Contrastive Learning Of Music Video Representations

Anonymous author(s)

Submission ID: 56

Abstract. Joint learning of visual and auditory representations for music video recommendation is a promising but relatively unexplored topic. In this work, we focus on exploring if we can learn music video representations in a contrastive manner. We create a dual encoder for the audio and video modalities, and train it using a contrastive loss. However, contrastive learning using an industry dataset of 550,000 music video segments did not manage to successfully pull embeddings from the two modalities closer together. A quantitative evaluation of the learned representations on the downstream tasks of genre classification and music tagging show that performance does not improve compared to several baselines from the literature. Through a qualitative analysis of the learned representations, we demonstrate that embedding the two modalities in the same space is not trivial. The dual encoders extract different kinds of features, which are not easy to unite using contrastive learning. Our code is made publicly available, as well as the pretrained model.

Keywords: Contrastive learning, Multimodal, Content based recommendation, Music videos, Representation learning

1 Introduction

Music videos are some of the most circulated items on the internet, with the most popular videos being watched billions of times [48]. They constitute another medium for musicians to express themselves artistically, and serve as a means to promote albums for record labels. Although music representation learning is a well studied topic (for an overview, see [18]), there is less research on representing music videos specifically. An important consideration for music videos is that they are most relevant shortly after their release, resulting in the need for an approach that works well for new items. A class of methods that are known to work well for new items are content based (CB) recommender systems. These methods recommend items that are similar to items that a user has previously liked, and in order to calculate the similarity to other items, it is essential to have a good item representations [44, 51]. In the use case if music videos, these representations should contain information from both the audio and video domains.

One method for multimodal representation learning that has been proven to be powerful recently is contrastive learning (CL) [6]. Rather than learning from labelled datapoints, these methods learn from pairs of datapoints from different modalities. The most renowned application of this methodology was in the text-image domain [34], but it has also been applied in the audio-video domain [24, 47]. In this paper, we examine whether CL can successfully be generalized to music videos. We train a neural network

on music videos using CL, and evaluate the quality of learned representations on the downstream tasks of genre classification and music tagging. Additionally, we examine whether similarity metrics based on these representations could be useful for CB recommendation.

This research was carried out in collaboration with a European online music video platform. The platform provided a large private dataset of music videos, as well as data infrastructure and in-depth collaboration on the development and implementation of our model. Although this private dataset provides us with the unique opportunity to test state-of-the-art (SOTA) methods on a large collection of real music videos, it unfortunately also means that we can not share our training data publicly, since it largely consists of copyrighted music videos. However, to keep this research as reproducible as possible, we evaluate our trained model on a subsection of the publicly available Million Song Dataset (MSD) that overlaps with the private dataset. Additionally, our code and trained models are made publicly available, along with a description of the private dataset and a comparison to the public MSD.

Research questions. In this paper, we devise a method that learns audio-visual representations for music videos based on their content using CL. We aim to answer the following research questions:

(RQ1) Can CL of audio-video representations be used to improve the quality of music video representations when evaluated on the downstream tasks of genre classification and music tagging compared to existing models?
(RQ2) Can our audio-visual representation be used to calculate music video similarity that is in line with subjective human judgement of similarity?

Contributions. The principal contributions of our research are the following:

- The adaptation of a contrastive deep learning method to learn multimodal music video representations.
- A quantitative evaluation of the learned representations on downstream tasks.
- A qualitative evaluation of the representations based on item similarity.
- To facilitate reproducibility of our work, we provide the code of our experiments, and a description of the private dataset on GitHub[1].

The remainder of this paper is structured as follows. First, we provide an overview of literature on multimodal machine learning for the audio and video modalities. Then, we describe our model, as well as our experimental setup. Finally, we present our results, and discuss their implications for music video representation.

2 Related work

Unimodal approaches. The most commonly used feature in music processing is the log mel spectogram (LMS), which consists of the cosine transform of the log Fourier transform of an audio waveform [33]. It has been used both for general audio related tasks [10], as well as music related tasks [22]. A popular approach to modelling in this

[1] `https://anonymous.4open.science/r/Multimodal-Musicvideo-Representation-6EAF/README.md`

field is to train a convolutional neural network (CNN) that convolves along the time and frequency domain of these spectogram features. This approach has been used for various music related tasks [1, 9, 49]. In a comparative study of models trained on the task of music tagging, Won et al. [49] found that a specific CNN called *musicnn* [32] could get good results with less data compared to different CNNs and other SOTA models.

CNNs are also a popular approach in the video domain. Tran et al. [42] attain SOTA performance on the task of action recogntion with their *R3D* model, which convolves along the temporal dimension as well as the two spatial dimensions. Additionally, the authors show that the three dimensional convolutions can be separated into separate 2D convolutions for space, and 1D convolutions for time. This *R(2+1)D* model reduces the complexity of the model significantly, without degrading performance.

