Interspeech 2022
18-22 September 2022, Incheon, Korea

Uncertainty Calibration for Deep Audio Classifiers

Tong Ye[1,2†], *Shijing Si*[1†], *Jianzong Wang*[1⋆], *Ning Cheng*[1], *Jing Xiao*[1]

[1]Ping An Technology (Shenzhen) Co., Ltd.
[2]University of Science and Technology of China

jzwang@188.com

Abstract

Although deep Neural Networks (DNNs) have achieved tremendous success in audio classification tasks, their uncertainty calibration are still under-explored. A well-calibrated model should be accurate when it is certain about its prediction and indicate high uncertainty when it is likely to be inaccurate. In this work, we investigate the uncertainty calibration for deep audio classifiers. In particular, we empirically study the performance of popular calibration methods: (i) Monte Carlo Dropout, (ii) ensemble, (iii) focal loss, and (iv) spectral-normalized Gaussian process (SNGP), on audio classification datasets. To this end, we evaluate (i–iv) for the tasks of environment sound and music genre classification. Results indicate that uncalibrated deep audio classifiers may be over-confident, and SNGP performs the best and is very efficient on the two datasets of this paper.

Index Terms: Model calibration, Audio classification, Deep neural networks, Gaussian Process, Bayesian Deep Learning

1. Introduction

Modern deep neural networks (DNNs) [1, 2, 3, 4] have been widely utilized in many audio classification tasks such as multimedia search and retrieval, urban sound monitoring, bioacoustic monitoring, and audio captioning. For example, [5] has shown that fully connected multi-layered perceptron (MLP), AlexNet [6], Inception [7], and ResNet [8] significantly outperforms raw features on the Audio Set [9] for Acoustic Event Detection (AED) classification task.

Despite their extraordinary performance, DNNs are often-criticized as being poorly calibrated and prone to be over-confident, thus leading to unsatisfied uncertainty estimation [10, 11, 12]. The process of adapting deep learning's output to be consistent with the actual probability is called uncertainty calibration, and has drawn a growing attention in recent years [13]. In practical applications, miscalibrated probability estimates can be misleading in the sense that the end user of these estimates has an incentive to mistrust (and therefore potentially misuse) them [14].

Many research have been devoted to calibrating deep models in machine learning, computer vision (CV) and natural language processing (NLP). [11] explored with several classical calibration fixes and found that simple post-hoc methods like Temperature Scaling [15] and Histogram Binning [16] are significantly effective for DNNs. [17, 18] proposed to learn linear and non-linear transformation functions to rescale the original output logits respectively. [19] proposed a mutual information maximization-based binning strategy to solve the severe sample-inefficiency issue in Histogram Binning. [20] showed that training models using the standard CE loss with label smoothing, instead of one-hot labels, has a very favourable effect on model calibration. [21] proposed to improve uncertainty calibration by replacing the conventionally used CE loss with the focal loss proposed in [22] when training DNNs.

However, model calibration for audio classifiers is still under-explored. Our goal is not only to understand whether deep audio classifiers are miscalibrated, but also to study what methods can alleviate this problem. As shown in Fig. 1, through an empirical study we find that audio classifiers using ResNet-50 is over-confident. In the topleft plot, the average confidence of all samples is about 0.91, but the accuracy is only 0.84. The topright shows the performance of a classifier calibrated by focal loss. Though the accuracy decreases to around 0.75, the model's confidence is consistent to its accuracy. We argue that it is not trivial to transfer expertise in CV and NLP areas to audio classification, due to the difference between various modalities. we compare various calibration methods on three popular network architectures, Inception, ResNet and DenseNet, and examine their performances on two audio classification datasets.

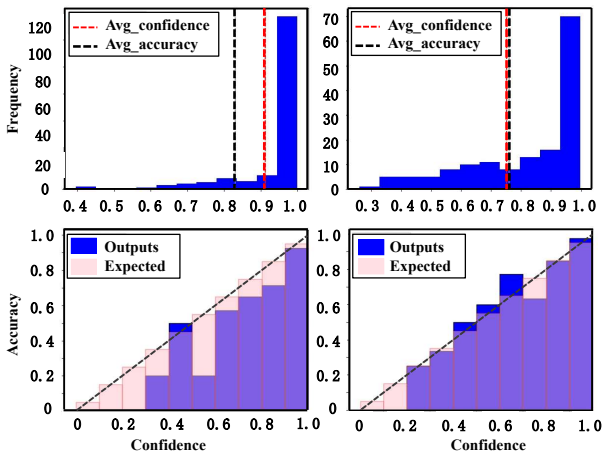

Figure 1: *Confidence histograms (top) and reliability diagrams (bottom) for a base ResNet-50 audio classifier (left) and a calibrated method (right) on the ESC-50 dataset.*

