MediaEval 2020 Emotion and Theme Recognition in Music Task: Loss Function Approaches for Multi-label Music Tagging

Dillon Knox, Timothy Greer, Benjamin Ma, Emily Kuo,
Krishna Somandepalli, Shrikanth Narayanan
Signal Analysis and Interpretation Laboratory (SAIL)
University of Southern California, USA
{dillonkn,timothdg,benjamjm,ekuo,somandep}@usc.edu,shri@ee.usc.edu

ABSTRACT

We present USC SAIL's submission to the 2020 Emotions and Themes in Music challenge: an ensemble-based convolutional neural network (CNN) model trained using various loss functions. In this work, we investigate the effect of different loss functions and resampling strategies on prediction performance, finding that using focal loss improves overall performance on the provided imbalanced, multi-label dataset. Additionally, we report results from varying the receptive field on our base classifier—a CNN-based architecture trained using Mel spectrograms—which also results in better model performance. We conclude that the choice of the loss function is paramount for improving on existing methods in music tagging, particularly in the presence of class imbalance.

1 INTRODUCTION

Content-based automatic music tagging is a challenging task: a robust system must accurately predict multi-label tags (such as mood, theme, or genre) associated with a piece of music, regardless of the frequency of such labels. Great strides in the field have been enabled recently by the release of large high-quality music datasets, like the MTG-Jamendo dataset [2]. In MediaEval's Emotions and Themes in Music challenge, participants are tasked with building models that maximize multi-label tag prediction performance on the autotagging-moodtheme subset of this dataset [3].

Models based on convolutional neural networks (CNNs) are an effective choice for a wide variety of audio-based tasks, including speech recognition [6], acoustic scene classification [12], and music-related tasks, like the 2019 MediaEval Emotions and Themes in Music challenge [8, 13]. Inspired by the success of CNNs in previous music tagging applications, we utilize a VGGish-based short-chunk CNN with residual connections, as implemented in [14], extending this work by experimenting with different loss functions designed to address label imbalance.

In sparse multi-label tasks, loss functions are susceptible to being overwhelmed by the large number of negative labels [4, 7]. We attempt to resolve this issue by using mixup [13], class-aware sampling [11], and novel loss functions. In our experiments, changing loss functions shows the greatest increase in performance, so we focus our reporting on various loss function approaches in this study.

MediaEval'20, December 14-15 2020, Online

We test *focal loss*, originally developed for computer vision models, as a way to train the model to emphasize improving predictions on samples that have lower confidence [9]. We also examine *class-balanced loss*, which increases loss penalties for under-represented classes [5]. Finally, we evaluate the recently-proposed *distribution-balanced loss*, which attempts to overcome the confounding issues of label co-occurrence and negative-class dominance by rebalancing weights with respect to class co-occurrence and incorporating negative-tolerant regularization [15]. Using these different loss functions, we improve on the performance achieved by last year's best model ensemble [8].

2 DATA PREPARATION

We use the provided subset of data from MTG-Jamendo and expand the training set by using instances from the Music4All dataset [10] that exactly match the challenge label set, resulting in an additional 5,666 instances. Additionally, we pretrain the low-level convolutional layers of our models using the Million Song Dataset (MSD) [1], in the manner presented by Won et al. [14].

For feature extraction, we first resample all audio to 16 kHz, and then extract 128-bin Mel spectrograms using a 512-sample FFT, with a window size of 32 ms and 50% window overlap.

3 APPROACH

3.1 Model Architecture

We use a modified short-chunk CNN with residual connections for our model, based on the architecture presented by Won et al. [14], with some modifications. We investigate increasing the receptive field of the CNN, given that our label set is generally composed of high-level music descriptors, such as emotions and themes. The original model has seven convolutional layers; we modify the kernel size of the final two layers to increase the receptive field along the temporal dimension from 3.69 seconds [14] to 4.6 seconds, which results in a better overall performance.

3.2 Loss Functions

We explore three loss functions aimed at increasing average class-wise performance on an imbalanced dataset. Where applicable, we modify the loss functions for multi-label classification.

