INTERPRETING SONG LYRICS WITH AN AUDIO-INFORMED PRE-TRAINED LANGUAGE MODEL

Yixiao Zhang[1] **Junyan Jiang**[2,3] **Gus Xia**[2,3]
Simon Dixon[1]

[1] Centre for Digital Music, Queen Mary University of London
[2] Music X Lab, NYU Shanghai   [3] MBZUAI

`yixiao.zhang@qmul.ac.uk, jj2731@nyu.edu, gxia@nyu.edu, s.e.dixon@qmul.ac.uk`

ABSTRACT

Lyric interpretations can help people understand songs and their lyrics quickly, and can also make it easier to manage, retrieve and discover songs efficiently from the growing mass of music archives. In this paper we propose BART-fusion, a novel model for generating lyric interpretations from lyrics and music audio that combines a large-scale pre-trained language model with an audio encoder. We employ a cross-modal attention module to incorporate the audio representation into the lyrics representation to help the pre-trained language model understand the song from an audio perspective, while preserving the language model's original generative performance. We also release the Song Interpretation Dataset, a new large-scale dataset for training and evaluating our model. Experimental results show that the additional audio information helps our model to understand words and music better, and to generate precise and fluent interpretations. An additional experiment on cross-modal music retrieval shows that interpretations generated by BART-fusion can also help people retrieve music more accurately than with the original BART. [1]

1. INTRODUCTION

Lyrics play a key role in the understanding and creation of songs, expressing emotions and delivering messages in the form of natural language [1]. Lyrics have both linguistic and musical characteristics: the field of Lyric Information Processing (LIP) can consequently be seen as a bridge between Music Information Retrieval (MIR) and Natural Language Processing (NLP), encompassing a range of new challenges such as lyric structure analysis [2], lyric semantic analysis [3], automatic lyric generation [4], and lyric understanding [5]. In this paper, we focus on the task of multimodal lyric interpretation, which requires the model

---

[1] Open-sourced code and pretrained models: `https://github.com/ldzhangyx/BART-fusion`.

Ray LaMontagne – Empty

Music Mel-Spectrogram

Lyrics

She lifts her skirt up to her knees / Walks through the garden rows / With her bare feet laughin' / I never learned to count my blessings / I choose instead to dwell in my disasters / I walk on down the hill

Through grass grown tall and brown and still / It's hard somehow to let go of my pain / On past the busted back of that old and rusted Cadillac / That sinks into this field collecting rain / …

Human Interpretation: I think this song is about a man who was completely in love with a woman. He sits and remembers their time together, and by the lyrics it seems as if what he's remembering most are the simple times they had together, but they may have been the most amazing. Like just watching her laugh, walking through a garden, making love while it's raining outside…
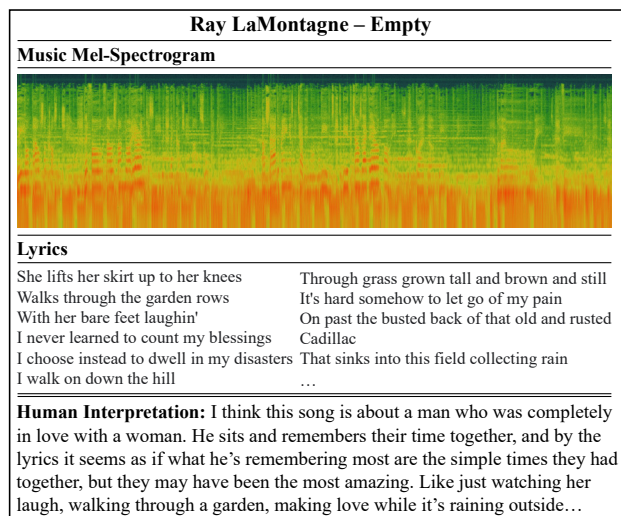
Figure 1. An example of lyrics and their interpretation in real-life, where the interpretation is written by a human. Information from the audio modality includes the representations of instruments, styles, chords, etc., which may help the model to understand the meaning of the lyrics.

to understand both the words and music of a song, and to produce a natural, concise and human-like description of its lyrics.

In real life, human interpretations of lyrics often contain both a general summary of the lyrical theme and a detailed analysis in relation to specific lines. Figure 1 shows such an example. Considering that the human interpretation of the lyrics contains subjective elements, the lyrics interpretation task is like an extension of the lyrics summarisation task. The task requires the model to be able to (1) select an excerpt from the lyrics, as in extractive text summarisation; (2) generate explanatory text from lyrics, which is similar to abstractive text summarisation. However, previous summarisation methods for general texts [6, 7] are not necessarily applicable in the context of lyrics, because song lyrics often contain rich metaphors, poetic themes, and a high degree of rhythm [2]. Previous studies have attempted to apply extractive summarisation to song lyrics using TextRank algorithms [8] and audio-text alignment algorithms [9]. The drawback of such approaches is that the summary by itself is not enough to explain the lyrics.

