas



References

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J. and Houlsby, N. (2020), 'An image is worth 16x16 words: Transformers for image recognition at scale', *CoRR* **abs/2010.11929**.
URL: *https://arxiv.org/abs/2010.11929*

Rafii, Z., Liutkus, A., Stöter, F.-R., Mimilakis, S. I. and Bittner, R. (2017), 'The MUSDB18 corpus for music separation'.
URL: *https://doi.org/10.5281/zenodo.1117372*