3.2.1 Focal Loss. We implement a multi-label version of focal loss [9]. The focal loss for a sample y is given as

$$FL(y) = -\sum_{c=1}^{C} \alpha_t (1 - p_c^t)^{\gamma} log(p_c^t) \quad (1)$$

Approach	PR-AUC	ROC-AUC
BCE Loss	0.150	0.766
Focal Loss	0.156	0.778
CB Focal Loss	0.153	0.773
DB Focal Loss	0.153	0.768
Ensemble	**0.161**	**0.781**
VGG-ish-Baseline	0.107	0.725
Popular-Baseline	0.031	0.500

Table 1: Test-set performance of our model trained using various loss functions. BCE stands for "binary cross-entropy"; CB for "class-balanced"; DB for "distribution-balanced."

where p_c^t and α_t are equal to p_c and α, respectively, if c is a label for the sample y. Conversely, p_c^t and α_t are equal to $1 - p_c$ and $1 - \alpha$ if c is not a label for y. In our experiments, we use $\alpha = 0.25$ and $\gamma = 2$, as recommended by [9].

Here, γ suppresses the contribution to loss from the relatively well-classified examples, focusing instead on harder-to-classify examples. In the case of our dataset, where no single label is present in a majority of instances, the negative classes may be easy for the model to learn. Thus, we instead want to focus on the harder cases where a given label is present. α_t provides an additional weight term for positive and negative classes.

3.2.2 Class-Balanced Loss. We also implement a class-balanced version of focal loss [5]. Here, the focal loss weight α_t is replaced by a ratio based on the number of samples containing a given label. Concretely:

$$CBL(y) = -\sum_{c=1}^{C} \frac{1-\beta}{1-\beta^{n_c}}(1-p_c^t)^{\gamma} log(p_c^t) \tag{2}$$

where n_c is the number of samples in the training set in which label c appears and β is a hyperparameter. We set β to 0.995 for our experiments.

3.2.3 Distribution-Balanced Loss. Lastly, we use distribution-balanced loss, which was first presented by Wu et al. [15]:

$$\begin{aligned} DBL(y) = \frac{1}{C}\sum_{c=1}^{C} r_c \Big(& y_c log(1 + e^{-(z_c - v_c)}) \\ + & \frac{1}{\lambda}(1-y_c) log(1 + e^{\lambda(z_c - v_c)}) \Big) \end{aligned} \tag{3}$$

Here, λ is a scale factor for the negative logits, controlling for the preponderance of negative labels, while r_c is a class-specific rebalancing weight that tries to close the gap between the expected number of samples and actual number of samples for a given class after resampling. Wu et al. used a BCE variant of this equation, as shown above; we implement a focal-loss-based function for our study.

4 SUBMISSIONS AND RESULTS

4.1 Submitted Models

We submitted three models to the challenge: the short-chunk CNN model described above using focal loss, an ensemble model which combines multiple CNN-based models trained using the above four different loss functions, and an identical ensemble model that is trained without any outside data (Million Song Dataset or Music4All).[1]

Approach	Head	Middle	Tail
BCE Loss	0.179	0.163	0.101
Focal Loss	0.174	0.171	**0.113**
CB Focal Loss	0.173	0.170	0.104
DB Focal Loss	0.173	0.170	0.105
Ensemble	**0.182**	**0.179**	0.109

Table 2: Class-wise subset performance of various loss functions in terms of PR-AUC.

4.2 Results

We display the PR-AUC and ROC-AUC test-set performance of each loss function approach against provided baselines in Table 1.

We find that the model trained using focal loss produces the best performance both in terms of PR-AUC and ROC-AUC. All variants of focal loss outperform binary cross-entropy, and our final ensemble of averaging the predictions from models trained using the four different loss functions achieves the highest performance.

Additionally, to further investigate the effects of using different loss functions on class-wise performance, we split the label set into *head*, *middle*, and *tail* classes, based on frequency in the training set. Head classes contain over 550 samples, middle classes contain between 200 and 550 samples, and tail classes contain at most 200 samples. The results in terms of PR-AUC are displayed in Table 2.

Indeed, we find that the focal loss-based methods perform better than BCE loss on the less-frequent classes (both middle and tail subsets). In our experiments, this comes with a slight penalty in performance on the head classes, but leads to overall better performance, as well as a better-performing ensemble model.

Lastly, we try class-aware sampling [11] for all models, but observe a performance degradation in each. We further experiment with mixup [16], which has been shown to lead to performance increases on this task [8]. We find that mixup does indeed increase performance using binary cross-entropy, but shows lower performance for focal loss and its variants.

5 CONCLUSION

We present an ensemble-based CNN model trained using focal loss for this year's submission to the 2020 Emotions and Themes in Music challenge. We find that focal loss helps predict labels that do not occur frequently in the dataset, and that emphasizing correct predictions of these labels results in better model performance. We posit that the choice of a loss function is an essential consideration when developing a prediction model for multi-label classification, particularly in music processing. Future work will determine if our approach generalizes to other methods for automatic music tagging.