To the best of our knowledge, this paper is the first study to use both extractive and abstractive methods to generate lyric interpretations.

Compared to unimodal lyric interpretation models, multimodal models can use information from the music audio domain, such as style, emotion and instrument representations, to reduce the difficulty of understanding lyrics and to improve the quality of the generated text. In recent years, Transformer-based multimodal generative pre-training models have performed well for tasks such as text understanding [10, 11] and text generation [12, 13]. Some related works have attempted to adapt existing pre-trained language models to multimodal tasks [14] or conditional generation tasks [13, 15]. The model we introduce in this paper is inspired by Yu *et al.* [14] and we choose to adapt it from a pre-trained language model (BART) [7]. Our model makes use of two modalities: the text of the song lyrics and the corresponding music audio. We add a convolutional encoder (CNN-SA) [16] to extract a representation of the audio and transfer it to the text domain by computing the similarity between the semantics of the song lyrics and the audio representation through a cross-modal attention mechanism, implemented by a multi-headed attention layer [17]. The transformed music audio representation is then fused into the semantic representation of the lyrics as an additional embedding. We discuss more details of the model in Section 2. To train and evaluate our model, we propose a new dataset, the Song Interpretation Dataset, which contains 27,834 songs with 490,000 corresponding user interpretations. This is the first large-scale open-source dataset for lyric interpretation. We describe the dataset in detail in Section 3.

We evaluate our model against the original BART model (as a baseline) on the Song Interpretation Dataset, as described in Section 4. On the lyric interpretation task, our model outperforms the baseline model according to the standard text summarisation metrics ROUGE, ME-TEOR, and BERT-Score. Ablation experiments show that our dataset filtering techniques also improve model performance. To show the value of our model for other tasks, we also present experimental results demonstrating that it performs better than the baseline on a cross-modal retrieval task.

The main contributions of this work can be summarised as follows:

1. We present BART-fusion, the first multimodal generative model for lyric interpretation. We investigate the integration of audio representations with lyric representations and show that audio representations can improve the performance of lyric interpretation models;

2. We contribute a large-scale multimodal dataset containing paired audio, lyrics, and lyric interpretations that can be used for music understanding tasks such as lyric interpretation.

2. METHOD

In this section, we first revisit the original BART model in Section 2.1. We then discuss the approach to extract musical features from an audio spectrogram in Section 2.2. Finally, we introduce the music-text representation fusion mechanism in Section 2.3. We identify text-domain features with the subscript t and music-domain features with m.

2.1 BART Model for the Generative Task

Transformer-based pre-trained encoder-decoder language models such as BART [7], MASS [18] and T5 [19] generalize BERT [20] (due to the bidirectional encoder) and GPT [21] (with the left-to-right decoder), achieving good results on sequence-to-sequence tasks such as text summarisation and machine translation. Our model takes advantage of the text generation ability of BART, the structure of which is shown on the right side of Figure 2.

The lyric text input is firstly tokenized and embedded. We assume the lyric text sequence has L_t tokens, and the embedding dimension is d_t, resulting in an embedding $X_t \in \mathbb{R}^{L_t \times d_t}$. Following Vaswani *et al.* [22], we add an absolute positional embedding E_{pe} to get the final input features H_t^0:

$$H_t^0 = X_t + E_{\mathrm{pe}}. \tag{1}$$

These input features are then passed to the encoder. The encoder has a stack of six layers, as illustrated in Figure 2, where a single Transformer layer is shown by a yellow box. Each Transformer layer contains a multi-head Self-Attention (SA) module and a Feed-Forward Network (FFN), each followed by a Layer Normalization (LN) module. For the i-th layer, the representation is calculated as:

$$\widetilde{H}_t^i = \mathrm{LN}(\mathrm{SA}(H_t^{i-1}W_Q, H_t^{i-1}W_K, H_t^{i-1}W_V)W_a + H_t^{i-1}) \tag{2}$$

$$H_t^i = \mathrm{LN}(\mathrm{FFN}(\widetilde{H}_t^i) + \widetilde{H}_t^i), \tag{3}$$

where $H_t^i \in \mathbb{R}^{L_t \times d_t}$, and $W_Q \in \mathbb{R}^{d_t \times d_a}$, $W_K \in \mathbb{R}^{d_t \times d_a}$ and $W_V \in \mathbb{R}^{d_t \times d_a}$ denote linear transformation matrices which map the representations to a common space. $W_a \in \mathbb{R}^{d_a \times d_t}$ linearly projects the attention value back to the desired dimensionality.

