Dynamically Mitigating Data Discrepancy with Balanced Focal Loss for Replay Attack Detection

One step towards securing speaker verification systems

Yongqiang Dou, Haocheng Yang, Maolin Yang,
Yanyan Xu, Dengfeng Ke

Beijing Forestry University & Chinese Academy of Sciences

Abstract

It becomes urgent to design effective anti-spoofing algorithms for vulnerable automatic speaker verification systems due to the advancement of high-quality playback devices. Current studies mainly treat anti-spoofing as a binary classification problem between bonafide and spoofed utterances, while lack of indistinguishable samples makes it difficult to train a robust spoofing detector. In this paper, we argue that for antispoofing, it needs more attention for indistinguishable samples over easily-classified ones in the modeling process, to make correct discrimination a top priority. Therefore, to mitigate the data discrepancy between training and inference, we propose to leverage a balanced focal loss function as the training objective to dynamically scale the loss based on the traits of the sample itself. Besides, in the experiments, we select three kinds of features that contain both magnitude-based and phase-based information to form complementary and informative features. Experimental results on the ASVspoof2019 dataset demonstrate the superiority of the proposed methods by comparison between our systems and top-performing ones. Systems trained with the balanced focal loss perform significantly better than conventional cross-entropy loss. With complementary features, our fusion system with only three kinds of features outperforms other systems containing five or more complex single models by 22.5% for min-tDCF and 7% for EER, achieving a min-tDCF and an EER of 0.0124 and 0.55% respectively. Furthermore, we present and discuss the evaluation results on real replay data apart from the simulated ASVspoof2019 data, indicating that research for anti-spoofing still has a long way to go.

Main Objectives

1. To introduce the data discrepancy problem in anti-spoofing.
2. To mitigate the discrepancy by leveraging balanced focal loss as a novel training objective for anti-spoofing, which enables the model to attend more to indistinguishable samples with dynamically scaled loss value.
3. To explore complementary features that are suitable for this task.
4. To show that the performance of current top-performing systems on real data are not as good as on the simulated ASVspoof2019 data, which is unexpected and considered very worthy of discussion.

Cost-sensitive Training — the Balanced Focal Loss

In Fig.1, an example for comparison between a bonafide utterance and its corresponding spoofed utterances with different attack types, illustrates that high-quality attack AA has only subtle differences from the bonafide one, yet taking up a small portion of the data. We term this phenomenon as data discrepancy in anti-spoofing, and propose a method to replace the conventional balanced cross-entropy loss (BCE) with the novel balanced focal loss (BFL) as the training objective. It is worth noting that hard and easy samples do not have strict boundaries, nor will there be a certain attack method as the dividing line, which demonstrates the necessity of making dynamic adjustments with BFL. The MGD-gram feature is used for visualization.