Multimodal learning. The main challenge for multimodal learning is that feature vectors for different modalities encode for different types of information, which is referred to as the heterogeneity gap [14]. The most common way to deal with this problem is to project the different modalities into a single subspace, while somehow still retaining information from the different modalities [3]. Where originally this was often done using feature engineering and statistical learning algorithms, most recent work on multimodal learning relies on deep neural networks [2, 14]. In case of a supervised task, this can be done using separate input layers for each modality that are fused into one layer deeper in the network, which is referred to as multimodal fusion [16, 27, 29]. Multimodal representations can also be learned in an unsupervised manner using autoencoders [26, 38]. However, research seems to have moved in other directions in recent years. Specifically, one of the research topics that has been gaining traction recently is CL [6].

Contrastive learning. Instead of learning from individual samples, CL methods learn from pairs of samples by pulling representations closer together for positive pairs of similar samples, and further apart for negative pairs of dissimilar samples [19]. Which pairs are positive and negative is determined by the researcher. In a unimodal setting, CL is often done by augmenting samples, and choosing pairs of original-augmented samples as positive pairs and pairs of samples from different originals as negative pairs. This method has been used to learn high quality representations of images [8], clips of music [40] and action recognition videos [23].

In a multimodal setting, positive pairs are matching samples from different modalities, and negative pairs are non matching samples from different modalities. A recent study in this field that attracted a lot of attention introduced Contrastive Language-Image Pretraining (CLIP) [34]. The authors of this paper were able to create a multimodal embedding space for images and text using CL on a large dataset of image-caption pairs. Using their network, they were able to attain very impressive performance on several computer vision tasks that it was not specifically trained for. Their approach works as follows. Both modalities have a separate modality-specific encoder network, which is initialised using a pretrained network for the relevant modality. Then, a projection head is added, which consists of some extra feedforward layers that make the output dimensions equal for the two modalities. The network is trained contrastively, by maximizing the similarity of the output layers for image-text pairs that belong together, and minimizing similarity for pairs that do not. This seemingly simple training

task results in very informative representations of the input data, evident from the high zero shot performance on an array of different tasks.

In the wake of this paper, several studies using variations on this method have been published. Guzhov et al. [15] show that the CLIP model can be improved upon by including audio, resulting in a tri-modal embedding space. Several other studies extend CL to the audio-video domain [1, 24, 47], training contrastive networks on video clips that of events that include audio, and narrated video clips. The contrastively learned representations in these papers resulted in increased performance on the downstream tasks of video and audio event recognition, as well as non-semantic tasks, such as music instrument classification. Since CL has proven to be such a useful way of learning representations in the audio-visual domain, it might also be a good way to learn representations for music videos.

Although earlier research has shown that CL can be successful in the audio-video domain, it has not previously been used for music videos. Compared to the music video use case, the two modalities were quite closely related in previous research. The most common datasets used in the field, which were also used in the aforementioned papers, are the kinetics human action dataset [17], the AudioSet dataset [13] and the HowTo100M dataset [25]. Respectively, these consist of YouTube video clips of actions and the sounds they produce, such as humans playing an instrument or shaking hands, YouTube video clips of audio events, like the sounds of a human voice, a musical instrument or a bell ringing, and video clips with narration, explaining what is going on in the video. Unlike previous research, we explore whether the constrastive learning approach can be generalised to the domain of music videos, in which audio and video are much less directly related. Although there are certainly correlations between the type of music and the type of music video that goes with it, this relationship is clearly much less direct that the relation between a video of an instrument and its sound, or a video and its narration.

3 Approach

In this paper, we adapt the multimodal CL approach discussed in Section 2 to the use case of music videos. Figure 1 presents an overview of the proposed architecture. Which is discussed in detail in the folowing paragraphs.

Audio encoding pipeline. the audio encoder pipeline consists of an encoder network f_a and a projection head g_a. We use the *musicnn* architecture as encoder [32], Initialising the network with the pretrained weights, and freezing these weights during training. The projection head consists of two dense feed forward layers. We use a hidden layer size of 512 and an embedding size of 256. These embedding sizes are relatively small compared to some other multimodal networks [11, 34] because we have relatively little training data compared to other studies. The hidden layer uses a ReLU activation function and is trained using a dropout rate of 0.3. The embedding layer uses a sigmoid activation function. To obtain a representation for an audio segment x_a, we pass the segment though the pretrained encoder network and the projection head respectively:

$$a = g_a(f_a(x_a)). \tag{1}$$

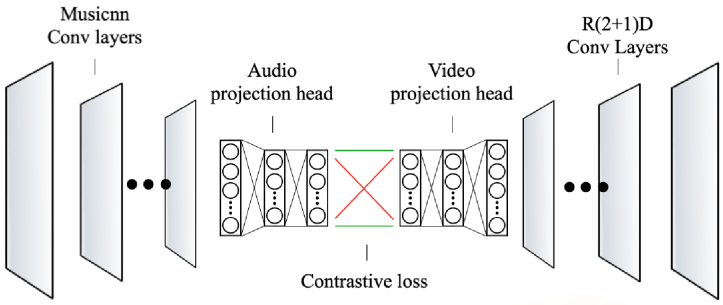

Fig. 1: Architecture of our model, comprised of the convolutional layers of the pretrained *musicnn* and R(2+1)D networks, and a dense projection head for each modality.