The decoder also consists of a stack of six Transformer layers, which is similar to the encoder. But the multi-head self-attention module in the decoder is masked to respect causality, and an additional multi-head encoder-decoder attention is introduced to incorporate the encoder representation.

2.2 Audio Encoder

For the multimodal lyric interpretation model, we expect the music audio modality to provide some additional semantic information, such as style, mood, instrumentation, etc., to help the model understand the lyrics better. We
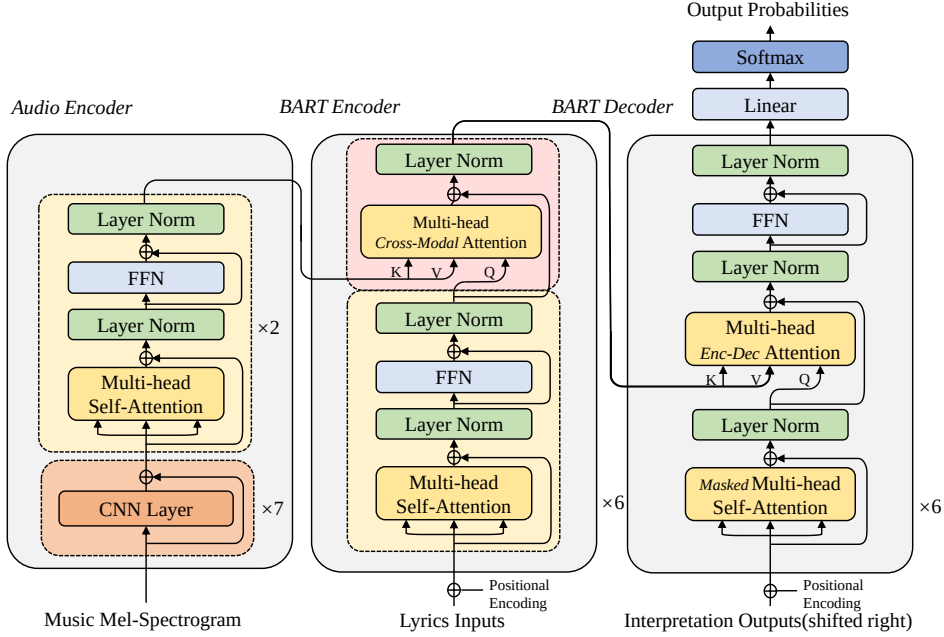
Figure 2. An overview of our proposed model. The model is divided into three parts from left to right: the audio encoder, the BART encoder, and the BART decoder. The fusion of the semantic representation of music audio and lyric text occurs in the upper part of the middle module (pink background). Only at the last two layers of the BART encoder, the music audio representation and the lyric text representation are semantically fused: the music audio representation and the lyric text representation are fed into the cross-modal attention module, and the result is added as an additional embedding to the original lyric text representation. The fused representation is fed into the BART decoder to generate an interpretation.

design an audio encoder to extract a representation following Won *et al.* [16]. The audio encoder uses a stack of CNN layers as a filter to extract local features, followed by a self-attention module to capture the global and temporal features of the audio.

The audio encoder receives an audio clip X_m and transforms it to a mel spectrogram H_m^0 as input. We compute the feature map of the i-th layer of the encoder as follows:

$$\widetilde{H}_m^i = \text{BN}(\text{CNN}_2(\text{ReLU}(\text{CNN}_1(H_m^{i-1})))) \qquad (4)$$

$$H_m^i = \widetilde{H}_m^i + \text{BN}(\text{CNN}_3(H_m^{i-1})), \qquad (5)$$

where BN is the Batch Normalization operation, and the CNNs are convolutional modules with different parameters. We add a residual connection to each CNN layer.

We then add two Transformer layers, identical to those in the BART encoder, to extract the final representation of music audio. Finally, we get the music audio representation $Z_m \in \mathbb{R}^{L_m \times d_m}$, where L_m and d_m denote the shape of the music audio feature map at the last CNN layer.

2.3 Representation Fusion

As shown in Figure 2, we insert a representation fusion module into the BART encoder to incorporate musical information. Inspired by Tsai *et al.* [17] and Yu *et al.* [14], we apply cross-modal attention to enable the music audio representation to be transferred to the lyric text domain. Jawahar *et al.* [23] have shown that BART encoders tend to extract semantic information in the last few layers, so we only fuse semantic representations at the final two Transformer layers.

