Zero-Shot Music Retrieval For Japanese Manga

Megha Sharma and Yoshimasa Tsuruoka

The University of Tokyo
{meghas, yoshimasa-tsuruoka}@g.ecc.u-tokyo.ac.jp

Abstract. This demo proposes a novel task for curating theme music for manga (Japanese comics). One of the biggest challenges in this field is the lack of available paired data for manga and music. Hence, we employ alignment properties of pre-trained models to infer the relationship between music and manga and retrieve music given an input manga page. We call this zero-shot, as we do not train on any explicit aligned music-manga dataset. Our preliminary results show potential in the task of music retrieval from manga when fine tuned on independent manga-text and music-text pairs compared to the original AudioCLIP model.

Keywords: Music Retrieval, Multi-Modal, Manga, Emotion-Aware Retrieval

1 Introduction

Storytelling shares fictional or non-fictional accounts for knowledge, entertainment, and even branding in modern society [1]. To bridge the gap between reality and stories, artists make use of additional modalities such as illustrations, onomatopoeia, sound and movement. Comic books are an example of story telling that employs illustrations and onomatopoeia. In a country like Japan, where manga or Japanese comics hold significance in history of popular culture [2], the evolution of comic story telling has experienced waves of digitisation and animation [3]. In an effort to enhance the reading experience, such adaptations also use sound effects and music with digital comics[1].

However, additional modalities can be expensive and require domain knowledge. Attempts have been made to curate background music for books using the soundtracks from the movie adaptations [4]. However, curating background music for comics remains a novel task. To our knowledge, our work remains one of the first attempts to curate music based on comics. One of our biggest challenges remains the lack of publicly available music datasets from manga or anime. Hence, we focus on extracting alignment relationships between music and manga books by training a shared embedding space of image, text and audio. In the current stage, our demo proposes a solution to retrieve music for manga based on a given page, by aligning the relationships implicitly through text. Our method is inspired from the recent success of ImageBind

[1] https://www.webtoons.com/en/

[6], which trained six modalities together by binding them with image embeddings. A shared embedding space is trained with independent datasets: AudioSet [7] (Music-Text) and Manga109 [8] (Manga-Text). In this paper, we report the progress of the training and results from our first iteration. We also comment on our current limitations and future work for the task. The code and some example runs are made available at `https://github.com/ms3744/Music-Manga-Retrieval`.

2 Implementation

At the current stage, we approach the problem as a manga-to-music retrieval task. Hence, our focus is on strengthening the embedding space between the modalities. An ESResNeXT encoder [12] is used for audio while the Res-Net and Transformer encoders from CLIP [11] are used as image and text encoders respectively. The three encoders share a multi-modal embedding space using the AudioCLIP architecture [10], which is an extension of the CLIP model with an audio encoder. We use the pre-trained weights from AudioCLIP as large-scale trained models exhibit emergent properties in modality alignment [6]. We first train the audio encoder on (audio-text) pairs, and then finetune the image and text encoders on (image-text) pairs.

Datasets. The model is fine-tuned on two datasets independently. The first is a subset of AudioSet [7]. We use the "Music Mood"[2] collection which contains music audio classified into seven classes, namely: *Happy, Funny, Sad, Tender, Exciting, Angry, and Scary*. We use the label of each audio as its corresponding text in the audio-text pairs. Due to the subjectivity of music retrieval, we focus on class-based descriptors instead of long-form descriptors as seen in MusicCap [9] to analyse the preliminary results in the current stage. We also use the Manga109 [8] dataset for manga-text pairs, which contains a collection of pages of 109 manga books from 12 genres. We use the genre of each manga as the corresponding text label for the manga-text pairs.

Model Hyper-parameters. The model is trained with an SGD optimizer, using a momentum of 0.9. The learning rate is 5e-5. While training on AudioSet, we trained the model for 30 epochs, used a batch size of 32, and applied audio augmentation techniques on the training set from [10]. For training on Manga109, we trained the model for 50 epochs, used a batch size of 64, and applied image normalisation techniques from CLIP [11] on the training and validation set . Since the AudioCLIP model is already pre-trained on AudioSet, we need fewer epochs to improve the performance on the music mood subset. We only saved the model with the best validation loss while training. For training the model with each dataset, we use a symmetric contrastive learning loss [13] for the (text, image) pairs with Manga109 and (text, audio) pairs with Audioset.

Evaluation Methods. Given the subjective nature of the task, our focus was on measuring the quantitative performance while training. We use the mean Average Precision (mAP) scores and mean accuracy scores for evaluating on validation sets of AudioSet and Manga109, respectively. For the qualitative performance, we report example retrievals based on input images and audios, and discuss our results here.

[2] https://research.google.com/audioset/ontology/music_mood_1.html

3 Results and Discussion

Table 1 highlights the validation results on the two datasets, before and after fine-tuning the model. We achieve a 4% improvement on the Music-Mood subset of AudioSet after training the model further for 30 epochs. Unlike the original training in AudioCLIP [10], we train only on audio and text modalities to strengthen the relationship between music and text. We achieved high accuracy on the Manga109 dataset after training the model for 50 epochs compared to the zero-shot accuracy. The zero-shot model tends to classify images of manga panels as "fantasy" genre label. However, the pre-trained encoders from CLIP [11] fine-tune well on the manga dataset, and the model achieves over 97% accuracy on classifying input images. On the other hand, the audio-head is a much slower learner, and does not seem to fine-tune as well on the audio dataset.