ACKNOWLEDGMENTS

This study is done at the Center for Computational Media Intelligence at University of Southern California's Signal Analysis and Interpretation Laboratory and supported by research awards from Google and the U.S. Chamber of Commerce Foundation.

[1] Our code can be found at https://github.com/usc-sail/media-eval-2020

REFERENCES

[1] Thierry Bertin-Mahieux, Daniel PW Ellis, Brian Whitman, and Paul Lamere. 2011. The Million Song Dataset. (2011).

[2] Dmitry Bogdanov, Won Minz, Philip Tovstogan, Alastair Porter, and Xavier Serra. 2019. The MTG-Jamendo Dataset for Automatic Music Tagging. In *Machine Learning for Music Discovery Workshop, International Conference on Machine Learning (ICML 2019)*. Long Beach, CA, United States. http://hdl.handle.net/10230/42015

[3] Dmitry Bogdanov, Alastair Porter, Philip Tovstogan, and Minz Won. 2020. MediaEval 2020: Emotion and Theme Recognition in Music Using Jamendo. Proc. of the MediaEval 2020 Workshop, Online, 14-15 December 2020.

[4] Mateusz Buda, Atsuto Maki, and Maciej A Mazurowski. 2018. A Systematic Study of the Class Imbalance Problem in Convolutional Neural Networks. *Neural Networks* 106 (2018), 249–259.

[5] Yin Cui, Menglin Jia, Tsung-Yi Lin, Yang Song, and Serge Belongie. 2019. Class-Balanced Loss Based on Effective Number of Samples. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 9268–9277.

[6] Md Amaan Haque, Abhishek Verma, John Sahaya Rani Alex, and Nithya Venkatesan. 2020. Experimental Evaluation of CNN Architecture for Speech Recognition. In *First International Conference on Sustainable Technologies for Computational Intelligence*, Ashish Kumar Luhach, Janos Arpad Kosa, Ramesh Chandra Poonia, Xiao-Zhi Gao, and Dharm Singh (Eds.). Springer Singapore, Singapore, 507–514.

[7] Grant Van Horn and Pietro Perona. 2017. The Devil is in the Tails: Fine-Grained Classification in the Wild. (2017). arXiv:cs.CV/1709.01450

[8] Khaled Koutini, Shreyan Chowdhury, Verena Haunschmid, Hamid Eghbal-zadeh, and Gerhard Widmer. 2019. Emotion and Theme Recognition in Music with Frequency-Aware RF-Regularized CNNs. Proc. of the MediaEval 2019 Workshop, Sophia Antipolis, France, 27-30 October 2019.

[9] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. 2017. Focal Loss for Dense Object Detection. In *Proceedings of the IEEE International Conference on Computer Vision*. 2980–2988.

[10] I. A. Pegoraro Santana, F. Pinhelli, J. Donini, L. Catharin, R. B. Mangolin, Y. M. e. G. da Costa, V. Delisandra Feltrim, and M. A. Domingues. 2020. Music4All: A New Music Database and Its Applications. In *2020 International Conference on Systems, Signals and Image Processing (IWSSIP)*. 399–404. https://doi.org/10.1109/IWSSIP48289.2020.9145170

[11] Li Shen, Zhouchen Lin, and Qingming Huang. 2016. Relay Backpropagation for Effective Learning of Deep Convolutional Neural Networks. *Computer Vision - ECCV 2016* 9911 (2016), 467–482.

[12] Sangwon Suh, Sooyoung Park, Youngho Jeong, and Taejin Lee. 2020. *Designing Acoustic Scene Classification Models with CNN Variants*. Technical Report. DCASE2020 Challenge.

[13] Manoj Sukhavasi and Sainath Adapa. 2019. Music Theme Recognition Using CNN and Self-Attention. Proc. of the MediaEval 2019 Workshop, Sophia Antipolis, France, 27-30 October 2019.

[14] Minz Won, Andres Ferraro, Dmitry Bogdanov, and Xavier Serra. 2020. Evaluation of CNN-based Automatic Music Tagging Models. 17th Sound and Music Computing Conference (SMC2020), 2020.

[15] Tong Wu, Qingqiu Huang, Ziwei Liu, Yu Wang, and Dahua Lin. 2020. Distribution-Balanced Loss for Multi-Label Classification in Long-Tailed Datasets. (2020). arXiv:cs.CV/2007.09654

[16] Hongyi Zhang, Moustapha Cisse, Yann N. Dauphin, and David Lopez-Paz. 2018. mixup: Beyond Empirical Risk Minimization. In *International Conference on Learning Representations*. https://openreview.net/forum?id=r1Ddp1-Rb