Project Deliverable D1.2

Team

Vladislav Kalinichenko v.kalinichenko@innopolis.university
Polina Korobeinikova p.korobeinikova@innopolis.university
Janna Ivanova j.ivanova@innopolis.university

Repository

https://github.com/poinka/NewWave.git

Project Topic

Natural Language Music Recommender System

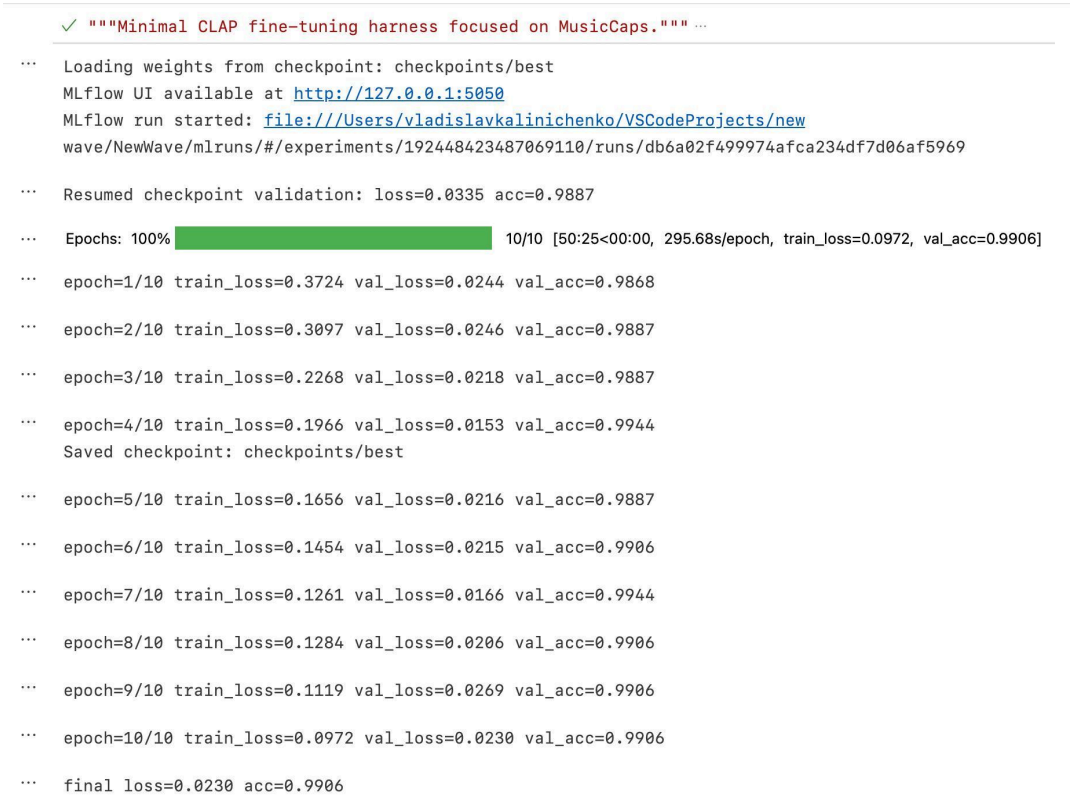
What has been done so far

- Prepared MusicCaps 10-second audio clips and built a simple 90/10 train–val split for CLAP training; about 5,300 of ~5,600 targeted clips were successfully cached due to some deletions and access limits on YouTube.
- Collected text data from two sources for the lyrics–description model: enrich-music4all (artist, tags, pseudo captions) and Song-Interpretation-Dataset (lyrics and interpretation comments) for pairing song text with metadata.
- Implemented and ran the CLAP training script on cached MusicCaps clips, logged metrics with MLflow, fine-tuned for 10 epochs, and saved the best validation checkpoint for use in the retrieval pipeline.
- Verified CLAP inference by ranking cached clips with a natural-language prompt to ensure the audio–text alignment behaves as expected on short segments.
- Polina tried using a cross-encoder, but it was too slow, so we decided to switch to a bi-encoder model instead.
- Built the lyrics–description bi-encoder with separate text encoders, started training, observed loss decreasing on early runs, and set up evaluation to report Recall@K and MRR on the held-out split.

Results

- **System overview**
  The system combines CLAP for matching audio with text, a lyrics encoder to capture song meaning, and a lightweight rewriter that turns free text into attributes the models can use, with performance to be checked by Recall@K and MRR on held-out data.
- **Lyrics–description model**
  A simple bi-encoder is used: one text encoder reads descriptions and another reads lyrics, and the model learns to bring matching pairs closer and push non-matches apart, with training already underway and validation set up for retrieval metrics later.
  Unfortunately, InnoDataHub did not provide us points to train that model for now, but we plan to finish post training in a couple of days
- **CLAP pipeline**
  CLAP is used as the backbone on MusicCaps 10-second clips to align audio with captions, showing strong baseline behavior before a short fine-tune and best-checkpoint selection for the rest of the system.

