MusCaps: Generating Captions for Music Audio

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

***Abstract*—Content-based music information retrieval has seen rapid progress with the adoption of deep learning. Current approaches to high-level music description typically make use of classification models, such as in auto-tagging or genre and mood classification. In this work, we propose to address music description via audio captioning, defined as the task of generating a natural language description of music audio content in a human-like manner. To this end, we present the first music audio captioning model, MusCaps, consisting of an encoder-decoder with temporal attention. Our method combines convolutional and recurrent neural network architectures to jointly process audio-text inputs through a multimodal encoder and leverages pre-training on audio data to obtain representations that effectively capture and summarise musical features in the input. Evaluation of the generated captions through automatic metrics shows that our method outperforms a baseline designed for non-music audio captioning. Through an ablation study, we unveil that this performance boost can be mainly attributed to pre-training of the audio encoder, while other design choices – modality fusion, decoding strategy and the use of attention – contribute only marginally. Our model represents a shift away from classification-based music description and combines tasks requiring both auditory and linguistic understanding to bridge the semantic gap in music information retrieval[1].**

I. Introduction

Current music information retrieval (MIR) approaches to music description typically rely on single- or multi-label classification. A prominent example is music auto-tagging [1]–[3], in which descriptive keywords are assigned to a music clip so as to convey high-level characteristics of the input such as genre, instrumentation and emotion. While this offers a reasonable starting point to modelling high-level representations of music, its output is limited to a predefined set of categorical labels. This in turns limits its usefulness in applications such as music search and recommendation, which would benefit from being able to both process and generate more human-like, detailed and nuanced descriptions. It is in fact through natural language that we often query music collections and search for known and unknown music content. A significant part of our daily language is also devoted to describing the musical world, offering an interface between music as mere audio signals and the set of abstractions used to describe it. While a wealth of music information is encoded in text, research in machine listening and MIR has traditionally overlooked the relationship between audio and natural language.

This work was jointly supported by UK Research and Innovation [grant number EP/S022694/1], Queen Mary University of London, and Universal Music Group.

[1]Code available at https://github.com/ilaria-manco/muscaps

In this study, we focus on the novel task of *music captioning*, which we define as the ability to extract, disentangle and reason about high-level musical concepts in music audio, and then map these to the text modality by generating syntactically and semantically correct sequences of words. This is significantly harder than other, more common music description tasks such as classification and recognition, but presents a solution to some of their limitations. Captioning systems not only need to recognise signal-level features such as instrumentation and high-level descriptors such as genre, but they must also encode the relationship between them, thus better capturing the nuances of musical content; they also produce fully formed, descriptive sentences, that more closely match human queries. Through its joint use and processing of audio and linguistic information, music captioning also provides a first step towards the development of audio-and-language models for music understanding.

Finally, music captioning has several useful applications, such as producing descriptions for items in large music catalogues or vast collections of amateur and user-generated content; automatically generating evocative descriptions of music in films and videos for deaf and hard-of-hearing people; enabling search and discovery of music through more human-like queries; and providing explanations for automatic music recommendations.

To the best of our knowledge, this is the first work on music audio captioning. In the absence of benchmark datasets and established research on this task, we build upon previous literature on image and audio captioning and compare our model performance to a baseline sequence-to-sequence model. Differently from pioneering work on neural architectures for audio captioning, typically composed of an audio encoder and a text decoder, we propose a multimodal encoder that learns a joint representation of both audio and text to better account for the need to capture high-level semantics and summarise information that emerges at different levels of granularity in the input. We also show that, while audio-text data in the music domain is hard to obtain, this data scarcity issue can be effectively alleviated in the context of music captioning by employing suitable pre-trained audio representations. To accomplish this, we leverage large-scale pre-training on a publicly available music dataset and investigate the role of this pre-training step when only a smaller corpus is available for the downstream task of music audio captioning.

The main contributions of this work can be summarised as follows: (i) we propose the first music captioning model for

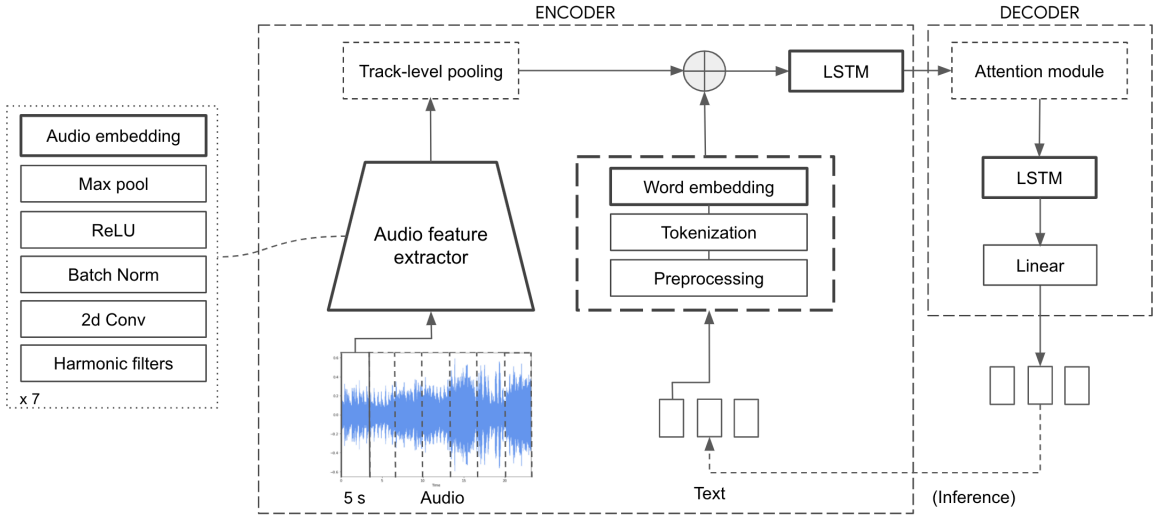

Fig. 1. Overview of the MusCaps architecture. Our model has a hybrid and modular design that allows us to join both convolutional and recurrent neural network architectures to process the audio-text input pairs, optionally learn a soft alignment between them and output a sequence of predicted text tokens conditioned on the input audio.