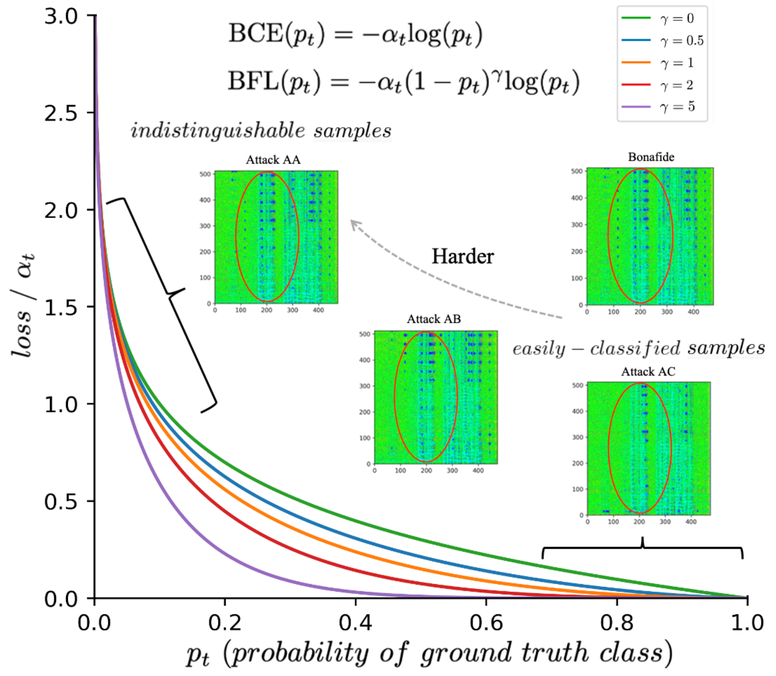

Figure 1: Illustration of the proposed method - Balanced Focal Loss

Results

As shown in Table 1, we experimented with mean-fusion and logistic regression (LR) fusion for models that use three kinds of features all trained with BCE or BFL. The *BFL + LR Fusion* achieves the best performance with only three single models. Better generalization ability of the proposed methods could be seen, i.e. reduced overfitting on the PA Dev Set.

Table 1: Overall Performance of different systems on the ASVspoof2019 PA eval set.

Method	System	# Models	PA Dev Set		PA Eval Set	
			t-DCF$^{min}_{norm}$	EER(%)	t-DCF$^{min}_{norm}$	EER(%)
Official Baseline	LFCC+GMM	-[a]	0.2554	11.96	0.3017	13.54
	CQCC+GMM	-	0.1953	9.87	0.2454	11.04
[1]	Fusion System	6	0.0064	0.24	0.0168	0.66
[2]	Fusion System	5	**0.0030**	**0.13**	0.0160	0.59
This work	BCE + Mean Fusion	3	0.0092	0.40	0.0153	0.62
	BCE + LR Fusion	3	0.0084	0.37	0.0151	0.61
	BFL + Mean Fusion	3	0.0075	0.35	0.0127	0.56
	BFL + LR Fusion	3	0.0077	0.35	**0.0124**	**0.55**

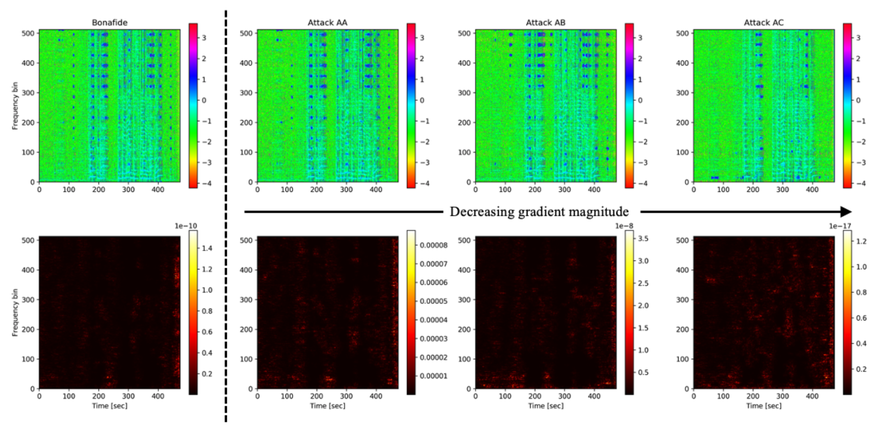

Figure 2: Visualization of original features (Top) vs. saliency maps (Bottom) via backpropagation to better understand critical parts for the network's making classification decisions.

References

[1] Weicheng Cai, Haiwei Wu, Danwei Cai, and Ming Li. The DKU Replay Detection System for the ASVspoof 2019 Challenge: On Data Augmentation, Feature Representation, Classification, and Fusion. In *Proc. Interspeech 2019*, pages 1023–1027, 2019.

[2] Cheng-I Lai, Nanxin Chen, Jesús Villalba, and Najim Dehak. ASSERT: Anti-Spoofing with Squeeze-Excitation and Residual Networks. In *Proc. Interspeech 2019*, pages 1013–1017, 2019.