Our contributions can be summarized as follows:

- We verify the existence of miscalibration of deep classifiers for audio datasets, which can raise the community's awareness of this uncertainty calibration problem.
- We empirically examine the performance of various calibration methods for audio classifiers, with SNGP performs the best and is efficient.

† Equal contribution.
⋆Corresponding author: Jianzong Wang, jzwang@188.com

10.21437/Interspeech.2022-11384

2. Background

2.1. Definition

Here we present some basic concepts for model calibration. In this paper, we consider the multi-classification problem for audio data, where we observe an audio (or its features) X and predict a categorical variable $Y \in \{1, 2, \ldots, K\}$. A predictor f as a function that maps every input instance X to a categorical distribution over K labels, represented using a vector $f(X)$ belonging to the $(K-1)$-dimensional simplex $\Delta = \{p \in [0,1]^K | \sum_{y=1}^{K} p_y = 1\}$.

Intuitively, a model f is well-calibrated if its output truthfully quantifies the predictive uncertainty. For example, if we take all data points x for which the model predicts $[f(x)]_y = 0.4$, we expect 40% of them to indeed have the label y. Formally, the model f is said to be calibrated if [23]

$$\forall p \in \Delta : P(Y = y | f(X) = p) = p_y. \tag{1}$$

The most common measure of the degree of miscalibration is the Expected Calibration Error (ECE), which computes the expected disagreement between confidence and accuracy. Typically we first bucket the predictions into m (usually $m = 10$) bins $B_1, \ldots, B_m$ based on their top predicted probability, and then takes the expectation over these buckets. Namely, if we are given a set of n i.i.d. samples $(x_1, y_1), \ldots, (x_n, y_n)$, then we assign each $j \in \{1, \ldots, n\}$ to a bucket B_i based on $\max f(x_j)$. Consequently, we compute in each bucket B_i the

$$\text{confidence}(B_i) = \frac{1}{|B_i|} \sum_{j \in B_i} \max f(x_j), \tag{2}$$

$$\text{accuracy}(B_i) = \frac{1}{|B_i|} \sum_{j \in B_i} 1[y_j \in \arg\max f(x_j)], \tag{3}$$

where the $1[y_j \in \arg\max f(x_j)]$ is an indicator function, taking 1 if $y_j \in \arg\max f(x_j)$ otherwise 0. Finally, ECE is evaluated by taking the expectation over the bins

$$\hat{\text{ECE}} = \sum_{i=1}^{m} \frac{|B_i|}{n} |\text{accuracy}(B_i) - \text{confidence}(B_i)|. \tag{4}$$

2.2. Popular Calibration Methods

Here we summarize popular uncertainty calibration methods widely used in the literature.

Monte Carlo Dropout [24] proposes a way to approximate Bayesian inference by employing dropout and generates a predictive distribution after a number of forward passes. Monte Carlo (MC) Dropout [25, 26] is easy to use, and has zero memory overhead compared to a single model. Unfortunately, it requires multiple forward passes which also result in a substantial obstacle given the prevalence of BERT and other large transformer architectures [27].

Ensemble method casts a set of models under the same architecture with different parameter initialization or other perturbations, encouraging independent member predictions. At test time, the ensemble prediction is the average of soft-max outputs of multiple individually trained models to evaluate the final accuracy. Independent trained identical models create diversity in ensembles due to differences in model initialization and mini-batch orderings [28], which results in different local optimal solutions. [29] proposed the Mix-n-Match calibration strategies which achieves remarkably better data-efficiency and expressive power while provably maintaining the classification accuracy of the original classifier. However, ensemble methods are parameter-efficient but still require multiple forward passes from the model, which consumes larger computing resources than other methods.