track-level audio and evaluate it on the captioning and audio retrieval tasks; (ii) we establish whether commonly available pre-trained music auto-tagging models can be usefully employed in a transfer learning setting for downstream audio-linguistic tasks; (iii) through an ablation study, we investigate the effect of modality fusion, temporal attention and beam search decoding on the model performance.

II. Related Work

A. Vision & Language

Within the field of machine perception, image caption generation has long been studied [4], [5]. Thanks to enormous progress in natural language processing (NLP) and computer vision, coupled with the increased availability of large-scale image-text datasets, research in this field has recently focussed on developing models to solve several vision-and-language (V&L) tasks. Most V&L models can be classified under two paradigms: CNN-RNN [4]–[6] and Transformer-based [7]–[10]. The former encompasses a family of fusion-based models in which convolutional neural network (CNN) architectures are employed as feature extractors, while recurrent neural networks (RNN), are employed as language models conditioned on the visual input. More recently, the success of BERT [11] has motivated research into its adaptation to V&L tasks, allowing to achieve state-of-the-art (SOTA) performance in all of them, both for still image and video [12], [13].

B. Audio & Language

While multimodal tasks have long been studied in the visual domain, audio-and-language research has only recently started to emerge, with the first audio captioning model proposed in [14]. Following a similar development to its visual counterpart, audio captioning has seen a rapid progress over the last few years [15]–[22], greatly encouraged by the recently introduced DCASE Challenge dedicated to the task[2]. Most prior audio captioning methods make use of encoder-decoder models, frequently including sequence modelling modules, such as RNNs [20] or variants like gated recurrent units (GRU) [14], [17] and long short-term memory (LSTM) networks [18], in their encoder to take care of the temporal structure of audio inputs. Most of these works also make use of attention mechanisms to align the audio and text modalities [14], [15], [18], [21]. More recently, following the success of self-attention in V&L models, a small body of work has also started exploring the use of Transformer-based models in audio captioning [19], [23].

Our work is inspired by CNN-RNN architectures developed for image and audio captioning, but focusses on how such approaches can be extended to the music domain for the first time. The only prior works that attempt a similar goal can be found in [24] and [25]. However, the method proposed in [24] fails to generate grammatically correct sentences, while [25] simplifies the task by reframing it as the generation of a sequence of tags. Similarly, prior work on audio-text representation learning also makes use of tags [26], while we stress that our approach focusses on natural language.

C. Transfer Learning for MIR Tasks

Finally, one of the underlying ideas of our study is to leverage pre-training on music data to alleviate the issue of data scarcity in tasks such as captioning where parallel data across modalities is required. Tasks such as music tagging have been shown to successfully extract salient characteristics and learn musically relevant concepts such as mood, genre, era, and emotional content [1], [3]. This suggests that the data representations learnt by networks designed for these tasks can act as useful descriptors for more complex tasks that are highly dependent on similar properties of the input data.

[2]http://dcase.community/challenge2020/task-automatic-audio-captioning

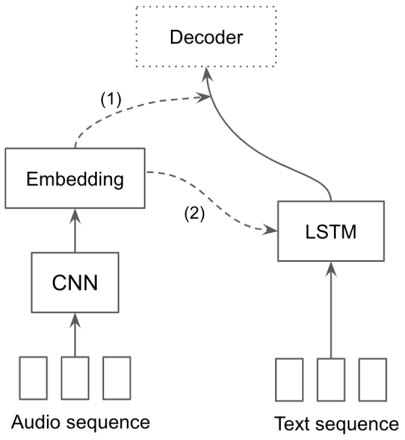

Fig. 2. Illustration of the two fusion strategies: (1) when using late fusion, the fixed-length representation of the whole input audio is only fused before being fed to the decoder; (2) in the early-fusion approach, the audio representation is concatenated to the word embedding and fed to the LSTM at each step.

Previous work has demonstrated the benefit of using these pre-trained features in a transfer learning setting for regression and classification tasks in the music domain [27], [28], but this has not yet been explored in a multimodal setting.

III. Proposed Method

Our model is composed of five main building blocks: a text embedding module, an audio feature extractor, a multimodal encoder, an attention mechanism and a natural language decoder. An overview of the architecture is presented in Fig. 1. In what follows, we offer a detailed description of each component.

A. Text Embedding

The text input is tokenized and encoded through an embedding matrix of dimensions $V \times d$, where V is the vocabulary size and d is the word embedding dimension. This is initialised with 300-dimensional GloVe word embeddings [29], pre-trained on Wikipedia 2014 and Gigaword 5 data, and kept frozen while training our model. Each sentence is thus encoded as a sequence $\mathcal{S} = \{\boldsymbol{w}_1, ..., \boldsymbol{w}_T\}, \boldsymbol{w}_t \in \mathbb{R}^d$, where T is the variable length of the sentence.

B. Audio Feature Extractor

The audio feature extraction module is designed to take a variable-length raw audio input, split it into fixed-length chunks and extract features to be fed to the encoder. In summary, its role is to encode audio into a fixed-length representation or a sequence thereof. Inspired by the established practice of using CNN-based image feature extraction for vision-and-language tasks [7], [9], [12] and further motivated by the development of similar feature extraction networks specifically designed for music audio data [3], [28], we make use of convolutional neural networks for music auto-tagging as our feature extractors.