Proof for training



- **Datasets**

  We used two data sources: MusicCaps audio clips for CLAP training, and two public datasets to pair song text with music metadata for retrieval and training

  CLAP audio (MusicCaps)

  MusicCaps entries with timestamps were turned into local 10-second audio clips where accessible, providing a reliable train/validation split that matches the CLAP training setup and supports later retrieval tests.

  Text for lyrics–description model

  [seungheondoh/enrich-music4all]

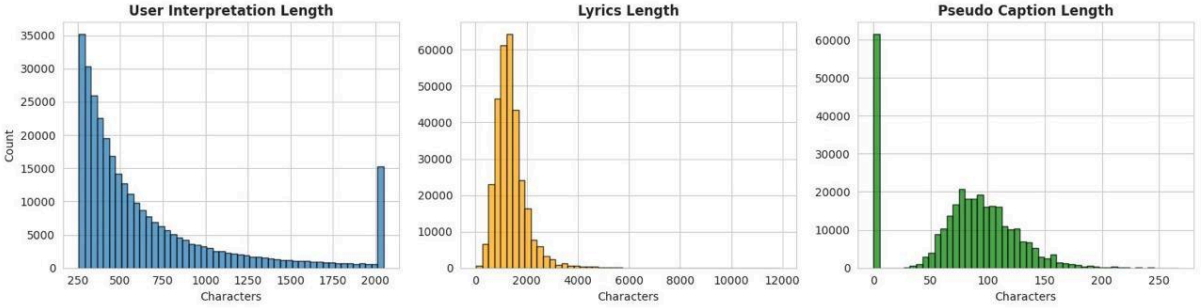
  A large metadata and text set for music retrieval with 108,363 rows and fields like track_id, tag_list, pseudo_caption, title, artist_name, release, m4a_genres, and m4a_tags. In training, artist_name and tag_list are used to sample hard negatives (tracks by the same artist or sharing tags), while pseudo_caption and user interpretation descriptions are used as the text describing each track.

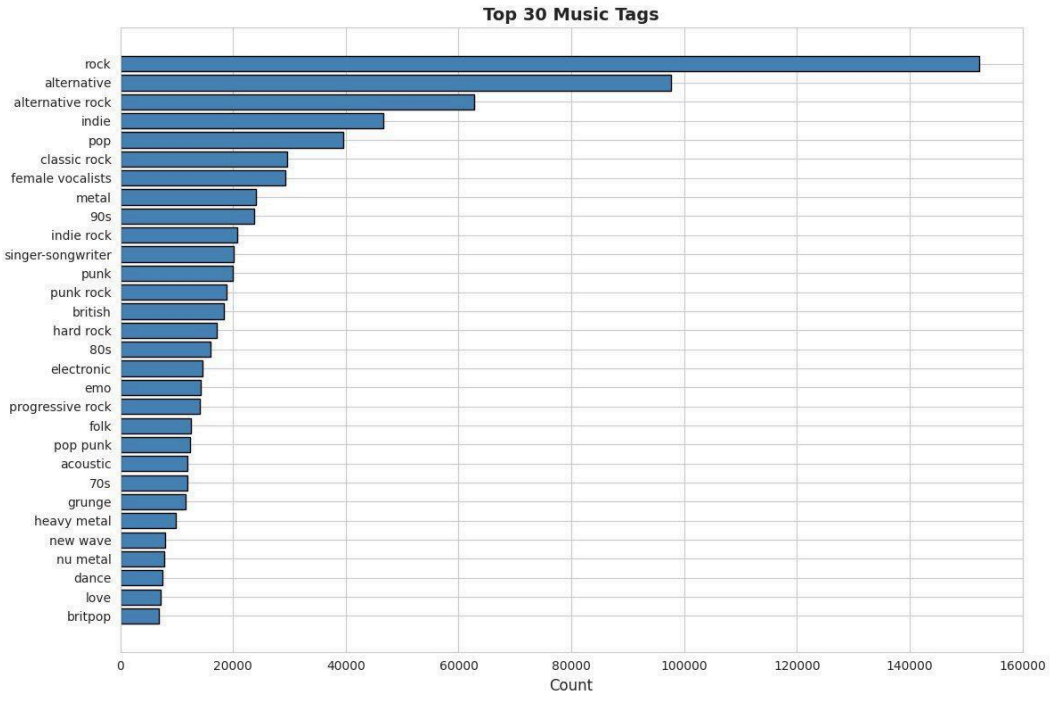
  [jamimulgrave/Song-Interpretation-Dataset]

  A dataset that links lyrics and user interpretations to Music4All via music4all_id; in our setup, each track's lyrics are matched against a combined description made from pseudo captions (from enrich-music4all) plus user interpretations to train and evaluate the lyrics–description bi-encoder.

