Music autotagging as captioning

Tian Cai
The Graduate Center
The City University of New York
New York, 10016, USA
tcai@gradcenter.cuny.edu

Michael I Mandel
Brooklyn College
The City University of New York
New York, 11210, USA
mim@sci.brooklyn.cuny.edu

Di He
The Graduate Center
The City University of New York
New York, 10016, USA
he15810251026@gmail.com

Abstract

Music autotagging has typically been formulated as a multi-label classification problem. This approach assumes that tags associated with a clip of music are an unordered set. With recent success of image and video captioning as well as environmental audio captioning, we we propose formulating music autotagging as a captioning task, which automatically associates tags with a clip of music in the order a human would apply them. Under the formulation of captioning as a sequence-to-sequence problem, previous music autotagging systems can be used as the encoder, extracting a representation of the musical audio. An attention-based decoder is added to learn to predict a sequence of tags describing the given clip. Experiments are conducted on data collected from the MajorMiner game, which includes the order and timing that tags were applied to clips by individual users, and contains 3.95 captions per clip on average.

1 Introduction

Music autotagging has been well studied in music information retrieval at ISMIR. From machine learning to deep learning, the community has witnessed progress over the past decade on this task, with new methods (Choi et al., 2016), new model architectures (Yan et al., 2015; Liu and Yang, 2016; Ibrahim et al., 2020; Wang et al., 2019), and new data sets (Law et al., 2009; Bogdanov et al., 2019). Most studies in content-based autotagging focus on automating the feature extraction to create better representations of music.

What seldom changes, however, is the formulation of the task as a multi-label classification problem (Tsoumakas and Katakis, 2009): treating tags associated with a clip of music as an unordered set. This formulation focuses on correlations between tags, but when a user listens to a clip and provides a sequence of tags, the user expresses his or her listening experience. What is the most "ear-catching" element? What is unexpected? Does this clip feature an instrument or style? These questions cannot be answered under the multi-label classification formulation for music autotagging.

One reason for this formulation is the datasets available for music tagging research, such as MagnaTagATune (Law et al., 2009) and the Million Song Dataset (Bertin-Mahieux et al., 2011). We base the current study on a new analysis of the data collected by the MajorMiner tagging game (Mandel and Ellis, 2008), which includes sequential information. In this game, players supply tags in a particular order and get immediate feedback about the relevance of their tags, further increasing the importance of understanding tag order.

Our switch from multi-label to sequential captions follows similar switches in image and video captioning (Staniute and Šešok, 2019; Chen et al., 2019) and deep learning for acoustic scene and

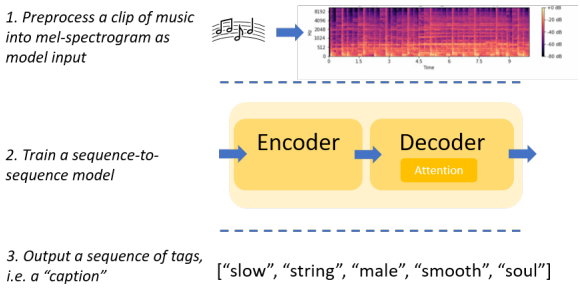


Figure 1: The system uses a sequence-to-sequence model to map mel spectrograms to sequences of tags. The encoder and decoder can be replaced with architectures such as 1D-CNN, 2D-CNN (Choi et al., 2016), MusiCNN (Pons et al., 2017), GRU, LSTM

		Training captions per clip			
		One		Multiple	
Encoder	Decoder	B1	B2	B1	B2
1D CNN	LSTM	9.5	02.2	38.5	38.5
2D CNN	LSTM	10.9	18.3	39.3	48.4
2D CNN	GRU	10.0	19.7	—	—
MusiCNN	LSTM	12.8	23.1	45.8	54.0
MusiCNN	GRU	12.9	23.1	—	—

Table 1: Results on the test set based on the best validation epoch of each model. B1 and B2 stand for BLEU1 and BLEU2, measured in percent.

event captioning. Drossos et al. (2017) presented the first work of audio captioning, focusing on identifying the human-perceived information in a general audio signal and expressing it through text using natural language. The current paper expands this audio captioning approach to the area of music autotagging.

