ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) | 978-1-6654-0540-9/22/$31.00 ©2022 IEEE | DOI: 10.1109/ICASSP43922.2022.9747736

TIME-BALANCED FOCAL LOSS FOR AUDIO EVENT DETECTION

Sangwook Park and Mounya Elhilali

Department of Electrical and Computer Engineering, Johns Hopkins University, Baltimore, MD, USA

ABSTRACT

Sound Event Detection (SED) tackles the challenge of identifying sound events in an audio recording by delimiting both their temporal boundaries as well as sound category. With recent advances in deep learning, current systems are able to leverage availability of large datasets to train sophisticated and highly effective SED models. Nonetheless, sound sources and acoustic characteristics of different classes vary greatly in their prevalence as well as representation in labeled datasets. The challenge with data imbalance in the case of SED stems not only from the representation (number of samples) across classes but also the natural asymmetry in time duration across different events varying from short transient events such as the clacking of dishes to more sustained events such as vacuuming. This variability results in an inherent disproportional representation of effective training samples. To address this compounded imbalance issue, this work proposes a balanced focal learning function that introduces a novel time-sensitive classwise weight. The proposed loss is applied to SED in the context of DCASE2021 challenge, and reports a notable improvement over the baseline, particularly in the case of shorter sound events.

Index Terms— Imbalanced data, focal loss, weighted loss, sound event detection, DCASE challenge

1. INTRODUCTION

Sound Event Detection (SED) is a critical technique in a number of applications spanning video analytics, multimedia tagging, baby monitoring, or other surveillance application [1, 2, 3]. In these applications, the SED model aims to identifying sounds of interest in terms of temporal boundaries as well as sound category which allow to understand the acoustic scene. During training, the SED model learns characteristics of sound events by leveraging labels that indicate the sound class and temporal boundaries for each sound interval. The model is then able to detect interesting sounds whenever they occur. However, this supervised learning causes a laborious work for assigning appropriate label to each audio sample.

Alternatively, synthetic soundscapes can be used to simulate training scenarios in order to alleviate efforts of building fully-labeled datasets. The SED task of Detection and Classification of Acoustic Scenes and Events (DCASE) challenge introduces a training framework incorporating synthetic audios in combination with extensive real audios which are either partially annotated or un-annotated [4, 5]. With statistics of co-occurrence rate for multiple sounds in real soundscapes, the synthetic audios are generated using the *Scaper* approach [6]. This technique is effective in mirroring the natural statistics of soundscapes under study; however it exacerbates a separate issue in terms of imbalance across target classes. Specifically, *Speech* events tend to be over-represented in many audio sets, which results in natural bias of trained models towards those events [4]. While this may be a desirable outcome because of the importance of speech in everyday communication, it is not a desirable performance in a general SED system that identifies likelihoods of a variety of sound events.

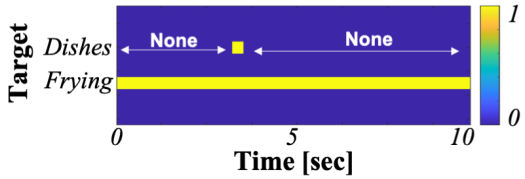

Fig. 1. Label of "*Y2KhtV4TsZ3M_120.000_130.000.wav*" from *Real:validation* of DESED database.

Class imbalance is a common issue that is faced by many classification models that deal with large and realistic datasets. A number of techniques have been proposed to tackle this imbalance problem [7], with most solutions proposing some combination of weighted sampling or adjusted loss that calibrate for the fact that some targets have more data samples compared to others in the training set. In the case of sound events however, this issue is compounded by the fact that sound classes may not only be represented differentially in the training set but that sound events themselves vary naturally in terms of their acoustic span. For instance, Fig 1 shows an audio recording of a kitchen scene where a sound of frying can be heard throughout the 10 seconds clip; while dishes clacking, which is are transient by nature are only present over a short period of time. This over-representation (and under-representation, respectively) of frying and dishes in this example highlights how the network could be biased towards positive or negative predictions for long-events and short-events.

This work was supported by NIH U01AG058532, ONR N00014-19-1-2014, and N00014-19-1-2689.

The current study alters the training objective to account for both the imbalance in sample size across classes as well as variability in event durations by proposing a Time-Balanced Focal Loss function (TBFL). In experiments under the context of DCASE challenge 2021, the proposed loss results a notable improvement over the baseline system.

