DCASE 2022 TASK 4: ACOUSTIC EVENT DETECTION INCLUDING ACTIVITY-IMBALANCE WEIGHTING AND MAX-BASED WEAK PREDICTIONS

Technical Report

Carlos Castorena[1]*, *Maximo Cobos*[1], *Francesc J. Ferri*[1] *and Fabio Antonacci*[2]

[1] Universitat de Valencia {carlos.castorena, maximo.cobos, francesc.ferri}@uv.es
[2] Politecnico di Milano, fabio.antonacci@polimi.it

ABSTRACT

This report proposes two activity-motivated strategies for audio event detection implemented over the DCASE 2022 Task 4 baseline system. The first one affects the way that weak predictions are generated from the system's frame-wise activity predictions, leading to a more robust and reasonable loss computation for weakly-labeled examples. The second one is based on a weighted binary cross-entropy loss driven by the resulting event activity imbalance in each training mini-batch. The combination of both strategies led to some performance improvements with little effort in system design and training.

***Index Terms*—** Audio Event Detection, Data Imbalance, Weighted Cross-Entropy Loss, Weak Labels

1. INTRODUCTION

The task of detecting sound events in domestic environments of the "Detection and Classification of Acoustic Scenes and Events 2022 (DCASE2022)" challenge, consists in designing systems that not only predict the presence or absence of the 10 domestic events considered, but also provide their temporal bounds. For training, there are 10,000 synthetic audios with a duration of 10 seconds each, which have their respective labeling with time marks (strong labeling) and, additionally, 1,578 audios that only have their weak label (only the presence of the sound is identified). There is also a set of 14,412 unlabeled audios [1].

Most systems participating in the challenge are based on deep neural networks, such as recurrent convolutional neural networks (CRNN), where multiple convolutional layers work as feature extractors and recurrent layers analyze output feature maps sequentially. The baseline for the year 2022 is based on a CRNN, taking the Mel spectrogram as input, and providing as output both strong and weak predictions. The binary cross entropy loss is used for those samples that have strong or weak labels, while the mean squared error is used for the predictions that do not have any labeling [2] following the baseline's proposed teacher-student model.

It is usual that there is a significant difference between the number of frames with activation and those that do not have any active event. This activity imbalance problem usually harms the performance of the neural network. Using loss functions that counteract the negative effects of activity imbalance issues usually lead to better overall results [3].

*We thank *Generalitat Valenciana* and the *Santiago Grisolía* program for financing this work (GRISOLIAP/2021/060, CPI-21-232). This work also received funding from Grant RTI2018-097045-B-C21 funded by MCIN/AEI/10.13039/501100011033 and "ERDF A way of making Europe". Mobility grant PRX21/00174 from Ministerio de Universidades of Spain funded also partially this work.

2. METHODOLOGY

In this work, two main modifications to the baseline are proposed. The first deals with the way in which weak predictions are calculated, while the second aims at including weighted loss mechanisms to reduce the effects of the imbalance between active and non-active frames. These modifications are described below.

2.1. Weak Prediction

The baseline model's output correponds to the temporal sequence of class activity likelihoods where, for each time frame, an activity probability is provided for each class, referred to as *strong predictions*. Note, however, that the model uses both strongly and weakly-labeled examples to update weights during training. For strongly-labeled audio examples, a frame-by-frame class activity sequence is available, which can be directly compared to the model's output. For weakly-labeled examples, only a binary label is available indicating class event presence along the whole example duration. For such examples, a *weak prediction* is generated from the model's frame-wise output by taking the average across all frames. Note that a weak-label can also be obtained from the strongly-labeled examples by generating a ground-truth of "1" whenever there is some frame with activity for any of the classes. Finally, unlabeled examples are treated following the baseline's student-teacher semi-supervised approach.