Following these typical captioning models, we use the encoder-decoder architecture with attention mechanism, as shown in Figure 1. We compared three encoders and two decoders, all combined using vanilla attention (Bahdanau et al., 2014). The 2D-CNN for autotagging is from (Choi et al., 2016) and the MusiCNN is from (Pons et al., 2017). We remove the final prediction layer and use the final embedding as the feature fed into the decoder. For the decoder, we compare the most common two choices in image captioning and video captioning: RNN-GRU and LSTM. All models are trained using teacher forcing with cross-entropy of the predicted tag as their loss. It is not attempted in this paper to propose new captioning architecture but to draw awareness of the potential and benefits to re-define music autotagging task leveraging the advancement in NLP.

2 Related Work

Several papers have explored the co-occurrence relationships between tags: Miotto et al. (2010) present one of the early works that explicitly used tag co-occurrence modeled by a Dirichlet mixture. Shao et al. (2018) modeled the tag co-occurrence pattern of a song via Latent Music Semantic Analysis (LMSA). Larochelle et al. (2012); Mandel et al. (2010, 2011a,b) utilized tags alone to build a conditional restricted boltzmann machine and hence demonstrated the value of tag-tag relationships in predicting tags.

Recent works such as (Choi et al., 2018) discussed the effect of tags from the perspective of mislabeling under the theme of multi-label classification.

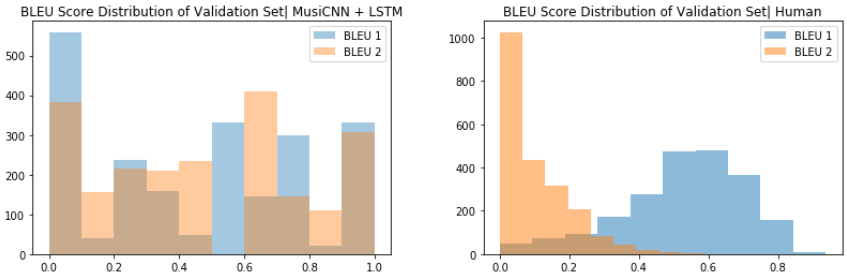
Following (Drossos et al., 2017), Gharib et al. (2018) also first applied domain adaptation techniques as used in NLP to scene classification. Drossos et al. (2019) added language modeling for sound event detection. Ikawa and Kashino (2019) used a captioning model to describe environmental audio. They proposed an extension to the standard sequence-to-sequence model in the captioning task by adding a controllable parameter, specifying the amount of context to provide in the caption.

Multi-label classification is a challenging and important task not only in music information retrieval but also in field such as document categorization, gene function classification and image labeling. In image labeling, Wang et al. (2016) has demonstrated the effectiveness of using RNNs to learn correlation among labels. However, what we propose is not only to learn the label correlations, but also capture user experience with music from the order of tags.

3 Dataset: MajorMiner

Guided by the goal of multi-label classification, most datasets do not retain or make available information about the ordering of tags by users. MajorMiner (Mandel and Ellis, 2008) is a web-based game[1] that naturally collects this information. Participants describe 10-second clips of songs and score points when their descriptions match those of other participants. Users are given the freedom to use any tag they want, but the rules were designed to encourage players to be thorough and the clip length was chosen to make judgments objective

[1] http://majorminer.org/



MusiCNN+LSTM Human

Figure 2: Bleu score distribution of validation data set using (a) the best epoch of MusiCNN + LSTM model and (b) inter-annotator BLEU score, both with multiple captions per clip.

and specific.

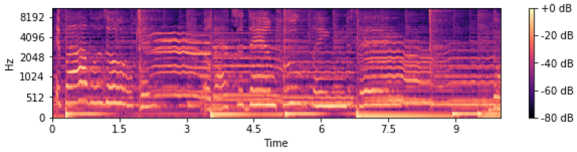
As required by the captioning task, one sample consists of a pair consisting of one audio clip and one corresponding caption provided by one user. The MajorMiner game is designed to collect sequences of tags describing a clip one tag at a time from a user. A sequence of tags is collected, which are ordered by time stamps, and act as a caption. By design, one clip is frequently heard by several different users (this is the only way that any of them may score it). Hence, one clip will receive several captions. This fits into the multi-reference scenario (Papineni et al., 2002) that is often encountered in NLP, for example, in machine translation, where one source sentence has many valid translations into another language.