2. RELATED WORKS

There are a number of approaches that have been proposed to explore data imbalance in detection and classification tasks. Of particular interest to the current work are two techniques that introduced calibration methods in terms of number of samples in binary and multi-class problems. Lin et al. proposed the concept of a Focal Loss (FL) to tackle the class-imbalance in binary classification problems [8]. This work introduced a cost-sensitive learning for binary classification where α-balanced FL is denoted as

$$\begin{aligned} \alpha FL(p,y) = &- \alpha y(1-p)^{\gamma} log(p) \\ &- (1-\alpha)(1-y)p^{\gamma} log(1-p) \end{aligned} \tag{1}$$

where p is a network prediction while y is a true label. γ is a preset parameter as a positive integer. And $\alpha \in (0,1)$ is an weight to adjust balance in between target and non-target. The αFL is the same with Binary Cross-Entropy (BCE) if it is given an equal weight as $\alpha = 0.5$ and $\gamma = 0$. When the network produces a high posterior nearly 1.0 due to over-training toward a major class, the prediction makes a small weight in the loss by means of $(1-p)^{\gamma}$.

For a class-balanced loss, Cui et al. introduced the idea of effective number of samples [9]. Intuitively, additional benefit of a newly added data sample will diminish if many samples are already available for network training. With this assumption, the authors formulate the effective number of samples ε_c as

$$\varepsilon_c = \frac{1-{\beta_c}^{n_c}}{1-\beta_c} \tag{2}$$

where n_c is the number of samples for class c and $\beta_c \in [0,1)$ is a hyper-parameter defined as $\beta_c = \frac{V_c - 1}{V_c}$ with data volume in class c, V_c. Note that the data volume is proportional to the number of data samples that are never overlaid with any others. With an assumption that all data samples are isolated from each other, the data volumes in each class are practically decided to total number of data N for all classes. Then, a Class-Balanced FL (CBFL) is defined as

$$\begin{aligned} &CBFL(p,y) = \\ &-\sum_c \frac{1}{\varepsilon_c}\{(1-p_c)^{\gamma} y_c log(p_c) + p_c^{\gamma}(1-y_c) log(1-p_c)\}, \end{aligned} \tag{3}$$

where y_c is a true label on class c.

3. PROPOSED METHOD

This paper proposes a Time Balanced Focal Loss (TBFL) to extend the idea of class-balanced focal loss for SED. This approach broadens the idea of a focal loss by incorporating both number of samples per class as well as event durations in a time-sensitive loss function. TBFL is denoted as

$$\begin{aligned} TBFL(p,y) = &-\sum_c w_c\{y_c(1-p_c)^{\gamma} log(p_c) \\ &+ (1-y_c)p_c^{\gamma} log(1-p_c)\}, \\ w_c \propto &\frac{1-\beta_c}{1-{\beta_c}^{\lfloor k \times r_c \rfloor}}, \qquad \sum_c w_c = C, \end{aligned} \tag{4}$$

where C is the number of target classes, $r_c = \frac{m_c}{\sum_c m_c}$ is a ratio of the number of frames, m_c in class c, and k is a hyper-parameter to convert from the ratio to the number of samples.

In scenario using synthetic audios for labeled data, data volume can be defined as the number of sound sources used in the generation. Although a simulator is able to generate an infinite number of audios, the synthetic audios result in a high redundancy representation of sound events because they are sampled from a finite sound source. Thus, the quantity of informative sound event would be limited to the number of sound sources with an assumption that the original sounds are different to each other. With this background, data volume is defined as the number of audio clips in the resource as in Table 1, hence resulting in a new definition of β_c. This consideration differs from the way the class-balanced loss in CBFL accounts for variability across classes [9].

In addition to class-sensitive definition of the cost function, we also consider variability across event durations by introducing an exponent of the parameter β_c because loss function during training is calculated on predictions of each time frame. If the number of frames is directly used for the exponent, the factor is ignored due to a very large number of frames in each class. Mathematically, $\lim_{\delta \to \infty} \frac{1-\beta}{1-\beta^{\delta}} = 1-\beta$ because of $\beta < 1$. Instead, the exponent is designed with a ratio of the number of frames across the targets r_c as in Table 1, then it is controlled by a hyper-parameter k. The effect

Table 1. Statistics of target sounds in terms of the number of sound events and average duration

	Resource		Strong labeled set	
class	V_c	[1] Avg. leng. (sec.)	r_c	[2] Avg. dur. (sec.)
A	190	1.58	0.0368	1.45
B	98	7.40	0.0526	4.14
C	88	1.22	0.0322	1.35
Di	109	0.55	0.0388	0.66
Do	136	1.01	0.0334	1.13
E	56	22.04	0.1165	8.67
F	64	21.09	0.1551	9.38
R	68	11.28	0.1002	6.66
S	128	1.26	0.3130	1.54
V	72	27.34	0.1212	9.46