Among several architectures, we select two in our study, based on their performance on the music tagging task and their musically informed design: *Musicnn* [3], which combines vertical and horizontal convolutional filters to capture both timbral and temporal features, and *Harmonic CNN* [30], which uses trainable filters to model harmonic structures in the audio. A comprehensive comparison of these two networks, among other architectures, is presented in [31]. We discard the classification layers of the networks, using only the front-end and convolutional layers in our feature extraction, and use pre-trained weights[3] from [31].

For each n-second chunk, where n is the input length of the feature extraction network (3s for *musicnn* and 5s for *Harmonic CNN*, both at a sampling rate of 16 kHz), we obtain a feature vector of dimension k, such that the whole input sequence is encoded by a variably sized set of L audio features $A = \{\boldsymbol{a}_1, ..., \boldsymbol{a}_L\}, \boldsymbol{a}_i \in \mathbb{R}^k$. When no attention mechanism (Section III-D) is included in the model, we then obtain track-level features by applying average pooling on the feature maps across the time dimension. This yields $\tilde{\boldsymbol{a}} \in \mathbb{R}^k$, a track-level vector representation of the input audio.

C. Multimodal Encoder

In its baseline variant, our multimodal encoder consists of a 1-layer unidirectional LSTM. Using the index $t = 1, ..., T$ to denote the t-th word in a caption, the t-th hidden state of the encoder LSTM is given by

$$\boldsymbol{h}_t^{enc} = \mathrm{LSTM}\left(\boldsymbol{x}_t^{enc}, \boldsymbol{h}_{t-1}^{enc}, \boldsymbol{m}_{t-1}\right), \tag{1}$$

where $\boldsymbol{x}_t^{enc}$ is the input and $\boldsymbol{m}_t$ is the cell state at step t.

In order to investigate the effect of using a multimodal input, we choose to perform modality fusion at different stages of the encoding procedure. In the *early fusion* approach, embeddings of the audio-text input pair are concatenated and passed to the LSTM as an input at each step:

$$\boldsymbol{x}_t^{enc} = [\boldsymbol{a}, \boldsymbol{w}_t], \tag{2}$$

where $\boldsymbol{a}$ is obtained by passing the track-level audio embedding $\tilde{\boldsymbol{a}}$ generated as described in the previous section through a fully connected layer. In the *late fusion* approach, the encoder LSTM only takes the text as input ($\boldsymbol{x}_t^{enc} = \boldsymbol{w}_t$) and its hidden state $\boldsymbol{h}_t^{enc}$ is then concatenated with $\boldsymbol{a}$ only prior to being passed to the decoder:

$$\boldsymbol{x}_t^{dec} = \left[\boldsymbol{a}, \boldsymbol{h}_{t-1}^{enc}\right]. \tag{3}$$

In this case, similarly to what was done for early fusion, $\boldsymbol{a}$ is obtained by passing the audio embeddings through an additional linear layer, matching the dimension of the encoder hidden state ($\boldsymbol{W}_a\tilde{\boldsymbol{a}} \in \mathbb{R}^{H_{enc}}$). As illustrated in Fig. 2, in the early-fusion approach the audio content is therefore available to the encoder LSTM, while in the late-fusion approach the audio input does not influence the sequence dynamics modelled by the encoder.

[3] https://github.com/minzwon/sota-music-tagging-models

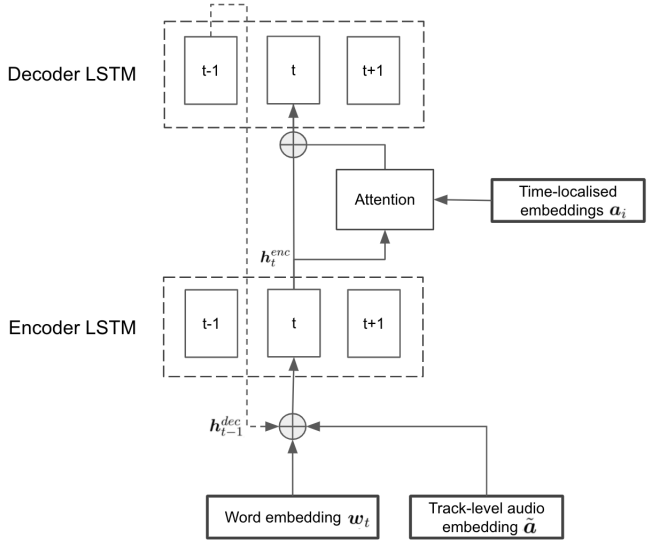

Fig. 3. Overview of the encoder-decoder architecture with the soft attention mechanism. This is similar to the top-down attention in [32].

D. Attention Mechanism

A key challenge in our multimodal encoder is to appropriately *align* and *summarise* the elements in the input audio to which an associated textual description is most sensitive to. We reduce this problem to tackling two distinct aspects: identifying the salient acoustic characteristics emerging from a global description of the input track and localising the temporal segments that are most strongly responsible for the associated description. An optimal approach would provide both and we note that the encoder described in the previous section may fall short of these requirements. In light of the shortcomings of compressing the whole audio input into a static representation, we adopt a soft attention mechanism to dynamically weigh different temporal sections in the audio input. For this, we obtain time-localised audio embeddings $\boldsymbol{a}_i$ by using the audio features produced by the audio encoder over n-second segments prior to average pooling, as described in Section III-B.