Caption data is pre-processed through case folding, removal of punctuation, and porter stemming. Sequences of tags, which get validity confirmed, are normalized and canonicalized. The longest tag sequence for a single clip is 30 tags. The total tag vocabulary is 984. Clips are randomly partitioned into train/valid/test set in the ratios of $75\% - 15\% - 10\%$.

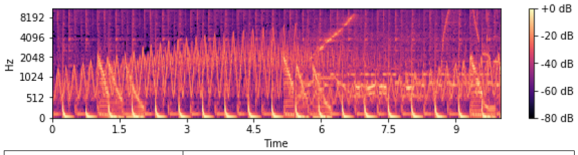
Log mel spectrograms with 96 mel bins are used as input for all models. With sample rate 12,000 Hz, the length of the FFT window is 512 samples (42 ms), and 256 samples between successive frames (21 ms). Each 10-second clip becomes a 469×96 matrix.

4 Experiments and Analysis

A series of experiments is carried out, pairing three encoders, 1D CNN, 2D CNN, MusiCNN, and two decoders, GRU and LSTM, under two settings, multiple captions per clip and one caption per clip, as shown in Table 1. BLEU1 and BLEU2 are used to evaluate each model's ability to capture tag orders.



Clip: (Singer/Album/Track)	Lee Ann Womack/I Hope You Dance/Lonely Too
Reference 1:	'folk', 'country', 'american', 'jones', 'norah jone', 'guitar', 'drum', 'slow', 'love', 'sad'
Reference 2:	'female', 'fool', 'country', 'guitar', 'drum', 'slow', 'quiet', 'lame'
Reference 3:	'country', 'fool', 'female', 'vocal', 'vocal', 'guitar', 'bass', 'drum', 'pedal steel', 'fiddle', 'twangy', 'slow', 'sad', 'boring'
Prediction:	'guitar', 'bass', 'drum'



Clip: (Singer/Album/Track)	Intermix/Future Primitives/Sonic Ritual
Reference 1:	'noise', 'effetcs', 'electronica', 'synth', 'mix', 'dj', 'repetitive', 'creepy'
Reference 2:	'electronic', 'electronic', 'house', 'techno', 'electronica', 'samples', 'mix'
Reference 3:	'electronica', 'techno', 'beat'
Reference 4:	techno', 'electronica', 'dance', 'happy'
Reference 5:	electronic', 'electronica', 'slow', 'synth'
Prediction:	'electronic', 'electronica', 'techno'

Figure 3: Prediction examples from the best epoch of the MusiCNN + LSTM model. Examples are from the validation set.

Other metrics that are standard in music autotagging will be used in future work. The reason to create two caption settings is that, while there are multiple captions per clip, this potentially complicates training a model. Thus, our initial experiments are restricted to a single caption per clip where that caption is selected at random from those applied to that clip. In the multi-caption-per-clip scenario, if a clip received four captions, it is presented in four caption-clip pairs in training, one with each caption. Yet, in the calculation of BLEU scores, all four reference sequences are used for the one clip.

MusiCNN provides both a waveform-based front end and spectrogram-based front end. We use the spectrogram-based front end to make fair comparisons with other encoders. MusiCNN used in this paper has the same configuration as in music autotagging papers (Pons et al., 2017). The 2D CNN used in this paper has six layers of 2D convolution, each followed by batch normalization and 2D max pooling. We also compare a 1D-CNN as an encoder in an attempt to retain more temporal information. This is out of the consideration that some tags appear only at some time steps. In the 1D-CNN, the frequency axis of the mel-spectrogram is taken as the "channel" so that convolutions are computed along the time axis only. The 1D CNN used in this paper has six layers of 1D convolution, each followed by batch normalization. Both 2D and 1D CNN use 256 filters and ReLu at each convolution layer. Both GRU and LSTM decoders have only one layer of RNN followed by two fully connected layers. All models use sparse categorical cross-entropy as loss function and Adam as optimizer.

