Temperature scaling for reliable uncertainty estimation: Application to automatic music genre classification

Hanna Lukashevich[0000−0002−9041−9475], Sascha Grollmisch[0000−0001−9703−110X], and Jakob Abeßer[0000−0003−4689−7944]

Fraunhofer IDMT, 98693 Ilmenau, Germany
hanna.lukashevich@idmt.fraunhofer.de
http://www.idmt.fraunhofer.de

Abstract. Music genre classification is a crucial task with applications in music recommendation systems and content organization. However, state-of-the-art neural network architectures often struggle to accurately estimate posterior class probabilities. In this paper, we investigate temperature scaling and deep ensembles to improve the reliability of the output predictions. We explore various metrics to find the optimal temperature value for calibration, aiming to align predicted probabilities with observed frequencies. Through experiments on the Free Music Archive small dataset, we demonstrate the effectiveness of temperature scaling, especially in a combination with deep ensembles. Interestingly, we find a discrepancy between optimal temperature values for the validation and test data, highlighting the importance of considering generalization capability and data distribution variations.

Keywords: Automatic music genre classification · Uncertainty estimation · Temperature scaling.

1 Introduction

The digital music boom created a surplus of data, prompting the need for efficient music annotation tools. Genre classification is vital in assigning representative labels like Rock, Pop, or Jazz. It aids in music recommendation systems, playlist generation, and retrieval. However, accurately classifying genres is tough due to subjective, ambiguous, and overlapping styles.

Deep learning-based classifiers have emerged as the leading approach for automatic genre classification [6], [8]. Despite their success, these classifiers often exhibit overconfidence in their predictions, posing challenges in interpreting the output values and compromising the reliability of the classification process [4], [5]. This study focuses on investigating the reliability of confidence values in neural network-based classifiers for music genre classification. We propose the weighted mean absolute error as a novel metric to assess the overconfidence or underconfidence of classifier outputs. Furthermore, we evaluate the effectiveness

of state-of-the-art methods, namely temperature scaling (TS) and deep ensembles, in obtaining more realistic posterior class probabilities. Temperature scaling involves rescaling classifier outputs, while deep ensembles utilize multiple independently trained neural networks to generate the final output. The ultimate objective of this research is to enhance music classification systems by developing more reliable classifiers suitable for real-world applications.

2 Experiments

Dataset We address music genre classification using the Free Music Archive (FMA) small dataset [2], consisting of 8,000 tracks across eight genres: Classical, Hip-Hop, Electronic, Folk, Rock, Experimental, International, and Instrumental. Each track has a duration of 30 seconds. The dataset includes predefined balanced splits for training, validation, and evaluation. Zhao et al. [8] achieved a 56.4% accuracy using a self-supervised pre-training approach with a Swin Transformer. Kostrzewa et al. [6] compared various deep learning network architectures and ensembles. The highest single-model accuracy of 51.6% was achieved with a Convolutional Neural Network (CNN) and ensembles of several CNNs increased the accuracy to 56.4%.

Experimental procedure Our main focus in this study is not to identify the optimal deep learning architecture for automatic music genre classification. Instead, we investigate the posterior class probabilities and explore the impact of temperature scaling and deep ensembles to obtain more realistic confidence outputs.

For our experiments, we consider two distinct network architectures: ResNet [3], which comprises of 420k parameters, and pre-trained OpenL3 embeddings[1] in combination with Multi-Layer Perceptron (MLP) as classifier, referred to as OpenL3-MLP. The ResNet is trained with random image augmentations applied to the mel spectrogram. For OpenL3 extraction, we use the audio branch trained with music data and a 512-dimensional embedding size. ResNet and OpenL3-MLP are trained on the corresponding subset of the FMA small dataset and evaluated on the respective evaluation and test subsets. Posterior class probabilities are computed for each patch and averaged over all patches within each file during the inference stage. We adopt the approach of training the models five times with random initialization to form a deep ensemble [7]. The posterior class probabilities for the ensemble are obtained by calculating the mean over the output of the single models. Temperature scaling is applied by multiplying the logits with a temperature value before passing them through the softmax activation function to obtain adjusted class probabilities.