The attention model used is a soft top-down attention mechanism, similar to approaches first proposed for the image captioning task in [5], [32]. This mechanism allows the decoder to attend to different sections of the input audio according to attention weights β_{ti}, which are based on an alignment score e_{ti} learned from both the audio features $\boldsymbol{a}_i$ and the hidden state of the encoder $\boldsymbol{h}_t^{enc}$:

$$\begin{aligned} e_{ti} &= \boldsymbol{w}_{att}^{\intercal} \tanh\left(\boldsymbol{W}_a^{att}\boldsymbol{a}_i + \boldsymbol{W}_h^{att}\boldsymbol{h}_t^{enc}\right) \\ \beta_{ti} &= \frac{\exp\left(e_{ti}\right)}{\sum_{k=1}^{N}\exp\left(e_{tk}\right)}, \end{aligned} \tag{4}$$

where $\boldsymbol{w}_{att} \in \mathbb{R}^{H_{dec}}$, $\boldsymbol{W}_a^{att} \in \mathbb{R}^{H_{dec}\times k}$, $\boldsymbol{W}_h^{att} \in \mathbb{R}^{H_{dec}\times H_{enc}}$ are parameters to be learnt and H_{enc} and H_{dec} are the number of hidden units in the encoder and decoder LSTM.

From (4), we obtain the attended vector $\hat{\boldsymbol{a}}_t$ as a weighted sum over the L n-second chunks forming the audio input:

$$\hat{\boldsymbol{a}}_t = \sum_{i=1}^{L}\beta_{ti}\boldsymbol{a}_i. \tag{5}$$

E. Language Model Decoder

The input to the decoder LSTM varies depending on the fusion approach used at the encoding stage and whether the attention module is used. In the case of early fusion, it simply consists of the hidden representations produced by the encoder LSTM, while in the late-fusion case it is given by the audio features concatenated with the hidden states as in (3).

Finally, if the attention module is added to the model architecture, the attended audio feature replaces the mean-pooled vector in (3):

$$\boldsymbol{x}_t^{dec} = \left[\hat{\boldsymbol{a}}_t, \boldsymbol{h}_{t-1}^{enc}\right]. \tag{6}$$

The overall architecture of the LSTM encoder, attention module and decoder is illustrated in Fig. 3.

A linear layer, with parameters $\boldsymbol{W}_d$ and biases and $\boldsymbol{b}_d$, is then appended to the recurrent layers, followed by a softmax nonlinearity, thus acting as a classifier, where each class represents one of the words in the vocabulary. Its output is $Y = \{\boldsymbol{y}_1, ..., \boldsymbol{y}_L\}, \boldsymbol{y}_t \in \mathbb{R}^V$, a sequence of vectors representing the probability distribution over the word vocabulary at each step t in a T-long sequence:

$$P(\boldsymbol{y}_t|\boldsymbol{y}_{t-1}) = \text{softmax}(\boldsymbol{W}_d\boldsymbol{h}_t^{dec} + \boldsymbol{b}_d), \tag{7}$$

such that the joint distribution

$$P(Y|A) = \prod_{t=1}^{T} P(\boldsymbol{y}_t|\boldsymbol{y}_{t-1}) \tag{8}$$

gives the probability distribution over full text sequences Y, given an audio input A.

During training, a caption is generated through greedy decoding by selecting the word with the highest probability at step $t = \{1, ..., L\}$. At inference time, beam search decoding is also explored as a post-processing step. In this case, a fixed number, equal to the beam size b, of most likely hypotheses up to step t is considered to generate the output at $t + 1$.

IV. Experiments

A. Dataset

In our experiments, we use a private production music[4] dataset consisting of audio-caption pairs. We denote such a dataset $\mathcal{D} = \{A_i, c_i\}_{i=1}^{P}$, where P is the number of pairs, A_i an audio track and c_i the corresponding caption. We filter out unsuitable examples by (i) keeping only tracks of length between 30 and 360 seconds; (ii) keeping only captions that contain between 3 and 22 tokens and that are not duplicated across the dataset. Through this cleaning procedure we retain

[4]*Production music* is written and recorded with the aim of being licensed for synchronisation in audio or audiovisual productions such as films and adverts. It is typically organised in catalogues and provided with metadata and descriptions to facilitate discovery of suitable content.

TABLE I
COMPARISON WITH THE BASELINE, EVALUATED ON OUR TEST SET. BOTH VARIANTS OF MUSCAPS, EACH USING A DIFFERENT PRE-TRAINED AUDIO FEATURE EXTRACTOR, PERFORM SIGNIFICANTLY BETTER THAN THE BASELINE MODEL ACROSS ALL METRICS.

MODEL	BLEU_1	BLEU_2	BLEU_3	BLEU_4	METEOR	ROUGE_L	CIDER	SPICE	SPIDER
DCASE BASELINE	13.8	5.8	3.6	2.0	4.3	13.7	19.6	5.6	12.6
MUSCAPS-MUSICNN	34.3	15.0	8.5	5.4	29.3	39.9	33.0	**24.2**	28.6
MUSCAPS-HCNN	**37.3**	**16.4**	**9.3**	**5.9**	**29.6**	**40.6**	**36.9**	23.5	**30.2**

$P = 6,035$ pairs. A random split was used to obtain training, validation and test sets with a 60/20/20 ratio.

The dataset is pre-processed by applying tokenization and encoding each token as a numerical ID. Special tokens, used for infrequent words ($\langle unk \rangle$), padding ($\langle pad \rangle$), start ($\langle sos \rangle$) and end of sentence ($\langle eos \rangle$) are then added to the vocabulary.