Video encoding pipeline. For the video modality, we use the (2+1)D CNN from Tran et al. [42] as encoder f_v. This encoder is initialized with the pretrained weights, and is frozen during training. We use the same projection head architecture for the video encoder as we did for the audio encoder. We obtain the representation for a video segment x_v by passing it though the video encoder and video respectively:

$$v = g_v(f_v(x_v)). \tag{2}$$

We also experiment with several variations to the architecture of our two projection heads, which are discussed in Section 4.

Loss Function. To train the dense layers of the two projection heads, we use the contrastive loss function from simCLR [8], adapted to the audio-video scenario. For a batch of N music videos, video to audio and audio to video losses for a positive pair of audio embedding a_j and video embedding v_j are defined respectively as:

$$l_j^{(v\to a)} = -log\frac{\exp(f_{sim}(a_j, v_j)/\tau)}{\sum_{k=1}^{N} \mathbb{1}_{[k\neq j]} \exp(f_{sim}(a_j, v_k)/\tau)}. \tag{3}$$

$$l_j^{(a\to v)} = -log\frac{\exp(f_{sim}(v_j, a_j)/\tau)}{\sum_{k=1}^{N} \mathbb{1}_{[k\neq j]} \exp(f_{sim}(v_j, a_k)/\tau)}. \tag{4}$$

Here $f_{sim}(x, y)$ is the cosine similarity function: $f_{sim}(\mathbf{q}, \mathbf{x}) = \frac{\mathbf{q}}{||\mathbf{q}||}\frac{\mathbf{x}}{||\mathbf{x}||}$; and $\mathbb{1}$ is an indicator function. Finally, τ is a temperature parameter that controls the strength of penalties on hard negative samples. Research has shown that a low temperature parameter helps to learn separable features, although setting it too low will result in penalising semantically similar negative samples too heavily [45]. We experiment with different values for τ, as detailed in Section 4. The overall loss is bidirectional, combining the values of the two aforementioned loss functions:

$$\mathcal{L} = \frac{1}{\beta}\sum_{j=1}^{\beta}\left(\ell_j^{(a\to v)} + \ell_j^{(v\to a)}\right). \tag{5}$$

Minimising this loss function pulls representations of audio and video from the same source closer together, while pushing representations of the audio and video from different sources further apart.

4 Experimental Setup

Datasets. We use two datasets in this paper. The first and main dataset is a company dataset. We use it for training and evaluation of our model. The dataset consists of a collection of more than 90,000 music videos published to the music video platform that we collaborate with. The second dataset is the MSD [4]. This dataset is a benchmark dataset on the task of music tagging. The main difference between these datasets is that the private dataset contains music videos, whereas the MSD is audio only. In terms of metadata, the music in the private dataset is more recent, since the MSD dataset was released in 2010, and has not been updated since. Also, music in the private dataset comes from a slightly more diverse set of origins, although both datasets consist mostly of western music. Finally, the MSD dataset has more rock and electronic music, whereas the private dataset contains more hip-hop and latin tracks. For a more detailed description and comparison of the two datasets, we refer to the GitHub page [2].

Evaluation method. There are two principal components of the paper: training the model, and evaluating the representations on downstream tasks. In the first component, we train our model on the private dataset. We experiment with several different configurations of the hyperparameters and architecture in this phase. Then, in the second component, we use the representations that we learned in the first step, and use them on the downstream tasks of genre classification and music tagging. This allows us to compare performance using the CL representations to baselines from the literature.

Metrics. Following Zolfaghari et al. [52], we use median rank of cross modal retrieval to evaluate the learned embeddings. Since our loss function is bidirectional, we report the audio to video median rank as well as the video to audio median rank. We also tracked these metrics alongside the value of the loss function to monitor training progress. To evaluate performance on the downstream tasks of genre classification and music tagging, we use area under the ROC curve (AUC), following Won et al. [49]. Additionally, we report F_1 score for downstream task performance.

Baselines. When evaluating the CL representations we use the pretrained backbone models as baselines. For the audio modality, we use *musicnn* [32] as a baseline, and for the video modality we use *R(2+1)D* [42]. This allows us to compare our representations to representations that were not fine tuned using CL.