For a specific Transformer layer i, we have the lyric text representation $H_t^i \in \mathbb{R}^{L_t \times d_t}$ and the music audio representation $Z_m \in \mathbb{R}^{L_m \times d_m}$, which is the same for each layer. We calculate the domain-adapted music audio representation with a multi-head Cross-Modal Attention (CMA) module:

$$H_{m \to t}^i = \text{CMA}(H_t^i W_Q', Z_m W_K', Z_m W_V')W_a', \qquad (6)$$

where $H_{m \to t}^i \in \mathbb{R}^{L_t \times d_t}$ and the symbol $m \to t$ denotes cross-modal attention from music audio to the lyric domain. $d_{a'}$ is the dimensionality of the attention module, and similar to Eq. 3, $W_Q' \in \mathbb{R}^{d_t \times d_{a'}}$, $W_K' \in \mathbb{R}^{d_m \times d_{a'}}$ and $W_V' \in \mathbb{R}^{d_m \times d_{a'}}$ denote linear transformation matrices which map the representations to a common space. $W_a' \in \mathbb{R}^{d_{a'} \times d_t}$ linearly projects the attention value back to the lyric text dimension.

We add the cross-domain representation as an additional embedding [24] to the lyric text representation to get the final representation for the last two layers of the BART encoder:

$$H_{\text{fusion}}^i = H_t^i + H_{m \to t}^i. \qquad (7)$$

3. SONG INTERPRETATION DATASET

The lack of suitable datasets has prevented deep learning models from learning to describe songs in natural lan-

guage. Related studies [25, 26] refer to some datasets, but they share two common problems: 1) the amount of data is small and 2) the datasets are not open-sourced. We propose a new dataset, the Song Interpretation Dataset, for the lyric interpretation task.[2]

The Song Interpretation Dataset combines data from two sources: (1) music and metadata from the Music4All Dataset [27], and (2) lyrics and user interpretations from SongMeanings.com[3]. We design a music metadata-based matching algorithm that aligns matching items in the two datasets with each other. In the end, we successfully match 25.47% of the tracks in the Music4All Dataset.

The dataset contains audio excerpts from 27,834 songs (30 seconds each, recorded at 44.1 kHz), the corresponding music metadata, about 490,000 user interpretations of the lyric text, and the number of votes given for each of these user interpretations. The average length of the interpretations is 97 words. Music in the dataset covers various genres, of which the top 5 are: Rock (11,626), Pop (6,071), Metal (2,516), Electronic (2,213) and Folk (1,760).

To the best of our knowledge, this is the first large-scale open-source dataset for lyric interpretation. A comparison with similar datasets is shown in Table 1.

Dataset	Music	Interpretation	Public
Choi *et al.* [25]	800	2000	✗
Manco *et al.* [26]	17,354	17,354	✗
Ours	**27,384**	**490,000**	✓

Table 1. A comparison of our dataset with previous music description datasets.

We observe three main issues with interpretations written by real users: (1) some interpretations are very short or very long; (2) interpretations can contain content unrelated to the lyrics themselves, and (3) some interpretations are of low quality. We therefore preprocess the dataset using two techniques:

1. We **remove overly short interpretations** with length less than 256 characters to improve data representativeness, since we find that sentences below this length are often meaningless interpretations. For interpretations longer than 2048 characters, we keep only the first 2048 characters, but ensure that the last word is complete.

2. We use a **voting-based filtering mechanism** to improve data quality. Every interpretation on SongMeanings.com has a voting result attached, indicating how much the community approves of it, so an interpretation with a higher vote is more likely to be a high-quality interpretation. We therefore create two subsets, keeping only interpretations with positive votes and interpretations with non-negative votes.

---

[2] The Song Interpretation Dataset is anonymously open for downloading: https://doi.org/10.5281/zenodo.7019124.
[3] https://songmeanings.com/

To enable the model to be comparable across datasets, we manually select 800 interpretations and use them as a test dataset after excluding them from the original dataset. We have specifically removed all songs that appeared in the test set from the training set to avoid data leakage issues. After the above preprocessing, the dataset has 3 different subsets with the information shown in Table 2.

Dataset Name	Train	Valid.	Test
Raw dataset	440,000	50,000	800
Dataset Full	279,283	31,032	800
Dataset w/vote ≥ 0	265,360	29,484	800
Dataset w/vote > 0	49,736	5,526	800

Table 2. A comparison of dataset sizes (in number of interpretations) with different filtering methods.

4. EXPERIMENTAL SETTINGS

4.1 Implementation Details

We pre-process the lyric text input data by truncating or padding text to 2048 tokens. All the audio signals are downsampled to 16,000 Hz sample rate and converted to short-time Fourier transform representations with a 512-point FFT and 50%-overlapping Hann window. Finally, we convert these to log mel spectrograms with 128 bins.

