Annotator Subjectivity in the MusicCaps Dataset

Minhee Lee[1], SeungHeon Doh[2] and Dasaem Jeong[3,*]

[1]Department of Computer Science and Engineering, Sogang University, South Korea
[2]Graduate School of Culture Technology, KAIST, South Korea
[3]Department of Art & Technology, Sogang University, South Korea

Abstract

Musical caption, when expressed in free-form text as opposed to more structured and limited musical tags, often encompasses the individual characteristics of the annotator, thereby injecting a degree of subjectivity into the resultant dataset. This study explores the impact of such annotator subjectivity within the MusicCaps dataset, a pioneering collection of human-annotated captions explaining 10-second music audio clips. We conducted three distinct analyzes to investigate the presence of this subjectivity. This includes examining the frequency distribution of tag categories (i.e., genre, mood, or instruments) among different annotators, a qualitative assessment of caption embeddings through UMAP visualizations, and a quantitative analysis where we train and compare cross-modal retrieval models using an annotator-specified training split. Our findings underscore the significant annotator subjectivity inherent in the MusicCaps dataset, emphasizing the need for its consideration when collecting free-form text annotations on music or developing machine-learning models using this type of dataset.

Keywords

Annotator subjectivity, Music caption, Music dataset

1. Introduction

As text-image multi-modal models such as CLIP [1] or text-to-image generation like DALL-E [2] and Latent Diffusion [3] evolve, the interest in multi-modality between music and text became more immense. This led the research community to introduce a new form of music dataset, a music caption dataset. Manco et al. [4] suggested collecting music caption data from the public, and Agostinelli et al. [5] introduced the first publicly available music caption dataset.

However, while free-form descriptions allow a full range of creative expression, they can be influenced by the annotator's personal characteristics. This paper analyzes the annotator subjectivity in a music caption dataset and its effect on audio-text joint embedding space. Our main contributions include: (1) identifying the subjectivity in the music caption dataset; (2) analyzing the impacts of this subjectivity in training audio-text joint embedding space.

2. MusicCaps Dataset

We use the MusicCaps dataset [5], which consists of 5.5k music-text pairs. Each dataset entry comprises a 10-second audio clip sourced from the AudioSet dataset [6], paired with a free-text

HCMIR23: 2nd Workshop on Human-Centric Music Information Research, November 10th, 2023, Milan, Italy
*Corresponding author.
mini@sogang.ac.kr (M. Lee); seungheondoh@kaist.ac.kr (S. Doh); dasaemj@sogang.ac.kr (D. Jeong)

Table 1
Proportions of specific musical characteristics described in captions, expressed as a percentage of each annotator's total annotations. Annotators who **most** frequently or least frequently included each feature are denoted in bold and underlined text, respectively. Note that column totals may not sum to 100% as multiple music categories can co-occur in a single caption, rendering the categories non-disjoint.

annotator	0	1	2	3	4	5	6	7	8	9
tempo	19.7	95.3	10.7	8.6	4.2	12.1	14.6	**95.6**	18.9	12.4
instrument	**96.0**	81.8	85.7	65.4	75.7	84.8	79.3	79.0	78.3	82.3
mood	33.6	65.2	11.6	48.5	**72.3**	51.5	12.3	69.2	14.4	40.9
genre	39.2	**77.9**	34.8	30.6	41.6	21.2	5.6	72.0	28.9	42.7
audio quality	23.8	48.3	15.2	29.2	**87.5**	9.1	29.3	16.1	11.1	25.3
theme	80.9	0.0	49.1	7.0	7.7	15.2	**94.4**	1.0	4.4	76.1
annotated examples	554	489	112	301	1383	33	895	708	180	866

caption and a list of music aspects. Note that the MusicCaps dataset uses the term *aspect* to represent tag-like annotations. The captions were annotated by one of ten annotators, all of whom were professional musicians. Each entry in the dataset includes metadata that indicates the respective annotator, represented by an author identifier ranging from 0 to 9. This dataset was initially designed for the evaluation of text-to-music generation. Nonetheless, given the dataset's versatility for training and evaluating various music and language models, including text-to-music retrieval [7, 8, 9, 10], generation [5, 11], and music captioning [12, 13], we analyze this dataset with various approaches that extend beyond its original intent.