Experiments. We run 3 experiments. In *Experiment 1* we use CL to learn representations for the dataset of music video segments. We start with a baseline configuration of our model, and experiment with several variations to this configuration:

1. $CL_{baseline}$: we use a two layer projection head with a hidden layer of size 512 and an embedding size of 256. We set τ to one, corresponding to using a loss function without a temperature parameter.
2. We increase the embedding size to 512.
3. We increase the depth of the projection head to four layers.
4. We experiment with using a single projection head for the video modality, which

[2] `https://anonymous.4open.science/r/Multimodal-Musicvideo-Representation-6EAF/README.md`

projects the video embeddings to the dimensions of the *musicnn* embeddings. This way the video representations are moved closer to the audio representations, rather than moving both closer to each other.
5. We lower the temperature parameter to .3, as suggested by Wang and Liu [45]. This penalises hard negatives more heavily, which helps learning separable embeddings.
6. We train autoencoders to initialize the weights of the projection heads for the two modalities. Previous research has shown that initializing weights in this way can improve performance on several deep learning tasks [12, 30].

In *Experiment 2*, we evaluate the representations on the downstream task of genre classification. For this task, there is no consensus on a benchmark dataset. The most used dataset in the field is the GTZAN dataset [43], but subsequent research has pointed out that the quality of the labels in this dataset is poor [31]. This lack of a good common dataset also makes it hard to compare different genre detection models, and to determine what methods are the SOTA [35]. In our study, we perform genre classification on the private dataset. This allows us to test our model in a realistic practical context, and additionally it means that we can study the effect of the video modality on classification. We fit three different models to predict genre based on our learned representations:

- CL_a: A model based on the contrastive embeddings calculated from the audio.
- CL_v: A model based on the contrastive embeddings calculated from the video.
- CL_{agg}: A model based on the aggregate of the contrastive audio and video embeddings.

To evaluate the effectiveness of our learned representations, we compare these models to three baseline models, which do classification based on the pretrained embeddings:

- *Musicnn*: A model based on the *musicnn* embeddings [32].
- *R(2+1)D*: A model based on the *R(2+1)D CNN* embeddings [42].
- *Musicnn+R(2+1)D*: A model based on the concatenated *musicnn* and *R(2+1)D CNN* embeddings.

Each of the models was chosen to be a simple three layer perceptron, with hidden layers sizes of 512 and 256 respectively. Comparing our models to these baselines allows us to evaluate whether learning representations contrastively improves performance on this task.

In *Experiment 3*, we evaluate the embeddings on the downstream task of music tagging. We use the same baselines and models as for genre classification. However, we use a subset of the public MSD [4], that overlaps with our private dataset. This allows us to keep our research as reproducible as possible. The comprehensive list of the 9416 MSD IDs of the overlapping tracks is available on Github [3]. The tagging task consists of predicting several tags that were manually annotated based on the audio. These tags contain genres (blues, pop), eras (70s, 80s), and moods (chill, happy) [9, 32]. Since our training set is relatively small, we only use the top ten most common tags. The models consist of a three layer perceptron with hidden layer sizes 512 and 256 respectively.

[3] `https://anonymous.4open.science/r/Multimodal-Musicvideo-Representation-6EAF/README.md`

Implementation Details. In order to run our experiments in relatively quick succession, we stored just six five-second non-overlapping segments per music video, rather than the entire music videos. We sampled the segments by dividing the track up into six equally sized sections, and sampling a random five second segment from each section, preventing overlap between the segments.

Consecutive video frames tend to have high redundancy [46], and there are different ways to sample a subset of frames from a video to process. One of the most simple and common approaches is to sample a random set of frames with equal temporal spacing between them [5, 39]. To achieve this effect, we save the video segments to an mp4 file at 432p and 10 frames per second. The lower frame rate allows to sample consecutive frames, which will naturally have an equal temporal spacing of a tenth of a second. We save 50 frames at a resolution of 768x432 pixels. We save the corresponding five seconds of audio to a wav file at a sampling rate of 16,000. The stereo channels are converted to mono by taking the average since the *musicnn* architecture is based on a single input channel. For model evaluation purposes, we split the data up into a training split of 80 percent, and a validation and testing split of 10 percent each.

When training the CL model, we implemented a constraint for music videos to be represented a maximum of once per batch, so that separate segments of the same music video can not appear in the same batch. In addition to the random sampling of segments, we applied random cropping to 112x112 in order to get the correct input shape for the network. To prevent the cropped videos to contain an unreasonably small fraction of the pixels in the original video, we also resize the videos to be smaller, so that the cropped videos still contain a large section of the original frames. This is resize-crop combination is commonly used as a data augmentation in visual deep learning [41].

During training, we use a batch size of 1000, with an initial learning rate of .01, and exponential learning decay with a gamma parameter of .95, which has been shown to improve improve training results in previous research [21]. To prevent overfitting, we implement an early stopping rule. Specifically, we stop training when the validation loss does not get any lower over a period of three epochs.

5 Results

Training experiments. In the first component of our research, we executed several experiments in order to enhance the training process. Table 1 presents the median rank of the corresponding segment for all of the experiments. The results show that the median rank stays close to 500 for all models, indicating that none of the models succeeded to unite audio and video embeddings for positive pairs.

