Soundtrack Retrieval for Film Production

Bill Wang[1] **Haven Kim**[1] **Leduo Chen**[1] **Minje Kim**[2] **Julian McAuley**[1]
[1]University of California San Diego
[2]University of Illinois Urbana-Champaign

Abstract

Music is an integral part of an enjoyable cinematic experience, elevating both the emotional depth and narrative. The growth of platforms that allow film producers to license soundtracks from extensive collections has enabled low-budget filmmakers to achieve high-quality productions. However, with the vast amount of content available, it becomes paramount to effectively retrieve soundtracks to help producers find suitable tracks without a significant time investment. Current soundtrack retrieval systems on these platforms rely heavily on the selection of tags, which can be time-consuming due to the large number tracks associated with each tag. In this work, we introduce a multi-modal transformer architecture with a cross-attention mechanism, trained using an image, plot summaries and relevant tags, through contrastive learning. Our objective evaluations demonstrate that our model effectively utilizes all three inputs to retrieve soundtracks that are fitting for a film.

1 Introduction

Music is widely recognized as essential components of cinematic storytelling [Szymczyk, 2023]. Independent filmmakers often purchase preexisting soundtracks from libraries like AudioJungle (`https://audiojungle.net/`) or NeoSounds (`https://www.neosounds.com/`) but current retrieval systems rely solely on tag-based searches based on genre, mood, and event type. This approach is time consuming due to large track volumes per tag and ignores how sound directors actually select music by considering plot, visual aesthetics, and soundtracks [Liu et al., 2020] [Lipscomb and Kendall, 1994] [Boltz, 2004].

We aim to address this concern by creating a retrieval system that suggests film soundtracks from a comprehensive library, using visual information and story lines, as well as tags. In this paper, we will show that each of these three input types play a role in identifying suitable soundtracks for films, with plot information demonstrating particularly strong influence on model performance. We also empirically show that the multi-modal input, which integrates visual textual information showcases the best performance, especially when plot information is included in the combination. To this end, we propose training multi-modal transformer architecture using contrastive learning with a cross-attention mechanism.

2 Related Work

Music retrieval systems aim to locate relevant musical content based on various input modalities. The foundational work, PICASSO [Stupar and Michel, 2011], is a retrieval algorithm that identifies suitable music tracks from text, video, or image queries by finding similar reference items in a dataset of fifty films and returning soundtracks with matching musical features. Other approaches include recommendation systems for user generated videos using support vector machines trained on visual features [Shah et al., 2014], and movie soundtrack recommendation via logistic regression modeling the relationship between film clips and soundtracks [Liu et al., 2020].

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: AI for Music.

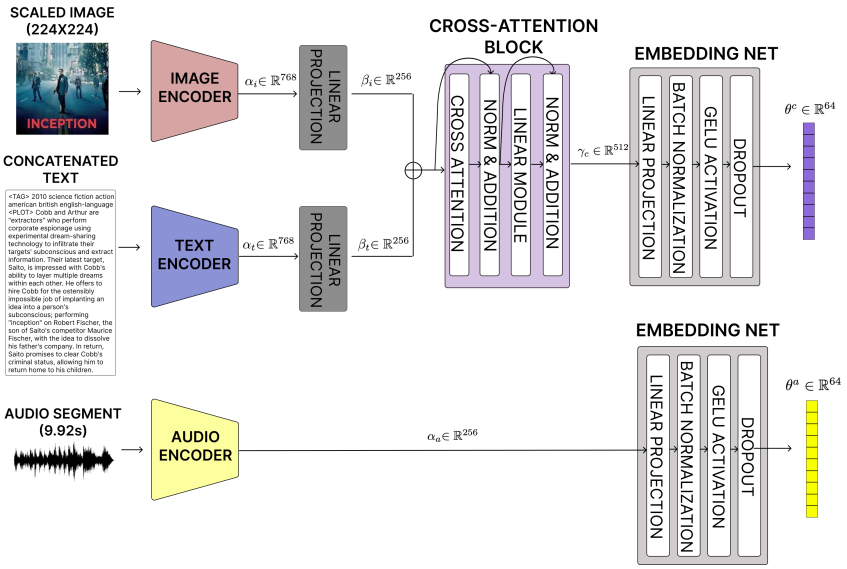

Figure 1: Model Architecture

Current systems face two key limitations: heavy reliance on hand-engineered features that constrain pattern discovery, and limited scalability due to small datasets, PICASSO used fifty films while recent work [Liu et al., 2020] used only two. We address these issues through automated large-scale data collection and contrastive learning approaches which have shown promising results in cross modal music retrieval Doh et al. [2022]

