CONTRASTIVE AUDIO-LANGUAGE LEARNING FOR MUSIC

Ilaria Manco[1,2] **Emmanouil Benetos**[1] **Elio Quinton**[2] **György Fazekas**[1]
[1] School of EECS, Queen Mary University of London, London, U.K
[2] Music & Audio Machine Learning Lab, Universal Music Group, London, U.K.
{i.manco, emmanouil.benetos, george.fazekas}@qmul.ac.uk, elio.quinton@umusic.com

ABSTRACT

As one of the most intuitive interfaces known to humans, natural language has the potential to mediate many tasks that involve human-computer interaction, especially in application-focused fields like Music Information Retrieval. In this work, we explore cross-modal learning in an attempt to bridge audio and language in the music domain. To this end, we propose MusCALL, a framework for Music Contrastive Audio-Language Learning. Our approach consists of a dual-encoder architecture that learns the alignment between pairs of music audio and descriptive sentences, producing multimodal embeddings that can be used for text-to-audio and audio-to-text retrieval out-of-the-box. Thanks to this property, MusCALL can be transferred to virtually any task that can be cast as text-based retrieval. Our experiments show that our method performs significantly better than the baselines at retrieving audio that matches a textual description and, conversely, text that matches an audio query. We also demonstrate that the multimodal alignment capability of our model can be successfully extended to the zero-shot transfer scenario for genre classification and auto-tagging on two public datasets.

1. INTRODUCTION

Developing effective methods for finding music is at the core of Music Information Retrieval (MIR). Over the years, many approaches have been proposed to browse, search and discover music through a variety of interfaces. Beyond simple search by metadata, existing music retrieval systems allow to express queries via lyrics [1,2], audio examples [3], videos [4] and humming [5], among others [6–8]. Although each of these query types has its merits, none of these systems supports another popular way of searching for music today: through free-form text. For example, we commonly look for songs by typing text into a search engine [9] or by asking online song naming communities to identify a piece of music we do not have bibliographic information about [10]. It becomes evident then that enabling MIR systems to interpret natural language queries can have far-reaching benefits. This idea is not entirely new to MIR, with prior work [11–13] suggesting similar research directions in the past. So far, however, multimodal systems that integrate natural language have not been widely adopted within the MIR community, possibly due to a lack of suitable datasets or to the practical limitations of NLP methods predating modern language models.

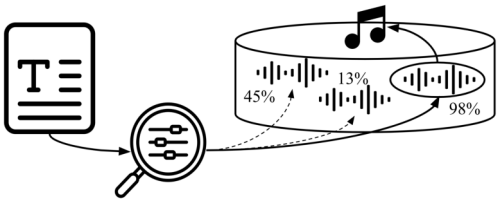

Figure 1: **Illustration of text-based music retrieval**, where audio items in a database are ranked according to their relevance to a free-form language query.

In light of recent breakthroughs in language modelling, we argue that audio-and-language learning has now the potential of closing the semantic gap in MIR [14], providing a bridge between computational representations of music signals and the high-level abstractions needed to use those representations in real-world scenarios. With this in mind, we propose MusCALL, a method for learning alignment between music-related audio and language data via multimodal contrastive learning. Our choice of a contrastive approach is inspired by the recent success of similar methods for joint visio-linguistic modelling (see Section 2.3). In designing MusCALL, we prioritise the ability to perform retrieval at scale and adopt a dual-encoder architecture, where modalities are processed independently. Compared to multimodal architectures with joint encoders and cross-modal attention mechanisms [15], this design allows to share embedding computations among pairs, resulting in a computationally more efficient model.

With this work, we aim to unify audio and language modelling, paving the way for MIR systems that can interpret language-based queries, as illustrated in Figure 1. Our primary contributions can be summarised as follows: (i) we explore multimodal contrastive learning in the context of music-related audio and language for the first time; (ii) we introduce a method for cross-modal retrieval of music, providing the first example of sentence-based music search; (iii) we perform an extensive set of experiments to systematically validate details of our approach and evaluate its performance on popular MIR tasks in a zero-shot setting. [1]

[1] Code and supplementary material are available at https://github.com/ilaria-manco/muscall

2. RELATED WORK

2.1 Natural Language Processing in MIR

Prior works in the MIR literature have explored leveraging natural language in the music domain from different angles. Early efforts focused on text as a modality in isolation, adopting NLP techniques to construct knowledge bases from music-related text corpora [16], build semantic graphs for artist similarity from biographies [17], or perform genre classification based on album reviews [18]. Recent efforts, more closely related to the present work, have instead started favouring multimodal approaches. These have explored deep learning with multimodal input data, typically audio combined with text such as reviews or lyrics, for applications as varied as music classification and recommendation [19], mood detection [20], music emotion recognition [21] and music captioning [22–24].

2.2 Audio-Text Cross-modal Learning

Another related line of work is cross-modal learning that leverages audio and tags as text input. Works in this area have proposed contrastive learning-based approaches to enrich audio representations via cross-modal alignment, either for general-purpose audio classification [25, 26] or for music-focused tasks [27]. Differently from our work, these approaches require finetuning on the downstream tasks and none of them directly uses cross-modal representations. Others have explored pre-trained word embeddings in triplet networks to perform tag-based music retrieval via a multimodal embedding space [28, 29]. At a high level, these works share a similar approach to ours and all aim to learn multimodal audio representations by leveraging text. Unlike our work, however, none of them makes use of natural language, using tags as text input instead.

Finally, similarly to our work, [30] also addresses the problem of matching audio and long-form text for music retrieval, but offers a fundamentally different approach, which relies on bridging the audio and text modalities via a common emotion embedding space.

