CLAMP: CONTRASTIVE LANGUAGE-MUSIC PRE-TRAINING FOR CROSS-MODAL SYMBOLIC MUSIC INFORMATION RETRIEVAL

Shangda Wu[1]   **Dingyao Yu**[2]   **Xu Tan**[2]   **Maosong Sun**[1,3]

[1] Central Conservatory of Music, China   [2] Microsoft Research Asia
[3] Tsinghua University, China

shangda@mail.ccom.edu.cn, {v-dingyaoyu, xuta}@microsoft.com, sms@tsinghua.edu.cn
https://ai-muzic.github.io/clamp

ABSTRACT

We introduce **CLaMP**: **C**ontrastive **La**nguage-**M**usic **P**re-training, which learns cross-modal representations between natural language and symbolic music using a music encoder and a text encoder trained jointly with a contrastive loss. To pre-train CLaMP, we collected a large dataset of 1.4 million music-text pairs. It employed text dropout as a data augmentation technique and bar patching to efficiently represent music data which reduces sequence length to less than 10%. In addition, we developed a masked music model pre-training objective to enhance the music encoder's comprehension of musical context and structure. CLaMP integrates textual information to enable semantic search and zero-shot classification for symbolic music, surpassing the capabilities of previous models. To support the evaluation of semantic search and music classification, we publicly release WikiMusicText (WikiMT), a dataset of 1010 lead sheets in ABC notation, each accompanied by a title, artist, genre, and description. In comparison to state-of-the-art models that require fine-tuning, zero-shot CLaMP demonstrated comparable or superior performance on score-oriented datasets. Our models and code are available at https://github.com/microsoft/muzic/tree/main/clamp.

1. INTRODUCTION

Symbolic Music Information Retrieval (MIR) is a field that deals with the automatic analysis and retrieval of music based on symbolic representations such as sheet music or MIDI files. Symbolic MIR has numerous practical applications, including music genre classification [1, 2], automatic music transcription [3, 4], and music recommendation systems [5]. However, traditional symbolic MIR approaches based on handcrafted features are often limited in their ability to capture the complex nature of music.

Deep learning has become increasingly popular in symbolic MIR [6–9] due to its ability to extract complex and
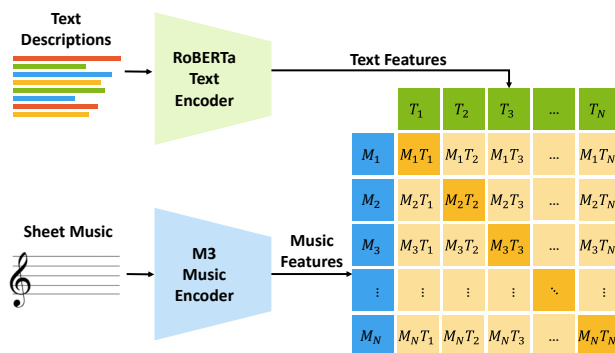
Figure 1. The architecture of CLaMP, including two encoders - one for music and one for text - trained jointly with a contrastive loss to learn cross-modal representations.

abstract music features from large datasets. However, obtaining sufficient labelled data can be costly and time-consuming, as most labelled symbolic music datasets are small in size [10–12]. To address this issue, semantic search and zero-shot classification techniques can be used to retrieve and label extensive unlabelled data. These techniques enable the search for music by a given open-domain query (e.g., "*upbeat music with a fast tempo*"), or the automatic identification of music characteristics based on customized labels without the need for training data.

To enable semantic search and zero-shot classification for symbolic music, it is necessary to establish a connection between music and language. This can be achieved through the use of contrastive learning [13–17] and pre-training [18–20]. Contrastive learning trains models to learn a feature space where similar sample pairs are grouped and dissimilar pairs are separated, while pre-training involves training a model on a large dataset that can be fine-tuned or directly applied to a specific task.

In this paper, we introduce a solution for cross-modal symbolic MIR that utilizes contrastive learning and pre-training. The proposed approach, **CLaMP**: **C**ontrastive **La**nguage-**M**usic **P**re-training, is inspired by the success of vision-language models [13]. Unlike prior models that rely solely on symbolic music [9, 12, 21], CLaMP learns semantically rich representations of musical concepts from both sheet music and natural language. The contributions of this paper are as follows:

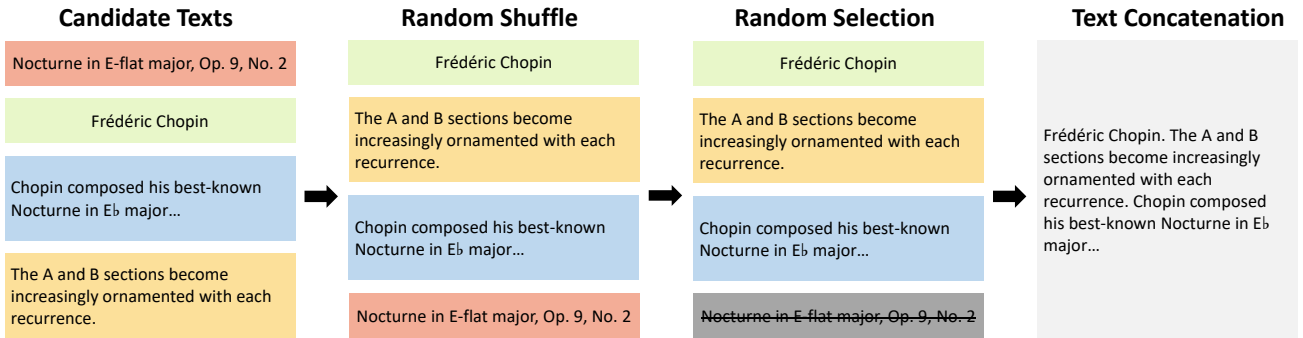| **Candidate Texts**	**Random Shuffle**	**Random Selection**	**Text Concatenation**
Nocturne in E-flat major, Op. 9, No. 2	Frédéric Chopin	Frédéric Chopin	
Frédéric Chopin	The A and B sections become increasingly ornamented with each recurrence.	The A and B sections become increasingly ornamented with each recurrence.	Frédéric Chopin. The A and B sections become increasingly ornamented with each recurrence. Chopin composed his best-known Nocturne in E♭ major...
Chopin composed his best-known Nocturne in E♭ major...	Chopin composed his best-known Nocturne in E♭ major...	Chopin composed his best-known Nocturne in E♭ major...	
The A and B sections become increasingly ornamented with each recurrence.	Nocturne in E-flat major, Op. 9, No. 2	~~Nocturne in E-flat major, Op. 9, No. 2~~	

Figure 2. Text dropout is a data augmentation technique that involves a process in which candidate texts are shuffled randomly and then selected to form a concatenated text. In this example, three candidate texts were randomly selected and concatenated to produce the input text.

- CLaMP is a cross-modal model for symbolic MIR, which is pre-trained on WebMusicText (WebMT), a dataset of 1.4 million music-text pairs. To the best of our knowledge, this is the first model of its kind and it achieves comparable or better performance than existing state-of-the-art models without training.

- We propose multiple techniques to improve contrastive language-music pre-training. Our proposed techniques include applying text dropout as a data augmentation method, utilizing bar patching for efficient music representation, and implementing the masked music model pre-training objective.

- The cross-modal pre-training empowers CLaMP to perform tasks beyond the capabilities of unimodal models. It possesses unique features such as semantic search for desired music using open-domain text queries and zero-shot classification for new music.

- To facilitate the evaluation of semantic search and music classification, we release the WikiMusicText (WikiMT) dataset, which consists of 1010 music-text pairs sourced from Wikifonia and Wikipedia.

2. METHODOLOGY

This section presents CLaMP and its cross-modal symbolic MIR abilities. Additionally, we describe the WebMT dataset, which we created to pre-train our model.

2.1 Model Design

2.1.1 Contrastive Learning Objective

CLaMP jointly trains music and text encoders to represent the structural and semantic aspects of both modalities in a shared feature space. This is achieved using a batch construction method and objective [22, 23], as illustrated in Fig. 1, whereby the correct pairings of a batch of N music-text pairs are predicted. The music and text encoders employ global average pooling to obtain corresponding features from the last hidden states.

The objective of CLaMP is to minimize the distance between N paired music-text examples while maximizing the distance between $N^2 - N$ unpaired examples. We denote a

batch of N music-text pairs as $(m_i, t_i)_{i=1}^N$, where m_i and t_i represent the i-th music and text inputs, respectively. The music and text encoders are represented as f_m and f_t. The contrastive loss for $(m_i, t_i)_{i=1}^N$ is defined as follows:

$$\mathcal{L}_{CL} = -\frac{1}{2N} \sum_{i=1}^{N} (\log \frac{\exp(f_m(m_i) \cdot f_t(t_i)/\tau)}{\sum_{j=1}^{N} \mathbb{K}_{i \neq j} \exp(f_m(m_i) \cdot f_t(t_j)/\tau)} + \\ \log \frac{\exp(f_m(m_i) \cdot f_t(t_i)/\tau)}{\sum_{j=1}^{N} \mathbb{K}_{i \neq j} \exp(f_m(m_j) \cdot f_t(t_i)/\tau)}), \quad (1)$$

where τ is a temperature hyper-parameter that controls the sharpness of the softmax distribution, and $\mathbb{K}_{i \neq j}$ is an indicator function that equals 1 if $i \neq j$, and 0 otherwise. The two terms in Eq. 1 consider either music-to-text or text-to-music logits.

2.1.2 Text Encoder

CLaMP includes a text encoder to extract musically relevant features from the input text. To achieve optimal performance, a pre-trained language model is used to initialize the text encoder. Furthermore, text dropout is employed as a data augmentation technique to prevent overfitting and improve the generalization ability of the text encoder.

Pre-trained Language Model RoBERTa [24] is a transformer-based language model pre-trained on a large corpus of English text using the Masked Language Modeling (MLM) objective [18]. This model is designed to be fine-tuned on downstream tasks and has demonstrated excellent performance as a text encoder for the contrastive language-audio pre-training [25]. To improve training efficiency, we used DistilRoBERTa [26] instead, which has fewer parameters (82M) compared to RoBERTa-base (125M) while achieving comparable performance.

Text Dropout Text dropout is a data augmentation technique that encourages models to learn robust features from input texts. This technique involves using a dataset consisting of multiple paired candidate texts from various sources for each musical composition. Similar to [27], for a given composition with L candidates, text dropout shuffles the set of candidate texts and randomly selects K texts, where K is uniformly and randomly sampled from integers ranging from 1 to L. These selected texts are concatenated to form a single input text for the text encoder, as shown in

Table 1. The average number of tokens per lead sheet in the WikiMT dataset with different encoding methods.