3 Methodology

In this section, we provide details on our methodology for acquiring and processing the data, as well as a description on training our retrieval system (Figure 1). The retrieval system consists of three encoders that convert the image, text, and audio modalities to their corresponding embeddings. The image and text encoders are to process the input modalities, while the audio modality is used to compute the target embedding to retrieve music. The text encoder is shared by two types of input sequences, tags and plot. [1]

3.1 Dataset Construction

Our objective is to develop a machine learning model that automatically retrieves soundtracks for films by leveraging visual and textual information. To achieve this, we gathered images, tags, and plot summaries for 7,771 movies as the multi-modal input to the system, along with their corresponding soundtracks as the target of the retrieval system during training.

All images, tags, and plots were extracted from `Wikipedia`. Thumbnails represent visual aspects of the films, while tags provide categorical information about the film's genre, nationality, and release era. Tags are sourced from the "categories" section of `Wikipedia`, specifically those categories ending in "films". Only tags that appear in at least two films are retained to ensure their relevance. Plot summaries are extracted from the "Plot" section of each film's `Wikipedia` page, offering an additional layer of textual context.

The dataset includes 70,215 soundtracks sourced from YouTube, corresponding to the films in the dataset. These soundtracks are used as the target for training the retrieval model.

[1] https://github.com/BillWang04/multimodal_soundtrack_retrieval

3.2 Data Preprocessing and Encoding

Images: We proportionally scaled and cropped the image to reach size of 224 x 224 and then encoded the image using a Vision Transformer (VIT) [Dosovitskiy et al., 2021] that processes them as 64 patches through six Transformer blocks, producing 768 embeddings ($\alpha_i \in \mathbb{R}^{768}$).

Text: We adopted `Bert-base-uncased` [Devlin et al., 2019] to encode concatenated tags and plot summaries, which were differentiated by special tokens `<TAG>` and `<PLOT>`. Plot summaries are randomly sampled by paragraph to handle length constraints and provide data augmentation. However, when evaluating these trained models, we selected the first paragraph of each plot, allowing for fair evaluation. The encoder outputs 768- dimensional embeddings ($\alpha_t \in \mathbb{R}^{768}$).

Audio: Music tracks are resampled to 22,050 Hz, segmented into 9.92-second fragments and converted to mel-spectrograms. A CNN-Transformer architecture adopted from Won et al. [2021] processes these spectrograms outputting 256-dimensional embeddings ($\alpha_a \in \mathbb{R}^{256}$).

3.3 Key Modules

Cross-attention Block: Given multi-modal inputs, we implement a cross-attention block to enable joint learning of visual and textual features. Following [Mercea et al., 2022], image and text encoder outputs ($\alpha_i, \alpha_t \in \mathbb{R}^{768}$) are linearly projected to smaller embeddings ($\beta_i, \beta_t \in \mathbb{R}^{256}$), concatenated ($\beta_i \oplus \beta_t \in \mathbb{R}^{512}$), and processed through multi-head cross-attention with residual connections and layer normalization. Unlike prior work that outputs separate modality vectors, our module produces a unified representation $\gamma_c \in \mathbb{R}^{512}$ summarizing both visual and textual film narratives.

Embedding Net: Final representations are mapped to a shared feature space via Embedding Net instances that convert the cross-attention output $\gamma_c \in \mathbb{R}^{512}$ and audio encoder output $\alpha_a \in \mathbb{R}^{256}$ into 64-dimensional embeddings θ^c and θ^a, respectively. Each instance comprises linear projection, batch normalization, GELU activation, and 0.3 dropout rate.