We use `BART-base` [7] as the pre-trained language model to construct BART-fusion, which has a 6-layer encoder and decoder. For the audio encoder (CNNSA), we use 3×3 kernels for all layers with [128, 128, 256, 256, 256, 256, 256] channels and [(2, 2), (2, 2), (2, 2), (2, 1), (2, 1), (2, 1), (2, 1)] strides. The cross-modal attention module has 1 head and 1 layer, with $d_a = 768$.

We use AdaFactor [28] as the optimizer. We set the learning rate to 6×10^{-4} and reduce it to 6×10^{-5} from the 11th epoch. For all experiments, we use a batch size of 8. We train all models for 20 epochs with early stopping of 3 epochs using the ROUGE-1 score on the validation set.

4.2 Evaluation Metrics

Since there are no existing metrics for the lyric interpretation task, we borrow several complementary metrics from similar tasks to evaluate the performance level of the model in a comprehensive manner. We use ROUGE, METEOR and BERT-Score to evaluate the generated interpretations. ROUGE is considered as the main metric for evaluation, because our task is closest to the text summarisation task.

4.2.1 ROUGE-{1, 2, L}

ROUGE [29] is a common metric for evaluating abstractive text summaries. It calculates the overlap of 1-gram phrases (R-1), 2-gram phrases (R-2) and their weighted results (R-L). In the context of lyrics, a higher ROUGE score indicates that the generated explanatory text leaves out less necessary information, which is better.

Training dataset	Method	Data size	R-1	R-2	R-L	METEOR	BERT-Score
Dataset w/random	BART	56,470	40.0	12.5	21.7	21.1	**83.7**
Dataset w/random	BART-fusion	56,470	**42.1***	**13.6***	**23.4***	**22.0***	83.3
Dataset w/voting > 0	BART	56,470	41.2$_+$	13.0$_+$	22.8$_+$	22.0$_+$	**83.6**
Dataset w/voting > 0	BART-fusion	56,470	**44.3**$^*_+$	**14.6**$^*_+$	**24.7**$^*_+$	**22.6**$^*_+$	83.3
Dataset Full	BART	316,478	44.1	14.0	24.5	22.5$_+$	**83.5**
Dataset Full	BART-fusion	316,478	**46.1***	**15.0***	**25.1***	**23.0***	**83.5**
Dataset w/voting ≥ 0	BART	300,712	44.8$_+$	14.9$_+$	24.7	22.7	83.9
Dataset w/voting ≥ 0	BART-fusion	300,712	**46.7**$^*_+$	**15.6**$^*_+$	**25.5**$^*_+$	**23.4***	**84.1**

Table 3. Evaluation results of BART-fusion and BART (baseline) on the Song Interpretation Dataset with different settings. $*$: BART-fusion outperforms BART with $p < 0.05$; $+$: the filtered dataset outperforms the unfiltered dataset with $p < 0.05$.

4.2.2 METEOR

METEOR [30] takes into account similar semantic information such as synonyms through WordNet and calculates the similarity score based on F-measure. It complements ROUGE by jointly measuring how semantically similar the model-generated lyric interpretation is to the reference text.

4.2.3 BERT-Score

BERT-Score [31] is mainly used to evaluate the naturalness and fluency of the generated text. We expect that incorporating the audio modality should not sacrifice the generative performance of the language model.

We use `rouge`[4], `nltk`[5] and `bert-score`[6] respectively to compute these metrics.

5. RESULTS AND ANALYSIS

5.1 Main Results

We train BART-fusion and the corresponding original BART on several different dataset settings. Results are shown in Table 3, where all values are the means of multiple independent repeated experiments. We perform a *paired Student's t-test* on the results of BART-fusion and BART, and the results on the filtered datasets and unfiltered datasets respectively, where p-value is 0.05.

Our first finding is that applying a voting-based filtering mechanism to the dataset significantly improves the performance of the models (both BART-fusion and baseline) on this task. For fairness, we take a random subset of *Dataset Full*, which we call *Dataset w/random*, to match the size of *Dataset w/voting > 0*. (We still use *Dataset Full* to compare *Dataset w/voting ≥ 0*.) The experimental results show that the performances of both BART-fusion and the baseline are significantly improved on the filtered datasets.

The experimental results also show that BART-fusion significantly outperforms the baseline, and generates more precise interpretion text for lyrics. For all dataset settings, our BART-fusion models show better performance on the ROUGE and METEOR scores.