One improvement direction is that regarding the way that weak predictions are treated in the baseline model. Currently, the weak prediction when the attention layer is not used is computed from the average of the strong predictions (Avg-Weak), so the binary cross-entropy (bce) computed as a cost function will push the predictions from all the frames of the same event towards 1, contradicting the natural frame-by-frame output of the model. For this reason, it is proposed to use the maximum value of the strong predictions as weak prediction (Max-Weak), replacing the average. This way, the corresponding weak predictions are expected to be 1 when there is at least one active frame for a given event class. Figure 1 shows the difference between both calculations. Event A is only active over a small number of frames, leading to a small Avg-Weak prediction and a high bce loss, while the Max-Weak prediction reflects better the fact that the event is actually present as deduced from the strong prediction. When an event is active over all frames (Event B) there is almost no different between Avg-Weak and Max-Weak. Finally,

Event C is not active but the strong prediction is incorrectly indicating class activity, which is properly translated to a significant loss value when Max-Weak is used, but not so for Avg-Weak.

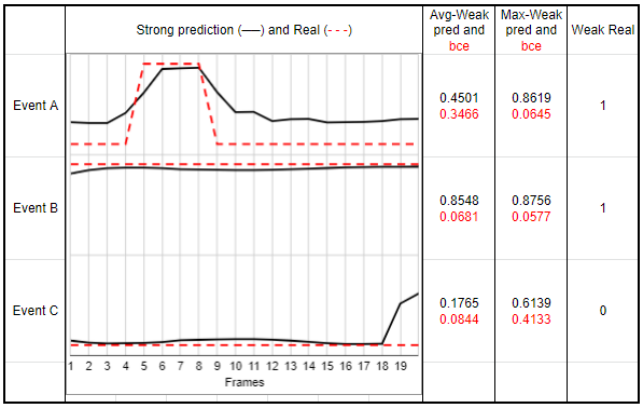

Figure 1: Weak prediction representation

2.2. Weighted Binary Cross Entropy

To balance the data, a weighted binary cross entropy ($wbce$) is used, applying a factor w which weights the error differently and proportionally as a function of the ground-truth label. The loss function will be given by:

$$wbce(y,x) = \frac{1}{S}\sum_{s=1}^{S}((1-w)\cdot y_s \cdot \log(x_s) + w\cdot(1-y_s)\cdot\log(1-x_s)), \quad (1)$$

where S is the number of samples in the batch and y_s and x_s are the ground-truth labels and corresponding predictions for sample s, respectively. Note that both x_s and y_s are either vectors in $\mathbb{R}^K$ when weak labels are used or matrices in $\mathbb{R}^{T\times K}$ when dealing with strong labels, where K is the number of event classes and T the number of considered time frames. The weight w is selected as the ratio of active events in the batch with respect to the total number of events in such batch. When y_s is a weak label, w is calculated by:

$$w = \frac{\sum_{s=1}^{S}\left(\frac{\sum_{k=1}^{K} y_s^k}{K}\right)+1}{S+1}. \quad (2)$$

Note that w represents the density of events in the batch: when w is close 0, this means that there are few samples in the batch with activated events, while when it is close to 1 , almost every sample has all events simultaneously active. On the other hand, when calculating the loss for a strong label, we use:

$$w = \frac{\sum_{s=1}^{S}\left(\frac{\sum_{k=1}^{K}\sum_{t=1}^{T} y_s^{(t,k)}}{TK}\right)+1}{S+1}, \quad (3)$$

where in this case w represents the proportion of frames activated in the batch. Note that both for strong and weak labels, an all-zero y_s leads to $w = 0$, which produces zero loss independently of the prediction x_s. To account for this situation, a one is added to the numerator and denominator.

2.3. System Configurations

Below is a brief description of the systems presented by showing their best results.

- Avg-Weak-Balanced: The baseline model is used without any modifications, but $wbce$ is used to account for activity imbalance over weak predictions.
- Max-Weak-Balanced: This system is an extension of the baseline without attention layer, substituting the way of predicting the weak labels, by Max-Weak and applying $wbce$.
- Strong and Max-Weak Balanced: Both the weak error (with Max-Weak predictions) and strong error use $wbce$.

