Aligning Audio-Lyric Embedding Space

using contrastive fine-tuning on CLAP

Aimon Benfield-Chand • Javon Hickmon • Donovan Clay

PAUL G. ALLEN SCHOOL OF COMPUTER SCIENCE & ENGINEERING

Introduction

Problem Statement:

We investigate using emotional sentiment as means of aligning an audio-lyric song embedding model.

Applications:
- Music Information Retrieval (MIR)
- Music Emotion Recognition (MER)
- Sentiment-conditioned song recommendation
- Prompt-based playlist creation
- Music Generation
  - Text-conditioned audio generation
  - Audio-conditioned lyric generation

Contributions:
- Robustness & generalizability of emotion
- Open-vocabulary encoder
- Versatile representation learning

Related Works

MER Models
- Lyric-based approaches: LLMs
- Audio-based approaches: AST [5]
- Fusion approaches [2,7,9,11]

Cross-modal Embedding Models
- CLAP [4], MuLan [6]
- Wav2CLIP [10]
- Spotify Multilingual Audio-Lyric Synchronization [3]

Music Generation Models
- Meta's MusicLM (uses MuLan) [1]

Dataset

CLAP Dataset: LAION-Audio-630K
- 128,010 audio and text pairs
- FSD50k, ClothoV2, AudioCaps, MACS

DALI [8]: Synchronized Audio, Lyrics, & Notes
- 5358 songs of real music
- Metadata for song genre, language, and musician
- YouTube link to audio MP4
- time-aligned lyrics at four levels of granularity: note, word, line and paragraphs

Preprocessing

DALI Pre-Processing:
1. Group lines of lyrics into "lyric segments", each corresponding to ~10 seconds of audio
2. Use Chat-GPT-3.5 to generate a sentence text description of each lyric segment
3. Download audio MP4 files from YouTube
4. Convert audio to WAV format
5. Generate log MEL-spectrograms

Methods

Model: Aligns audio and lyrics using emotional sentiment as a proxy
1. Generate a *text description* of the emotional sentiment of each song segment's lyric using Chat-GPT-3.5
2. Embed the *audio spectrogram* and *sentiment description* using CLAP
3. Finetune CLAP's audio encoder using contrastive loss

Ablation Baseline: No emotional sentiment → Align using lyrics directly
1. ~~Generate a text description of the emotional sentiment of each song segment's lyric using Chat-GPT-3.5~~
2. Embed the *audio spectrogram* and *lyrical segment* using CLAP
3. Finetune CLAP's audio encoder using contrastive loss

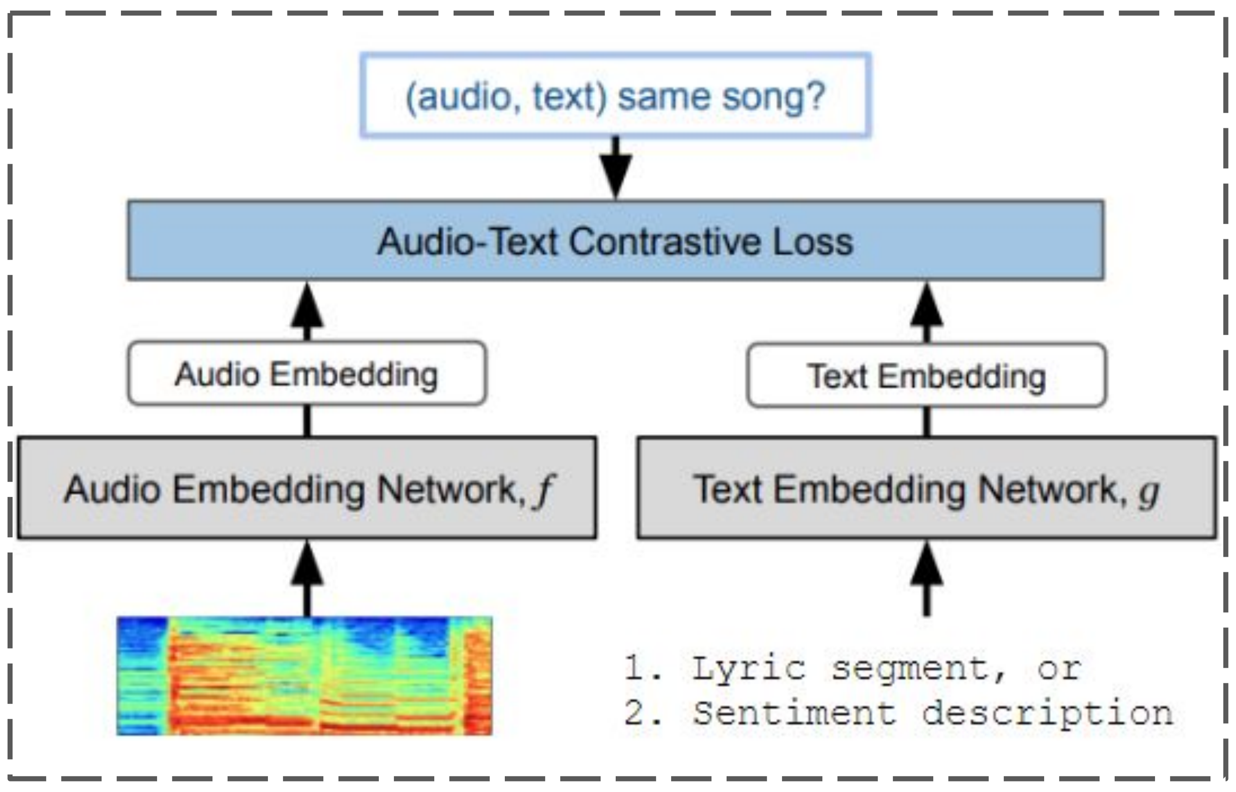


Figure 2. Contrastive Fine-Tuning Pipeline

CLAP Encoders

Audio: CNN14
- 80M parameters
- Embedding dim: 2048
- Pretrained on AudioSet-2M

Text: BERT
- 110M parameters
- Embedding dimension: 768
- [CLS] token as sentence embedding