Encoding	Bar Patching	ABC Notation	OctupleMIDI [9]
Tokens	**47.07±21.60**	749.16±379.56	469.09±256.43

Fig. 2. Text dropout offers a wider range of possible text combinations and allows the model to learn more complex and diverse textual features.

2.1.3 Music Encoder

The CLaMP music encoder is designed to understand the complex musical structure and context within ABC notation. As a text-based format for symbolic music, ABC notation incorporates a wide range of musical symbols commonly used in sheet music. To keep all musical information while shortening sequence length, the encoding process utilizes the bar patching technique. To optimize performance, the music encoder is specifically designed for symbolic music understanding based on bar patching.

Bar Patching The bar in musical notation groups phrases by defining a fixed number of beats and each bar can be read and played as a single unit. It is separated by vertical lines, providing reference points for locating positions within a score.

Previous models [28–31] for ABC notation utilized character-based tokenization, resulting in sequences that are too lengthy to process efficiently. On the other hand, MeasureVAE [32] demonstrated the feasibility of encoding scores at the bar-level for music generation. To improve the efficiency of processing, we proposed bar patching, inspired by patch-based techniques in computer vision [33].

Bar patching divides a score into several small segments corresponding to bars or headers (i.e. meta-information) in ABC notation. In our implementation, each patch is assigned a maximum of 64 characters, covering 98.8% of the headers or bars in the pre-training dataset. We add an [END] token at the end of each patch to indicate the end of the sequence. Patches with fewer than 64 are padded with [PAD] tokens, while those with over 64 characters are truncated. For the vocabulary, 95 ASCII printable characters and three special tokens (i.e., [PAD], [MASK], and [END]) are considered, resulting in a total of 98 tokens. Thus, each patch can be represented as a 64×98 matrix. These patches are then flattened and projected into 768 dimensions embeddings and used as input tokens, as illustrated in Fig. 3.

Bar patching effectively reduces the average sequence length of the encoded music to less than 10% of the original ABC notation, as shown in Table 1. This technique improves the efficiency of representing music and facilitates faster computation while preserving all musical information in the notation.

Masked Music Model The Masked Music Model (M3) is a self-supervised model for symbolic MIR based on bar patching representation. The primary concept of M3 is to introduce random noise to certain patches of the input music, and then reconstruct the characters in the noise-
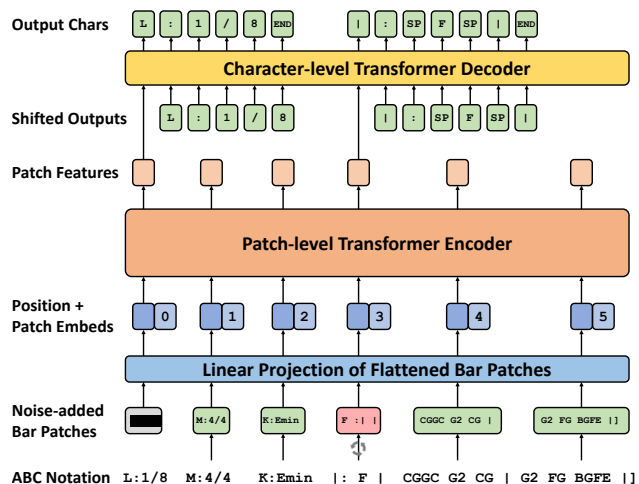

Figure 3. The masked music model architecture, where the encoder takes in a sequence of patches, and the decoder reconstructs character information of noise-added patches.

added bar patches based on the context. This pre-training enables M3 to learn from unlabelled musical data, making it useful for initializing the CLaMP music encoder.

M3 is based on an asymmetric encoder-decoder architecture, similar to MAE [34], as shown in Fig. 3. It uses an encoder to extract contextualized features of individual patches, along with a decoder, which is lightweight and autoregressively reconstructs the characters for each patch. After pre-training, the decoder is discarded and the encoder is used to initialize the music encoder of CLaMP.

The pre-training objective is inspired by MLM [18]. We first randomly select $M\%$ of the bar patches in the input music, and then the noise is added in three different ways:

- *Masking:* 80% of the selected bar patches are replaced with a special patch filled with [MASK] tokens. This encourages the model to learn to fill in missing information and understand the relationship between different musical elements.

- *Shuffling:* 10% of the selected bar patches are randomly shuffled internally. For example, a bar patch "|: F |" may be randomly shuffled to "F :| |" as shown in Fig. 3. This forces the model to learn the patterns and structures within bar patches.

- *Unchanged:* 10% of the selected bar patches are left unchanged. This can narrow down the gap between pre-training and fine-tuning.

