# MuLan: A Joint Embedding of Music Audio and Natural Language

Qingqing Huang, Aren Jansen, Joonseok Lee, Ravi Ganti, Judith Yue Li, Daniel P. W. Ellis

ISMIR 2022 • Bengaluru, India

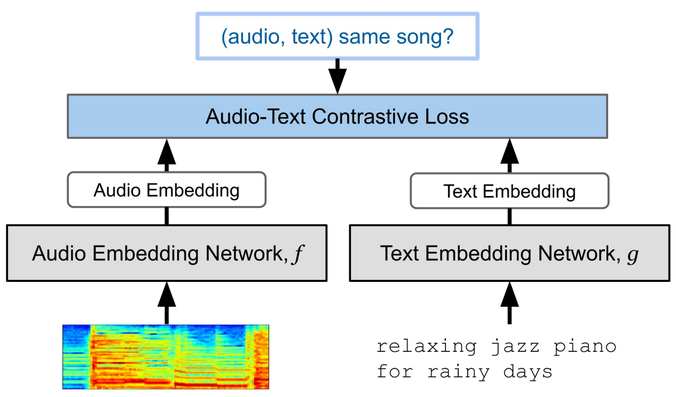
## ABSTRACT

This paper presents MuLan: a first attempt at a new generation of acoustic models that link music audio directly to unconstrained natural language music descriptions. MuLan takes the form of a two-tower, joint audio-text embedding model trained using 44 million music recordings (370K hours) and weakly-associated, free-form text annotations. Through its compatibility with a wide range of music genres and text styles (including conventional music tags), the resulting audio-text representation subsumes existing ontologies while graduating to true zero-shot functionalities. We demonstrate the versatility of the MuLan embeddings with a range of experiments including transfer learning, zero-shot music tagging, language understanding in the music domain, and cross-modal retrieval applications.

## INTRODUCTION

- **Goal**
  - Build a new audio foundation model that provides a natural language interface to support arbitrary music tagging and retrieval tasks
- **The Strategy**
  - Following success of image-text representation pretraining models (e.g. CLIP, ALIGN), we train a joint embedding of music audio and natural language using a very large collection of associated audio-text pairs mined from the internet

## LEARNING FRAMEWORK




- **Loss Function**

$$\sum_{i=1}^{B} -\log\left[\frac{h[f(\mathbf{x}^{(i)}), g(\mathbf{t}^{(i)})]}{\sum_{j\neq i} h[f(\mathbf{x}^{(i)}), g(\mathbf{t}^{(j)})] + h[f(\mathbf{x}^{(j)}), g(\mathbf{t}^{(i)})]}\right]$$

- **Audio Embedding Model Architecture**
  - Input: log-mel spectrogram of 10 second music clip
  - Encoder: Resnet-50, AudioSpectrogramTransformer
  - Output: 128-dimensional embedding
- **Text Embedding Model Architecture**
  - Warmstart from public BERT-base-uncased encoder

## TRAINING DATA

- Music audio: 370K hours of music videos
- Music descriptive text: Short-form and long-form labels. Optionally filtered to music related text
  - Examples

**Table 1**. Text annotation examples.

| Type | Examples |
|---|---|
| Short-form (SF) | tags like genre, mood, instrument, artist name, song title, album name |
| Long-form (LF) | 'Hip-hop features rap with an electronic backing.' 'The melody is so nostalgic and unforgettable.' |
| Playlist (PL) | 'Feel-good mandopop indie', 'Latin workout' 'Salsa for broken hearts', 'Piano for study' |

  - Statistics

**Table 2**. Statistics for text data sources. Tokens counts (in billions) are across all 44M videos. APV is the average number of text annotations (i.e. separate free-form strings) per video, including those with none.

| | Pre-filter | | Post-filter | |
|---|---|---|---|---|
| **Type** | Tokens (B) | APV | Tokens (B) | APV |
| Short-form | 31.2 | 42.9 | 5.4 | 29.6 |
| Long-form | 30.7 | 70.7 | 0.2 | 0.4 |
| Playlists | - | - | 2.5 | 24.3 |