2.3 Learning from Language Supervision

In the previous sections, we have highlighted some research efforts towards bringing together audio and language, but, ultimately, multimodal learning still occupies a marginal role in MIR and has yet to fully enjoy the benefits of modern language models. In adjacent fields such as computer vision, jointly modelling vision and language has instead become a very active area of research, with several successful attempts at using multimodality to develop task-agnostic models that can easily adapt to novel tasks [31–34]. The key insight behind these models is that language captures many of the abstractions humans use to navigate the world and can therefore act as a rich supervisory signal for general-purpose learning, even in tasks that are not directly based on language [35]. Among these works, CLIP [33] is a particularly influential example, having demonstrated for the first time that supervision from natural language can induce highly generic visual representations at scale.

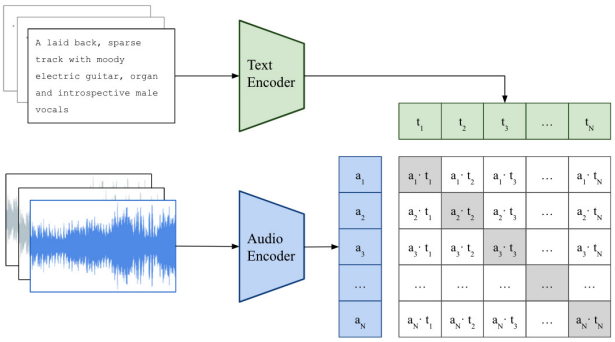

Figure 2: **Overview of MusCALL**. An audio encoder and a text encoder are trained via a contrastive loss to maximise the similarity between representations of N aligned (audio, text) pairs within a mini-batch. At test time, the similarity between embeddings in the learnt multimodal space is used to rank database items and perform cross-modal retrieval.

These breakthroughs suggest that natural language supervision has a large potential beyond the image domain. This has recently prompted the adoption of similar approaches in machine listening, where applications to both music [15] and non-music audio [36, 37] have started to emerge. A subset of works in this area have proposed to directly extend CLIP by incorporating audio as an additional modality [38–40]. Although similar in spirit to our work, these approaches exploit audio-visual correspondences in video, thus requiring large-scale visio-linguistic pre-training. Like these works, we also borrow from CLIP's contrastive dual-encoder approach, but we adapt this framework to the audio modality without any pre-training, while also addressing the specific set of challenges that arise from joint audio-linguistic modelling in the music domain.

3. MUSCALL

3.1 Multimodal Contrastive Learning (MCL)

Contrastive learning has emerged as a powerful paradigm in self-supervised representation learning for images [41] and audio [42–45]. At their core, all flavours of contrastive learning rely on constructing different views of the data and forcing representations of positive pairs of such views to have, on average, higher similarity than those of negative pairs. This encourages a model to learn representations that are *discriminative* with respect to changes to the content and *invariant* to nuisance transformations. In representation learning, this is a useful property, since it results in more robust and effective representations [46–48]. Most formulations of contrastive learning in unimodal scenarios use data augmentations to create multiple views of the inputs. When switching to the multimodal setting, however, different views naturally co-occur in the data, making this a natural testbed for contrastive learning [33, 49, 50].

3.2 Extending MCL to Music and Language

In this work, we explore multimodal contrastive learning (MCL) in the context of audio-linguistic data. More specifically, we consider pairs of music audio tracks and their

associated textual descriptions and aim to learn a model that is able to predict when items from a given pair match, i.e. when they are semantically aligned. An overview of our approach is shown in Figure 2. In practice, our goal is to learn two encoders, $f_a(\cdot)$ for the audio modality and $f_t(\cdot)$ for the text modality, such that for any given (audio, text) pair formed by the tuple (a_i, t_i), the resulting L2-normalised embeddings $\boldsymbol{z}_{a,i} = f_a(a_i)$ and $\boldsymbol{z}_{t,i} = f_t(t_i)$ lie closely in the joint embedding space, with respect to some distance metric, only if a_i and t_i represent similar content. Given our problem definition, we achieve this by optimising a type of N-pair contrastive loss, known as InfoNCE [51]. Based on this, our audio-to-text loss can be defined as follows:

$$\mathcal{L}_{a\to t} = -\frac{1}{N}\sum_{i}^{N}\log\frac{\exp\left(\boldsymbol{z}_{a,i}\cdot\boldsymbol{z}_{t,i}^{+}/\tau\right)}{\sum\limits_{\boldsymbol{z}\in\left\{\boldsymbol{z}_{t,i}^{+},\boldsymbol{z}_{t,i}^{-}\right\}}\exp\left(\boldsymbol{z}_{a,i}\cdot\boldsymbol{z}/\tau\right)}, \quad (1)$$

where N is the batch size, $\boldsymbol{z}_{t,i}^{+}$ and $\boldsymbol{z}_{t,i}^{-}$ are embeddings of positive and negative text samples for item a_i, and τ is a temperature parameter used to scale the similarity scores. By noting that $\mathcal{L}_{t\to a}$ can be defined symmetrically to Eq. (1), the total loss over all (audio, text) pairs in a mini-batch is simply obtained by summing the two losses together:

$$\mathcal{L}_{a\leftrightarrow t} = \mathcal{L}_{a\to t} + \mathcal{L}_{t\to a}. \quad (2)$$

Given a dataset of tuples $\mathcal{D} = \{(a_i, t_i)\}_{i=1:D}$, there are several plausible strategies for constructing positive and negative pairs. One such way, and arguably the most intuitive, is to follow the implicit alignment found in the data. In this case, for each sample i, the positive and negative sets within a mini-batch become: $\boldsymbol{z}_{x,i}^{+} = \{\boldsymbol{z}_{y,i}\}$ and $\boldsymbol{z}_{x,i}^{-} = \{z_{y,j} \mid \forall j \in \{1,\ldots,N\}, j \neq i\}$, where x and y denote the two different modalities. We refer to this sampling strategy as *instance discrimination* [52].

3.2.1 Content-Aware Loss Weighting

Constructing positive and negative samples by instance discrimination is a design choice that relies on two implicit assumptions: firstly, that all items in the dataset are correctly aligned, i.e. that each (audio, text) pair is formed by the most semantically close items amongst all possible pairings; secondly, that all non-aligned pairs represent sufficiently different content and are therefore equally valid as negatives. In a real-world dataset, this is unlikely to hold true in all cases. Intuitively, within a randomly sampled mini-batch, some tracks will share similarities with other tracks, and so will their respective captions.