[1] Average length of audio clips
[2] Average duration of sound events

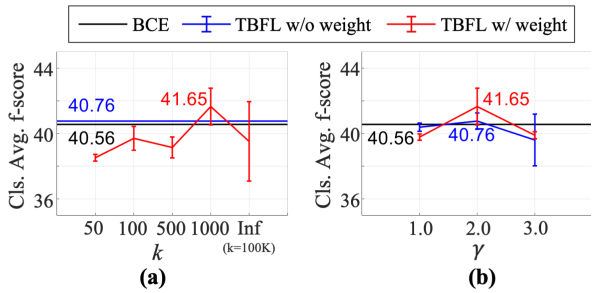

Fig. 2. Class averaging event based f-score on different parameter in (a) k ($\gamma = 2.0$) (b) γ ($k = 1000$ for TBFL)

of k will be explored in experiment.

4. EXPERIMENT

4.1. Database

The proposed method is demonstrated with DESED database that contains 10-sound events as the target: Alarm/bell/ringing (A), Blender (B), Cat (C), Dishes (Di), Dog (Do), Electric shaver/toothbrush (E), Frying (F), Running Water (R), Speech (S), and Vacuum Cleaner (V) [10]. The training set is composed of strong labeled, weakly labeled, and unlabeled sets. The strong labeled set consists of 10,000 synthetic audios produced by the *Scaper* while the other two subsets are composed of real recordings. Table 1 shows statistics of targets in the resource for generation and strong labeled set used in network training. Note that sound sources corresponding to long events are much longer than strong labeled audios because they are cropped during the generation of the 10 sec audio segments. The assessment is performed on the validation set.

4.2. Experimental setting

For performance comparison, the baseline of the SED task in DCASE2021 challenge, which is developed with Binary Cross-Entropy (BCE) loss, is considered as a counterpart to the proposed method [5]. In the baseline system, the loss function is composed of a classification loss for labeled data and a regularization for all data defined as:

$$
\begin{aligned}
&L = L^{cls}_{bce}(p, y) + \lambda L^{reg}(p, \hat{p}), \\
&L^{cls}_{bce}(p, y) = BCE_{x \in S}(p_x, y^s_x) + BCE_{x \in W}(E_m[p_x], y^w_x), \\
&L^{reg}(p, y) = MSE_{x \in S,W,U}(p_x, \hat{p}_x),
\end{aligned} \tag{5}
$$

where S, W and U are set of strong labeled, weakly labeled, and unlabeled set, respectively. $BCE(p, y)$ and Mean Squared Error, $MSE(p, y)$ result a scalar value by averaging over the classes and frames. y^s_x and y^w_x is strong label and weakly label for input x, respectively. p_x and $\hat{p}_x$ are network prediction for the x by student and teacher network, respectively. E_m is an expectation operator over the frame.

In the proposed method, the classification BCE loss is replaced with TBFL as $L^{cls}_{tbfl}(p, y) = TBFL_{x \in S}(p_x, y^s_x) +$ $TBFL_{x \in W}(E_m[p_x], y^w_x)$ where $TBFL(p, y)$ yields an average value over the classes and frames like the $BCE(p, y)$. Other than this adjustment of the loss function, all settings, including the structure and training parameters (optimizer, batch size, and learning rate) are exactly set the same in both methods.

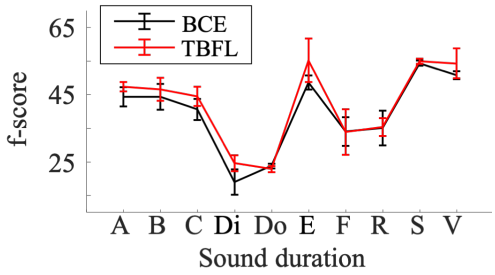

Fig. 3. Classwise f-scores of SED models trained with BCE and TBFL loss, respectively

4.3. Evaluation

In post processing composed of thresholding and smoothing, multiple thresholds (0.01 to 0.99 with 0.02 step) are applied to find best performance while a median filter (with 0.45 sec length) is used for smoothing. Assessment is performed with event based f-score and Polyphonic Sound Event Detection Score (PSDS) [11]. For f-score, a detected interval will be decided to true positive if it has matched to truth in time boundaries within 200 ms margin as well as sound class. As a moderate metric compared to the f-score, two different criteria are considered for PSDS. PSDS1 is focusing on time accuracy of the intervals, on the other hand, PSDS2 is interested in classification among the targets rather than time accuracy. All experiments are performed at least 3 times, and the results are summarized to mean and standard deviation over the iteration.