3. Subjectivity in Tag Categories Distribution

The first thing we analyzed was the difference in distribution of tag categories that annotators focus on while captioning. We calculated the distribution by counting the number of captions that contain the list of selected tags for each category. We chose musical keywords from the aspect list over the caption since the aspects offer a clearer keyword representation than the terms extracted directly from the caption. To identify the most representative keywords that are frequently used, we first collated the top 50 of the most frequently annotated aspects per annotator. Then we sorted the collated aspects following two criteria in order: first, by the number of annotators that have included the aspect in their top 50, and second, by the sum of the percentages of occurrences of the aspect. We get the top 50 aspects from the sorted results and divide them into five categories - *tempo*, *instrument*, *mood*, *genre*, and *audio quality*, referring to the explanation of the aspect list in [5]. We also paid attention to *theme* descriptions, which describe potential uses of the music, like "*This folk song can be played in a movie scene set in a Moroccan market.*" We used regular expression to detect the thematic descriptions, as they are often described through a specific phrase such as 'could be used' or 'may be playing' without including distinct musical aspect keywords.

Table 1 shows the counted result. We can see some extreme deviations, such as *theme* annotated in 94.4% of annotator 6's samples, but none and 1.0% samples of annotators 1 and

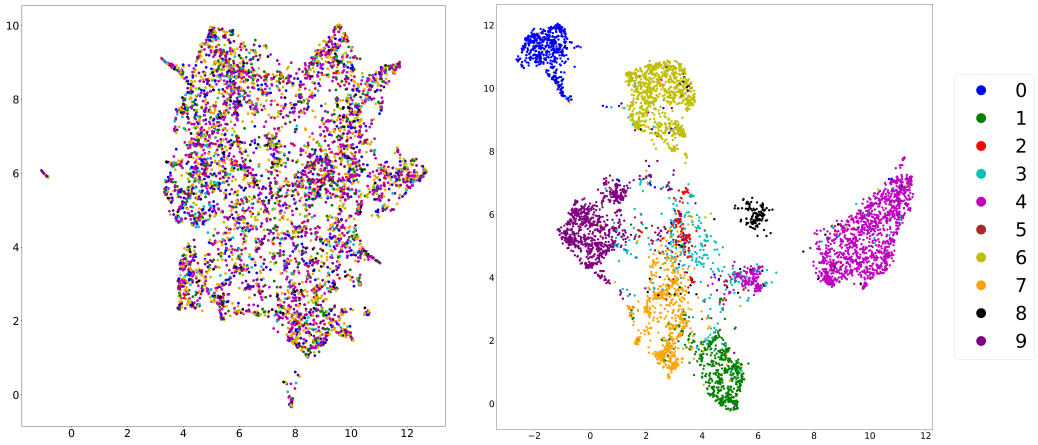

Figure 1: Visualization of audio (left) and caption (right) embeddings using UMAP. Colors represent different annotating annotators. Unlike audio, caption embeddings form clusters by the annotator.

7. Similarly, *tempo* was annotated in 96.5% and 95.3% samples of annotators 7 and 1, but only 4.2% among the samples annotated by annotator 4. This indicates that consideration of tag categories when captioning significantly varies by annotators, thus affecting the resulting captions' characteristics.

4. Subjectivity Illustrated in Semantics

We extracted feature embeddings to analyze the semantics of each music and caption quantitatively. For audio feature embeddings, we used VGGish model [14] pre-trained with the AudioSet dataset, and averaged across the time axis to get a single embedding for each 10-sec audio. We used 5,480 music examples, of which the audio source was obtainable. For text captions, we use a pre-trained BERT model [15] and extract the embeddings from the last hidden state of [CLS] token to capture the comprehensive semantics of the entire sentence. The resulting embeddings for audio and text are 128-dimensional and 768-dimensional, respectively.