We create a naïve baseline for the multi-captioning setting by predicting the top k most popular tags for all clips. This evaluates the amount of information our models are learning beyond frequency. Figure 4 shows that this baseline's best performance is to use top three most frequent tags, which achieves a BLEU1 of 0.49 and a BLEU2 of 0.086. This BLEU1 is comparable to our model, but the BLEU2 is much lower. A better baseline for BLEU2 might apply the most common sequence of bigrams. This will be evaluated in future work.

5 Results

The upper bound on the performance of our model is the inter-annotator agreement. Thus, we measure the BLEU score of our ground truth captions

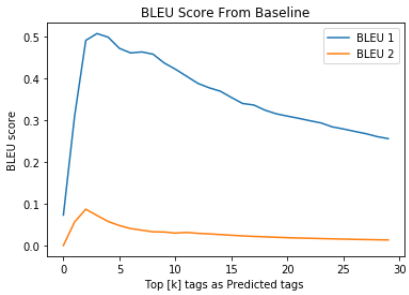


Figure 4: Average validation BLEU score for baseline selecting the k most popular tags for all clips with multiple captions per clip.

with relation to the other captions of the same clip. We find that the average BLEU1 in this case is 0.53 and the average BLEU2 is 0.09. Surprisingly, this is far below that of our best model MusiCNN+LSTM. Beyond the mean, the distribution of inter-annotator BLEU scores is shown in Figure 2(b). Comparing with Figure 2(a), the BLEU score distribution for human tag sequences is smoother and unimodal. This helps to understand why our training/validation BLEU score curve improves very slowly.

To further analyze our results, Figure 2(a) shows a histogram of BLEU1 and BLEU2 scores for the MusiCNN+LSTM model. As can be seen, there is high variability in performance across clips with some having very high scores and some very low scores. This phenomenon may be because tags such as "guitar" are quite frequent and heavily influence the model training, leading to better performance on samples where those tags are relevant. Dealing with imbalances in word frequencies is a common issue in NLP, but we leave it for future work.

Figure 3 shows example annotations, spectrograms, and predictions from the MusiCNN+LSTM model on two example clips. It shows that the model is able to capture general genre information but lacks the nuance of the human annotations.

6 Conclusion and Future Work

The paper demonstrates the promise of formulating music autotagging as a captioning task. It also opens up new possibilities for music autotagging. More advanced NLP techniques such as Transformers (Vaswani et al., 2017; Zhou et al., 2018; Yu et al., 2019) and Masked Language Model pretraining (Devlin et al., 2018) could be utilized to enhance the performance of a language model for

music. There is still more information in the sequence of tags that are applied to a clip that we are not using, such as the temporal locality of tags such as "clap." As pointed out in the recent audio captioning work (Çakır et al., 2020), the distribution of words in captions is a significant challenge. Future work will also address the issue of very frequent yet less useful tags.

References

Dzmitry Bahdanau, Kyunghyun Cho, and Y. Bengio. 2014. Neural machine translation by jointly learning to align and translate. *ArXiv*, 1409.

Thierry Bertin-Mahieux, Daniel P.W. Ellis, Brian Whitman, and Paul Lamere. 2011. The million song dataset. In *Proceedings of the 12th International Conference on Music Information Retrieval (ISMIR 2011)*.

D Bogdanov, M Won, P Tovstogan, A Porter, and X. Serra. 2019. The mtg-jamendo dataset for automatic music tagging.

Emre Çakır, Konstantinos Drossos, and Tuomas Virtanen. 2020. Multi-task regularization based on infrequent classes for audio captioning. *arXiv preprint arXiv:2007.04660*.

Shaoxiang Chen, Ting Yao, and Yu-Gang Jiang. 2019. Deep learning for video captioning: A review. pages 6283–6290.

Keunwoo Choi, George Fazekas, and Mark Sandler. 2016. Automatic tagging using deep convolutional neural networks.

Keunwoo Choi, György Fazekas, Kyunghyun Cho, and Mark Sandler. 2018. The effects of noisy labels on deep convolutional neural networks for music tagging. *IEEE Transactions on Emerging Topics in Computational Intelligence*, 2:139–149.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding.