The above observations suggest that a more informed sampling procedure or an alternative way to define our cross-modal loss may better suit the structural properties of our data. Due to its simplicity, we focus on the latter option and test this hypothesis by introducing some modifications to Eq. (1). By observing that similar captions are likely to correspond to similar audio, we can leverage this side information to estimate the *relevance* between non-paired (negative) items in a mini-batch. Since more relevant items should contribute more to the learning process, similarly to [53], we can introduce a relevance-based weight:

$$w_i = \exp\left(\frac{\frac{1}{N}\sum_{j}^{N}\mathrm{sim}(t_i, t_j)}{\kappa}\right), \quad (3)$$

where κ is a temperature hyperparameter and $\mathrm{sim}(t_i, t_j)$ a similarity score between text sample t_i and text sample t_j. We can then redefine our audio-to-text loss in Eq. (1) as follows:

$$\mathcal{L}^{*}_{a\to t} = -\frac{1}{N}\sum_{i}^{N} w_i \log\frac{\exp\left(\boldsymbol{z}_{a,i}\cdot\boldsymbol{z}_{t,i}^{+}/\tau\right)}{\sum\limits_{\boldsymbol{z}\in\left\{\boldsymbol{z}_{t,i}^{+},\boldsymbol{z}_{t,i}^{-}\right\}}\exp\left(\boldsymbol{z}_{a,i}\cdot\boldsymbol{z}/\tau\right)} \quad (4)$$

and obtain the total loss by summing this to its symmetrical counterpart $\mathcal{L}^{*}_{t\to a}$. We dub this procedure *content-aware loss weighting*, or *loss weighting* for short.

3.3 Combining MusCALL with Audio Self-Supervision

At its core, the design of our approach is centred around audio-text matching. However, we are also interested in learning representations that can be transferred to other tasks, especially in a zero-shot way. Prior work has demonstrated the benefit of using self-supervised learning (SSL) [54] to improve representation quality in a similar setting in the image domain [55]. We hypothesise that our model may also benefit from this and experiment with combining our multimodal contrastive objective with self-supervision on the audio modality. Similarly to the approach proposed in [55], we do this via multi-task learning, using an adaptation of SimCLR [48] to the audio modality as the self-supervised component. Two correlated views of the audio input are produced via a data augmentation pipeline and passed through a shared audio encoder. The SSL objective is then computed on the embeddings resulting from the two views and added to our cross-modal objective as follows:

$$\mathcal{L} = \lambda\mathcal{L}_{SSL} + (1-\lambda)\mathcal{L}_{a\leftrightarrow t}, \quad (5)$$

where $\mathcal{L}_{SSL}$ is the *NT-Xent* loss used in SimCLR [48] and $\lambda \in [0, 1]$ is a scalar weight.

We refer to the SSL-enhanced variant by $\text{MusCALL}_{\text{SSL}}$ to distinguish it from the base variant $\text{MusCALL}_{\text{BASE}}$.

3.4 Audio & Text Encoding

As our audio backbone, we choose ResNet-50 [57] operating on melspectrogram representations of the input and adopt the same architectural modifications introduced in CLIP [33]: 3 stem convolutions followed by average pooling instead of max pooling and anti-aliased blur pooling. We obtain fixed-length audio representations via an attention pooling mechanism: we append the average-pooled audio feature to the backbone output and compute multi-head self-attention, taking the output corresponding to the average-pooled feature as the global audio representation. Also analogously to CLIP, our text encoder is a Transformer [58, 59]. To avoid overfitting on our dataset, which is $\sim$ 2K times smaller than the dataset used to train CLIP,

Method	Text → Audio					Audio → Text				
	R@1	R@5	R@10	mAP10	MedR ↓	R@1	R@5	R@10	mAP10	MedR ↓
DCASE [56]	2.3	10.4	17.4	5.5	50	1.1	5.6	10.1	3.0	84
DCASE + CL	3.9	12.4	18.1	6.8	81.5	2.0	8.6	16.4	4.5	64
MusCALL (ours)	**25.9**	**51.9**	**63.3**	**36.0**	**5**	**25.8**	**53.0**	**63.0**	**35.9**	**5**

Table 1: **Cross-modal retrieval results**. MusCALL improves performance on all metrics by a large margin, compared to two variants of the baseline: one with the original triplet ranking loss (DCASE) and one with our loss (DCASE + CL).

we downsize the network and use 4 hidden layers, as we empirically find this to be the optimal depth (see Figure 3).

Two learned linear projections are applied to map the audio and text features produced by the audio and text backbones onto a 512-dimensional multimodal embedding space respectively. The resulting features are then L2-normalised before their dot product is calculated in the contrastive loss.

4. EXPERIMENTAL DESIGN

4.1 Dataset

We train and evaluate our model on a dataset of 250k (audio, text) pairs created from a production music library. This consists of full-length audio tracks, covering a broad range of genres, and a piece of text describing the overall musical content of each track, as shown in Table 3. For training, validation and testing, we use a random 80/10/10 split.

4.2 Implementation Details

For each audio track, we take a $20s$ random crop at training time, and the central $20s$ segment at testing time. Unless otherwise specified, we then apply a stochastic data augmentation pipeline, where each transformation is applied with an independent probability p. We adopt some of the same transformations, such as noise injection and pitch shift, as prior work on music audio representation learning [45, 60]. For each audio caption in the dataset, we take the text input in full and tokenise it following the same procedure as in CLIP, based on byte pair encoding [61] with a 49K token vocabulary and maximum sequence length of 77.

As an estimate for text-text similarity in the calculation of our loss scaling weight w_i in Eq. (4), we use the cosine distance between L2-normalised embeddings produced by a pre-trained Sentence-BERT [62]. We set the loss weighting temperature parameter κ to 0.005, following [53].