CLAP embedding dim: 1024

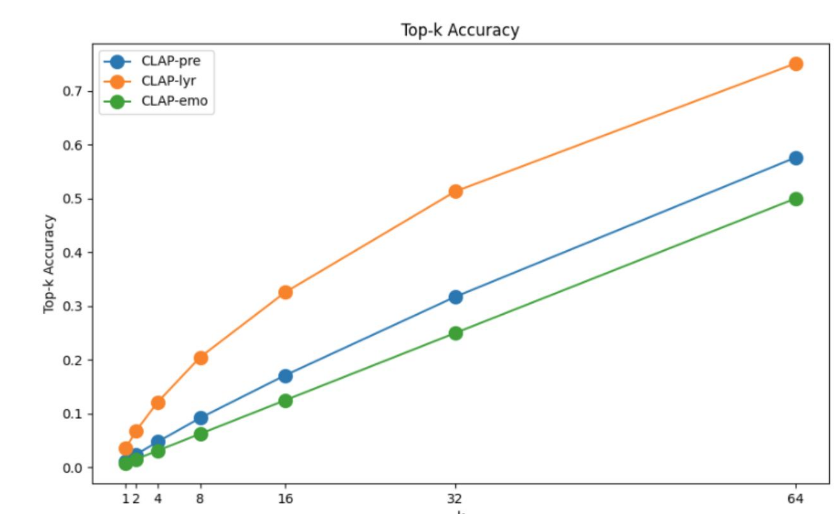
Experiments

Evaluation Task: Cross-modal audio-lyric Retrieval (MIR)
- Classify an audio query amongst a set of lyric segments
- Compare model with ablated baseline and non-fine-tuned model

Metrics:
1. Classification Accuracy
2. KL-Divergence: use cosine similarity between lyric embeddings as gold

Results

Baseline (pre-trained w/o sentiment)	KL-Divergence	Accuracy (top-1)
Baseline (fine-tuned w/ sentiment)	2.1878	3.49%
Model (fine-tuned w/ sentiment)	2.3137	3.12%



Discussion

- It's important to to note that certain training and evaluation decisions were made based on the project timeline for this course.
- Longer training times, more data, and a variety of evaluation metrics will be needed to fully confirm or deny our hypothesis.
- We decided to evaluate on cross-modal audio-lyric retrieval, but we also aim to evaluate performance for Music Emotion Recognition

References

[1] Andrea Agostinelli, Timo I. Denk, Zal´an Borsos, Jesse Engel, Mauro Verzetti, Antoine Caillon, Qingqing Huang, Aren Jansen, Adam Roberts, Marco Tagliasacchi, Matt Sharifi, Neil Zeghidour, and Christian Frank. Musiclm: Generating music from text, 2023.

[2] Remi Delbouys, Romain Hennequin, Francesco Piccoli, Jimena Royo-Letelier, and Manuel Moussallam. Music mood detection based on audio and lyrics with deep neural net, 2018.

[3] Simon Durand, Daniel Stoller, and Sebastian Ewert. Contrastive learning-based audio to lyrics alignment for multiple languages. *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5, 2023.

[4] Benjamin Elizalde, Soham Deshmukh, Mahmoud Al Ismail, and Huaming Wang. Clap learning audio concepts from natural language supervision. In *ICASSP 2023 – 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5, 2023.

[5] Yuan Gong, Yu-An Chung, and James Glass. Ast: Audio spectrogram transformer, 2021.

[6] Qingqing Huang, Aren Jansen, Joonseok Lee, Ravi Ganti, Judith Yue Li, and Daniel P. W. Ellis. MuLan: A joint embedding of music audio and natural language, 2022.

[7] Gaojun Liu and Zhiyuan Tan. Research on multi-modal music emotion classification based on audio and lyric. In *2020 IEEE 4th Information Technology, Networking, Electronic and Automation Control Conference (ITNEC)*, volume 1, pages 2331–2335, 2020.

[8] Gabriel Meseguer-Brocal, Alice Cohen-Hadria, and Geoffroy Peeters. Dali: A large dataset of synchronized audio, lyrics and notes, automatically created using teacher-student machine learning paradigm. 2018.

[9] Loreto Parisi, Simone Francia, Silvio Olivastri, and Maria Stella Tavella. Exploiting synchronized lyrics and vocal features for music emotion detection, 2019.

[10] Ho-Hsiang Wu, Prem Seetharaman, Kundan Kumar, and Juan Pablo Bello. Wav2clip: Learning robust audio representations from clip. In *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4563–4567, 2022.

[11] Junfeng Zhang, Lining Xing, Zhen Tan, Hongsen Wang, and Kesheng Wang. Multi-head attention fusion networks for multi-modal speech emotion recognition. *Computers Industrial Engineering,* 168:108078, 2022.