4.4. Results

4.4.1. Parameter optimization

To explore the effect of hyper-parameter k, the network is trained using TBFL with different values of k. Also, no weighted TBFL ($w_c = 1.0$) is considered in this test. With event based f-score, the results are summarized in Fig. 2(a). Note that γ is set to 2.0 in this test. When k goes to ∞, the weight is determined by the data volume of resources only. However, it is important to note that training loss is calculated on the synthetic training set not the original resource. In addition, the data distribution in the training and resource datasets do vary, and the result with infinite k reflects this effect. As noted in the results, the case of $k = 1000$ shows the best f-score, and outperforms the BCE and no weighted TBFL as well. For a fixed $k = 1000$, Fig. 2(b) explores the effect of the parameter γ; and shows that the best performance is achieved with $k = 1000, \gamma = 2.0$. These parameters are used for the next investigations.

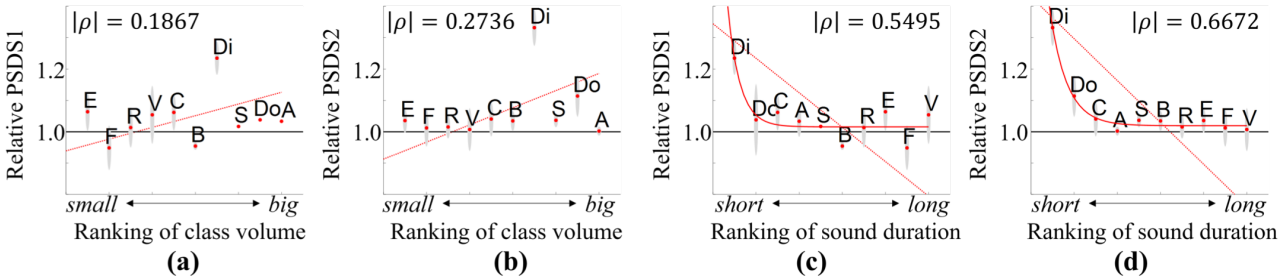

Fig. 4. Relative improvement compared to BCE loss: (a) PSDS1 sorted by data volume, (b) PSDS2 sorted by data volume, (c) PSDS1 sorted by event duration, (d) PSDS2 sorted by event duration

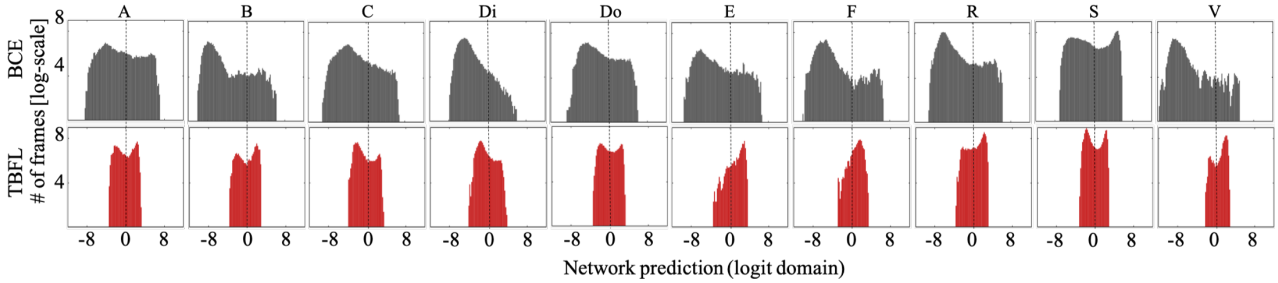

Fig. 5. Histogram of network prediction on each class: Horizontal axis represents the predictions in logit domain (inverse sigmoid) and the number of frames is onto the vertical axis.

4.4.2. Performance comparison

In classwise f-score, the TBFL leads the improvement in most of classes (Fig. 3). To investigate the improvement with a moderate metric, relative PSDS calculated by $\frac{PSDS_{TBFL}}{PSDS_{BCE}}$ are sorted in two perspectives: data volume (Fig. 4(a-b)) and event duration (Fig. 4(c-d)). In the figure, the gray region represents standard deviation over the iteration while the mean is denoted as red point for each class. From the results, the improvement seems to be related to event duration rather than data volume, and TBFL is more effective for short-length events such as *Dishes* and *Dog*. In case of long-events, time accuracy (as in PSDS1) has been improved in *Electric shaver/toothbrush* and *Vacuum cleaner* with a comparable PSDS2 with the baseline. This is consistent with the improvement of f-score in those targets.