Reliability diagrams and reliability metrics In the next step, we analyze the disparities between posterior class probabilities and the expected accuracy values for each specific decision. This analysis is based on so-called reliability

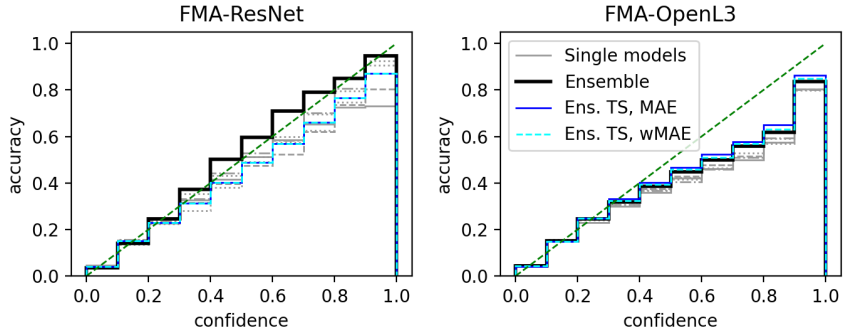

Fig. 1. Reliability diagrams comparing models: individual models (light gray), ensemble of five models (bold black line), and ensemble after temperature scaling (blue and cyan lines). Optimal temperature determined by MAE (blue) or wMAE (cyan).

diagrams, where the accuracy values of the data are distributed into ten confidence buckets. The reliability curves obtained from these diagrams are then compared to the expected accuracy values, and the mean absolute error (MAE) is calculated as a reliability metric. To further enhance the assessment, a novel metric called weighted mean absolute error (wMAE) is proposed. The wMAE incorporates weights determined by the frequencies of winning items within each confidence bucket, providing a more comprehensive evaluation.

Results The current study focuses on the FMA small dataset and presents reliability diagrams for the test subset, as depicted in Figure 1. The diagrams showcase light gray lines representing individual models, while the bold black line represents the ensemble of five models. Additionally, the blue and cyan lines show the ensemble after temperature scaling. The optimal temperature is determined on the basis of MAE (blue line) or wMAE (cyan line). Figure 2 illustrates how MAE changes with different temperature values for both the validation and test subsets and both models. The results indicate that the optimal temperature for the validation subset does not perfectly match that of the test subset, particularly for the OpenL3-MLP architecture. This shows that the validation needs to represent the real-world data as accurate as possible for determining valid temperature values. Furthermore, we compare the absolute differences between optimal temperature values obtained using MAE and wMAE metrics and observed that these differences are minimal for the ResNet architecture, especially for the ensembles, and are slightly higher for the OpenL3-MLP architecture, yet still marginal.

Acknowledgements This research was supported by the H2020 EU project AI4Media—A European Excellence Centre for Media, Society and Democracy—under Grand Agreement 951911 and by the German Research Foundation (AB 675/2-2).

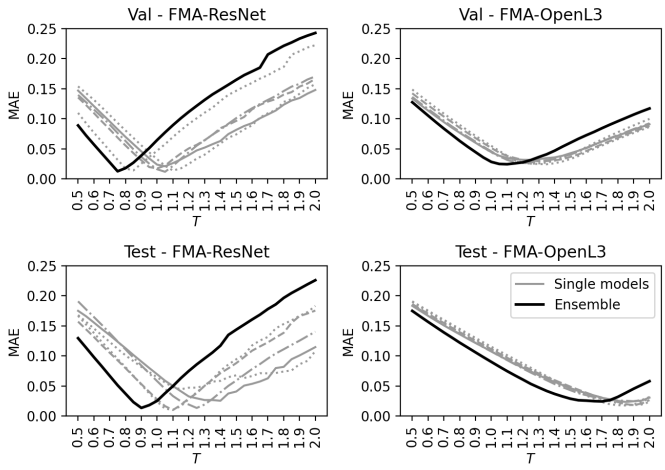

Fig. 2. Mean absolute error (MAE) values dependency on temperature scaling (T)

References

1. Cramer, J., Wu, H.H., Salamon, J., Bello, J.P.: Look, listen, and learn more: Design choices for deep audio embeddings. In: Proc. of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 3852–3856 (2019)
2. Defferrard, M., Benzi, K., Vandergheynst, P., Bresson, X.: FMA: A dataset for music analysis. In: Proc. of the 18th International Society for Music Information Retrieval Conference (2017)
3. Grollmisch, S., Cano, E.: Improving semi-supervised learning for audio classification with fixmatch. Electronics **10**(15), 1807 (2021)
4. Guo, C., Pleiss, G., Sun, Y., Weinberger, K.Q.: On calibration of modern neural networks. In: Proc. of International conference on machine learning (ICML). pp. 1321–1330 (2017)
5. Hein, M., Andriushchenko, M., Bitterwolf, J.: Why ReLU networks yield high-confidence predictions far away from the training data and how to mitigate the problem. In: Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2019). https://doi.org/10.1109/CVPR.2019.00013
6. Kostrzewa, D., Kaminski, P., Brzeski, R.: Music genre classification: Looking for the perfect network. In: Proc. of the 21st International Conference in Computational Science (ICCS). pp. 55–67 (2021)
7. Lakshminarayanan, B., Pritzel, A., Blundell, C.: Simple and scalable predictive uncertainty estimation using deep ensembles. Advances in neural information processing systems **30** (2017)
8. Zhao, H., Zhang, C., Zhu, B., Ma, Z., Zhang, K.: S3T: Self-supervised pre-training with swin transformer for music classification. In: Proc. of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 606–610 (2022)