M3 is trained to predict the original characters in the noise-added bar patches based on contextualized patch features. The model is optimized using the cross-entropy loss, which compares the predicted characters with the ground truth characters. The final objective is to minimize the average loss over all the noise-added bar patches in the training set. By denoising these bar patches, M3 learns to capture the dependencies and relationships between different musical elements and structures, allowing it to extract meaningful features from ABC notation.
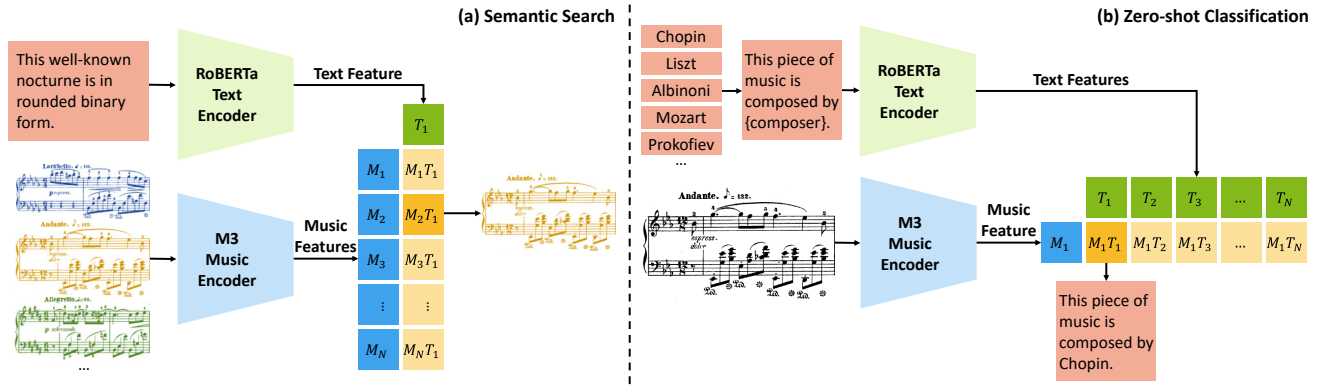
Figure 4. The processes of CLaMP performing cross-modal symbolic MIR tasks, including semantic search and zero-shot classification for symbolic music, without requiring task-specific training data.

2.2 Cross-Modal Symbolic MIR

CLaMP is capable of aligning symbolic music and natural language, which can be used for various cross-modal retrieval tasks, including semantic search and zero-shot classification for symbolic music.

Semantic search is a technique for retrieving music by open-domain queries, which differs from traditional keyword-based searches that depend on exact matches or meta-information. This involves two steps: 1) extracting music features from all scores in the library, and 2) transforming the query into a text feature. By calculating the similarities between the text feature and the music features, it can efficiently locate the score that best matches the user's query in the library.

Zero-shot classification refers to the classification of new items into any desired label without the need for training data. It involves using a prompt template to provide context for the text encoder. For example, a prompt such as `"This piece of music is composed by {composer}."` is utilized to form input texts based on the names of candidate composers. The text encoder then outputs text features based on these input texts. Meanwhile, the music encoder extracts the music feature from the unlabelled target symbolic music. By calculating the similarity between each candidate text feature and the target music feature, the label with the highest similarity is chosen as the predicted one.

2.3 WebMusicText Dataset

To facilitate the learning of relationships between natural language and symbolic music, we developed a dataset named WebMusicText (WebMT) by crawling an extensive collection of music-text pairs from the web. Our dataset comprises 1,448,750 pairs of music-text data, where all music files are in score-oriented formats (e.g., MusicXML, LilyPond, and ABC notation). To reduce the disparity between scores in different notations, we first converted all music files to MusicXML and then to ABC notation [1] . In addition, to avoid information leakage, we removed any natural language (e.g., titles, composers, and lyrics) in ABC notation. The text parts of each pair were obtained

[1] https://wim.vree.org/svgParse/xml2abc.html

from corresponding meta-information (e.g., title and composer) or user comments, and are all in English. WebMT features diverse musical compositions, from monophonic folk music to polyphonic orchestral music, which enables the model to learn a wide range of musical information.

3. EXPERIMENTS

3.1 Settings

3.1.1 Models

- **MusicBERT [9]**: This model combines unsupervised pre-training with supervised fine-tuning, which achieved state-of-the-art results. MusicBERT is available in two settings: *MusicBERT-S/1024* (MusicBERT$_{small}$), and *MusicBERT-B/1024* (MusicBERT$_{base}$). MusicBERT-S/1024 consists of 4 layers and was pre-trained on the small-scale Lakh MIDI Dataset (LMD, 148,403 pieces) [35], while MusicBERT-B/1024 has 12 layers and was pre-trained on the large-scale Million MIDI Dataset (MMD, 1,524,557 pieces). Both models have a maximum length of 1024.

- **M3**: Our proposed music encoder is used to compare the performances of unimodal and multimodal models trained on the same dataset (i.e., WebMT). M3 comes with two settings: *M3-S/512* and *M3-S/1024*, with maximum lengths of 512 and 1024, respectively. In the following experiments, both settings use the 6 encoder layers only.

- **CLaMP**: Several variants were tested to verify the effectiveness of the proposed techniques for improving contrastive language-music pre-training. These include *CLaMP-S/512* which is the full model, *CLaMP-S/512 (w/o TD)* which removes text dropout, *CLaMP-S/512 (w/o M3)* which has a randomly initialized music encoder, and *CLaMP-S/512 (w/o M3, BP)* which removes both M3 and bar patching, and uses char-level tokenization to encode raw ABC notation instead. *CLaMP-S/1024* was included to verify the effectiveness of an extended maximum length.

3.1.2 Pre-training

The text encoder was initialized using DistilRoBERTa [26], with a maximum length of 128, and the music encoder was initialized using two settings: M3-S/512 and M3-S/1024. A length of 512 resulted in truncating 17.29% of compositions in WebMT, while a length of 1024 reduced truncation to 7.7%. Both models were trained for 40 epochs with 6 encoder layers and 3 decoder layers, an embedding size of 768, and a noise ratio of 45%. Based on these two M3 encoders, we developed CLaMP-S/512 and CLaMP-S/1024. Both of them were trained for 20 epochs, using the AdamW optimizer [36] with $\beta_1 = 0.9$, $\beta_2 = 0.999$, $\epsilon = 10^{-8}$, and a weight decay coefficient of 0.01. The batch size is set to 640, and the temperature $\tau = 0.2$. The training process was accelerated and memory was saved by using mixed precision [37].