## RESULTS

### MUSIC TAGGING

- **Eval sets** (not included in training data)
  - Music splits of AudioSet; MagnaTagATune
- **Zero-shot tagging**
  - Score(track, text label) = Cosine similarity(audio emb, text emb)
  - For each track, rank the scores for all labels
- **Linear probe**
  - Train linear classifiers using frozen audio embeddings as inputs
- **Result**

| Model | AudioSet Gen-25 | AudioSet Mu-141 | MTAT Top-50 | MTAT All-188 |
|---|---|---|---|---|
| **(a) Zero-shot** (Trained w/ ASET + SF + LF + PL) | | | | |
| M-AST | 0.840 | 0.909 | 0.778 | **0.776** |
| M-Resnet-50 | 0.840 | 0.899 | **0.782** | 0.772 |
| **(b) Text ablation** (using M-Resnet-50 Zero-shot) | | | | |
| ASET + SF + LF | 0.839 | 0.907 | 0.760 | 0.756 |
| ASET + SF | 0.839 | 0.885 | 0.754 | 0.747 |
| ASET | **0.886** | **0.942** | 0.753 | 0.771 |
| SF/LF Unfiltered | 0.845 | 0.908 | 0.774 | 0.766 |
| **(c) Linear probe** | | | | |
| M-AST | 0.906 | **0.942** | 0.925 | 0.953 |
| M-Resnet-50 | **0.910** | 0.940 | **0.927** | **0.954** |
| *Baselines:* | | | | |
| Hybrid [25] | 0.904 | 0.920 | 0.915 | 0.941 |
| JukeBox [15, 23] | - | - | 0.915* | - |
| MuLaP [32] | - | - | 0.893* | - |
| CLMR [22] | - | - | 0.866* | - |
| **(d) End-to-end training baselines** | | | | |
| AST [10] | 0.888 | **0.949** | - | - |
| SC-CNN [42] | - | - | 0.913* | - |

* indicates that the number is brought from the original paper.

### TEXT TO MUSIC RETRIEVAL

- **Method**
  Score(track, query) = Cosine similarity(audio emb, query emb)
  For each query, rank the scores of all candidate tracks
- **Eval set** 7K expert-curated playlists of a total 100K songs (not included in training data). Queries are playlist titles/descriptions, target audio are songs contained in the playlist.
- **Result**

**Table 5**. Text query music retrieval evaluation results. Text ablation/unfiltered models use M-Resnet-50.

| | Title | | Description | |
|---|---|---|---|---|
| **Model** | AUC | mAP | AUC | mAP |
| M-AST | **0.933** | 0.110 | **0.903** | **0.090** |
| M-Resnet-50 | 0.931 | 0.104 | 0.901 | 0.084 |
| *Text Ablation:* | | | | |
| ASET+SF+LF | 0.917 | 0.101 | 0.892 | 0.077 |
| ASET+SF | 0.913 | 0.089 | 0.867 | 0.060 |
| ASET | 0.626 | 0.005 | 0.688 | 0.009 |
| *SF/LF Unfiltered* | **0.933** | **0.111** | 0.897 | 0.081 |

### TEXT TO TEXT RETRIEVAL

- **Method** Evaluate the text embeddings on a triplet classification task
- **Eval set** We use both the AudioSet ontology and Playlist titles/descriptions to sample such triplets.
- **Examples**

| Eval Set | Anchor / Positive / Negative |
|---|---|
| Ontology | Steelpan / Sounds of a tuned percussion instrument originally constructed from steel oil drums by hammering out small patches on the head to produce separate pitches. / The sound of a musical instrument that produces sound by vibration of air in a tubular resonator in sympathy with the vibration of the player's lips. |
| Playlist | Relaxing Korean Pop / Lets make your chill mood with a collection of easy-going sounds from Korean artists. / These fun and upbeat songs from the alternative side of the pop music spectrum will keep you energized while you exercise. |

**Result**

**Table 6**. Text triplet classification accuracy AudioSet ontology evaluation and Playlist title to description evaluation. Text ablation/unfiltered models use M-Resnet-50.

| Model | Playlist | AudioSet |
|---|---|---|
| M-AST | **0.959** | **0.962** |
| M-Resnet-50 | 0.945 | 0.951 |
| *Text Ablation:* | | |
| ASET + SF + LF | 0.935 | 0.952 |
| ASET + SF | 0.910 | 0.938 |
| ASET | 0.693 | 0.818 |
| *SF/LF Unfiltered* | 0.949 | 0.959 |
| *Baselines:* | | |
| SimCSE [45] | **0.950** | 0.938 |
| SBERT [46] | 0.942 | 0.889 |
| USE [47] | 0.918 | **0.946** |
| BERT [38] | 0.850 | 0.847 |