The rest of the systems tested correspond to different combinations of the two proposed strategies and their name indicates the applied methods.

3. RESULTS

Figure 2 shows the results obtained for the different systems under test. On the x-axis, the Polyphonic Sound Event Detection Score (PSDS) for scenario 1 is shown, which basically focuses on how fast an activation is detected. On the y-axis, the PSDS for scenario 2, which measures the confusion between events. Among the proposed methods, the best results when looking at the PSDS-scenario1 metric, are obtained for systems using Max-Weak, occupying the first 5 positions. However, for PSDS-scenario2, its performance drops considerably, except when balancing both the weak and strong predictions ("Strong and Max-Weak Balanced").

The systems presented to the Challenge competition for obtaining the best results are indicated in red. The 3 methods are based on the proposed activity balance weighting. The "Avg-Weak-Balanced" system presents a slight increase in the PSD-scenario2 metric with respect to the baseline. The system "Max-Weak-Balanced" has a good performance with respect to the baseline with attention layer in the PSDS-scenario1 metric and, finally, "Strong and Max-Weak Balanced" shows a competitive result for both metrics.

When using imbalance weighting, the results are better only when the attention layer is not included. If attention is included, the proposed strategies are not really effective, as noticed for the systems "Attention Max-Weak-Balanced" or "Attention Weak-Balanced".

4. CONCLUSION

Two activity-motivated changes have been proposed to the baseline model for DCASE2022 Task 4. The proposed strategies, despite being very simple, led to a slight improvement in the performance of the baseline system. The results show that balancing strategies can lead to some performance gains with little effort within the model training stage.

5. REFERENCES

[1] R. Serizel, N. Turpault, A. Shah, and J. Salamon, "Sound event detection in synthetic domestic environments," in *ICASSP 2020 - 45th International Conference on Acoustics, Speech, and Signal Processing*, Barcelona, Spain, 2020. [Online]. Available: https://hal.inria.fr/hal-02355573

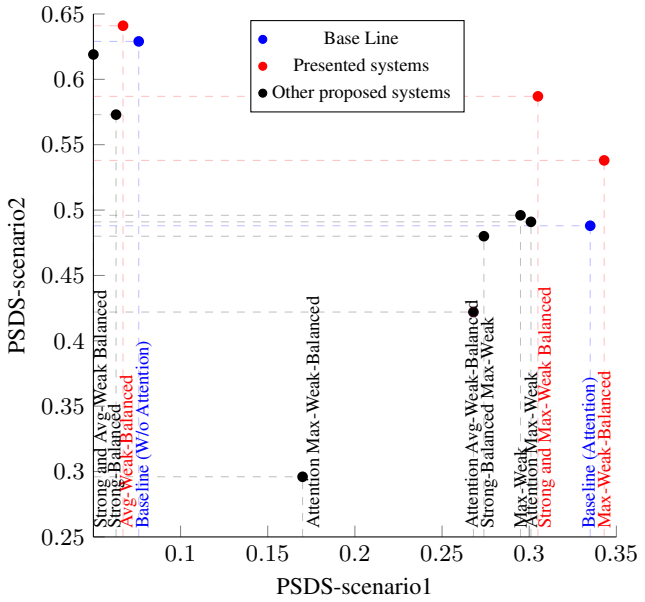

Figure 2: Validation results

[2] N. Turpault, R. Serizel, A. Parag Shah, and J. Salamon, "Sound event detection in domestic environments with weakly labeled data and soundscape synthesis," in *Workshop on Detection and Classification of Acoustic Scenes and Events*, New York City, United States, October 2019. [Online]. Available: https://hal.inria.fr/hal-02160855

[3] K. Imoto, S. Mishima, Y. Arai, and R. Kondo, "Impact of sound duration and inactive frames on sound event detection performance," 2021. [Online]. Available: https://arxiv.org/abs/2102.01927