B. Experimental Setting

Our caption generation model is trained using *teacher forcing* [33]. At each step t, the $t-1$ element of the target caption is supplied as text input during training, while at test time this is replaced by the previous decoder output. In practice this is achieved by providing the special start token $\langle sos \rangle$ at $t = 0$ and stopping when the predicted output corresponds to the special end token $\langle eos \rangle$ or when the maximum length (22 in our experiments) is reached.

All our captioning models are trained by minimising a Cross Entropy Loss between the probability distribution over the candidate sentence and the ground-truth caption.

We set the number of hidden units in both the encoder (H^{enc}) and the decoder (H^{dec}) to 256. As part of the training procedure we use Adam as an optimiser, with a simple learning rate schedule to linearly reduce the learning rate from an initial value of 10^{-4}. We train for a maximum of 200 epochs with a batch size of 16, making use of dropout with rate 0.25 and early stopping with a patience of 10 epochs for regularisation. Training takes roughly one day on a single RTX 2080 Ti GPU.

C. Evaluation

1) Baseline: We compare our method to the publicly available baseline system used for the audio captioning task of the DCASE 2020 Challenge[5], trained and evaluated on our music captioning dataset. This is a sequence-to-sequence model consisting of three bidirectional GRUs as the encoder, operating on the audio input, and one bidirectional GRU as the decoder, with no alignment mechanism between the two. Since the model is tuned to take as input audio samples between 15s and 30s at a sampling rate of 16 kHz, we randomly select a 30s segment from each of our audio tracks, using the same data splits as in the training and evaluation of our model. To ensure a fair comparison, we empirically verify that training our model with smaller, randomly selected audio chunks does not significantly affect its performance.

2) Captioning: In line with previous work in image and audio captioning, we provide an evaluation of the generated captions using standard automatic metrics. These can be divided into two groups: the first, comprising of BLEU [34], METEOR [35] and ROUGE_L [36], is a set of metrics first proposed to evaluate text generated through machine translation; the second, including CIDEr [37], SPICE [38] and SPIDEr [39], was instead introduced for the evaluation of image captioning models. BLEU is computed from the geometric mean of the n-gram-based precision, using either only unigrams (BLEU_1), or combinations of n-grams up to 4, while METEOR and ROUGE_L are instead computed from an F-score based on matches between the candidate sentence and the reference one. Among the second set of metrics, CIDEr computes the cosine similarity between tf-idf weighted n-grams and is observed to better correlate with human judgement compared to previous metrics [37], SPICE is based on a comparison of semantic similarity of scene graphs parsed from the candidate and reference caption, and SPIDEr is a linear combination of the two. We also note that these metrics have been found to be more accurate when more than one reference sentence is provided [37], while only one caption per example is available in our dataset.

3) Text-based Music Retrieval: We also evaluate our model on music audio retrieval based on a natural language query. The goal of this task is to retrieve the correct audio among a pool of candidate tracks, given its corresponding textual description. We note that our model was not trained on retrieval and its objective is not explicitly designed to optimise text-audio ranking, and therefore does not allow for a direct way to retrieve audio given a caption (or vice versa). However, the model can be easily adapted to perform retrieval by ranking each candidate audio track according to the conditional probability of generating the text query given the audio input, corresponding to the joint probability over the T words in the query sentence, as can be seen in (8). For this experiment, we use the 1207 audio tracks in our test set and 100 test queries from the set of ground-truth captions. Following prior work on image description [4], [6], we provide results for the following retrieval metrics: Recall @ K (higher is better) with $k = \{1, 5, 10\}$, the percentage of correctly retrieved items within the top-K results, and the median rank (lower is better) of the ground-truth items across all queries.

D. Results & Analysis

In this section, we discuss our experimental results for the captioning and retrieval tasks, comparing several variants of our model. We also perform an ablation study (Table II) to tease apart the relative contributions of the main design choices. In order to account for variance due to random initialization, we repeat each experiment in the ablation study three times and provide confidence intervals on our results.

[5]https://github.com/audio-captioning/dcase-2020-baseline/

TABLE II

CAPTIONING PERFORMANCE OF THE ABLATION MODELS. PT: PRE-TRAINING; BS: BEAM SEARCH (WITH BEAM SIZE $b = 5$ FOR EARLY FUSION, $b = 3$ IN ALL OTHER CASES); ATT: ATTENTION. WE HIGHLIGHT IN BOLD SCORES WITHIN TWO STANDARD DEVIATIONS OF THE BEST RESULT FOR EACH METRIC.

FUSION	PT	BS	ATT	$BLEU_1$	$BLEU_2$	$BLEU_3$	$BLEU_4$	METEOR	$ROUGE_L$	CIDER	SPICE	SPIDER
EARLY				35.3	14.4	7.6	4.6	29.3	39.3	29.2	21.8	25.5
EARLY	✓			**37.8**	**16.5**	**9.2**	**5.8**	29.4	40.4	37.6	23.3	**30.4**
	✓	✓		33.4	15.0	8.7	**5.7**	29.3	40.8	**38.8**	23.9	**31.4**
LATE	✓			35.8	15.8	9.0	**5.8**	**29.6**	**41.0**	36.8	**24.3**	30.6
	✓	✓		32.0	14.4	8.5	5.6	29.5	**41.2**	35.9	**24.7**	**30.3**
–	✓		✓	36.7	16.2	**9.2**	**5.8**	**29.7**	**41.0**	37.1	23.2	30.2
–	✓	✓	✓	33.1	14.7	8.6	5.6	**29.5**	**41.1**	36.7	23.5	30.1