We visualized 2D UMAP [16] embedding space in Figure 1. The plot for caption embedding shows clear cluster by the annotator. This means that the neighbors of each caption in the BERT embedding space are mainly decided by who wrote the caption, not by which music it described. On the other hand, audio embeddings do not show author-wise clusters, which implies that the audio samples were not assigned to the annotators by their preferences or choices.

To quantify how distinctive those embeddings are, we conducted an experiment to predict the annotator from the embedding using a random forest classifier with 100 decision trees. The experiment yielded an average F1 score of 0.76 across all annotators, underlining the influence of annotator-specific subjectivity in text embeddings. To assess whether these clusters could result from each annotator being assigned to specific music or favoring music they are familiar with, we carried out the same prediction experiment on audio embeddings. The result was an F1 score of 0.08 on average, which implies that the audio samples were randomly assigned to the annotators regardless of their musical preferences. Therefore, we can presume that the

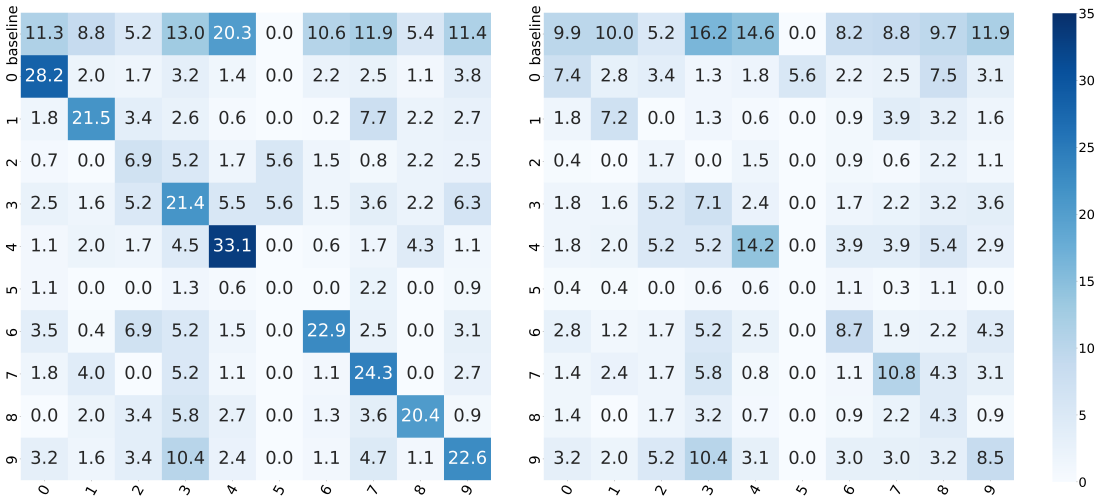

Figure 2: R@10 results for audio-to-text retrieval (left) and text-to-audio retrieval (right). Each row represents a specific model differentiated by its training set, while each column signifies the corresponding test set employed.

clear distinction between each annotator's caption embedding originates from the annotator subjectivity.

5. Influence of Subjectivity on Joint Embedding Space

We also explored the impact of the subjectivity observed within the feature embedding space on the other applications. A common area of research in audio and text is cross-modal retrieval tasks that involve the use of audio-text joint embedding spaces [7, 8, 9, 10].

Following the previous works [7, 10], we trained a dual-encoder using the symmetric version of InfoNCE Loss [17, 1]. Each encoder comprises a pre-trained backbone model (VGGish, BERT) and a linear projection layer without bias. During training, we kept the backbone model frozen and only updated the linear projection layer. The audio and text projection layers were jointly trained to maximize the similarity between N positive pairs while minimizing the similarity for $N \times (N-1)$ negative pairs. All models were optimized using Adam [18] with a learning rate of 1.5e-4. We used a batch size of 32, and the models were trained for 100 epochs. To prevent overfitting, 25% of the training samples were reserved as a validation set, and early stopping based on validation loss was implemented.