Downstream tasks. Table 2 shows the average area under the curve and F1 score for each baseline and model on the downstream task of music tagging. The scores represent macro-averages, indicating that we first calculate the scores per label, and then take the unweighted mean over tags. Overall, these results present two main findings. Firstly, the networks that are trained on embeddings that include information about audio clearly outperform models that are only based on video. This is in line with expectations, as the MSD tags are based solely on the audio of a song, and do not take into account the music video that goes with the audio. Secondly, the performance of the baseline

Table 1: Effect of different variations in *experiment 1* on the median rank of cross modal retrieval.

Configuration	$MedianRank_{a \to v}$	$MedianRank_{v \to a}$
CL_{base}	507	502
Embedding size 512	494	498
Four projection layers	503	501
Single projection head	499	497
Temperature .3	497	503
Autoencoder initiated	501	500

models is better than the performance of the models that were fine-tuned using CL. We expected the model based on contrastively learned embeddings to outperform the baselines, but given the failure to unite embeddings from the different modalities, these results are less surprising.

The results for the downstream task of genre classification are similar to the results for music tagging. Table 2 presents the same two statistics for the task of genre classification. Similar to the results above, we see that models including audio data outperform models that are solely based on video, and more importantly, that all three baselines outperform their CL counterparts.

Table 2: Performance our models on the downstream tasks of genre classification (*Experiment 2*) and music tagging (*Experiment 3*), evaluated on AUC and F_1. The best performance is highlighted in bold.

Embedding type	**Music tagging**		**Genre classification**	
	AUC	F_1	**AUC**	F_1
Musicnn	0.78	0.22	0.79	0.18
R(2+1)D	0.66	0.09	0.68	0.05
Musicnn+R(2+1)D	**0.78**	**0.24**	**0.82**	**0.21**
CL_a	0.70	0.10	0.69	0.08
CL_v	0.61	0.09	0.63	0.04
CL_{agg}	0.71	0.11	0.68	0.07

6 Qualitative analysis of representations

To better understand our results, we also analysed our representations qualitatively, both on the level on entire music videos and individual samples.

Music videos. Initially, we aggregated embeddings over different segments and modalities. We first took the average over segments, to come to a single audio and a single video representation, and then took the mean to get to a single multimodal representation per music video. Then, we selected 25 random seed videos, and retrieved the three most similar music videos to each seed. We planned to have a group of human judges compare the most similar videos according to the contrastive representations to the most similar music videos according to the metadata. However, it soon became clear that the

similarity metric based on contrastive representations was inferior to the point that such a user study was redundant. In general, the most similar music videos did not appear to be similar to the seed video at all in terms of both audio and video. One exception to this rule were live performances, which are often most similar to other live performances. When taking into account only the video modality, rather than the aggregate of audio and video embeddings, some other patterns were also present: Some seed videos that were all black and white, as well as some videos of people dancing or playing instruments, would result in similar videos being retrieved. One thing that these exceptions have in common, is that they are very consistent in terms of video across different parts of the music video. This could point to the fact that aggregating representations of different segments of the same music video is not a valid way to get a representation of the music video, especially when different parts of the music video are very inconsistent. This is particularly often the case for the video modality, as music videos often consist of a wide variety of shots and scenes.

Segments. In order to investigate this potentially negative effect of aggregating the different segments of a music video, we also examined similarity on the level of individual segments. We took the same approach as before, this time selecting 20 random seed segments, and retrieving the three most similar segments to this seed segment. As expected, the results were somewhat better for single segments. There were more cases in which the retrieved videos were similar to the seed, although in general the retrieved videos were still not similar. For example, some seeds for which the retrieved segments were similar were close-ups of women singing, shots of people playing guitar, segments with a dark background with bright neon accents, and shots of men talking. Overall similar segment retrieval was poor, but there were more exceptions in which there was some specific similarity between the videos.

7 Discussion and limitations

Our results demonstrate the difficulty of representing the audio and video of music videos in a common space. Our audiovisual representations did not improve performance on either of the downstream task, answering RQ1. Additionally, with respect to RQ2, results indicated that our representations can not be used to infer music video similarity. In this section, we discuss some of the potential reasons for these issues, as well as some limitations to our approach.