For the SimCLR module in $\text{MusCALL}_{\text{SSL}}$, following [45] we use a 256-dimensional non-linear projection layer and set the temperature parameter to 0.5. The loss scaling parameter λ in Eq. (5) is set to 0.3 as we find this to yield best results.

Together with the parameters for the audio and text encoders and multimodal projections, we also learn the temperature parameter τ in Eq. (1) to avoid tuning it as a hyperparameter. We train using the Adam optimizer, with weight decay 0.2, batch size of 256, initial learning rate of 5e-5, reduced throughout training following a cosine schedule. After training for a maximum of 150 epochs, we select the best model based on the R@10 score (see Section 5.1) computed on the validation set. To save GPU memory, we perform training with automatic mixed precision.

5. EXPERIMENTS & RESULTS

In this section we first describe our experimental setup and report our main results on cross-modal retrieval (Section 5.1), which constitute the focus of the paper. We then explore transferring our model to zero-shot music classification, highlighting key findings (Section 5.2).

5.1 Cross-Modal Retrieval

In cross-modal retrieval, given a query item of modality A, our goal is to identify the matching item of modality B. We can do this by casting the retrieval as a ranking problem: for a given query item of modality A, we rank all candidate items of modality B by the cosine similarity between the query embedding and each candidate embedding. The retrieval performance is then evaluated by computing the Recall at K (R@K) over the testing set as the percentage of correctly retrieved items within the top-K items for each query, with $\text{K} = \{1, 5, 10\}$. In line with previous work on cross-modal retrieval, we also report mean average Precision at 10 (mAP10) and Median Rank (MedR) in our main results. We construct our testing set by randomly sampling a subset of 1000 (audio, text) pairs from our testing split. This is chosen to be consistent in size with other datasets for sentence-based audio retrieval [63, 64].

Results Table 1 shows the performance of MusCALL on the cross-modal retrieval task. We compare this to the baseline system for the *Language-Based Audio Retrieval* subtask of Task 6 in the 2022 DCASE Challenge, [2] trained and evaluated on our dataset. This is a simplified version of the approach proposed in [56] and consists of a pre-trained *word2vec* model [65] as the text encoder and a convolutional recurrent neural network as the audio encoder. Average pooling is used to obtain global representations for each modality from the output of the encoders. These are then jointly trained via a triplet ranking loss [66]. We also consider a variant of this baseline where we use our loss instead of the triplet ranking loss, to provide a closer comparison to our approach.

Our results show that MusCALL significantly outperforms the baseline on both text-to-audio and audio-to-text retrieval. Enhancing the baseline with our loss slightly reduces this margin, bringing a 14.2% and 60.7% average

[2] `https://dcase.community/challenge2022`

Method	Prompt	Genre	Tagging	
		Acc.	ROC	PR
$\text{MusCALL}_{\text{BASE}}$	✗	55.5	**78.0**	28.3
$\text{MusCALL}_{\text{BASE}}$	✓	52.0	72.0	21.0
$\text{MusCALL}_{\text{SSL}}$	✗	58.2	**77.4**	**29.3**
$\text{MusCALL}_{\text{SSL}}$	✓	**62.0**	73.4	23.2

Table 2: **Zero-shot transfer results**. We report $\text{MusCALL}_{\text{BASE}}$ and $\text{MusCALL}_{\text{SSL}}$ accuracy on GTZAN (genre classification), and ROC-AUC and PR-AUC on MTAT (auto-tagging). *Prompt* indicates whether a template was used to wrap the class label.

improvement over the vanilla baseline for text-to-audio and audio-to-text respectively. However, the use of our contrastive loss alone does not account for the full difference, indicating that the design of our text and audio encoders and the use of linear projection layers, play a crucial role.

5.2 Zero-shot Transfer

By design, our cross-modal learning approach endows the model with the ability to interpret arbitrary text inputs. This powerful property can be exploited to perform new tasks based on textual descriptions via a transfer learning paradigm known in the literature as *zero-shot transfer* [33, 34, 67]. Similarly to zero-shot learning, zero-shot transfer aims to learn a model that can generalise to unseen classes or tasks without further training. But, in contrast to zero-shot learning, it somewhat relaxes the requirement that target classes must be completely unseen. Instead, the model is usually pre-trained in a task-agnostic fashion on large amounts of natural language text, and may be exposed to information relevant to the target tasks through this, although no supervised examples are provided.

We explore this paradigm in our work and investigate whether MusCALL exhibits zero-shot transfer capabilities on two tasks, genre classification and auto-tagging. We evaluate this on the most popular public datasets for these tasks, GTZAN [68] and MagnaTagATune (MTAT) [69]. Treating audio clips in each dataset as queries, we obtain classification predictions by considering the similarity scores between audio embeddings and text embeddings of the target labels, similarly to the procedure described for cross-modal retrieval in Section 5.1. Based on this, we calculate accuracy for GTZAN, and area under the receiver operating characteristic curve (ROC-AUC) and area under the precision-recall curve (PR-AUC) for the MTAT dataset.

In order to reduce the distribution shift between pre-training and downstream text input when doing zero-shot transfer, it is common practice to wrap the labels in templates. For example, the label "*rock*" may be wrapped in the sentence "*This is a rock song with electric guitars*", making it much closer to a typical caption encountered in training. Based on evidence that it can improve performance [33], we explore this in our evaluation by passing "*A* [LABEL] *track*" as the text input, but do not further tune the prompts to our model, leaving this for future work.

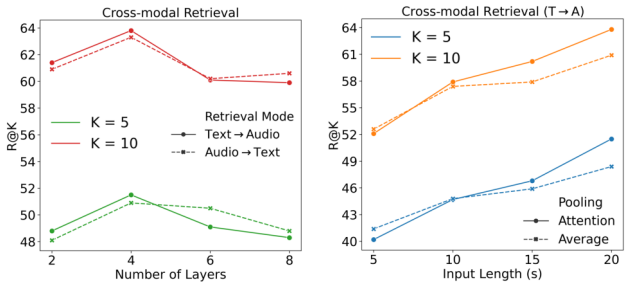

Figure 3: **The effect of varying** (left) the **depth** of the text encoder and (right) the **audio input length** on retrieval.