3.4 Contrastive Learning

We employ contrastive learning to effectively capture the semantic relationship between two embedding vectors: the audio embedding and the multi-modal embedding from the same movie (θ^c and θ^a) should be similar, while those from different movies should be dissimilar. In the learning phase, models are optimized to enhance the similarity across N positive pairs within a min-batch while simultaneously reducing the similarity between non-matching pairs. We formulate our cross-entropy loss function L by utilizing the InfoNCE loss as a foundation [van den Oord et al., 2019]. It operates on the softmax of scaled pairwise similarities for each anchor $\{\theta_1^c, \ldots, \theta_N^c\}$ and each $\{\theta_1^a, \ldots, \theta_N^a\}$. In particular, for any given anchor θ_i^c, the softmax operation on scaled pairwise similarities yields the probability where θ_i^a is the class to which θ_i^c belongs. Similarly, for any anchor θ_i^a, applying softmax of scaled pairwise similarities generates the probability assigned to θ_i^c. In summary, our loss function L is formulated as below, where the logarithm turns the softmax probabilities into the negative log-likelihood loss, and τ is a trainable parameter that controls the smoothness of probability distributions.

$$L = -\frac{1}{2N}\sum_{i=1}^{N}\left(\log\frac{\exp(\theta_i^c \cdot \theta_i^a)/\tau}{\sum_{j=1}^{N}\exp(\theta_i^c \cdot \theta_j^a)/\tau} + \log\frac{\exp(\theta_i^c \cdot \theta_i^a)/\tau}{\sum_{j=1}^{N}\exp(\theta_j^c \cdot \theta_i^a)/\tau}\right) \tag{1}$$

3.5 Data Split and Training Details

To simulate real-world scenarios where models must infer from past data, we split our dateset chronologically based on the release year. Films released in 2021 or earlier are used as the training set while those from 2022 formed the validation set. Films released in 2023 or later are set aside for testing purposes which included 1964 soundtracks.

We optimize the model using Adam optimizer [Kingma and Ba, 2017], setting the learning rate at 1e-5. We use a batch size of 16 and conducted validation at the end of each epoch. If the validation loss fail to decrease over three consecutive epochs, we apply early stopping. For testing, we use models that demonstrated the lowest validation loss.

Model	R@1↑	R@5↑	R@10↑	R@20↑	R@50 ↑	R@100↑	MedR↓
All	**0.41**	**1.58**	2.65	**5.55**	**14.61**	26.17	**227**
Baseline	0.05	0.41	1.07	2.19	5.45	9.88	527.5
Image (Im)	0.15	0.51	1.02	1.83	4.79	10.74	658
Plot (Plo)	0.25	1.37	2.19	3.77	8.50	14.41	405
Tag (Ta)	0.15	0.41	0.61	1.22	3.51	7.54	668
Image+Plot (ImPlo)	0.15	1.32	**2.75**	5.24	13.29	23.12	272.5
Image+Tag (ImTa)	0.15	0.81	1.37	3.05	7.33	12.73	461.5
Tag+Plot (TaPlo)	0.25	1.32	2.55	5.14	12.83	**26.73**	**227**

Table 1: Performance comparison across different modality configurations. Recall values are reported as percentages.

4 Evaluation

We trained and tested the Mercea et al. [2022] model as our baseline, adapted for retrieval by replacing video embeddings with image embeddings. For evaluation, we used the outputs of θ_a (audio), θ_w (text), and θ_i (image) embeddings and employed the triplet loss for training as described in Mercea et al. [2022]. Table 1 exhibits that our model outperforms the baseline significantly in all metrics.

To examine how different types of input affect model performance, we tested seven models with similar architectures but varying inputs. Im, Ta, and Plo are unimodal models that process only one input type: images, tags, and plot summaries respectively. ImTa, ImPlo, and TaPlo are bimodal models that handle two inputs, combining images with tags, images with plots, and tags with plots respectively. The All model is trimodal, integrating all three inputs: images, tags, and plot summaries.

For models with only one type of textual input (Ta and Plo), we did not add special tokens `<TAG>` or `<PLOT>`. For unimodal models (Im, Ta, Plo) and the TaPlo model, we replaced the cross-attention module with a self-attention module that processes 256-dimensional vectors while maintaining the same output dimensionality.