Table 1. Classification results from the validation set. We used a train-validation split of 80-20. *Original number of epochs for training the AudioCLIP on AudioSet [10]. **Results from zero-shot classification using CLIP [11]

Dataset	Metric	Epochs	Score
AudioSet (Music Mood Subset)	mAP	60	40.8%
		30*	36.8%
Manga109	Accuracy	50	97.2%
		0**	6.9%

To understand the emergent trends of the model on the unaligned music and manga, we queried the model on 200 samples for image (manga) to music and vice versa. We then calculated the average confidence for each music mood retrieved given all genres of manga and vice versa. Our qualitative results imply some subjective emergence of relation between the moods of the music and the genre of the Manga. When querying images of genre "historical drama", *Sad* (54% confidence) and *Tender* (45% confidence) are the most common moods of the retrieved audio files. Genres like "romantic comedy" retrieve *Happy* (37% confidence) while "suspense" retrieves *Scary* (34.4% confidence) music. However, certain genres like "horror" have very low confidence, with *Tender* music (19.6% confidence) being the strongest case. We note that the relationship music to manga retrieval is not as strong. "animal" is a common manga genre that is most likely to be retrieved when querying *Angry, Happy, Scary and Tender* music. We plan to continue experiments to understand the reason for this behaviour in future work.

Although there is no explicit baseline in our model, we compare our results to original AudioCLIP [10]. Interestingly, when we query any manga image, the model does not achieve confidence beyond 20%. The highest confidence is *Sad* music (15.6%) for "historical drama" manga. In general, the model has 5% confidence, and retrieved music with poor subjective compatibility, such as retrieving *Angry* music for "humour" and "romantic comedy" manga. Similarly, music to manga retrieval performs inconsistently, with 72% confidence for retrieving "romantic comedy" manga for *Angry* music. Overall, the model seems to achieve better zero-shot retrieval after training the model. This

implies the potential of the model for building a music generation system for Manga without aligned music and manga datasets.

4 Future Work and Conclusion

We propose a task for retrieving music from pages of Japanese manga and demonstrate how strong multi-modal embeddings have the potential to solve the novel task through emergent properties. We plan to demonstrate this capability to a larger audience, and understand the behaviour of these emergent relationships. We view our work as preliminary to future work including, improving training efficiencies for the audio encoder, incorporating long form text such as manga dialogues and music descriptions, and conducting a qualitative test of the model. Ultimately, we plan to build a decoder for the model to generate novel music for any given Manga.

References

1. Fog, K., Budtz, C., Yakaboylu, B.: Storytelling. Springer (2005)
2. Bouissou, J. M.: Manga: A historical overview. Manga: An anthology of global and cultural perspectives. 17-33 (2010)
3. Wong, T. T., Igarashi, T., Xu, Y. Q., Shi, D..: The Computational manga and anime. In SIGGRAPH Asia 2013 Courses. 1–52 (2013)
4. Shriram, J., Tapaswi, M., Alluri, V.: Sonus Texere! Automated Dense Soundtrack Construction for Books using Movie Adaptations. In: Proceedings of the 23rd International Society for Music Information Retrieval Conference. (2022)
5. Brenner, Robin E.: Understanding manga and anime. Greenwood Publishing Group (2007)
6. Girdhar, R., El-Nouby, A., Liu, Z., Singh, M., Alwala, K. V., Joulin, A., Misra, I.: Imagebind: One embedding space to bind them all. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 15180–15190. (2023)
7. Gemmeke, J.F., Ellis, D.P., Freedman, D., Jansen, A., Lawrence, W., Moore, R.C., Plakal, M. Ritter, M.: Audio set: An ontology and human-labeled dataset for audio events. In: 12017 IEEE international conference on acoustics, speech and signal processing (ICASSP), pp. 776–780. IEEE (2017)
8. Fujimoto, A., Ogawa, T., Yamamoto, K., Matsui, Y., Yamasaki, T., Aizawa, K.: Manga109 dataset and creation of metadata. In: Proceedings of the 1st international workshop on comics analysis, processing and understanding, pp. 1–5. IEEE (2016)
9. Agostinelli, A., Denk, T.I., Borsos, Z., Engel, J., Verzetti, M., Caillon, A., Huang, Q., Jansen, A., Roberts, A., Tagliasacchi, M., Sharifi, M.: Musiclm: Generating music from text. arXiv preprint arXiv:2301.11325 (2023)
10. Guzhov, A., Raue, F., Hees, J., Dengel, A.: Audioclip: Extending clip to image, text and audio. In: ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 976–980. IEEE (2022)
11. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G.: Learning transferable visual models from natural language supervision. In: International conference on machine learning, pp. 8748–8763. PMLR (2021)
12. Guzhov, A., Raue, F., Hees, J., Dengel, A.: Esresne (x) t-fbsp: Learning robust time-frequency transformation of audio. In: 2021 International Joint Conference on Neural Networks (IJCNN), pp. 1–8. IEEE (2021)
13. Oord, A.V.D., Li, Y., Vinyals, O.: Representation learning with contrastive predictive coding. arXiv preprint arXiv:1807.03748 (2018)