3.1.3 Evaluation Datasets

We introduce WikiMusicText (WikiMT) [2], a new dataset for the evaluation of semantic search and music classification. It includes 1010 lead sheets (melodies with harmonies) in ABC notation sourced from Wikifonia, each accompanied by a title, artist, genre, and description. The title and artist information is extracted from the score, whereas the genre labels are obtained by matching keywords from the Wikipedia entries and assigned to one of the 8 classes that loosely mimic the GTZAN genres [38]. The description is obtained by utilizing BART-large [39] to summarize and clean the corresponding Wikipedia entry. Additionally, following WebMT, the natural language information within the ABC notation is removed.

In addition to WikiMT, we use two other datasets to evaluate music classification: VGMIDI and Pianist8. VGMIDI [11] includes 204 score-oriented MIDI arrangements that were classified according to the valence-arousal model. Pianist8 [12] contains symbolic piano performances of 411 pieces from 8 composers with distinct styles, which were automatically transcribed from audio using a model presented in [8].

3.1.4 Metrics

We use the following three metrics to evaluate the effectiveness of models in various downstream tasks:

- Mean Reciprocal Rank (MRR) is used to evaluate ranking systems. This metric calculates the average of the reciprocal ranks of the correct answers, which measures the effectiveness of the ranking.

- Hit Ratio at K (HR@K) measures the accuracy of the model by checking if the correct item is among the top K recommendations, which is often used in recommendation systems.

- F1-macro score is a metric that assesses the overall effectiveness of a classification model. It is computed using the arithmetic mean (i.e., unweighted mean) of all the per-class F1 scores.

Table 2. Semantic search performance of CLaMP on WikiMT (1010 music-text pairs) under different settings.

Setting	MRR	HR@1	HR@10	HR@100
S/512	**0.2561**	**0.1931**	**0.3693**	**0.7020**
S/1024	0.2016	0.1436	0.3109	0.6554
S/512 (w/o TD)	0.1841	0.1248	0.2911	0.6188
S/512 (w/o M3)	0.1262	0.0802	0.1960	0.5119
S/512 (w/o M3, BP)	0.0931	0.0525	0.1584	0.4426

3.2 Results

3.2.1 Semantic Search

In the semantic search evaluation, we assessed different versions of CLaMP for semantic search, aiming to test the efficacy of contrastive language-music pre-training techniques. The pre-training dataset WebMT and the evaluation dataset WikiMT have no overlap, thus guaranteeing the validity of our evaluation results. In addition, as semantic search requires no additional training for this dataset, it demonstrates the generalizability of CLaMP.

Table 2 shows that our full model (CLaMP-S/512) outperforms all other models across all metrics. Interestingly, we discovered that increasing the maximum sequence length to 1024 (CLaMP-S/1024) did not lead to an improvement in performance. We attribute this to the fact that all lead sheets in the WikiMT dataset, once encoded with bar patching, have a length smaller than 512, which limits the potential advantages of the longer sequence length of CLaMP-S/1024. We also observed that the removal of the proposed techniques from CLaMP had a considerable negative impact on semantic search performance. Notably, the removal of M3 pre-training had the greatest effect on model performance, followed by text dropout and bar patching.

In conclusion, our evaluation of CLaMP on WikiMT shows that CLaMP-S/512 with all proposed contrastive language-music pre-training techniques is the most effective for the semantic search task. This highlights the importance of these techniques for effective pre-training and semantic search tasks. Additionally, increasing the sequence length (CLaMP-S/1024) did not improve the model's performance. These results emphasize the significance of using appropriate pre-training techniques in language-music models and suggest that a longer sequence length may not necessarily result in better outcomes.

3.2.2 Music Classification

The goal of the classification evaluation is to assess how well the zero-shot CLaMP models perform compared to other fine-tuned models. In addition, to evaluate pre-trained models, linear probes are used to train a linear classifier for the classification based on the features from pre-trained models. Despite being less powerful and relying on pre-trained model features, linear classifiers offer a valuable means of quantitatively assessing feature quality [40].

Table 3. Classification performance of different models on three datasets: WikiMT (1010 pieces, 8 genres), VGMIDI (204 pieces, 4 emotions), and Pianist8 (411 pieces, 8 composers).

Model	WikiMT		VGMIDI [11]		Pianist8 [12]	
	F1-macro	Accuracy	F1-macro	Accuracy	F1-macro	Accuracy
Linear Probe MusicBERT-S/1024	0.2401	**0.3507**	0.4662	0.5350	0.8047	0.8102
Linear Probe MusicBERT-B/1024	0.1746	0.3219	0.5127	0.5850	**0.8379**	**0.8413**
Zero-shot CLaMP-S/512	**0.2660**	0.3248	**0.5217**	**0.6176**	0.2180	0.2512
Zero-shot CLaMP-S/1024	0.2248	0.3406	0.4678	0.5049	0.1509	0.2390
Linear Probe M3-S/512	0.2832	0.3990	0.5991	0.6667	0.6773	0.6909
Linear Probe M3-S/1024	0.3079	0.4020	0.5966	0.6522	0.6844	0.6958
Linear Probe CLaMP-S/512	**0.3452**	0.4267	**0.6453**	**0.6866**	0.7067	0.7152
Linear Probe CLaMP-S/1024	0.3449	**0.4416**	0.6345	0.6720	**0.7271**	**0.7298**

WikiMT was converted into the MIDI format using music21 [41] to be compatible with MusicBERT. In contrast, for VGMIDI and Pianist8, we employed MuseScore3's batch conversion tool [3] to convert the scores into the MusicXML format, which were then converted into ABC notation for use with M3 and CLaMP.