Results In Table 2 we compare zero-shot performance of two model variants, $\text{MusCALL}_{\text{BASE}}$ and $\text{MusCALL}_{\text{SSL}}$. We find that the self-supervised learning objective yields, on average, better performance, with a 5.4% improvement over $\text{MusCALL}_{\text{BASE}}$. This is not too surprising, as the SSL objective is designed to improve the representation quality of our audio branch and is therefore expected to have a positive effect on generalisation [70]. However, this improvement is not consistent across datasets and tasks, a result that may be attributed to different degrees of similarity between pre-training and downstream datasets, as found in prior work [71]. We also find that zero-shot performance is sensitive to the choice of text prompt. This confirms a well-known phenomenon reported in the literature [33, 72] and suggests that techniques such as prompt tuning and ensembling [33] may lead to further improvements.

Additionally, there are some important differences that should be considered when comparing $\text{MusCALL}_{\text{BASE}}$ and $\text{MusCALL}_{\text{SSL}}$, since the SSL module introduces more activations and overall learnable parameters. This has two implications: firstly, to offset the increased memory footprint while keeping the batch size unchanged and still satisfying our memory constraints, we reduce the size of the input audio to $10s$. We verify that this slightly degrades performance on cross-modal retrieval even in the absence of the SSL module, as shown in Figure 3. Assuming that this trend would be reflected in the zero-shot scenario, we conclude that $\text{MusCALL}_{\text{SSL}}$ would benefit from using longer audio clips. Secondly, the higher number of parameters makes $\text{MusCALL}_{\text{SSL}}$ prone to overfitting, a factor that we believe may also be limiting its performance.

6. ANALYSIS & DISCUSSION

We now discuss the qualitative characteristics of our model (Section 6.1) and examine the contributions of the main design choices through an ablation study (Section 6.2).

6.1 Qualitative Analysis

Similarity score distribution In Figure 4 (left) we show the kernel density estimation of the distributions of pairwise similarity scores in the joint embedding space for positive and negative pairs in our testing set. From this we can see that aligned pairs have distinctly higher scores compared to random ones, confirming that the model distinguishes positive and negative examples with a good level of confidence.

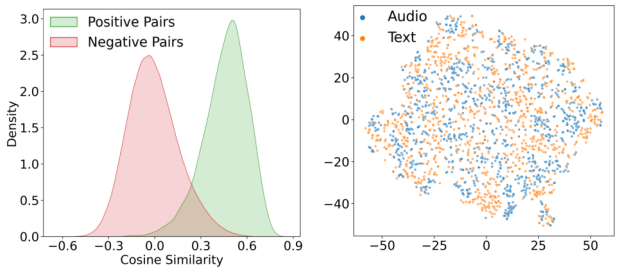

Figure 4: **Feature visualisation**. (Left) distributions of pair-wise similarity scores of aligned and non-aligned (audio, text) pairs in our testing set. (Right) t-SNE visualisation of audio and text features in the multimodal output space.

Feature visualisation Figure 4 (right) shows a t-SNE plot of audio and text embeddings in our testing set. We observe no separation between modalities in the output space, with representations of both audio and text inputs well mixed together, confirming that MusCALL learns to adequately map both input modalities to a common space.

Failure analysis In order to provide more context to our cross-modal retrieval results beyond metric-based evaluation, we also perform a qualitative error analysis by examining the relevance of the retrieved items in cases of incorrect retrieval. As highlighted in the examples in Table 3, in some cases, while the ground-truth item is not ranked first, MusCALL is still capable of retrieving audio tracks whose associated caption is semantically related to the query.

6.2 Ablation Study

The effect of data augmentations & random crop Table 4 shows that removing random cropping (RC) considerably degrades performance (-13.7% compared to $\text{MusCALL}_{\text{BASE}}$), while removing the audio augmentations (AA) only marginally alters results (-1.5%). When considering both together, we find that, in the absence of RC, the benefit of performing augmentations is amplified, and removing both has more drastic effect (-31.5%). This is expected, since both techniques effectively increase the variety of samples seen in training. To avoid costly hyperparameter tuning, we do not exhaustively explore audio transformations and their composition, and note that tailoring the AA pipeline, for example via additional frequency-domain transformations, may lead to better results.

The effect of loss weighting As also shown in Table 4, using our content-aware loss weighting (Section 3.2.1) results in comparable retrieval performance to our vanilla contrastive loss. To more closely observe the effect of using loss weighting, we compare the pairwise similarity distributions of positive and negative pairs produced by MusCALL with and without loss weighting (shown in the supplementary material). This reveals that similarity scores of positive pairs have lower variance and a higher mean when using loss weighting, suggesting that this technique nudges the learning process towards a better discrimination of positives and negatives, but not enough to produce significant improvements in the cross-modal retrieval task.

Query Text	Text of the Top-1 Audio
An atmospheric and introspective orchestral track featuring strings, piano, and synth.	*An inspirational and moody orchestral track featuring strings and choir.*
Deep chilled out space jazz with crisp beats and lush electronics.	*Jaunty swing featuring trumpet.*
Up tempo, pumping dance pop with female vocals.	*Quirky, fun, positive disco party music.*

Table 3: **Failure analysis** of text-to-audio retrieval.

LW	RC	AA	AP	Text → Audio		
				R@1	R@5	R@10
✗	✗	✗	✓	14.7	33.9	47.2
✗	✗	✓	✓	18.5	43.5	57.4
✗	✓	✓	✗	24.9	49.3	58.9
✗	✓	✗	✓	24.3	51.3	62.2
✗	✓	✓	✓	24.6	51.5	63.8
✓	✓	✓	✓	25.9	51.9	63.3

Table 4: **Ablations**: loss weighting (LW), random cropping (RC), audio augmentations (AA), attention pooling (AP).

