DISCOGS-VINET-MIREX

R. Oğuz Araz[1] **Joan Serrà**[2] **Xavier Serra**[1]
Yuki Mitsufuji[2] **Dmitry Bogdanov**[1]
[1] Universitat Pompeu Fabra, Music Technology Group, Barcelona
[2] Sony AI
recepoguz.araz@upf.edu

ABSTRACT

This technical report presents our submission to the cover song identification task for the 2024 edition of the Music Information Retrieval Evaluation eXchange (MIREX). For this submission, we enhanced our Discogs-VINet model by changing the definition of an epoch, incorporating automatic mixed precision (AMP) during both training and inference, and sampling four versions per clique during triplet mining (which became possible with AMP). Due to this enhanced model's performance on the Discogs-VI test set, we trained a new model from scratch using the entire Discogs-VI dataset, rather than just the training partition used in Discogs-VINet (a 45% increase in the number of versions). This enhanced and retrained model is named Discogs-VINet-MIREX.

1. INTRODUCTION

Version identification (VI), also known as cover song identification (CSI), aims to identify the different versions of a musical work from a collection of tracks [1]. The identification process relies on generating digital audio representations of tracks, where the representations of versions of the same musical work are designed to be closer to each other compared to non-version tracks. In contemporary VI approaches, closeness is typically measured using a vector space operation such as the cosine similarity or Euclidean distance. Audio representations are obtained by training neural networks on datasets composed of multiple sets of versions, known as cliques. During the retrieval phase, the pre-trained neural network creates such representations, which are used for identifying versions.

Datasets such as Da-TACOS [2] or SHS100K [3] were commonly used to train neural networks for VI. However, the relatively small size of these datasets became a limiting factor in advancing VI models. To address this challenge, the Discogs-VI-YT dataset was recently introduced [4]. This new dataset offers a significant improvement, containing more than four times the number of versions found in the SHS100K dataset and approximately ten times the amount of cliques.

With the release of the Discogs-VI-YT dataset, we also introduced the Discogs-VINet model [4], built upon the CQT-Net [5] architecture. Discogs-VINet was designed to be trainable on a single commercial-grade GPU such as the NVIDIA RTX2080, and to exclude any data augmentations during training, highlighting the dataset's potential without additional pre-processing techniques. Given the presence of about 80,000 training set cliques, training the model with a classification objective was impractical on a commercial GPU. As a result, the model was trained exclusively using the triplet loss. In this submission, we improve the model in several aspects.

2. PROPOSED SYSTEM

2.1 Model

For this submission, we implemented several key modifications to the Discogs-VINet model. First, we used non-deterministic CUDA operations to cut back on the considerably long training time to handle the large amount of data in the Discogs-VI-YT dataset. Then, we introduced automatic mixed precision (AMP) training with the same goal. Importantly, besides increasing computation speed, AMP enables larger batch sizes, which are important for effective triplet mining strategies. During each training iteration, we randomly sample 54 cliques, with four versions selected from each clique, resulting in a total batch size of 216. If a clique contains fewer than four versions, we duplicate versions as needed. Each version in the batch serves as an anchor sample during triplet mining.

We utilize online triplet mining with random positives and hard negatives. An epoch is defined as one pass over the entire training set where each version has been used as the anchor sample once. We train the model for 40 epochs with this new and improved definition. For each version, we extract 7,600 constant-Q transform frames (about 176.4 s with 22,050 Hz sampling rate and a hop size of 512 samples). The learning rate was scheduled using cosine annealing with five warm-up steps, starting at 0.0001, reaching 0.01 at the end of the warm-up, and annealing down to 0.00001 by the end of training. The model generates 512-dimensional embeddings, and we apply L2-normalization after the last layer to ensure the embeddings lie on the unit hypersphere. This version of the

Attribution: R. O. Araz, J. Serrà, X. Serra, Y. Mitsufuji, and D. Bogdanov, "Discogs-VINet-MIREX", in *Proc. of the 25th Int. Society for Music Information Retrieval Conf.,* San Francisco, United States, 2024.

model was trained on the entire Discogs-VI-YT dataset, which contains approximately 493,000 versions of about 98,000 compositions. We also increased the number of parameters to 8 million from 5 million by using multiples of 40 channels in the convolutional layers instead of 32. The triplet loss margin was set to 0.3, with all other parameters following the Discogs-VINet configuration.

2.2 Retrieval

We use maximum inner product search for retrieval. Since our model creates L2-normalized representations, this is equivalent to maximum cosine similarity search.

2.3 Evaluation

We trained Discogs-VINet-MIREX on the entire Discogs-VI-YT dataset, and since Discogs-VI-YT contains most of the cliques of both SHS100K and Da-TACOS datasets, we did not evaluate the model on any external datasets. However, during training we monitored the validation set performance to guarantee over-fitting, and the model achieved about 0.91 mAP and 4 MR1 performance by the end of training. This final model was benchmarked in the competition.

3. ACKNOWLEDGMENTS

This work is supported by the pre-doctoral program AGAUR-FI ajuts (2024 FI-3 00065) Joan Oró, funded by the Secretaria d'Universitats i Recerca del Departament de Recerca i Universitats de la Generalitat de Catalunya; and the Cátedras ENIA program "IA y Música: Cátedra en Inteligencia Artificial y Música" (TSI-100929-2023-1), funded by the Secretaría de Estado de Digitalización e Inteligencia Artificial and the European Union-Next Generation EU.

4. REFERENCES

[1] F. Yesiler, G. Doras, R. M. Bittner, C. J. Tralie, and J. Serrà, "Audio-Based Musical Version Identification: Elements and challenges," *IEEE Signal Processing Magazine*, vol. 38, no. 6, pp. 115–136, 2021.

[2] F. Yesiler, C. Tralie, A. Correya, D. F. Silva, P. Tovstogan, E. Gómez, and X. Serra, "Da-TACOS: A Dataset for Cover Song Identification and Understanding," in *Proc. of the 20th Int. Soc. for Music Information Retrieval Conf. (ISMIR)*, 2019.

[3] X. Xu, X. Chen, and D. Yang, "Key-Invariant Convolutional Neural Network Toward Efficient Cover Song Identification," in *IEEE Int. Conf. on Multimedia and Expo (ICME)*, 2018.

[4] R. O. Araz, X. Serra, and D. Bogdanov, "Discogs-VI: A musical version identification dataset based on public editorial metadata," in *Proc. of the 25th Int. Soc. for Music Information Retrieval Conf. (ISMIR)*, 2024.

[5] Z. Yu, X. Xu, X. Chen, and D. Yang, "Learning a Representation for Cover Song Identification Using Convolutional Neural Network," in *IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, 2020.