Finally, we find that BART-fusion preserves the generative performance of the pre-trained language model while improving the generation accuracy. The performance of BART-fusion is essentially equal to that of baseline on the BERT-Score metric, which is the metric of text quality. It means that the introduction of audio modal information affects the model mainly semantically and does not affect the naturalness or fluency.

5.2 Case Study and Error Analysis

We observe that adding music modality information usually brings the benefits of accurate understanding of the theme, selecting highlighted lyric lines, and emotive sentences from the generated samples. In the case study, we select a representative example showing the generation results from different models with the same lyric input, as shown in Table 4. In the first lines, BART-fusion explains the theme of the lyrics more accurately than BART. Then, BART-fusion selects the highlighted lyric lines and gives further detailed interpretation, while the text generated by BART lacks a clear explanation of them. We find that BART-fusion talks about the mood of the song at the end, which is not present in the original BART example. [7]

We have noticed that when the lyrics are about complex topics, such as religious and philosophical topics, BART-fusion and BART sometimes fail to understand the meaning correctly and generate text that is only superficially correct, which we interpret as a lack of common sense often observed in this class of deep learning models.

5.3 Cross-Modal Retrieval Analysis

In addition, we design a cross-modal retrieval experiment to test the retrieval capability of our models. We randomly extract a sentence from the real interpretation in the test set as a query string, use this to search over a database of lyric text and musical audio on a semantic level, and return a ranked list of all possible results. We expect the correct result corresponding to the query to be as close to the top of this ranking as possible.

To accomplish this task, the model generates interpretations for all song lyrics in the database, calculates their semantic features using Sentence-BERT [32], and stores

---

[4] https://github.com/pltrdy/rouge
[5] https://github.com/nltk/nltk
[6] https://github.com/Tiiiger/bert_score

[7] We provide a demo page for more examples: https://sites.google.com/view/bart-fusion-ismir2022.

Lyrics: I cannot find a way to describe it / It's there inside, all I do is hide / I wish that it would just go away / What would you do, you do, if you knew? / What would you do? / / All the pain I thought I knew / All the thoughts lead back to you / Back to what was never said / Back and forth inside my head / I can't handle this confusion / I'm unable, come and take me away / / I feel like I am all alone / All by myself, I need to get around this / My words are cold, I don't want them to hurt you / If I show you, I don't think you'd understand / 'Cause no one understands / / All the pain I thought I knew / All the thoughts lead back to you / Back to what was never said / Back and forth inside my head / I can't handle this confusion / I'm unable, come and take me away / ...

Reference: I think this song is about a crush that the singer is afraid to tell the person they are in love with. She wonder if maybe the other person feels the same, but is also too afriad to tell her. Thinking "what if". Since, she won't admit it she'll never know and this not knowing ends up eating her up inside. Each day she wonders if she could be with this person who she really loves, but she just gets frustrated and eventually wishes that she never even fell in love with the person since it would make her life so much easier. It builds up so much that she just wishes someone would "come and take me away". That could mean two things. It could be death/suicide because she decides it's not worth it or she could just wish for someone else to date so she can just say she's taken and not have to worry about this crush.

BART: I think this song is about a girl who is in love with a guy, but she doesn't know how to tell him because she's afraid he'll hurt her. She's trying to convince herself that he loves her, but he doesn't understand how she feels about him. She wants him to come and take her away from all the pain she's caused and make her feel better about herself. This is what I think of when I listen to this song. I think it's about someone who is trying to tell someone they love them, but they don't want to tell them because they're afraid that they'll hurt them. I love this song, it's one of my favorites on the album.

BART-fusion: I think this song is about a girl who is in love with a guy but she is afraid to tell him because she doesn't want to ruin the friendship. In the first verse she is talking about how she can't find a way to describe how she feels and how she wishes it would just go away. The second verse is about how the guy is telling her how he feels but she isn't sure if he really feels the same way. The chorus is her saying how she is confused about what to say and how to tell the guy. "I'm unable, come and take me away" means that she is unable to tell her feelings and she is trying to get the guy out of her head. The last verse is her telling the guy that she feels like she is going nowhere in her life, and she just wants him to take her away from all of this. That's just my take on it. I can totally relate to this song. It's a very powerful song.

Table 4. A generation example for Avril Lavigne - *Take Me Away*.

features in advance. At query time, we also use Sentence-BERT to compute the semantic feature for the query string. We return the song whose feature is most similar to the query feature by computing the cosine similarities between the query feature and features in the database.