Discussion. As mentioned previously, the main challenge for multimodal learning consists of bridging the heterogeneity gap [14]. CL is one approach to overcome this problem, and some notable successful applications involve learning common representations for images and their captions [34], videos and their descriptions [52] and videos and their audio streams [47]. Compared to our music video use case, the heterogeneity gap in these studies is relatively small. Although an image and its caption are represented in different modalities, they are often closely related, as the caption tries to describe exactly what is in the image: the two modalities both refer to the same concept. This is also true for the use case of videos and their descriptions, and a video of an event and the sound of that video. However, for a piece of music and the video that goes with it, the relationship between the two is much less direct. As opposed to the other

examples, the two modalities in a music video are not both directly referring to a single concept. Rather than describing the music, the video modality serves as an extra means to entertain the user and to convey emotion. This large heterogeneity gap is reflected in our qualitative analysis of similarity. Whereas the music network mainly encodes for features relating to the types of (digital) instruments used, like distorted guitars or loud electronic bass drums, the video network encodes for features like the color of the video as well as specific actions like singing or dancing. The relationship between these features is not always immediately apparent. Different videos that use a black and white video can be very different in what type of instruments they use, and the fact that a music video contains people dancing, will not tell you much about the instruments that are being used in the song. This is quite different from the video-audio use case from previous research, where different different videos of similar events are likely to have similar audio streams, since the event in the video is often the direct source of the audio [47]. We believe that the fact that the modalities in our research are considerably less tightly connected could be an important cause of the difficulties we encountered trying to unite embeddings from the two modalities.

Limitations. The most important limitation of our research is that the median rank of corresponding segments did not manage to fall below 500 consistently. Taking into account that we are using a batch size of 1000, this indicates that the training process is failing to pull embeddings for corresponding audio and video segments closer to each other, relative to non-corresponding segments. This is a serious problem for our project, as uniting embeddings from different modalities is the fundamental idea of CL. A possible explanation for this result is that the heterogeneity gap is too big in our use case, making it hard to represent audio and video in the same space.

A second limitation of our approach concerns the input size of the encoder networks. Due to the high dimensionality and complexity of the input data, combined with our limited computational capacity, we were limited to training on segments of only three seconds. This makes the task of retrieving corresponding segments from the other modality significantly harder. Even for a human, retrieving the corresponding piece of music based on a three second videoclip is difficult. Additionally, this short input length makes it harder to compute similarity between complete music videos, which is necessary for CB recommendation. We tried aggregating representations of segments to come to a single representation per music video, but our qualitative analysis of similarity revealed that this aggregation was degradatory to the performance, especially so when the input data was inconsistent across segments, which is usually the case for video data from music videos. It might be more fruitful to use a larger input size, or to extract features that are more consistent across the duration of the music video.

A final limitation concerns the fact that we freeze the weights of the pretrained networks during training. Due to our limited computational capacity, we did not fine-tune the encoder networks. Possibly, freezing these layers does not leave enough freedom for the network to learn a feature space in which audio and video embeddings can be united. Although many previous applications of CL do train all layers of the network [1, 34, 47], there are also various earlier studies that freeze the pretrained layers and come to successful results [50, 52]. Although tuning the entire network could potentially help bridge the large heterogeneity gap, we believe that it is unlikely that this

would fix the problem, since fine-tuning generally just leads to a moderate improvement in performance. Another option would be to train the entire model form scratch, but this would require a dataset that is several orders of magnitude larger than ours.

8 Conclusions

In this work, we trained a contrastive deep learning model on the audio and video modalities of music videos. We evaluated the representations learned by our model quantitatively on the downstream tasks of music tagging and genre classification, and qualitatively by investigating the cosine similarity between music video embeddings.

We found that the contrastive loss was not able to pull embeddings from the two modalities closer together. Models based on the contrastively learned embeddings performed worse at both genre classification and music tagging compared to models that were using pretrained networks without contrastive fine-tuning. Additionally, our qualitative analysis revealed that music videos with dissimilar content will often have relatively similar representations. Similarity on the level of segments, without aggregation, resulted in more cases in which the similarity makes sense, although in general it still did not work well.

These results show that it is hard to unite embeddings from the two modalities, which could possibly be due to the large discrepancy between songs and their music videos, as well as the short input size of the encoder networks combined with the inconsistency of the data throughout the duration of a music video.

Based on these results, we have two main suggestions for future work. Firstly, given the large heterogeneity gap, it would be useful for future research to look into multimodal alternatives to CL that rely less heavily on the association between features in different modalities. Multimodal fusion might for example be more suitable, as these models combine features from different modalities based on a supervised task, rather than solely based on the relationship between features in the different modalities [27]. A potential supervised learning task learning representations for recommendation could be predicting latent matrix factorisation scores [28].

Secondly, we discussed the small input size of the encoder networks as a potential reason for the poor results. Future research could try to increase this input size in order to learn embeddings that are more representative for entire music videos. However, current research on video processing is based on short videoclips of single events [36], so learning meaningful representations for longer, more diverse videos might be more challenging than just increasing the input length of the network. Ideally, the encoder network for video would learn features relating to the general style of the video, rather than features related to concrete events, as the style is usually relatively consistent over the music video, as opposed to actions, which differ a lot across shots and scenes. Although most research on deep learning for video data consists of classifying action or events, detecting faces or estimating poses [37], there is also a class of models concerned with style transfer for video data [7, 20]. These methods could be relevant to future research on multimodal music video recommendation, as they aim to learn features related to style rather than content, which is more consistent throughout a music video, and might also be more closely related to the musical content of the song.