We conducted 5-fold cross-validation with the same folds to assess all linear probe models, using identical fine-tuning settings and a batch size of 10 to ensure consistency, given the limited size of the evaluation datasets. The linear probe CLaMP models used the music encoder only, while the text encoder was discarded. In the zero-shot classification setting, CLaMP had no previous exposure to these evaluation datasets during pre-training. We utilized manually designed prompts for the zero-shot CLaMP models.

The top half of Table 3 presents the comparison of the performance between linear probe MusicBERT and zero-shot CLaMP. The results found that the zero-shot CLaMP models demonstrated comparable or even superior performance compared to the linear probe MusicBERT models on WikiMT and VGMIDI datasets. Interestingly, the smaller zero-shot CLaMP-S/512 outperformed the larger linear probe MusicBERT-B/1024, indicating that the pre-training of CLaMP has enabled it to learn more generalizable features that are useful for zero-shot music classification. However, this trend was not observed on Pianist8, where MusicBERT models performed much better than zero-shot CLaMP models. This difference in performance can be attributed to the source of the datasets, as WikiMT and VGMIDI primarily focus on score information, whereas Pianist8 contains performance MIDI data derived from audio. Since both CLaMP and M3 were trained exclusively on score information, they lack knowledge of performance MIDI. However, we noticed that the performances of linear probe CLaMP models on Pianist8 significantly improved after fine-tuning compared to the zero-shot ones. This suggests that incorporating ABC notation from performance MIDI into the pre-training of CLaMP may enhance its ability to comprehend such data.

---

[3] https://musescore.org/en/project/batch-convert

The linear probe CLaMP models show better performance compared to the linear probe M3 models, as indicated in the bottom half of Table 3, despite being pre-trained on the same dataset with the same architecture. This is attributed to the use of contrastive learning, which aligns the music encoder of CLaMP with the text modality, thus implicitly introducing textual information to the music encoder. Furthermore, we found that CLaMP-S/1024 performed better on Pianist8 than CLaMP-S/512, suggesting that a larger maximum length is beneficial for models to learn performance MIDI.

In summary, our evaluation demonstrates that zero-shot CLaMP performs comparably to state-of-the-art models in music classification. Furthermore, the incorporation of contrastive learning and textual information enhances the music encoder's performance, resulting in better classification accuracy when compared to M3 which employed the same architecture. These results highlight the potential of CLaMP as a pre-training framework for symbolic MIR.

4. CONCLUSIONS

This paper introduces CLaMP, a pre-trained model that utilizes contrastive language-music pre-training techniques to build cross-modal representations between natural language and symbolic music. The model was trained on a dataset containing 1.4 million music-text pairs and has demonstrated unique abilities of semantic search and zero-shot classification for symbolic music. Compared to state-of-the-art models that require fine-tuning, zero-shot CLaMP exhibits comparable or superior performance in score-oriented music classification tasks without any training. However, the current version of CLaMP has limited comprehension of performance MIDI, and still has room for improvement. Future research will aim to expand its capabilities by scaling it up and pre-training it on larger datasets that incorporate a wider range of symbolic music formats beyond score-oriented ones. We expect that its cross-modal representations will facilitate research on new topics in music analysis, retrieval, and generation, and provide a foundation for the development of innovative systems and applications that integrate music and language.

5. REFERENCES

[1] I. Karydis, "Symbolic music genre classification based on note pitch and duration," in *Advances in Databases and Information Systems, 10th East European Conference, ADBIS 2006, Thessaloniki, Greece, September 3-7, 2006, Proceedings*, ser. Lecture Notes in Computer Science, Y. Manolopoulos, J. Pokorný, and T. K. Sellis, Eds., vol. 4152. Springer, 2006, pp. 329–338. [Online]. Available: https://doi.org/10.1007/11827252_25

[2] C. Kofod and D. O. Arroyo, "Exploring the design space of symbolic music genre classification using data mining techniques," in *2008 International Conferences on Computational Intelligence for Modelling, Control and Automation (CIMCA 2008), Intelligent Agents, Web Technologies and Internet Commerce (IAWTIC 2008), Innovation in Software Engineering (ISE 2008), 10-12 December 2008, Vienna, Austria*, M. Mohammadian, Ed. IEEE Computer Society, 2008, pp. 43–48. [Online]. Available: https://doi.org/10.1109/CIMCA.2008.223

[3] J. P. Bello, G. Monti, and M. B. Sandler, "Techniques for automatic music transcription," in *ISMIR 2000, 1st International Symposium on Music Information Retrieval, Plymouth, Massachusetts, USA, October 23-25, 2000, Proceedings*, 2000. [Online]. Available: http://ismir2000.ismir.net/papers/bello_paper.pdf

[4] C. Raphael, "Automatic transcription of piano music," in *ISMIR 2002, 3rd International Conference on Music Information Retrieval, Paris, France, October 13-17, 2002, Proceedings*, 2002. [Online]. Available: http://ismir2002.ismir.net/proceedings/02-FP01-2.pdf

[5] C. Walshaw, "A statistical analysis of the abc music notation corpus: Exploring duplication," 2014.