We use Mean Reciprocal Rank (MRR) [33], a common metric for information retrieval tasks, to measure the performance of the models on this task. Formally, for a set of query strings $Q = \{q_1, q_2, ..., q_m\}$ and the corresponding database $S = \{s_1, s_2, ..., s_n\}$, the model outputs a list of songs sorted by probability for the i-th query, where the correct song is ranked k_i-th, and the model is scored as:

$$\text{score}_i = \frac{1}{k_i}, \text{where } k_i \leq n. \quad (8)$$

The final MRR is calculated by averaging the scores:

$$\text{MRR} = \frac{1}{m}\sum_{i=1}^{m}\text{score}_i = \frac{1}{m}\sum_{i=1}^{m}\frac{1}{k_i}. \quad (9)$$

From Table 5, we find that BART-fusion outperforms the original BART on all 4 dataset settings, i.e., the interpretations generated by BART-fusion can help users find music more accurately. If we return a random ranking result, the MRR value is about 0.9%, which is much lower than the model performance.

6. CONCLUSION AND FUTURE WORK

In this paper, we have proposed a novel model to generate interpretations from song lyrics and musical audio. Our modelis based on a pre-trained language model and generates better lyric interpretations by fusing text and music semantic representations. We also have proposed a new

Dataset	Method	MRR (%)
Dataset w/random	BART	25.3
Dataset w/random	BART-fusion	**27.5***
Dataset w/voting > 0	BART	26.1$_+$
Dataset w/voting > 0	BART-fusion	**28.2**$^*_+$
Dataset Full	BART	26.2
Dataset Full	BART-fusion	**30.5***
Dataset w/voting ≥ 0	BART	26.4
Dataset w/voting ≥ 0	BART-fusion	**32.5**$^*_+$

Table 5. Evaluation results for the music retrieval task. $*$: BART-fusion outperforms BART with $p < 0.05$; $+$: the filtered dataset outperforms the unfiltered dataset with $p < 0.05$.

dataset and explored the process of dataset creation, and have investigated how different treatments of the dataset can affect the performance of the model. We have designed an additional experiment that shows that our model outperforms BART on cross-modal music retrieval tasks. Our work has a range of potential applications, such as helping people better understand English lyrics (especially for non-native English speakers) and natural language based music discovery.

The current model still has shortcomings, such as difficulty in understanding complex topics and lack of general common sense. In future work, we will try to improve the model by adding metadata and knowledge base inputs. We also plan to extend the model to describe the musical content of a song in natural language. In addition, large-scale language models may be biased, reinforce stereotypes and introduce harms, which deserves our attention in the future.

7. ACKNOWLEDGEMENT

8. REFERENCES

[1] K. Watanabe and M. Goto, "Lyrics information processing: Analysis, generation, and applications," in *Proceedings of the 1st Workshop on NLP for Music and Audio (NLP4MusA)*, 2020, pp. 6–12.

[2] M. Fell, Y. Nechaev, E. Cabrio, and F. Gandon, "Lyrics segmentation: Textual macrostructure detection using convolutions," in *COLING*, 2018, pp. 2044–2054.

[3] R. Delbouys, R. Hennequin, F. Piccoli, J. Royo-Letelier, and M. Moussallam, "Music mood detection based on audio and lyrics with deep neural net," *arXiv preprint arXiv:1809.07276*, 2018.

[4] Z. Sheng, K. Song, X. Tan, Y. Ren, W. Ye, S. Zhang, and T. Qin, "Songmass: Automatic song writing with pre-training and alignment constraint," *arXiv preprint arXiv:2012.05168*, 2020.

[5] A. Tsaptsinos, "Lyrics-based music genre classification using a hierarchical attention network," *arXiv preprint arXiv:1707.04678*, 2017.

[6] K. Al-Sabahi, Z. Zuping, and M. Nadher, "A hierarchical structured self-attentive model for extractive document summarization (HSSAS)," *IEEE Access*, vol. 6, pp. 24 205–24 212, 2018.

[7] M. Lewis, Y. Liu, N. Goyal, M. Ghazvininejad, A. Mohamed, O. Levy, V. Stoyanov, and L. Zettlemoyer, "BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension," *arXiv preprint arXiv:1910.13461*, 2019.

[8] J. Son and Y. Shin, "Music lyrics summarization method using textrank algorithm," *Journal of Korea Multimedia Society*, vol. 21, no. 1, pp. 45–50, 2018.

[9] M. Fell, E. Cabrio, F. Gandon, and A. Giboin, "Song lyrics summarization inspired by audio thumbnailing," in *RANLP*, 2019.

[10] W. Li, C. Gao, G. Niu, X. Xiao, H. Liu, J. Liu, H. Wu, and H. Wang, "UNIMO: Towards unified-modal understanding and generation via cross-modal contrastive learning," *arXiv preprint arXiv:2012.15409*, 2020.

