ETH
Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

Distributed Computing

Prof. R. Wattenhofer

Music Captioning

In this thesis, we address a key issue in the generative music domain: the lack of descriptive captions for music. Text-to-Image models, such as DALL-E, MidJourney, and Stable Di usion have been trained on billions of image-caption pairs. In the music domain, however, the largest music-caption dataset consists of only 5.5k samples.

We propose to develop a machine learning model capable of generating accurate and concise music captions, encapsulating elements like genre, style, and mood. To achieve this, we will leverage state-of-the-art models such as EnCodec, CLAP, and various LLMs, while borrowing methodologies from successful image captioning models, such as BLIP, to adapt them to the audio domain. We anticipate the primary challenge to be the shortage of music-caption pair data. Consequently, this thesis will not only focus on creating an e ective music captioning model, but also on generating a larger dataset for music-caption pairs.

Requirements: Knowledge in Python and Machine Learning. Experience with LLMs and Pytorch is an advantage.

We will have weekly meetings to address questions, discuss progress and think about future ideas.

Contact

- Luca Lanzendorfer : lanzendoerfer@ethz.ch, ETZ G93