[6] C. Hawthorne, I. Simon, R. Swavely, E. Manilow, and J. H. Engel, "Sequence-to-sequence piano transcription with transformers," in *Proceedings of the 22nd International Society for Music Information Retrieval Conference, ISMIR 2021, Online, November 7-12, 2021*, J. H. Lee, A. Lerch, Z. Duan, J. Nam, P. Rao, P. van Kranenburg, and A. Srinivasamurthy, Eds., 2021, pp. 246–253. [Online]. Available: https://archives.ismir.net/ismir2021/paper/000030.pdf

[7] J. Gardner, I. Simon, E. Manilow, C. Hawthorne, and J. H. Engel, "MT3: multi-task multitrack music transcription," in *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net, 2022. [Online]. Available: https://openreview.net/forum?id=iMSjopcOn0p

[8] Q. Kong, B. Li, X. Song, Y. Wan, and Y. Wang, "High-resolution piano transcription with pedals by regressing onset and offset times," *IEEE ACM Trans. Audio Speech Lang. Process.*, vol. 29, pp. 3707–3717, 2021. [Online]. Available: https://doi.org/10.1109/TASLP.2021.3121991

[9] M. Zeng, X. Tan, R. Wang, Z. Ju, T. Qin, and T. Liu, "Musicbert: Symbolic music understanding with large-scale pre-training," in *Findings of the Association for Computational Linguistics: ACL/IJCNLP 2021, Online Event, August 1-6, 2021*, ser. Findings of ACL, C. Zong, F. Xia, W. Li, and R. Navigli, Eds., vol. ACL/IJCNLP 2021. Association for Computational Linguistics, 2021, pp. 791–800. [Online]. Available: https://doi.org/10.18653/v1/2021.findings-acl.70

[10] A. Ferraro and K. Lemström, "On large-scale genre classification in symbolically encoded music by automatic identification of repeating patterns," in *Proceedings of the 5th International Conference on Digital Libraries for Musicology, DLfM 2018, Paris, France, September 28, 2018*, K. R. Page, Ed. ACM, 2018, pp. 34–37. [Online]. Available: https://doi.org/10.1145/3273024.3273035

[11] L. N. Ferreira and J. Whitehead, "Learning to generate music with sentiment," 2019.

[12] Y. Chou, I. Chen, C. Chang, J. Ching, and Y. Yang, "Midibert-piano: Large-scale pre-training for symbolic music understanding," *CoRR*, vol. abs/2107.05223, 2021. [Online]. Available: https://arxiv.org/abs/2107.05223

[13] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever, "Learning transferable visual models from natural language supervision," in *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, ser. Proceedings of Machine Learning Research, M. Meila and T. Zhang, Eds., vol. 139. PMLR, 2021, pp. 8748–8763. [Online]. Available: http://proceedings.mlr.press/v139/radford21a.html

[14] W. Kim, B. Son, and I. Kim, "Vilt: Vision-and-language transformer without convolution or region supervision," in *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, ser. Proceedings of Machine Learning Research, M. Meila and T. Zhang, Eds., vol. 139. PMLR, 2021, pp. 5583–5594. [Online]. Available: http://proceedings.mlr.press/v139/kim21k.html

[15] J. Li, D. Li, C. Xiong, and S. C. H. Hoi, "BLIP: bootstrapping language-image pre-training for unified vision-language understanding and generation," in *International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA*, ser. Proceedings of Machine Learning Research, K. Chaudhuri, S. Jegelka, L. Song, C. Szepesvári, G. Niu, and S. Sabato, Eds., vol. 162. PMLR, 2022, pp. 12 888–12 900. [Online]. Available: https://proceedings.mlr.press/v162/li22n.html

[16] B. Elizalde, S. Deshmukh, M. A. Ismail, and H. Wang, "CLAP: learning audio concepts from natural language supervision," *CoRR*, vol. abs/2206.04769, 2022. [Online]. Available: https://doi.org/10.48550/arXiv.2206.04769

[17] Q. Huang, A. Jansen, J. Lee, R. Ganti, J. Y. Li, and D. P. W. Ellis, "Mulan: A joint embedding of music audio and natural language," *CoRR*, vol. abs/2208.12415, 2022. [Online]. Available: https://doi.org/10.48550/arXiv.2208.12415

[18] J. Devlin, M. Chang, K. Lee, and K. Toutanova, "BERT: pre-training of deep bidirectional transformers for language understanding," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, J. Burstein, C. Doran, and T. Solorio, Eds. Association for Computational Linguistics, 2019, pp. 4171–4186. [Online]. Available: https://doi.org/10.18653/v1/n19-1423

[19] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. M. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, and D. Amodei, "Language models are few-shot learners," in *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, Eds., 2020. [Online]. Available: https://proceedings.neurips.cc/paper/2020/hash/1457c0d6bfcb4967418bfb8ac142f64a-Abstract.html

[20] 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," *J. Mach. Learn. Res.*, vol. 21, pp. 140:1–140:67, 2020. [Online]. Available: http://jmlr.org/papers/v21/20-074.html

[21] Z. Wang and G. Xia, "Musebert: Pre-training music representation for music understanding and controllable generation," in *Proceedings of the 22nd International Society for Music Information Retrieval Conference, ISMIR 2021, Online, November 7-12, 2021*, J. H. Lee, A. Lerch, Z. Duan, J. Nam, P. Rao, P. van Kranenburg, and A. Srinivasamurthy, Eds., 2021, pp. 722–729. [Online]. Available: https://archives.ismir.net/ismir2021/paper/000090.pdf