For this experiment, we employed an annotator-specific training/test split to investigate the generalizability of a joint embedding model across different annotators' audio-caption pairs. Each separated dataset is represented as (A_i, T_i), where A_i and T_i denote the sets of audio and text examples annotated by the annotator i. We trained an annotator-specified model model_i with $(A_{i,\text{train}}, T_{i,\text{train}})$ for each annotator. The baseline model for the comparison was trained with the whole train set $(A_{i,\text{train}}, T_{i,\text{train}}) : i \in \{0, 1, 2, ..., 9\}$. We evaluate the performance of the baseline model and model_i by computing Recall at K (R@K) of $(A_{i,\text{test}}, T_{i,\text{test}})$ over the whole test set $(A_{i,\text{test}}, T_{i,\text{test}}) : i \in \{0, 1, 2, ..., 9\}$, for both audio-to-text and text-to-audio retrieval

tasks.

The train-evaluation split provided by MusicCaps metadata, inherited from the AudioSet dataset, yields imbalanced partitions across the datasets of individual annotators. Importantly, no examples of annotators 2 and 5 were included in the MusicCaps training set. As an alternative, we employed our own partitioning strategy, maintaining the original train-evaluation ratio of 48:52 but applying it in an annotator-specific manner. Note that there exists an imbalance in the sample distribution across annotators. In particular, annotator 5 has the smallest test sample size with only 18 examples, followed by annotator 2 with 58 test samples.

The result is presented in Figure 2. As shown in the result, the annotator-specified models outperform the baseline model for the target annotator in audio-to-text retrieval, even though it used a smaller training set. Conversely, the annotator-specific models exhibit significantly reduced accuracy compared to the baseline when applied to other annotators' data subsets. When trained to align audio with captions from a specific annotator (in domain), it does not generalize well to audios paired with captions from other annotators (out of domain). We assume that this is due to annotator subjectivity.

In contrast, the baseline model showed better results in the text-to-audio retrieval task. The annotator-specified model showed severely degraded performance even on the target annotator compared to the audio-to-text retrieval. In audio-to-text retrieval, the clear differences in the description of the caption might help to narrow down the candidates to a specific annotator's caption. For instance, a model trained on the captions from annotator 6—who includes theme descriptions in 94.4% and genre descriptions in only 5.6% of the captions—would prioritize retrieving captions that include theme descriptions but not genre descriptions. This can largely narrow down the retrieval candidates, thus increasing the accuracy of the target-specified model.

Nonetheless, in text-to-audio retrieval tasks, the same difference in the text caption failed to enhance the performance, as evidenced by the result. This limitation arises because the stylistic differences of the query text are insufficient to refine the pool of audio candidates for retrieval. The enhanced accuracy achieved through mixed training sets, as opposed to annotator-specific sets, suggests that incorporating captions from multiple annotators can improve text-to-audio retrieval performance. This implies that the captions still share commonness in how they describe the audio, regardless of annotator subjectivity.

The observed asymmetry between audio-to-text and text-to-audio retrieval outcomes indicates that annotator subjectivity exerts a more pronounced impact when text serves as a retrieval candidate rather than a query.

6. Conclusion and Future Work

In this paper, we have shown the clear presence of annotator subjectivity in the music caption dataset and the consequent impact on the embedding space and its applications. In our future work, we aim to figure out what leads to this subjectivity and how we might be able to lessen its effects. Its effect on other tasks, such as text-based music generation, must also be investigated, as the style of the captions can largely influence the generation result. We hope that our work helps the research community to be aware of and consider annotator subjectivity when working

with music caption datasets.