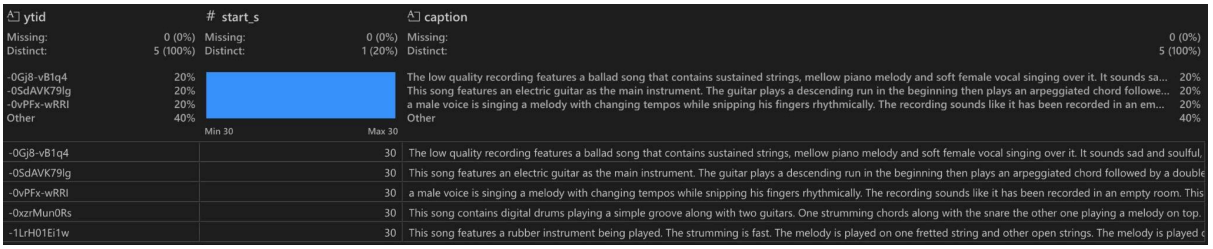
Screenshots from Exploratory Data Analysis:

```
============================================================
FINAL DATASET SUMMARY:
============================================================
Total tracks: 310315
Unique artists: 3241

Average lengths:
  - User interpretation: 666 chars
  - Lyrics: 1362 chars
  - Pseudo caption: 78 chars

Description availability:
  - Has both descriptions: 248784 (80.2%)
  - Has pseudo_caption only: 0 (0.0%)
  - Has user_interpretation only: 61531 (19.8%)

Tag statistics per split:
  - Train: 894761 total tags
  - Val:   147449 total tags
  - Test:  147151 total tags
```





That’s how CLAP’s dataset looks like:



ytid	start_s	caption
Missing: 0 (0%) Distinct: 5 (100%)	Missing: 0 (0%) Distinct: 1 (20%)	Missing: 0 (0%) Distinct: 5 (100%)
-0Gj8-vB1q4 20% -0SdAVK79lg 20% -0vPFx-wRRI 20% Other 40%	Min 30 Max 30	The low quality recording features a ballad song that contains sustained strings, mellow piano melody and soft female vocal singing over it. It sounds sa... 20% This song features an electric guitar as the main instrument. The guitar plays a descending run in the beginning then plays an arpeggiated chord followe... 20% a male voice is singing a melody with changing tempos while snipping his fingers rhythmically. The recording sounds like it has been recorded in an em... 20% Other 40%
-0Gj8-vB1q4	30	The low quality recording features a ballad song that contains sustained strings, mellow piano melody and soft female vocal singing over it. It sounds sad and soulful,
-0SdAVK79lg	30	This song features an electric guitar as the main instrument. The guitar plays a descending run in the beginning then plays an arpeggiated chord followed by a double
-0vPFx-wRRI	30	a male voice is singing a melody with changing tempos while snipping his fingers rhythmically. The recording sounds like it has been recorded in an empty room. This
-0xzrMun0Rs	30	This song contains digital drums playing a simple groove along with two guitars. One strumming chords along with the snare the other one playing a melody on top.
-1LrH01Ei1w	30	This song features a rubber instrument being played. The strumming is fast. The melody is played on one fretted string and other open strings. The melody is played

And It’s preview:

Work distribution

Polina

Wrote and ran the training code for the lyrics–description part, saw the loss go down in early runs, and continues training and integration work as planned earlier.

Vladislav

Built the audio/text path on top of CLAP and used the MusicCaps dataset with 10-second clips to match the model’s setup, then did a short fine-tuning and kept the best validation checkpoint for use in the pipeline.

Janna

Prepared datasets, downloaded and cut raw audio segments for MusicCaps entries from YouTube where possible, and compiled the report text for this submission.

Plan for the Next Weeks

Vlad plans to extend CLAP to handle longer musical contexts beyond short clips to better capture song-level structure and dynamics.
Polina plans to finish training, validate and fix errors if any, and integrate components into a cohesive pipeline.
Janna plans to prepare two other datasets to expand the context of CLAP's work.
Later we plan to try a new architecture with three separate encoders for music, lyrics, and textual descriptions, where embeddings from all three are mapped into the same dimensional space and compared directly there.