[22] K. Sohn, "Improved deep metric learning with multi-class n-pair loss objective," in *Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016,*

December 5-10, 2016, Barcelona, Spain, D. D. Lee, M. Sugiyama, U. von Luxburg, I. Guyon, and R. Garnett, Eds., 2016, pp. 1849–1857. [Online]. Available: https://proceedings.neurips.cc/paper/2016/hash/6b180037abbebea991d8b1232f8a8ca9-Abstract.html

[23] A. van den Oord, Y. Li, and O. Vinyals, "Representation learning with contrastive predictive coding," *CoRR*, vol. abs/1807.03748, 2018. [Online]. Available: http://arxiv.org/abs/1807.03748

[24] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, "Roberta: A robustly optimized BERT pretraining approach," *CoRR*, vol. abs/1907.11692, 2019. [Online]. Available: http://arxiv.org/abs/1907.11692

[25] Y. Wu, K. Chen, T. Zhang, Y. Hui, T. Berg-Kirkpatrick, and S. Dubnov, "Large-scale contrastive language-audio pretraining with feature fusion and keyword-to-caption augmentation," *CoRR*, vol. abs/2211.06687, 2022. [Online]. Available: https://doi.org/10.48550/arXiv.2211.06687

[26] V. Sanh, L. Debut, J. Chaumond, and T. Wolf, "Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter," *ArXiv*, vol. abs/1910.01108, 2019.

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

[28] B. L. Sturm, J. F. Santos, O. Ben-Tal, and I. Korshunova, "Music transcription modelling and composition using deep learning," *CoRR*, vol. abs/1604.08723, 2016. [Online]. Available: http://arxiv.org/abs/1604.08723

[29] C. Geerlings and A. Meroño-Peñuela, "Interacting with gpt-2 to generate controlled and believable musical sequences in abc notation," in *NLP4MUSA*, 2020.

[30] S. Wu and M. Sun, "Tunesformer: Forming tunes with control codes," *CoRR*, vol. abs/2301.02884, 2023. [Online]. Available: https://doi.org/10.48550/arXiv.2301.02884

[31] ——, "Exploring the efficacy of pre-trained checkpoints in text-to-music generation task," in *The AAAI-23 Workshop on Creative AI Across Modalities*, 2023. [Online]. Available: https://openreview.net/forum?id=QmWXskBhesn

[32] A. Pati, A. Lerch, and G. Hadjeres, "Learning to traverse latent spaces for musical score inpainting," in *Proceedings of the 20th International Society for Music Information Retrieval Conference, ISMIR 2019, Delft, The Netherlands, November 4-8, 2019*, A. Flexer, G. Peeters, J. Urbano, and A. Volk, Eds., 2019, pp. 343–351. [Online]. Available: http://archives.ismir.net/ismir2019/paper/000040.pdf

[33] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16x16 words: Transformers for image recognition at scale," in *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021. [Online]. Available: https://openreview.net/forum?id=YicbFdNTTy

[34] K. He, X. Chen, S. Xie, Y. Li, P. Dollár, and R. B. Girshick, "Masked autoencoders are scalable vision learners," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*. IEEE, 2022, pp. 15 979–15 988. [Online]. Available: https://doi.org/10.1109/CVPR52688.2022.01553

[35] C. Raffel, "Learning-based methods for comparing sequences, with applications to audio-to-midi alignment and matching," Ph.D. dissertation, PhD Thesis, 2016.

[36] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," in *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net, 2019. [Online]. Available: https://openreview.net/forum?id=Bkg6RiCqY7

[37] P. Micikevicius, S. Narang, J. Alben, G. F. Diamos, E. Elsen, D. García, B. Ginsburg, M. Houston, O. Kuchaiev, G. Venkatesh, and H. Wu, "Mixed precision training," in *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net, 2018. [Online]. Available: https://openreview.net/forum?id=r1gs9JgRZ

[38] B. L. Sturm, "An analysis of the GTZAN music genre dataset," in *Proceedings of the second international ACM workshop on Music information retrieval with user-centered and multimodal strategies, MIRUM '12, Nara, Japan, October 29 - November 02, 2012*, C. C. S. Liem, M. Müller, S. K. Tjoa, and G. Tzanetakis, Eds., 2012, pp. 7–12. [Online]. Available: https://doi.org/10.1145/2390848.2390851

[39] 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," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, D. Jurafsky, J. Chai, N. Schluter, and J. R. Tetreault, Eds. Association for Computational Linguistics, 2020, pp. 7871–7880. [Online]. Available: https://doi.org/10.18653/v1/2020.acl-main.703

[40] Y. M. Asano, C. Rupprecht, and A. Vedaldi, "A critical analysis of self-supervision, or what we can learn from a single image," in *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net, 2020. [Online]. Available: https://openreview.net/forum?id=B1esx6EYvr

[41] M. S. Cuthbert and C. Ariza, "Music21: A toolkit for computer-aided musicology and symbolic music data," in *Proceedings of the 11th International Society for Music Information Retrieval Conference, ISMIR 2010, Utrecht, Netherlands, August 9-13, 2010*, J. S. Downie and R. C. Veltkamp, Eds. International Society for Music Information Retrieval, 2010, pp. 637–642. [Online]. Available: http://ismir2010.ismir.net/proceedings/ismir2010-108.pdf