We evaluate all models using a consistent protocol with cosine similarity between L2-normalized embeddings. For the baseline AVCA model, we combine normalized text and image embeddings as $\theta_c = \texttt{norm}((\texttt{norm}(\theta_w)+\texttt{norm}(\theta_i))/2)$ since we need to combine the two embeddings for comparison. For other models, we use their learned combined embedding θ_c. Retrieval ranking is based on cosine similarity between θ_c and audio embedding θ_a. We report Recall@K (K=1,5,10,20,100) and Median Rank (MedR).

Among single modality approaches, plot information demonstrates the strongest performance, with Plo outperforming both Image and Tag models across all metrics. This superiority of plot information is further evidenced in the multi-modal results: both ImPlo and TaPlo, which incorporate plot summaries, significantly outperform ImTa, which relies solely on visual and tag information. Notably, TaPlo achieves the best Recall@100 (26.73%) and ties for the lowest median rank (227), while ImPlo has the highest Recall@10 (2.75%).

5 Conclusion

In this paper, we proposed a multi-modal transformer system for automatic soundtrack retrieval in film production, leveraging visual, textual, and narrative information through contrastive learning. Our evaluation demonstrates that plot information is particularly effective for soundtrack retrieval, with multi-modal approaches incorporating plot summaries significantly outperforming single modality and plot-free combinations. While our approach offers a practical solution for efficiently navigating large soundtrack libraries, it is limited by its reliance on film-level representations rather than scene-specific analysis. Future work perhaps should explore incorporating video sequences to enable scene-level soundtrack recommendations.

References

Marilyn G. Boltz. The cognitive processing of film and musical soundtracks. *Memory & Cognition*, 32(7):1194–1205, 2004. doi: 10.3758/BF03196892.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In Jill Burstein, Christy Doran, and Thamar Solorio, editors, *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1423. URL https://aclanthology.org/N19-1423/.

SeungHeon Doh, Minz Won, Keunwoo Choi, and Juhan Nam. Toward universal text-to-music retrieval, 2022. URL https://arxiv.org/abs/2211.14558.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale, 2021. URL https://arxiv.org/abs/2010.11929.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization, 2017. URL https://arxiv.org/abs/1412.6980.

Scott D. Lipscomb and Roger A. Kendall. Perceptual judgement of the relationship between musical and visual components in film. *Psychomusicology: A Journal of Research in Music Cognition*, 13 (1-2):60–98, 1994. doi: 10.1037/h0094101.

Shan Liu, Jiayuan Zhang, Mingyang Liu, Yu Guo, Jiaqi Guo, and Jiacheng Cai. Soundtrack matching and recommendation system of film and tv series. In *2020 13th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics (CISP-BMEI)*, pages 591–596, 2020. doi: 10.1109/CISP-BMEI51763.2020.9263604.

Otniel-Bogdan Mercea, Lukas Riesch, A. Sophia Koepke, and Zeynep Akata. Audio-visual generalised zero-shot learning with cross-modal attention and language, 2022. URL https://arxiv.org/abs/2203.03598.

Rajiv Ratn Shah, Yi Yu, and Roger Zimmermann. Advisor - personalized video soundtrack recommendation by late fusion with heuristic rankings. 11 2014. doi: 10.1145/2647868.2654919.

Aleksandar Stupar and Sebastian Michel. Picasso: automated soundtrack suggestion for multi-modal data. In *Proceedings of the 20th ACM International Conference on Information and Knowledge Management*, CIKM '11, page 2589–2592, New York, NY, USA, 2011. Association for Computing Machinery. ISBN 9781450307178. doi: 10.1145/2063576.2064027. URL https://doi.org/10.1145/2063576.2064027.

Michael Szymczyk. *Independent Filmmaking 101: A 60 Minute Crash Course on the Basics of No to Low Budget Filmmaking*. VIDART, 2023. Kindle Edition.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding, 2019. URL https://arxiv.org/abs/1807.03748.

Minz Won, Keunwoo Choi, and Xavier Serra. Semi-supervised music tagging transformer, 2021. URL https://arxiv.org/abs/2111.13457.