The effect of attention pooling Our ablation study confirms that removing attention pooling hurts performance: on average, the Recall results drop by 4.9% compared to simple average pooling. We note that the advantage of using attention pooling becomes more pronounced as we increase the audio input length, as shown in Figure 3. This is particularly relevant in the music domain, since music signals exhibit structure over longer ranges compared to other types of audio signals like environmental sounds.

7. CONCLUSION

We presented MusCALL, a method for multimodal contrastive learning of audio-linguistic representations. By leveraging aligned (audio, text) pairs, MusCALL successfully learns to perform cross-modal retrieval, allowing to search for music via natural language queries and vice versa. Extending this multimodal alignment capability to the zero-shot setting, MusCALL can also be transferred to music classification tasks by simply providing target labels as text inputs. Through an extensive set of experiments, we validated the main design choices in our core approach and explored two variants. Through the first variant, we demonstrated the viability of using a text-based similarity metric to weigh the loss contribution of each negative sample, providing a starting point for improving multimodal contrastive learning on real-world music data. Through the second variant, we explored including a self-supervised objective to improve the audio representation quality. Both variants show promising results and provide opportunities for further research. Future work will focus on assessing their performance in more depth, particularly in the zero-shot scenario, including human evaluations alongside automatic metrics and considering a wider set of tasks.

8. ACKNOWLEDGEMENTS

This work was supported by UK Research and Innovation [grant number EP/S022694/1] and Universal Music Group. We would also like to thank The Alan Turing Institute for the support provided during the first author's time as an Enrichment Student.

9. REFERENCES

[1] M. Müller, F. Kurth, D. Damm, C. Fremerey, and M. Clausen, "Lyrics-Based Audio Retrieval and Multimodal Navigation in Music Collections," *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 4675 LNCS, pp. 112–123, 2007.

[2] K. Tsukuda, "Lyric Jumper: A Lyrics-Based Music Exploratory Web Service by Modeling Lyrics Generative Process." in *ISMIR*, 2017, pp. 544–551.

[3] J. Lee, N. J. Bryan, J. Salamon, Z. Jin, and J. Nam, "Disentangled Multidimensional Metric Learning for Music Similarity," in *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing*, 8 2020.

[4] B. Li and A. Kumar, "Query by Video: Cross-Modal Music Retrieval," in *20th International Society for Music Information Retrieval Conference (ISMIR)*, 2019, pp. 604–611.

[5] A. Ghias, J. Logan, D. Chamberlin, and B. C. Smith, "Query by humming," in *Proceedings of the third ACM international conference on Multimedia*. Association for Computing Machinery (ACM), 1995, pp. 231–236.

[6] M. Müller, A. Arzt, S. Balke, M. Dorfer, and G. Widmer, "Cross-Modal Music Retrieval and Applications: An Overview of Key Methodologies," *IEEE Signal Processing Magazine*, vol. 36, no. 1, pp. 52–62, 1 2019.

[7] P. Knees, M. Schedl, and M. Goto, "Intelligent User Interfaces for Music Discovery," *Transactions of the International Society for Music Information Retrieval*, vol. 3, no. 1, pp. 165–179, 10 2020.

[8] K. Watanabe and M. Goto, "Query-by-Blending: a Music Exploration System Blending Latent Vector Representations of Lyric Word, Song Audio, and Artist," in *Proceedings of the 20th International Society for Music Information Retrieval Conference*, 2019.

[9] C. Hosey, L. Vujović, B. St. Thomas, J. Garcia-Gathright, and J. Thom, "Just give me what I want: How people use and evaluate music search," in *Conference on Human Factors in Computing Systems - Proceedings*. Association for Computing Machinery, 5 2019.

[10] J. Ha, L. J. Stephen, D. Sally, and J. Cunningham, "Challenges in cross-cultural/multilingual music information seeking." in *ISMIR*, 2005.

[11] C. C. Liem, M. Müller, D. Eck, G. Tzanetakis, and A. Hanjalic, "The need for music information retrieval with user-centered and multimodal strategies," in *MM'11 - Proceedings of the 2011 ACM Multimedia Conference and Co-Located Workshops - MIRUM 2011 Workshop, MIRUM'11*, 2011, pp. 1–6.

[12] P. Knees, "Search & Select - Intuitively Retrieving Music from Large Collections," in *ISMIR*, 2007.

[13] B. Whitman and R. Rifkin, "Musical query-by-description as a multiclass learning problem," in *Proceedings of 2002 IEEE Workshop on Multimedia Signal Processing, MMSP 2002*. Institute of Electrical and Electronics Engineers Inc., 2002, pp. 153–156.

[14] A. van den Oord, S. Dieleman, and B. Schrauwen, "Deep content-based music recommendation," in *Advances in Neural Information Processing Systems*, 2013.

[15] I. Manco, E. Benetos, E. Quinton, and G. Fazekas, "Learning music audio representations via weak language supervision," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022.

[16] S. Oramas, L. Espinosa-Anke, M. Sordo, H. Saggion, and X. Serra, "Information extraction for knowledge base construction in the music domain," *Data and Knowledge Engineering*, 2016.

[17] S. Oramas, M. Sordo, L. Espinosa-Anke, and X. Serra, "A semantic-based approach for artist similarity," in *ISMIR*, 2015.

[18] S. Oramas, L. Espinosa-Anke, A. Lawlor, X. Serra, and H. Saggion, "Exploring Customer Reviews for Music Genre Classification and Evolutionary Studies," in *17th International Society for Music Information Retrieval Conference*, 2016.

[19] S. Oramas, F. Barbieri, O. Nieto, and X. Serra, "Multimodal Deep Learning for Music Genre Classification," *Transactions of the International Society for Music Information Retrieval*, vol. 1, no. 1, pp. 4–21, 9 2018.