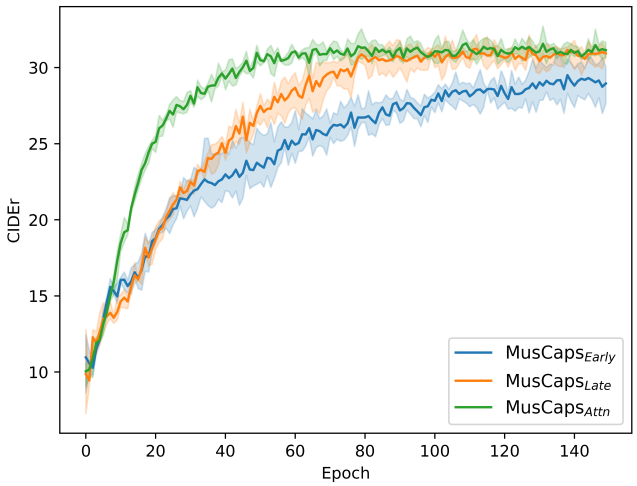

Fig. 4. CIDEr performance of the ablation models on the validation set in the first 150 training epochs. Each curve is averaged across 3 runs and the shaded regions indicate 95% confidence intervals.

For the captioning task, we look at: (1) comparison with the baseline; (2) the effects of large-scale pre-training of the audio feature extractor module; (3) the choice of modality fusion in the encoder and the role of the attention mechanism in providing temporal alignment between the audio embeddings and the representations learnt by the decoder; (4) additionally, we provide a comparison of model performance and caption quality with two different decoding strategies.

1) Comparison with the Baseline: In our first experiment, we focus on the comparison between different audio feature extractors (*musicnn* and *Hcnn*), pre-trained on the MagnaTagATune dataset [40] and compare the two resulting variants of our model, denoted as MusCaps-Musicnn and MusCaps-Hcnn. All experiments were run with the same settings, as detailed in Section IV-B, using early fusion and keeping the pre-trained audio feature extractor module frozen during training. For a fair comparison to the baseline, which does not include an attention mechanism, we only include results of our attention-free model in this section. Our experimental results, shown in Table I, indicate that both variants of our model significantly outperform the DCASE baseline across all metrics, suggesting that the use of a musically informed audio feature extractor brings a considerable performance boost. Illustrative examples of the qualitative differences in output captions between our model and the baseline are also reported in Table IV.

2) Effect of Pre-training: To further support our claim that pre-training the audio feature extractor module on the music auto-tagging task is indeed crucial to obtaining a good model performance, we empirically verify this by running experiments without initialising the audio encoder parameters with pre-trained weights, training them instead from scratch with the rest of the model. Due to memory constraints, we randomly select $20s$ chunks from the full track as our audio input. As hypothesised, our experimental results demonstrate that removing pre-training of the CNN component results in a significant performance drop, as shown in Table II, and slower convergence. This can be attributed to the lack of sufficient training data to successfully accomplish both feature extraction and sentence generation end-to-end, which quickly leads to overfitting, and to a possible difference in the rates at which these two model components generalise. Similarly to what was done when comparing our model with the baseline, we also train MusCaps with pre-training using shorter, randomly selected chunks as the audio input, to ensure that this does not significantly affect performance, and find that only a small decrease in some of the scores is observed in this case.

3) Effect of Fusion & Attention: We next experiment with the two fusion strategies described in Section III-C and the use of a simple attention mechanism for temporal alignment (Section III-D), denoting the respective models as MusCaps$_{Early}$, MusCaps$_{Late}$ and MusCaps$_{Attn}$. Overall, we do not observe a significant difference across these three variants. Remarkably, we find that, even when including the attention mechanism, the model performs similarly. We identify two main explanations for this. Firstly, the attention mechanism used assumes an alignment between audio segments and word tokens at a consistent temporal scale. In music captioning, however, an exact alignment between discrete units in the audio and text modalities may be too strong of an assumption and, if present, is likely to occur at different timescales. Secondly, as shown in Fig. 4, the learning curves of the ablation models reveal that MusCaps$_{Attn}$ converges faster than all other variants. This behaviour is more pronounced for CIDEr, but a similar trend is observed on all metrics. This suggests that the model may be overfitting and ultimately warrants further investigation.

4) Effect of Decoding Strategies: Finally, we compare performance when using greedy decoding and beam search. We find that a beam size of 5 brings the highest performance gain on the captioning metrics for MusCaps$_{Early}$, while in

TABLE III
RETRIEVAL PERFORMANCE. WE HIGHLIGHT IN BOLD SCORES WITHIN TWO STANDARD DEVIATIONS OF THE BEST RESULT FOR EACH METRIC.

MODEL	RECALL@1 ↑	RECALL@5 ↑	RECALL@10 ↑	MEDIAN RANK ↓
DCASE BASELINE	1.2	5.2	10.3	41.8
MUSCAPS W/O PRE-TRAINING	1.7	6.0	9.0	107.2
MUSCAPS$_{Early}$	2.3	12.0	21.3	55.2
MUSCAPS$_{Attn}$	2.7	**15.0**	24.3	45.7
MUSCAPS$_{Late}$	**5.7**	**18.0**	**27.3**	**37.8**

TABLE IV
EXAMPLES OF PREDICTED CAPTIONS ILLUSTRATING THE IMPROVED SENTENCE QUALITY ACHIEVED THROUGH BEAM SEARCH (BEAM SIZE $b = 5$). CASES OF REPETITIONS GENERATED THROUGH GREEDY DECODING THAT DO NOT OCCUR WHEN USING BEAM SEARCH ARE HIGHLIGHTED IN BOLD.