Focal loss is originally proposed to address the class imbalance problem in object detection [22]. It reshapes the standard CE loss through weighting loss components of all samples according to how well the model fits them. Therefore it focuses on fitting hard samples and prevents the easy samples from overwhelming the training procedure. [21] verified the effectiveness of focal loss for uncertainty calibration. [30] studied how to recover the true class-posterior probability from the outputs of the focal risk minimizer.

Spectral-normalized Neural Gaussian Process (SNGP) This method employs a Gaussian process, boosting the model's ability to properly quantify the distance of a testing example from the training data manifold and enable a DNN to achieve high-quality uncertainty estimation [31]. Specifically, on top of modern DNNs, it adds a weight normalization step during training and replacing the output layer with a Gaussian Process.

3. Experiments

We implement various calibration methods in Python. Our code is publicly available at a github repository[1].

3.1. Datasets

We conduct extensive experiments on two commonly used datasets: ESC-50 and GTZAN. Details on these two datasets are presented as follows.

ESC-50 is a collection of short environmental recordings available in a unified format (5-second-long clips, 44.1 kHz, single channel, Ogg Vorbis compressed @ 192 kbit/s). It consists of a labeled set of 2000 environmental recordings (50 classes, 40 clips per class). We split the whole dataset into training, validation and testing sets in the ratio of 8:1:1, while keeeping the labels balanced.

GTZAN[32] The GTZAN dataset consists of 1000 music clips each of length 30s. There are 10 distinct genre classes. The music clips are sampled at a rate of 22.5 kHz. There is no official training and validation split of the dataset. Therefore we split the whole dataset into training, validation and testing sets in the ratio of 6:2:2, while keeeping the labels balanced.

3.2. Preprocessing

The input audio signal is re-sampled to 22.5 kHz at the preprocessing step. Re-sampling is applied to reduce dimensionality of the input signal. In addition, every sample is padded with zeros to guarantee uniformity in input data. Each audio is transformed as a 2-dimensional feature map representing frequencies with respect to time [33]. Since mel-spectrograms with different window sizes and hop lengths in each channel yield varied classification performance. The mel-spectrograms were obtained using 128 mel bins and then log scaled. For ESC-50, we use the input of size (128, 250), whereas, for GTZAN, we use the input of size (128, 1500).

3.3. Experimental configuration

A well-calibrated deep learning model should: 1.) produce confidence scores close to its accuracy; and 2.) exhibit higher

[1] https://github.com/shijing001/Unicertainty_calibration_audio_classifiers

uncertainty on inputs far away from training data. To empirically evaluate the performance of calibration methods, our experiments are divided into two parts: in-distribution calibration and out-of-distribution detection. In-distribution calibration measures how well a model's predicted confidence aligns with observed accuracy. Out-of-distribution detection measures the ability of a model to reject OOD inputs.

For the in-distribution calibration, we train classifiers with various methods, and take the output of softmax as predicted probabilities, and then evaluate the ECE scores. For evaluating out-of-distribution detection, we conduct our experiments in the similar approach as introduced by [34]. In these experiments, a neural network is first trained on some ESC-50 data, which represents the in-distribution examples. Out-of-distribution examples are represented by music audio examples from GTZAN that contain classes different from those found in the in-distribution dataset. For each sample in the in-distribution test set, and each out-of-distribution example, a confidence score is produced, which will be used to predict which distribution the samples come from. Finally, several different evaluation metrics are used to measure and compare how well different confidence estimation methods can separate the two distributions.

3.4. Deep Learning models

We employ three popular CNN architectures as the backbone in our experiments: Inception, ResNet and DenseNet.

Inception An Inception Layer [35] is a combination of all the layers namely, 1×1 Convolutional layer, 3×3 Convolutional layer, 5×5 Convolutional layers with their output filter banks concatenated into a single output vector. Here we used Inception-V3 backbone.

ResNet [8] The residual block has two 3×3 convolutional layers with the same number of output channels. Each convolutional layer is followed by a batch normalization layer and a ReLU activation. A skip connection is added which skips these two convolution operations and adds the input directly before the final ReLU activation. Here we used ResNet-50 backbone.

DenseNet[36] Dense Convolutional Network (DenseNet), connects each layer to every other layer in a feed-forward fashion. For each layer, the feature-maps of all preceding layers are used as inputs, and its own feature-maps are used as inputs into all subsequent layers. Traditional convolutional networks with L layers have L connections one between each layer and its subsequent layer a dense network has $L(L+1)/2$ direct connections. We used DenseNet-201 backbone for the experiments.