[11] Y. Zhang, Z. Wang, D. Wang, and G. Xia, "BUTTER: A representation learning framework for bi-directional music-sentence retrieval and generation," in *Proceedings of the 1st Workshop on NLP for Music and Audio (NLP4MusA)*, 2020, pp. 54–58.

[12] L. Zhou, H. Palangi, L. Zhang, H. Hu, J. Corso, and J. Gao, "Unified vision-language pre-training for image captioning and VQA," in *AAAI*, vol. 34, no. 07, 2020, pp. 13 041–13 049.

[13] J. Chen, H. Guo, K. Yi, B. Li, and M. Elhoseiny, "VisualGPT: Data-efficient adaptation of pretrained language models for image captioning," *arXiv preprint arXiv:2102.10407*, 2021.

[14] T. Yu, W. Dai, Z. Liu, and P. Fung, "Vision guided generative pre-trained language models for multimodal abstractive summarization," *arXiv preprint arXiv:2109.02401*, 2021.

[15] S. Dathathri, A. Madotto, J. Lan, J. Hung, E. Frank, P. Molino, J. Yosinski, and R. Liu, "Plug and play language models: A simple approach to controlled text generation," *arXiv preprint arXiv:1912.02164*, 2019.

[16] M. Won, S. Chun, and X. Serra, "Toward interpretable music tagging with self-attention," *arXiv preprint arXiv:1906.04972*, 2019.

[17] Y.-H. H. Tsai, S. Bai, P. P. Liang, J. Z. Kolter, L.-P. Morency, and R. Salakhutdinov, "Multimodal transformer for unaligned multimodal language sequences," in *ACL*, vol. 2019. NIH Public Access, 2019, p. 6558.

[18] K. Song, X. Tan, T. Qin, J. Lu, and T.-Y. Liu, "MASS: Masked sequence to sequence pre-training for language generation," *arXiv preprint arXiv:1905.02450*, 2019.

[19] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. J. Liu, "Exploring the limits of transfer learning with a unified text-to-text transformer," *arXiv preprint arXiv:1910.10683*, 2019.

[20] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[21] A. Radford, K. Narasimhan, T. Salimans, I. Sutskever *et al.*, "Improving language understanding by generative pre-training," 2018.

[22] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in Neural Information Processing Systems*, 2017, pp. 5998–6008.

[23] G. Jawahar, B. Sagot, and D. Seddah, "What does BERT learn about the structure of language?" in *ACL*, 2019.

[24] C. Li, X. Gao, Y. Li, B. Peng, X. Li, Y. Zhang, and J. Gao, "OPTIMUS: Organizing sentences via pre-trained modeling of a latent space," *arXiv preprint arXiv:2004.04092*, 2020.

[25] K. Choi, J. H. Lee, X. Hu, and J. S. Downie, "Music subject classification based on lyrics and user interpretations," *AIST*, vol. 53, no. 1, pp. 1–10, 2016.

[26] I. Manco, E. Benetos, E. Quinton, and G. Fazekas, "Muscaps: Generating captions for music audio," *CoRR*, vol. abs/2104.11984, 2021. [Online]. Available: https://arxiv.org/abs/2104.11984

[27] I. A. P. Santana, F. Pinhelli, J. Donini, L. Catharin, R. B. Mangolin, V. D. Feltrim, M. A. Domingues *et al.*, "Music4all: A new music database and its applications," in *IWSSIP*. IEEE, 2020, pp. 399–404.

[28] N. Shazeer and M. Stern, "Adafactor: Adaptive learning rates with sublinear memory cost," in *ICML*. PMLR, 2018, pp. 4596–4604.

[29] C.-Y. Lin, "ROUGE: A package for automatic evaluation of summaries," in *Text Summarization Branches Out: Proceedings of the ACL-04 Workshop*, 2004, pp. 74–81.

[30] S. Banerjee and A. Lavie, "METEOR: An automatic metric for MT evaluation with improved correlation with human judgments," in *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, 2005, pp. 65–72.

[31] T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger, and Y. Artzi, "BERTScore: Evaluating text generation with BERT," *arXiv preprint arXiv:1904.09675*, 2019.

[32] N. Reimers and I. Gurevych, "Sentence-BERT: Sentence embeddings using Siamese BERT-networks," *arXiv preprint arXiv:1908.10084*, 2019.

[33] O. Chapelle, D. Metlzer, Y. Zhang, and P. Grinspan, "Expected reciprocal rank for graded relevance," in *Proceedings of the 18th ACM Conference on Information and Knowledge Management*, 2009, pp. 621–630.