MODEL	EXAMPLE 1	EXAMPLE 2	EXAMPLE 3	EXAMPLE 4
GROUND TRUTH	*Relaxed feelgood $\langle unk \rangle$ guitars*	*Rousing and emotive epic adventure piece*	*Dark brooding orchestral motifs with ghostly ethnic woodwind effects*	*Haunting vocal textures over earthy tribal drums and epic orchestra*
DCASE BASELINE	*Warm and and with acoustic guitar*	*Sweeping and and with strings strings and and*	*Heart-wrenching and and with with*	*Dark and and with to and and and*
MUSCAPS W/ GREEDY DECODING	*Upbeat acoustic guitar **guitar***	*Powerful and dramatic opening to $\langle unk \rangle$ $\langle unk \rangle$ and $\langle unk \rangle$ $\langle unk \rangle$ at $\langle unk \rangle$*	*Dark and **dramatic** introduction to dramatic orchestral theme*	*Epic **epic** theme featuring choir and **choir** and $\langle unk \rangle$ $\langle unk \rangle$*
MUSCAPS W/ BEAM SEARCH	*Upbeat acoustic guitar **tune***	*Powerful and ominous introduction builds to dramatic epic theme*	*Dark and **atmospheric** introduction builds to dramatic theme*	***Haunting** introduction with female vocals and haunting strings*

most other variants performance saturates quickly and even degrades with a beam size larger than 3. This is consistent with observations from previous studies [41], which attribute the performance degradation to overfitting or discrepancy between optimisation of the metrics and of the learning objective.

To better understand how caption quality if affected by the use of different decoding algorithms beyond what is captured by automatic metrics, we extract some statistics from the predicted captions. When greedily decoding the text output, we observe that, although the generated sentences generally capture the audio content well, they often present undesirable features, such as a particularly high occurrence of the $\langle unk \rangle$ token and the frequent repetition of n-grams ("*and strings and strings*", "*rock rock*", "*bass bass and bass*"). When comparing the output of greedy decoding to that obtained through beam search, we observe a significant reduction in both of these. On MusCaps$_{Early}$, beam search decoding with $b = 5$ produces 20% fewer $\langle unk \rangle$ tokens and 37% fewer repetitions. A similar trend is observed in other architectural variants, such as MusCaps$_{Attn}$, although the improvement is more modest (-11% and -18% respectively). As shown in Table IV, this results in more readable captions and we therefore argue that, although not consistently reflected in the evaluation metrics, beam search improves structure and is therefore beneficial to caption generation overall.

5) Retrieval Performance: In Table III we report the audio retrieval performance of the 4 model variants analysed in the ablation study, comparing them to the baseline. The results generally mirror what is observed in the captioning task: pre-training of the audio feature extractor brings the highest performance gain, while changes due to design choices such as fusion strategy and the inclusion of the attention mechanism are less substantial. Surprisingly, unlike in the captioning results, the recall performance of the DCASE baseline model is comparable to MusCaps without pre-training, while its median rank is similar to the top-performing MusCaps variants. Overall, although our experiments demonstrate that the model can be used for text-based retrieval, the scores are not particularly high and it remains unclear whether this is due to genuinely incorrect audio-text matching or to limitations of the ranking procedure. It is worth noting, for example, that the task itself may be partially ill-defined in this context, since multiple textual descriptions may be acceptable for a given audio clip (and vice versa), while our dataset, and therefore our evaluation procedure, does not account for this. A more thorough evaluation protocol may include similarity measures between the retrieved and ground-truth items and user studies.

V. CONCLUSIONS & FUTURE WORK

In this paper, we presented the first music audio captioning model, MusCaps, a simple encoder-decoder network consisting of a multimodal CNN-LSTM encoder with temporal attention and an LSTM decoder. We formulate the problem in a transfer learning setting and highlight the benefit of leveraging music audio feature extractors pre-trained on large-scale, publicly available datasets. Our experiments show that our model can successfully generate descriptive sentences of a music track and outperforms a general-purpose baseline model for audio captioning, even when training on a small audio-text corpus. This confirms that both a musically informed architecture and the use of music tagging as a pretext task are beneficial to music captioning. Although our results are encouraging, our work offers several avenues for further study. Among these, data-related limitations warrant particular attention and we emphasise that additional data collection or data augmentation techniques may be required to obtain several reference captions per example and thus make metric-based evaluation more reliable.

REFERENCES

[1] K. Choi, G. Fazekas, and M. Sandler, "Automatic tagging using deep convolutional neural networks," in *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*, 2016.

[2] T. Kim, J. Lee, and J. Nam, "Sample-level CNN Architectures for Music Auto-tagging Using Raw Waveforms," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018.

[3] J. Pons, O. Nieto, M. Prockup, E. Schmidt, A. Ehmann, and X. Serra, "End-to-end learning for music audio tagging at scale," in *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*, 2018, pp. 637–644.

[4] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan, "Show and Tell: A Neural Image Caption Generator," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.

[5] K. Xu, J. L. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhutdinov, R. S. Zemel, and Y. Bengio, "Show, attend and tell: Neural image caption generation with visual attention," in *Proceedings of the 32nd International Conference on Machine Learning*, 2015, pp. 2048–2057.

[6] A. Karpathy and L. Fei-Fei, "Deep Visual-Semantic Alignments for Generating Image Descriptions," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 4, p. 664–676, 2017.

[7] 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*, vol. 32, 2019, pp. 13–23.

[8] W. Su, X. Zhu, Y. Cao, B. Li, L. Lu, F. Wei, and J. Dai, "Vl-bert: Pre-training of generic visual-linguistic representations," in *International Conference on Learning Representations*, 2019.