[20] R. Delbouys, R. Hennequin, F. Piccoli, J. Royo-Letelier, and M. Moussallam, "Music mood detection based on audio and lyrics with deep neural net," in *Proceedings of the 19th International Society for Music Information Retrieval Conference, ISMIR 2018*, 2018.

[21] B. Jeon, C. Kim, A. Kim, D. Kim, J. Park, and J. W. Ha, "Music emotion recognition via end-To-end multimodal neural networks," in *CEUR Workshop Proceedings*, 2017.

[22] K. Choi, G. Fazekas, M. Sandler, B. Mcfee, and K. Cho, "Towards Music Captioning: Generating Music Playlist Descriptions," *arXiv preprint arXiv:1608.04868*, 2016.

[23] C. Tian, M. Michael, and H. Di, "Music autotagging as captioning," in *Proceedings of the 1st Workshop on NLP for Music and Audio (NLP4MusA)*. Association for Computational Linguistics, 2020, pp. 67–72.

[24] I. Manco, E. Benetos, E. Quinton, and G. Fazekas, "MusCaps: Generating Captions for Music Audio," in *2021 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 2021.

[25] X. Favory, K. Drossos, V. Tuomas, and X. Serra, "COALA: Co-Aligned Autoencoders for Learning Semantically Enriched Audio Representations," in *International Conference on Machine Learning (ICML), Workshop on Self-supervised learning in Audio and Speech*, 2020.

[26] X. Favory, K. Drossos, T. Virtanen, and X. Serra, "Learning Contextual Tag Embeddings for Cross-Modal Alignment of Audio and Tags," in *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021.

[27] A. Ferraro, X. Favory, K. Drossos, Y. Kim, and D. Bogdanov, "Enriched Music Representations with Multiple Cross-modal Contrastive Learning," *IEEE Signal Processing Letters*, vol. 28, pp. 733–737, 4 2021.

[28] J. Choi, J. Lee, J. Park, and J. Nam, "Zero-shot learning for audio-based music classification and tagging," in *Proceedings of the 20th International Society for Music Information Retrieval Conference, ISMIR 2019*, 2019.

[29] M. Won, S. Oramas, O. Nieto, F. Gouyon, and X. Serra, "Multimodal Metric Learning for Tag-based Music Retrieval," in *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021.

[30] M. Won, J. Salamon, N. J. Bryan, G. J. Mysore, and X. Serra, "Emotion Embedding Spaces for Matching Music to Stories," in *International Society for Music Information Retrieval Conference (ISMIR)*, 11 2021.

[31] J. Lu, D. Batra, D. Parikh, and S. Lee, "ViLBERT: Pretraining Task-Agnostic Visiolinguistic Representations for Vision-and-Language Tasks," in *Advances in Neural Information Processing Systems*, 2019, pp. 13–23.

[32] Y. C. Chen, L. Li, L. Yu, A. El Kholy, F. Ahmed, Z. Gan, Y. Cheng, and J. Liu, "UNITER: UNiversal Image-TExt Representation Learning," in *European Conference on Computer Vision*, 2020.

[33] 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 *International Conference on Machine Learning*. PMLR, 2021.

[34] C. Jia, Y. Yang, Y. Xia, Y.-T. Chen, Z. Parekh, H. Pham, Q. V. Le, Y. Sung, Z. Li, and T. Duerig, "Scaling Up Visual and Vision-Language Representation Learning With Noisy Text Supervision," in *International Conference on Machine Learning*. PMLR, 2021, pp. 4904–4916.

[35] J. Andreas, D. Klein, and S. Levine, "Learning with Latent Language," in *NAACL HLT 2018 - 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, vol. 1. Association for Computational Linguistics (ACL), 11 2018, pp. 2166–2179.

[36] X. Liu, X. Mei, Q. Huang, J. Sun, J. Zhao, H. Liu, M. D. Plumbley, V. Kılıç, and W. Wang, "Leveraging Pre-trained BERT for Audio Captioning," *arXiv preprint arXiv:2203.02838*, 2022.

[37] A.-M. Oncescu, A. S. Koepke, J. F. Henriques, Z. Akata, and S. Albanie, "Audio Retrieval with Natural Language Queries," in *Interspeech*, 5 2021.

[38] H.-H. Wu, P. Seetharaman, K. Kumar, and J. P. Bello, "Wav2CLIP: Learning Robust Audio Representations From CLIP," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 10 2022.

[39] Y. Zhao, J. Hessel, Y. Yu, X. Lu, R. Zellers, and Y. Choi, "Connecting the Dots between Audio and Text without Parallel Data through Visual Knowledge Transfer," *arXiv preprint arXiv:2112.08995*, 2021.

[40] A. Guzhov, F. Raue, J. Hees, and A. Dengel, "AudioCLIP: Extending CLIP to Image, Text and Audio," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022.

[41] C. Chen, D. Han, and J. Wang, "Multimodal Encoder-Decoder Attention Networks for Visual Question Answering," *IEEE Access*, pp. 1–1, 2 2020.

[42] D. Niizumi, D. Takeuchi, Y. Ohishi, N. Harada, and K. Kashino, "BYOL for Audio: Self-Supervised Learning for General-Purpose Audio Representation," in *2021 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 3 2021.

[43] A. Saeed, D. Grangier, and N. Zeghidour, "Contrastive Learning of General-Purpose Audio Representations," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 10 2021.

[44] H. Al-Tahan, Y. M. I. C. On, and U. 2021, "CLAR: Contrastive Learning of Auditory Representations," in *International Conference on Artificial Intelligence and Statistics*. PMLR, 2021.

[45] J. Spijkervet and J. A. Burgoyne, "Contrastive Learning of Musical Representations," in *ISMIR*, 2021.

[46] E. Fonseca, D. Ortego, K. McGuinness, N. E. O'Connor, and X. Serra, "Unsupervised Contrastive Learning of Sound Event Representations," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 11 2021.