Bibliography

[1] Akbari H, Yuan L, Qian R, Chuang WH, Chang SF, Cui Y, Gong B (2021) Vatt: Transformers for multimodal self-supervised learning from raw video, audio and text. Advances in Neural Information Processing Systems 34

[2] Atrey PK, Hossain MA, El Saddik A, Kankanhalli MS (2010) Multimodal fusion for multimedia analysis: a survey. Multimedia systems 16(6):345–379

[3] Baltrušaitis T, Ahuja C, Morency LP (2018) Multimodal machine learning: A survey and taxonomy. IEEE transactions on pattern analysis and machine intelligence 41(2):423–443

[4] Bertin-Mahieux T, Ellis DP, Whitman B, Lamere P (2011) The million song dataset

[5] Carreira J, Zisserman A (2017) Quo vadis, action recognition? a new model and the kinetics dataset. In: proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp 6299–6308

[6] Chai W, Wang G (2022) Deep vision multimodal learning: Methodology, benchmark, and trend. Applied Sciences 12(13):6588

[7] Chen D, Liao J, Yuan L, Yu N, Hua G (2017) Coherent online video style transfer. In: Proceedings of the IEEE International Conference on Computer Vision, pp 1105–1114

[8] Chen T, Kornblith S, Norouzi M, Hinton G (2020) A simple framework for contrastive learning of visual representations. In: International conference on machine learning, PMLR, pp 1597–1607

[9] Choi K, Fazekas G, Sandler M (2016) Automatic tagging using deep convolutional neural networks. arXiv preprint arXiv:160600298

[10] Darji MC (2017) Audio signal processing: A review of audio signal classification features. International Journal of Scientific Research in Computer Science, Engineering and Information Technology 2(3):227–230

[11] Faghri F, Fleet DJ, Kiros JR, Fidler S (2017) Vse++: Improving visual-semantic embeddings with hard negatives. arXiv preprint arXiv:170705612

[12] Ferreira MF, Camacho R, Teixeira LF (2020) Autoencoders as weight initialization of deep classification networks for cancer versus cancer studies. arXiv preprint arXiv:200105253

[13] Gemmeke JF, Ellis DP, Freedman D, Jansen A, Lawrence W, Moore RC, Plakal M, Ritter M (2017) Audio set: An ontology and human-labeled dataset for audio events. In: 2017 IEEE international conference on acoustics, speech and signal processing (ICASSP), IEEE, pp 776–780

[14] Guo W, Wang J, Wang S (2019) Deep multimodal representation learning: A survey. IEEE Access 7:63373–63394

[15] Guzhov A, Raue F, Hees J, Dengel A (2021) Audioclip: Extending clip to image, text and audio. arXiv preprint arXiv:210613043

[16] Jiang YG, Wu Z, Wang J, Xue X, Chang SF (2017) Exploiting feature and class relationships in video categorization with regularized deep neural networks. IEEE transactions on pattern analysis and machine intelligence 40(2):352–364

[17] Kay W, Carreira J, Simonyan K, Zhang B, Hillier C, Vijayanarasimhan S, Viola F, Green T, Back T, Natsev P, et al. (2017) The kinetics human action video dataset. arXiv preprint arXiv:170506950
[18] Kim J, Urbano J, Liem C, Hanjalic A (2020) One deep music representation to rule them all? a comparative analysis of different representation learning strategies. Neural Computing and Applications 32(4):1067–1093
[19] Le-Khac PH, Healy G, Smeaton AF (2020) Contrastive representation learning: A framework and review. IEEE Access 8:193907–193934
[20] Li X, Liu S, Kautz J, Yang MH (2019) Learning linear transformations for fast image and video style transfer. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp 3809–3817
[21] Li Z, Arora S (2019) An exponential learning rate schedule for deep learning. arXiv preprint arXiv:191007454
[22] Liu X, Chen Q, Wu X, Liu Y, Liu Y (2017) Cnn based music emotion classification. arXiv preprint arXiv:170405665
[23] Lorre G, Rabarisoa J, Orcesi A, Ainouz S, Canu S (2020) Temporal contrastive pretraining for video action recognition. In: Proceedings of the IEEE/CVF winter conference on applications of computer vision, pp 662–670
[24] Ma S, Zeng Z, McDuff D, Song Y (2020) Active contrastive learning of audio-visual video representations. arXiv preprint arXiv:200909805
[25] Miech A, Zhukov D, Alayrac JB, Tapaswi M, Laptev I, Sivic J (2019) Howto100m: Learning a text-video embedding by watching hundred million narrated video clips. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp 2630–2640
[26] Ngiam J, Khosla A, Kim M, Nam J, Lee H, Ng AY (2011) Multimodal deep learning. In: ICML
[27] Nojavanasghari B, Gopinath D, Koushik J, Baltrušaitis T, Morency LP (2016) Deep multimodal fusion for persuasiveness prediction. In: Proceedings of the 18th ACM International Conference on Multimodal Interaction, pp 284–288
[28] Van den Oord A, Dieleman S, Schrauwen B (2013) Deep content-based music recommendation. Advances in neural information processing systems 26
[29] Ortega JD, Senoussaoui M, Granger E, Pedersoli M, Cardinal P, Koerich AL (2019) Multimodal fusion with deep neural networks for audio-video emotion recognition. arXiv preprint arXiv:190703196
[30] Paine TL, Khorrami P, Han W, Huang TS (2014) An analysis of unsupervised pre-training in light of recent advances. arXiv preprint arXiv:14126597
[31] Pálmason H, Jónsson BÞ, Schedl M, Knees P (2017) Music genre classification revisited: An in-depth examination guided by music experts. In: International Symposium on Computer Music Multidisciplinary Research, Springer, pp 49–62
[32] Pons J, Serra X (2019) musicnn: Pre-trained convolutional neural networks for music audio tagging. arXiv preprint arXiv:190906654
[33] Purwins H, Li B, Virtanen T, Schlüter J, Chang SY, Sainath T (2019) Deep learning for audio signal processing. IEEE Journal of Selected Topics in Signal Processing 13(2):206–219
[34] Radford A, Kim JW, Hallacy C, Ramesh A, Goh G, Agarwal S, Sastry G, Askell A, Mishkin P, Clark J, et al. (2021) Learning transferable visual models from

