Clap Exploration and Fine Tuning

Dom Urso

What is CLAP?

- Dual-encoder: audio encoder + text encoder
- Trained with contrastive InfoNCE loss
- Embeds matching audio/text pairs together, pushes apart mismatches
- Enables unified audio–text semantic space

Zero Shot Capabilities

No task-specific training needed for new labels

Build prompts like “This audio is a jazz song.”

Encode prompts + audio; classify by cosine similarity

~90 % zero-shot accuracy on ESC-50

Dataset & Constraints

Free Music Archive: 100 K+ Creative-Commons tracks

Rich metadata: title, artist, genre, tags

GPU limits → use first 500 tracks, core 100 for experiments

Highlights small-data fine-tuning

Metadata Extraction

Flattened `tracks.csv` header to fields:

- title, artist name, genre_top, user tags

Built lookup map keyed by “title||artist” (lowercased)

Fast genre/tag retrieval during preprocessing

Prompt Engineering

Defined 5 varied templates, e.g.:

- ""{title}" by {artist}. Genre: {genre}. Tags: {tags}."
- "A {genre} track called "{title}" by {artist}."

At preprocess, randomly sample one template

Prevents over-fitting, improves generalization

Audio Preprocessing

Decode MP3 with FFmpeg → 48 kHz waveform

Pad/truncate to 5 s (240 k samples)

Store as `input_values` float32 array

Fine-Tuning Setup

Model: `laion/clap-htsat-unfused` + `ClapProcessor`

Split: 80 % train / 20 % val on 100 clips

Batch size 16, LR 3×10^{-5}, 3 epochs

Symmetric InfoNCE loss (audio→text + text→audio)

Training Results

Epoch 2 (best):

- Recall@1 = 0.35
- Recall@5 = 0.70
- Recall@10 = 0.85

Random baseline on 20 clips: R@1=0.05, R@5=0.25, R@10=0.50

Overfitting by Epoch 3 → use early stopping

Qualitative Examples

Clip 0 (“Relaxing”)

- Epoch 1: wrong Folk track
- Epoch 2: correct “Relaxing” (score 8.83)
- Epoch 3: overfitted to a different Folk track

Shows value of prompt variety & epoch selection

Retrieval Demo

Query: “Folk country music”

Top 5 matches:

1. “Ohio” by Alec K. Redfearn & the Eyesores (0.47)
2. “Castle Of Stars” by Ed Askew (0.45)
3. “Song For R” by Ed Askew (0.43)
4. “Inis Meain” by So Cow (0.40)
5. “This World” by AWOL (0.38)

Applications & Future Work

Natural-language search in streaming apps (Spotify, etc.)

Automated playlist generation by mood/genre

Sampling tools for producers (“warm lo-fi guitar”)

Next: scale to full FMA, integrate audio features, user-in-loop feedback

Acknowledgments

Thanks to the LAION team for open-sourcing CLAP

Thanks to the Free Music Archive creators for their dataset