3.5. Calibration Methods

Here are some details on how we implement popular calibration methods.

Focal loss makes the model focus on hard training examples, paying less attention to easy examples. In this experiment, we set the tuning parameters $\alpha = 0.25$ and $\gamma = 2$ for focal loss.

MC Dropout: We implement the dropout with 10 dropout samples for all CNN layers with probability 0.1.

Ensemble We trained $M = 5$ independent models to predict audio classification scores, using the same architecture, with different initialization values. At test time, the ensemble prediction is the average of soft-max outputs of these M individually trained models to evaluate the final accuracy.

SNGP Following [31], we implement SNGP methods for three network architectures, and employ Laplace approximation for inference.

Table 1: *Accuracy and ECE for In-distribution calibration of the base architecture (no calibration) and four calibration methods (focal loss, MC dropout, Ensemble and SNGP) on two datasets*

Archit.	Method	ESC-50		GTZAN	
		Acc↑	ECE↓	Acc↑	ECE↓
ResNet	+base	0.835	0.106	0.734	0.195
	+focal	0.765	0.049	0.643	0.127
	+Dropout	0.830	0.093	0.764	0.121
	+Ensemble	0.831	0.091	0.738	0.184
	+SNGP	**0.845**	**0.048**	**0.784**	**0.069**
DenseNet	+base	0.905	0.059	0.829	0.077
	+focal	0.886	0.055	0.822	0.057
	+Dropout	0.915	0.053	**0.849**	**0.054**
	+Ensemble	0.895	0.051	0.844	0.071
	+SNGP	**0.930**	**0.034**	0.839	0.075
Inception	+base	0.715	0.138	0.754	0.158
	+focal	0.644	0.106	0.758	**0.054**
	+Dropout	0.720	0.073	0.748	0.121
	+Ensemble	0.728	0.122	0.750	0.149
	+SNGP	**0.785**	**0.054**	**0.779**	0.086

3.6. Evaluation Metrics

ECE:[37] is used to evaluate calibration metric from in-distribution classification. We group all samples into $m = 10$ equally interval bins with respect to their confidence scores, then calculating the expected difference between the accuracy and average confidence, shown in Eq. (4). Smaller ECE scores means better performance.

AUROC: measures the Area Under the Receiver Operating Characteristic curve. It can be interpreted as the probability that a positive example (in-distribution) will have a higher detection score than a negative example (out-of-distribution).

AUPR: measures the Area Under the Precision-Recall (PR) curve. The PR curve is made by plotting $precision = TP/(TP + FP)$ versus $recall = TP/(TP + FN)$. In our tests, AUPR indicates that out-of-distribution examples are used as the positive class. Both AUROC and AUPR are used to evaluate the performance of out-of-distribution detection, and larger values meaning better performance.

3.7. Results and Analysis

Here we present the results and analysis of our experiments.

In-distribution calibration We begin by considering ECE on two datasets: ESC-50 and GTZAN. Table 1 shows in-distribution ECE and accuracy of the three base architectures (Inception-V3, ResNet-50, and DenseNet-201) and four calibration methods. As shown in Table 1, for predictive accuracy, SNGP consistently performs the best for both datasets across three network architectures. For calibration error (ECE), SNGP clearly outperforms the other approaches on ESC-50 dataset and is also competitive on GTZAN dataset. The performance of other methods vary significantly across different architectures and datasets, but are significantly better than the uncalibrated base model in terms of ECE. Among three architectures, DenseNet-201 achieves better accuracy and ECE than ResNet-50 and Inception-V3. This is mainly because it has much more layers (201) than others. Therefore, network architecture also affects the performance of calibration methods.