4.4.3. Effect on posterior distribution

To further investigate the improvement, 10-histograms of the prediction on each class are built with weakly labeled audio clips (10 sec length) annotated to each class on pretrained networks (Fig. 5). Ideally, two clusters: one is in positive region for target frames and the other is in negative region for non-target frames can be found in the histogram because event duration is typically less than whole audio length. For long events such as *Electric shaver/toothbrush, Frying* and *Vacuum cleaner*, their distributions are likely to be biased toward positive side because their duration is generally longer than a half of whole length. Similarly, the distribution of short sounds might be biased toward negative side. In BCE loss, the distributions are so different to each other. In cases of *Blender*, *Electric shaver/toothbrush* and *Vacuum cleaner*, their distributions seem to be biased toward negative side. Besides, the distribution of *Dishes* show a single cluster in negative region. On the other hand, the proposed loss shows pretty similar across the targets in the two categories depending on duration. For *Alarm/bell/ringing* and *Cat*, the distributions show two clusters for target and non-target. For *Blender*, *Electric shaver/toothbrush*, and *Vacuum cleaner*, it shows a cluster in positive region. These effect might be related to the improvement of those sounds in f-score (Fig. 3). For *Dishes* and *Dog*, both posteriors have been changed, however *Dishes* has been significantly improved in f-score while the other is comparable with the BCE loss. Because of trade-off in between precision and recall, the TBFL shows a comparable f-score in *Dog* while it makes an improvement in both precision and recall for *Dishes*.

5. CONCLUSIONS

Imbalance in training set is a general issue in machine learning. While sample size across data classes is a universal challenge across classification problems, SED tasks face the additional concern of difference in temporal coverage for each sound class. The current study proposed a time-sensitive loss function that introduces novel weights to consider both class and time coverage. Relative to a cross-entropy loss that is agnostic to class variability, the proposed TBFL results in notable improvement in multiple event detection metrics using f-score as well as PSDS. The approach shows results in more precise network predictions across classes but reveals marked benefits for shorter events such as *Dishes* and *Dog* which tend to be under-represented in baseline training methods.

6. REFERENCES

[1] Yizhar Lavner, Rami Cohen, Dima Ruinskiy, and Hans Ijzerman, "Baby cry detection in domestic environment using deep learning," in *IEEE International Conference on the Science of Electrical Engineering,*. 2017, IEEE.

[2] Sangwook Park, Younglo Lee, David K Han, and Hanseok Ko, "Subspace projection cepstral coefficients for noise robust acoustic event recognition," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, New Orleans, LA, USA, 2017, pp. 761–765.

[3] Kashif Ahmad and Nicola Conci, "How deep features have improved event recognition in multimedia: A survey," *ACM Trans. on Multimedia Compu., Comm. and App.*, vol. 15, no. 2, 2019.

[4] Nicolas Turpault, Romain Serizel, Justin Salamon, and Ankit Parag Shah, "Sound Event Detection in Domestic Environments with Weakly Labeled Data and Soundscape Synthesis," 2019.

[5] Nicolas Turpault and Romain Serizel, "Training Sound Event Detection On A Heterogeneous Dataset," in *DCASE workshop*, 2020.

[6] Justin Salamon, Duncan MacConnell, Mark Cartwright, Peter Li, and Juan Pablo Bello, "Scaper: A library for soundscape synthesis and augmentation," in *2017 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*. 10 2017, vol. 2017-Octob, pp. 344–348, IEEE.

[7] Justin M. Johnson and Taghi M. Khoshgoftaar, "Survey on deep learning with class imbalance," *Journal of Big Data*, vol. 6, no. 27, 2019.

[8] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollar, "Focal Loss for Dense Object Detection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 42, no. 2, pp. 318–327, 2 2020.

[9] Yin Cui, Menglin Jia, Tsung-Yi Lin, Yang Song, and Serge Belongie, "Class-Balanced Loss Based on Effective Number of Samples," in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 6 2019, vol. 2019-June, pp. 9260–9269, IEEE.

[10] Romain Serizel, Nicolas Turpault, Ankit Shah, and Justin Salamon, "Sound Event Detection in Synthetic Domestic Environments," in *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 5 2020, pp. 86–90, IEEE.

[11] Emre Cakir, Toni Heittola, Heikki Huttunen, and Tuomas Virtanen, "Polyphonic sound event detection using multi label deep neural networks," in *2015 International Joint Conference on Neural Networks (IJCNN)*, 2015, pp. 1–7.