natural language supervision. In: International Conference on Machine Learning, PMLR, pp 8748–8763
[35] Ramírez J, Flores MJ (2020) Machine learning for music genre: multifaceted review and experimentation with audioset. Journal of Intelligent Information Systems 55(3):469–499
[36] Ren Q, Bai L, Wang H, Deng Z, Zhu X, Li H, Luo C (2019) A survey on video classification methods based on deep learning. DEStech Transactions on Computer Science and Engineering (cisnrc)
[37] Santos GNd, de Freitas PV, Busson AJG, Guedes ÁL, Milidiú R, Colcher S (2019) Deep learning methods for video understanding. In: Proceedings of the 25th Brazillian Symposium on Multimedia and the Web, pp 21–23
[38] Silberer C, Lapata M (2014) Learning grounded meaning representations with autoencoders. In: Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp 721–732
[39] Simonyan K, Zisserman A (2014) Two-stream convolutional networks for action recognition in videos. Advances in neural information processing systems 27
[40] Spijkervet J, Burgoyne JA (2021) Contrastive learning of musical representations. arXiv preprint arXiv:210309410
[41] Takahashi R, Matsubara T, Uehara K (2018) Ricap: Random image cropping and patching data augmentation for deep cnns. In: Asian conference on machine learning, PMLR, pp 786–798
[42] Tran D, Wang H, Torresani L, Ray J, LeCun Y, Paluri M (2018) A closer look at spatiotemporal convolutions for action recognition. In: Proceedings of the IEEE conference on Computer Vision and Pattern Recognition, pp 6450–6459
[43] Tzanetakis G, Cook P (2002) Musical genre classification of audio signals. IEEE Transactions on speech and audio processing 10(5):293–302
[44] Volkovs M, Yu GW, Poutanen T (2017) Content-based neighbor models for cold start in recommender systems. In: Proceedings of the Recommender Systems Challenge 2017, pp 1–6
[45] Wang F, Liu H (2021) Understanding the behaviour of contrastive loss. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 2495–2504
[46] Wang L, Xiong Y, Wang Z, Qiao Y, Lin D, Tang X, Gool LV (2016) Temporal segment networks: Towards good practices for deep action recognition. In: European conference on computer vision, Springer, pp 20–36
[47] Wang L, Luc P, Recasens A, Alayrac JB, Oord Avd (2021) Multimodal self-supervised learning of general audio representations. arXiv preprint arXiv:210412807
[48] Wikipedia contributors (2022) List of most-viewed youtube videos — Wikipedia, the free encyclopedia. URL `https://en.wikipedia.org/wiki/List_of_most-viewed_YouTube_videos`, [Online; accessed 2-February-2022]
[49] Won M, Ferraro A, Bogdanov D, Serra X (2020) Evaluation of cnn-based automatic music tagging models. arXiv preprint arXiv:200600751
[50] Xu H, Ghosh G, Huang PY, Okhonko D, Aghajanyan A, Metze F, Zettlemoyer L, Feichtenhofer C (2021) Videoclip: Contrastive pre-training for zero-shot video-text understanding. arXiv preprint arXiv:210914084

[51] Zhu Y, Lin J, He S, Wang B, Guan Z, Liu H, Cai D (2019) Addressing the item cold-start problem by attribute-driven active learning. IEEE Transactions on Knowledge and Data Engineering 32(4):631–644

[52] Zolfaghari M, Zhu Y, Gehler P, Brox T (2021) Crossclr: Cross-modal contrastive learning for multi-modal video representations. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp 1450–1459