Figure 2 displays the reliability diagrams of the base ResNet-50 (no calibration) and four calibration methods on the two datasets: ESC-50 (top row) and GTZAN (bottom row). On

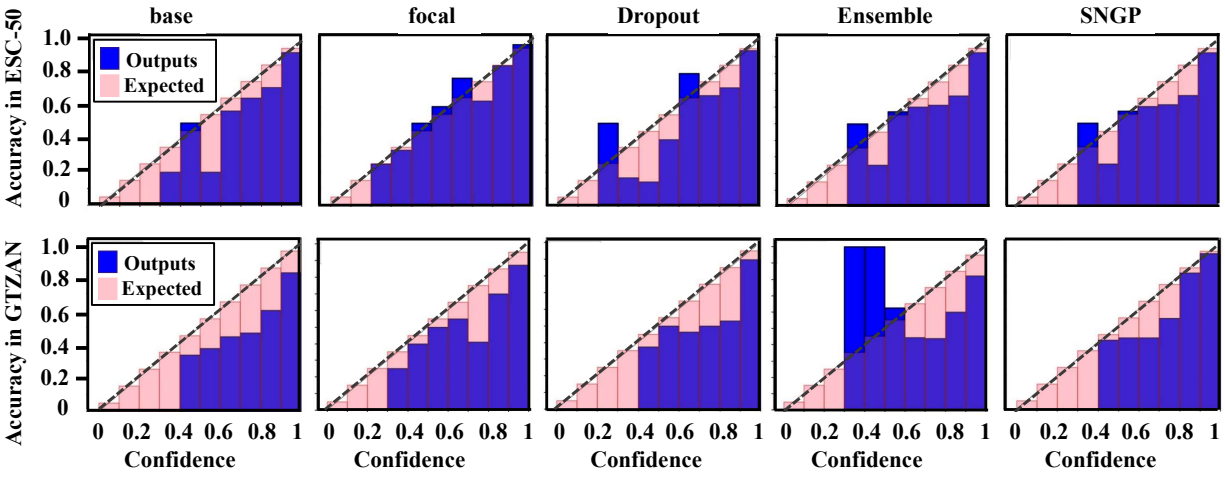

Figure 2: *Reliability diagrams of the ResNet-50 architecture and four calibration methods (focal loss, MC dropout, Ensemble and SNGP) on two datasets: ESC-50 (top row) and GTZAN (bottom row). Less gap between the expected (pink bars) and the output (blue bars) means better performance.*

Table 2: *Performance of the base architecture and four calibration methods (focal loss, MC dropout, Ensemble and SNGP) for out-of-distribution (OOD) detection*

Archit.	Method	AUROC↑	AUPR↑
ResNet	+base	0.828	0.848
	+focal	0.756	0.788
	+Dropout	0.834	0.858
	+Ensemble	0.835	0.853
	+SNGP	**0.849**	**0.881**
DenseNet	+base	0.879	0.894
	+focal	0.885	0.906
	+Dropout	0.878	0.893
	+Ensemble	0.885	0.900
	+SNGP	**0.928**	**0.944**
Inception	+base	0.713	0.763
	+focal	0.643	0.661
	+Dropout	0.724	0.760
	+Ensemble	0.733	0.778
	+SNGP	**0.788**	**0.811**

each plot, less gap between the output bars (blue) and the expected bars (pink) means better performance. From this figure, both focal loss and SNGP yields less gap than other methods.

Out-of-distribution detection To evaluate how suitable the learned confidence estimates are for separating in- and out-of-distribution examples, we conduct out-of-distribution detection and compare the performance of various calibration methods. Table 2 exhibits the performance of three base architectures and four methods on out-of-distribution detection. For this task, SNGP performs the best across all datasets and architectures, followed by the ensemble method. In general, calibration methods perform no worse than the uncalibrated base method, except the focal loss, which performs the worst on ResNet-50 and Inception. This means that focal loss could lead to a classifier bad at discriminating in- and out-of-distribution samples.

Computing Efficiency To compare the efficiency of calibration methods, Figure 3 shows the the number of parameters (by millions) and inference time for one single sample (by milliseconds). This figure takes ResNet-50 as the base model. The ensemble method has largest number of parameters, almost 5 times more than the others. In terms of inference time, ensemble consumes the most, followed by MC dropout. SNGP and focal loss are very efficient, close to the uncalibrated baseline.