Konstantinos Drossos, Sharath Adavanne, and Tuomas Virtanen. 2017. Automated audio captioning with recurrent neural networks. In *2017 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, pages 374–378. IEEE.

Konstantinos Drossos, Shayan Gharib, Paul Magron, and Tuomas Virtanen. 2019. Language modelling for sound event detection with teacher forcing and scheduled sampling. *CoRR*, abs/1907.08506.

Shayan Gharib, Konstantinos Drossos, Emre Cakir, Dmitriy Serdyuk, and Tuomas Virtanen. 2018. Unsupervised adversarial domain adaptation for acoustic scene classification. In *Proceedings of the Detection and Classification of Acoustic Scenes and Events 2018 Workshop (DCASE2018)*, pages 138–142.

Karim Ibrahim, Jimena Royo-Letelier, Elena Epure, Geoffroy Peeters, and Gael Richard. 2020. Audio-based auto-tagging with contextual tags for music. pages 16–20.

Shota Ikawa and Kunio Kashino. 2019. 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)*, pages 99–103, New York University, NY, USA.

Hugo Larochelle, Michael Mandel, Razvan Pascanu, and Y. Bengio. 2012. Learning algorithms for the classification restricted boltzmann machine. *The Journal of Machine Learning Research*, 13:643–669.

Edith Law, Kris West, Michael Mandel, Mert Bay, and J. Downie. 2009. Evaluation of algorithms using games: The case of music tagging. pages 387–392.

Jen-Yu Liu and yi-hsuan Yang. 2016. Event localization in music auto-tagging. pages 1048–1057.

Michael Mandel, Douglas Eck, and Y. Bengio. 2010. Learning tags that vary within a song. pages 399–404.

Michael Mandel, Razvan Pascanu, Douglas Eck, Y. Bengio, Luca Aiello, Rossano Schifanella, and Filippo Menczer. 2011a. Contextual tag inference. *ACM Transactions on Multimedia Computing, Communications, and Applications - TOMCCAP*, 7S:1–18.

Michael Mandel, Razvan Pascanu, Hugo Larochelle, and Yoshua Bengio. 2011b. Autotagging music with conditional restricted boltzmann machines.

Michael I. Mandel and Daniel P. W. Ellis. 2008. A web-based game for collecting music metadata. *Journal of New Music Research*, 37(2):151–165.

Riccardo Miotto, Luke Barrington, and Gert Lanckriet. 2010. Improving auto-tagging by modeling semantic co-occurrences. pages 297–302.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation.

Jordi Pons, Oriol Nieto, Matthew Prockup, Erik Schmidt, Andreas Ehmann, and Xavier Serra. 2017. End-to-end learning for music audio tagging at scale.

Xi Shao, Zhiyong Cheng, and Mohan Kankanhalli. 2018. Music auto-tagging based on the unified latent semantic modeling. *Multimedia Tools and Applications*, 78.

Raimonda Staniute and Dmitrij Šešok. 2019. A systematic literature review on image captioning. *Applied Sciences*, 9(10.3390/app9102024):2024.

Grigorios Tsoumakas and Ioannis Katakis. 2009. Multi-label classification: An overview. *International Journal of Data Warehousing and Mining*, 3:1–13.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need.

Jiang Wang, Yi Yang, Junhua Mao, Zhiheng Huang, Chang Huang, and Wei Xu. 2016. Cnn-rnn: A unified framework for multi-label image classification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2285–2294.

Qianqian Wang, Feng Su, and Yuyang Wang. 2019. A hierarchical attentive deep neural network model for semantic music annotation integrating multiple music representations. pages 150–158.

Q. Yan, C. Ding, J. Yin, and Y. Lv. 2015. Improving music auto-tagging with trigger-based context model. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 434–438.

Jun Yu, Jing Li, Zhou Yu, and Qingming Huang. 2019. Multimodal transformer with multi-view visual representation for image captioning. *IEEE Transactions on Circuits and Systems for Video Technology*, PP:1–1.

Luowei Zhou, Yingbo Zhou, Jason Corso, Richard Socher, and Caiming Xiong. 2018. End-to-end dense video captioning with masked transformer. pages 8739–8748.