Acknowledgement

This work was supported by Sogang University Research Grant of 202110035.

References

[1] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, et al., Learning transferable visual models from natural language supervision, in: International conference on machine learning, PMLR, 2021, pp. 8748–8763.

[2] A. Ramesh, M. Pavlov, G. Goh, S. Gray, C. Voss, A. Radford, M. Chen, I. Sutskever, Zero-shot text-to-image generation, in: International Conference on Machine Learning, PMLR, 2021, pp. 8821–8831.

[3] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, B. Ommer, High-resolution image synthesis with latent diffusion models, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2022, pp. 10684–10695.

[4] I. Manco, B. Weck, P. Tovstogan, M. Won, D. Bogdanov, Song describer: a platform for collecting textual descriptions of music recordings, in: Ismir 2022 Hybrid Conference, 2022.

[5] A. Agostinelli, T. I. Denk, Z. Borsos, J. Engel, M. Verzetti, A. Caillon, Q. Huang, A. Jansen, A. Roberts, M. Tagliasacchi, et al., Musiclm: Generating music from text, arXiv preprint arXiv:2301.11325 (2023).

[6] J. F. Gemmeke, D. P. Ellis, D. Freedman, A. Jansen, W. Lawrence, R. C. Moore, M. Plakal, M. Ritter, Audio set: An ontology and human-labeled dataset for audio events, in: 2017 IEEE international conference on acoustics, speech and signal processing (ICASSP), IEEE, 2017, pp. 776–780.

[7] I. Manco, E. Benetos, E. Quinton, G. Fazekas, Contrastive audio-language learning for music, arXiv preprint arXiv:2208.12208 (2022).

[8] Y. Wu, K. Chen, T. Zhang, Y. Hui, T. Berg-Kirkpatrick, S. Dubnov, Large-scale contrastive language-audio pretraining with feature fusion and keyword-to-caption augmentation, in: ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, 2023, pp. 1–5.

[9] J. Choi, J. Lee, J. Park, J. Nam, Zero-shot learning for audio-based music classification and tagging, arXiv preprint arXiv:1907.02670 (2019).

[10] S. Doh, M. Won, K. Choi, J. Nam, Toward universal text-to-music retrieval, in: ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, 2023, pp. 1–5.

[11] H. Liu, Z. Chen, Y. Yuan, X. Mei, X. Liu, D. Mandic, W. Wang, M. D. Plumbley, Audioldm: Text-to-audio generation with latent diffusion models, arXiv preprint arXiv:2301.12503 (2023).

[12] I. Manco, E. Benetos, E. Quinton, G. Fazekas, Muscaps: Generating captions for music

audio, in: 2021 International Joint Conference on Neural Networks (IJCNN), IEEE, 2021, pp. 1–8.
[13] S. Doh, K. Choi, J. Lee, J. Nam, Lp-musiccaps: Llm-based pseudo music captioning, arXiv preprint arXiv:2307.16372 (2023).
[14] S. Hershey, S. Chaudhuri, D. P. Ellis, J. F. Gemmeke, A. Jansen, R. C. Moore, M. Plakal, D. Platt, R. A. Saurous, B. Seybold, et al., Cnn architectures for large-scale audio classification, in: 2017 ieee international conference on acoustics, speech and signal processing (icassp), IEEE, 2017, pp. 131–135.
[15] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding, arXiv preprint arXiv:1810.04805 (2018).
[16] L. McInnes, J. Healy, J. Melville, Umap: Uniform manifold approximation and projection for dimension reduction, arXiv preprint arXiv:1802.03426 (2018).
[17] A. v. d. Oord, Y. Li, O. Vinyals, Representation learning with contrastive predictive coding, arXiv preprint arXiv:1807.03748 (2018).
[18] D. P. Kingma, J. Ba, Adam: A method for stochastic optimization, arXiv preprint arXiv:1412.6980 (2014).