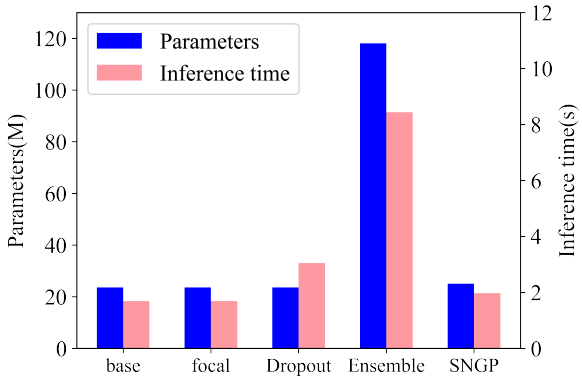

Figure 3: *The number of parameters (by Million) and inference time (by milli-seconds) in five methods.*

In summary, SNGP method performs the best on uncertainty calibration and also very efficient to implement. Through focal loss can produce good calibration, it performs bad at out-of-distribution detection. Ensemble performs well at out-of-distribution detection, but it is not efficient.

4. Conclusion

Audio classification has witnessed rapid improvement as an increasing number of deep learning models are deployed. However, calibration for audio classifiers is still under-explored. In this work, we investigate the performance of calibration methods for deep audio classifiers, verifying the effectiveness of SNGP and ensemble to audio classifiers. This will raise this community's awareness to the uncertainty calibration issue.

5. Acknowledgements

This paper is supported by the Key Research and Development Program of Guangdong Province under grant No.2021B0101400003. Corresponding author is Jianzong Wang from Ping An Technology (Shenzhen) Co., Ltd (jzwang@188.com).

6. References

[1] M. Scarpiniti, D. Comminiello, A. Uncini, and Y.-C. Lee, "Deep recurrent neural networks for audio classification in construction sites," in *2020 28th European Signal Processing Conference (EUSIPCO)*. IEEE, 2021, pp. 810–814.

[2] S. Si, J. Wang, H. Sun, J. Wu, C. Zhang, X. Qu, N. Cheng, L. Chen, and J. Xiao, "Variational information bottleneck for effective low-resource audio classification," in *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, 2021, p. 31.

[3] B. Bahmei, E. Birmingham, and S. Arzanpour, "Cnn-rnn and data augmentation using deep convolutional generative adversarial network for environmental sound classification," *IEEE Signal Processing Letters*, 2022.

[4] S. Si, J. Wang, J. Peng, and J. Xiao, "Towards speaker age estimation with label distribution learning," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 4618–4622.

[5] S. Hershey, S. Chaudhuri, D. P. Ellis, J. F. Gemmeke, A. Jansen, R. C. Moore, M. Plakal, D. Platt, R. A. Saurous, B. Seybold *et al.*, "Cnn architectures for large-scale audio classification," in *ICASSP*. IEEE, 2017, pp. 131–135.

[6] W. Yu, K. Yang, Y. Bai, T. Xiao, H. Yao, and Y. Rui, "Visualizing and comparing alexnet and vgg using deconvolutional layers," in *ICML*, 2016.

[7] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. A. Alemi, "Inception-v4, inception-resnet and the impact of residual connections on learning," in *Thirty-first AAAI conference on artificial intelligence*, 2017.

[8] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *CVPR*, 2016, pp. 770–778.

[9] J. F. Gemmeke, D. P. Ellis, D. Freedman, A. Jansen, W. Lawrence, R. C. Moore, M. Plakal, and M. Ritter, "Audio set: An ontology and human-labeled dataset for audio events," in *ICASSP*. IEEE, 2017, pp. 776–780.

[10] G. Pleiss, M. Raghavan, F. Wu, J. Kleinberg, and K. Q. Weinberger, "On fairness and calibration," *NeurIPS*, vol. 30, 2017.

[11] C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger, "On calibration of modern neural networks," in *ICML*. PMLR, 2017, pp. 1321–1330.

[12] S. Thulasidasan, G. Chennupati, J. A. Bilmes, T. Bhattacharya, and S. Michalak, "On mixup training: Improved calibration and predictive uncertainty for deep neural networks," *NeurIPS*, vol. 32, 2019.

[13] M. Minderer, J. Djolonga, R. Romijnders, F. Hubis, X. Zhai, N. Houlsby, D. Tran, and M. Lucic, "Revisiting the calibration of modern neural networks," *NeurIPS*, vol. 34, 2021.

[14] K. R. M. Fernando and C. P. Tsokos, "Dynamically weighted balanced loss: class imbalanced learning and confidence calibration of deep neural networks," *IEEE Transactions on Neural Networks and Learning Systems*, 2021.

[15] J. Platt *et al.*, "Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods," *Advances in large margin classifiers*, vol. 10, no. 3, pp. 61–74, 1999.