[47] P. H. Le-Khac, G. Healy, and A. F. Smeaton, "Contrastive Representation Learning: A Framework and Review," *IEEE Access*, vol. 8, pp. 193 907–193 934, 10 2020.

[48] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A Simple Framework for Contrastive Learning of Visual Representations," in *International conference on machine learning*. PMLR, 2020.

[49] J.-B. Alayrac, A. Recasens, R. Schneider, R. Arandjelović, J. Ramapuram, J. De Fauw, L. Smaira, S. Dieleman, and A. Zisserman, "Self-Supervised MultiModal Versatile Networks," *Advances in Neural Information Processing Systems*, vol. 33, pp. 25–37, 6 2020.

[50] M. Patrick, Y. M. Asano, P. Kuznetsova, R. Fong, J. F. Henriques, G. Zweig, and A. Vedaldi, "Multi-modal Self-Supervision from Generalized Data Transformations," *arXiv preprint arXiv:2003.04298v2*, 2020.

[51] A. van den Oord, Y. Li, and O. Vinyals, "Representation learning with contrastive predictive coding," *arXiv preprint arXiv:1807.03748*, 2018.

[52] N. Saunshi, P. Orestis, S. Arora, K. Mikhail, and K. Hrishikesh, "A Theoretical Analysis of Contrastive Unsupervised Representation Learning," in *International Conference on Machine Learning*. PMLR, 2019.

[53] M. Zolfaghari, Y. Zhu, P. Gehler, and T. Brox, "CrossCLR: Cross-modal Contrastive Learning For Multimodal Video Representations," in *International Conference on Computer Vision (ICCV)*. Institute of Electrical and Electronics Engineers (IEEE), 9 2021, pp. 1430–1439.

[54] L. Ericsson, H. Gouk, C. C. Loy, and T. M. Hospedales, "Self-Supervised Representation Learning: Introduction, Advances and Challenges," *IEEE Signal Processing Magazine*, vol. 39, no. 3, pp. 42–62, 10 2021.

[55] N. Mu, A. Kirillov, D. Wagner, and S. Xie, "SLIP: Self-supervision meets Language-Image Pre-training," *arXiv preprint arXiv:2112.12750*, 2021.

[56] H. Xie, O. Räsänen, K. Drossos, and T. Virtanen, "Unsupervised Audio-Caption Aligning Learns Correspondences between Individual Sound Events and Textual Phrases," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022.

[57] K. He, Y. Wang, and J. Hopcroft, "A powerful generative model using random weights for the deep image representation," in *Advances in Neural Information Processing Systems*, 2016.

[58] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in Neural Information Processing Systems*, vol. 2017-Decem. Neural information processing systems foundation, 6 2017, pp. 5999–6009.

[59] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever, "Language Models are Unsupervised Multitask Learners," 2019.

[60] M. Won, K. Choi, and X. Serra, "Semi-Supervised Music Tagging Transformer," in *International Society for Music Information Retrieval Conference (ISMIR)*, 11 2021.

[61] R. Sennrich, B. Haddow, and A. Birch, "Neural Machine Translation of Rare Words with Subword Units," in *54th Annual Meeting of the Association for Computational Linguistics, ACL 2016 - Long Papers*, vol. 3. Association for Computational Linguistics (ACL), 2016, pp. 1715–1725.

[62] N. Reimers and I. Gurevych, "Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks," in *EMNLP-IJCNLP 2019 - 2019 Conference on Empirical Methods in Natural Language Processing and 9th International Joint Conference on Natural Language Processing, Proceedings of the Conference*. Association for Computational Linguistics, 8 2019, pp. 3982–3992.

[63] K. Drossos, S. Lipping, and T. Virtanen, "Clotho: An Audio Captioning Dataset," in *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, vol. 2020-May. IEEE, 5 2020, pp. 736–740.

[64] C. D. Kim, B. Kim, H. Lee, and G. Kim, "AudioCaps: Generating captions for audios in the wild," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Association for Computational Linguistics (ACL), 2019, p. 119–132.

[65] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," in *1st International Conference on Learning Representations, ICLR 2013 - Workshop Track Proceedings*. International Conference on Learning Representations, ICLR, 2013.

[66] J. Bromley, I. Guyon, Y. Lecun, E. Sickinger, R. Shah, A. Bell, and L. Holmdel, "Signature verification using a" siamese" time delay neural network," *Advances in neural information processing systems*, vol. 6, 1993.

[67] X. Zhai, X. Wang, B. Mustafa, A. Steiner, D. Keysers, A. Kolesnikov, and L. Beyer, "LiT: Zero-Shot Transfer with Locked-image Text Tuning," *arXiv preprint arXiv:2111.07991*, 2021.

[68] G. Tzanetakis and P. Cook, "Musical genre classification of audio signals," *IEEE Transactions on Speech and Audio Processing*, vol. 10, no. 5, pp. 293–302, 7 2002.

[69] E. Law, K. West, M. Mandel, M. Bay, and J. Stephen Downie, "Evaluation of algorithms using games: The case of music tagging," in *Proceedings of the 10th ISMIR Conference*, 2009.

[70] L. Wang, P. Luc, Y. Wu, A. Recasens, L. Smaira, A. Brock, A. Jaegle, J.-B. Alayrac, S. Dieleman, J. Carreira, and A. van den Oord, "Towards Learning Universal Audio Representations," in *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 5 2022, pp. 4593–4597.

[71] L. A. Hendricks, J. Mellor, R. Schneider, J.-B. Alayrac, and A. Nematzadeh, "Decoupling the Role of Data, Attention, and Losses in Multimodal Transformers," *Transactions of the Association for Computational Linguistics*, vol. 9, pp. 570–585, 1 2021.

[72] B. Lester, R. Al-Rfou, N. Constant, and G. Research, "The Power of Scale for Parameter-Efficient Prompt Tuning," in *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics (ACL), 12 2021, pp. 3045–3059.