[9] L. Harold Li, M. Yatskar, D. Yin, C.-J. Hsieh, and K.-W. Chang, "VisualBERT: A Simple and Performant Baseline for Vision and Language," 2019.

[10] L. Zhou, H. Palangi, L. Zhang, H. Hu, J. Corso, and J. Gao, "Unified vision-language pre-training for image captioning and vqa," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 07, 2020, pp. 13 041–13 049.

[11] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 2019, pp. 4171–4186.

[12] B. Korbar, F. Petroni, R. Girdhar, and L. Torresani, "Video Understanding as Machine Translation," *arXiv preprint*, 6 2020.

[13] C. Sun, A. Myers, C. Vondrick, K. Murphy, and C. Schmid, "VideoBERT: A joint model for video and language representation learning," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2019, pp. 7463–7472.

[14] K. Drossos, S. Adavanne, and T. Virtanen, "Automated Audio Captioning with Recurrent Neural Networks," in *2017 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, 2017, pp. 374–37.

[15] M. Wu, H. Dinkel, and K. Yu, "Audio Caption: Listen and Tell," in *2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019.

[16] E. Çakır, K. Drossos, and T. Virtanen, "Multi-task Regularization Based on Infrequent Classes for Audio Captioning," in *Workshop on Detection and Classification of Acoustic Scenes and Events (DCASE2020)*, 2020.

[17] A. Eren and M. Sert, "Audio Captioning using Gated Recurrent Units," *arXiv preprint*, 2020.

[18] 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)*, 2019, p. 119–132.

[19] A. Tran, K. Drossos, and T. Virtanen, "WaveTransformer: A Novel Architecture for Audio Captioning Based on Learning Temporal and Time-Frequency Information," *arXiv preprint*, 2020.

[20] S. Ikawa and K. Kashino, "Neural Audio Captioning Based on Conditional Sequence-to-Sequence Model," in *Proceedings of the Detection and Classification of Acoustic Scenes and Events 2019 Workshop (DCASE2019)*, 2019, pp. 99–103.

[21] S. Perez-Castanos, J. Naranjo-Alcazar, P. Zuccarello, and M. Cobos, "Listen carefully and tell: an audio captioning system based on residual learning and gammatone audio representation," in *Workshop on Detection and Classification of Acoustic Scenes and Events (DCASE2020)*, 2020, pp. 150–154.

[22] Y. Koizumi, Y. Ohishi, D. Niizumi, D. Takeuchi, and M. Yasuda, "Audio Captioning using Pre-Trained Large-Scale Language Model Guided by Audio-based Similar Caption Retrieval," *arXiv preprint*, 2020.

[23] Y. Koizumi, R. Masumura, K. Nishida, M. Yasuda, and S. Saito, "A Transformer-Based Audio Captioning Model with Keyword Estimation," in *Proc. Interspeech 2020*, 2020, pp. 1977–1981.

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

[25] C. Tian, M. Michael, and H. Di, "Music autotagging as captioning," in *Proceedings of the 1st Workshop on NLP for Music and Audio (NLP4MusA)*, 2020.

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

[27] A. Van Den Oord, S. Dieleman, and B. Schrauwen, "Transfer learning by supervised pre-training for audio-based music classification," in *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*, 2014.

[28] K. Choi, G. Fazekas, M. Sandler, and K. Cho, "Transfer Learning for Music Classification and Regression Tasks," in *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*, 2017.

[29] J. Pennington, R. Socher, and C. D. Manning, "GloVe: Global vectors for word representation," in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2014.

[30] M. Won, S. Chun, O. Nieto, and X. Serrc, "Data-Driven Harmonic Filters for Audio Representation Learning," in *2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 536–540.

[31] M. Won, A. Ferraro, D. Bogdanov, and X. Serra, "Evaluation of CNN-based Automatic Music Tagging Models," in *Proceedings of the 17th Sound and Music Computing Conference*, 2020, pp. 331–337.

[32] P. Anderson, X. He, C. Buehler, D. Teney, M. Johnson, S. Gould, and L. Zhang, "Bottom-Up and Top-Down Attention for Image Captioning and Visual Question Answering," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 6077–6086.

[33] R. J. Williams and D. Zipser, "A learning algorithm for continually running fully recurrent neural networks," *Neural Computation*, vol. 1, no. 2, pp. 270–280, 1989.

[34] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "Bleu: a method for automatic evaluation of machine translation," in *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, 2002.

[35] A. Lavie and A. Agarwal, "METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments," *Proceedings of the Second Workshop on Statistical Machine Translation*, pp. 65–72, 2007.

[36] C. Y. Lin, "Rouge: A package for automatic evaluation of summaries," *Proceedings of the workshop on text summarization branches out (WAS 2004)*, pp. 74–81, 2004.

[37] R. Vedantam, C. L. Zitnick, and D. Parikh, "CIDEr: Consensus-based image description evaluation," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015, pp. 4566–4575.

[38] P. Anderson, B. Fernando, M. Johnson, and S. Gould, "Spice: Semantic propositional image caption evaluation," in *European conference on computer vision*, 2016, pp. 382–398.

[39] S. Liu, Z. Zhu, N. Ye, S. Guadarrama, and K. Murphy, "Improved Image Captioning via Policy Gradient optimization of SPIDEr," in *Proceedings of the IEEE International Conference on Computer Vision*, 2016, pp. 873–881.

[40] 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 International Society for Music Information Retrieval Conference (ISMIR)*, 2009.

[41] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan, "Show and Tell: Lessons Learned from the 2015 MSCOCO Image Captioning Challenge," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 4, pp. 652–663, 4 2017.