[16] B. Zadrozny and C. Elkan, "Obtaining calibrated probability estimates from decision trees and naive bayesian classifiers," in *ICML*, vol. 1. Citeseer, 2001, pp. 609–616.

[17] M. Kull, M. Perello Nieto, M. Kängsepp, T. Silva Filho, H. Song, and P. Flach, "Beyond temperature scaling: Obtaining well-calibrated multi-class probabilities with dirichlet calibration," *NeurIPS*, vol. 32, 2019.

[18] A. Rahimi, A. Shaban, C.-A. Cheng, R. Hartley, and B. Boots, "Intra order-preserving functions for calibration of multi-class neural networks," *NeurIPS*, vol. 33, pp. 13 456–13 467, 2020.

[19] K. Patel, W. Beluch, B. Yang, M. Pfeiffer, and D. Zhang, "Multi-class uncertainty calibration via mutual information maximization-based binning," in *ICLR*, 2021.

[20] R. Müller, S. Kornblith, and G. E. Hinton, "When does label smoothing help?" *NeurIPS*, vol. 32, 2019.

[21] J. Mukhoti, V. Kulharia, A. Sanyal, S. Golodetz, P. Torr, and P. Dokania, "Calibrating deep neural networks using focal loss," *NeurIPS*, vol. 33, pp. 15 288–15 299, 2020.

[22] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2980–2988.

[23] J. Bröcker, "Reliability, sufficiency, and the decomposition of proper scores," *Quarterly Journal of the Royal Meteorological Society: A journal of the atmospheric sciences, applied meteorology and physical oceanography*, vol. 135, no. 643, pp. 1512–1519, 2009.

[24] Y. Gal and Z. Ghahramani, "Dropout as a bayesian approximation: Representing model uncertainty in deep learning," in *ICML*. PMLR, 2016, pp. 1050–1059.

[25] G. Penha and C. Hauff, "On the calibration and uncertainty of neural learning to rank models for conversational search," in *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, 2021, pp. 160–170.

[26] D. Cohen, B. Mitra, O. Lesota, N. Rekabsaz, and C. Eickhoff, "Not all relevance scores are equal: Efficient uncertainty and calibration modeling for deep retrieval models," in *SIGIR*, 2021, pp. 654–664.

[27] N. Durasov, T. Bagautdinov, P. Baque, and P. Fua, "Masksembles for uncertainty estimation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 13 539–13 548.

[28] B. Lakshminarayanan, A. Pritzel, and C. Blundell, "Simple and scalable predictive uncertainty estimation using deep ensembles," *NeurIPS*, vol. 30, 2017.

[29] J. Zhang, B. Kailkhura, and T. Y.-J. Han, "Mix-n-match: Ensemble and compositional methods for uncertainty calibration in deep learning," in *ICML*. PMLR, 2020, pp. 11 117–11 128.

[30] N. Charoenphakdee, J. Vongkulbhisal, N. Chairatanakul, and M. Sugiyama, "On focal loss for class-posterior probability estimation: A theoretical perspective," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 5202–5211.

[31] J. Liu, Z. Lin, S. Padhy, D. Tran, T. Bedrax Weiss, and B. Lakshminarayanan, "Simple and principled uncertainty estimation with deterministic deep learning via distance awareness," *NeurIPS*, vol. 33, pp. 7498–7512, 2020.

[32] G. Tzanetakis and P. Cook, "Musical genre classification of audio signals," *IEEE Transactions on speech and audio processing*, vol. 10, no. 5, pp. 293–302, 2002.

[33] P. K. Ajmera, D. V. Jadhav, and R. S. Holambe, "Text-independent speaker identification using radon and discrete cosine transforms based features from speech spectrogram," *Pattern Recognition*, vol. 44, no. 10-11, pp. 2749–2759, 2011.

[34] S. Liang, Y. Li, and R. Srikant, "Enhancing the reliability of out-of-distribution image detection in neural networks," in *International Conference on Learning Representations*, 2018.

[35] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 2818–2826.

[36] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 4700–4708.

[37] M. P. Naeini, G. Cooper, and M. Hauskrecht, "Obtaining well calibrated probabilities using bayesian binning," in *Twenty-Ninth AAAI Conference on Artificial Intelligence*, 2015.