SOUND EVENT DETECTION WITH SOFT LABELS USING SELF-ATTENTION MECHANISMS FOR GLOBAL SCENE FEATURE EXTRACTION

Technical Report

Nhan Tri-Do[1], Param Biyani*[2], Zhang Yuxuan*[3], Andrew Koh Jin Jie[3], Chng Eng Siong[3],*

[1] University of Science, Vietnam National University, {dotrinhan99}@gmail.com
[2] BITS Pilani Goa Campus, {parambiyani8}@gmail.com
[3] Nanyang Technological University - NTU Singapore, {yzhang253, andr0081, aseschng}@e.ntu.edu.sg

ABSTRACT

This paper presents our approach to Task 4b of the Detection and Classification of Acoustic Scenes and Events (DCASE) 2023 Challenge, which focuses on Sound Event Detection with Soft Labels. Our proposed method builds upon a CRNN backbone model and leverages the benefits of data augmentation techniques to improve model robustness. Furthermore, we introduce self-attention mechanisms to capture global context information and enhance the model's ability to predict soft label segments more accurately. Our experiments demonstrate that incorporating soft labels and self-attention mechanisms result in significant performance gains compared to traditional methods on data varying across different scenarios.

***Index Terms*—** Sound event detection, Soft labels, Self-attention, convolutional recurrent neural network (CRNN)

1. INTRODUCTION

The detection of sound events is a fundamental task in audio signal processing with various applications. One of the main challenges in sound event detection is the availability of labeled training data, which is often difficult and expensive to obtain. The Detection and Classification of Acoustic Scenes and Events (DCASE) 2023 Challenge Task 4b aims to evaluate the performance of sound event detection systems that use soft labels for training, participants are expected to develop systems that leverage various machine learning techniques and architectures.

Soft labels [1] provide more informative annotations compared to traditional binary labels by characterizing the certainty of human annotators for the presence of a sound event at a specific time. The provided soft labels in Task 4b are continuous values between 0 and 1, with a temporal resolution of 1 second. Systems developed in Task 4b will be evaluated against hard labels, which are obtained by thresholding the soft labels at 0.5. Anything above 0.5 is considered 1 (sound active), and anything below 0.5 is considered 0 (sound inactive).

2. DATASET

The Detection and Classification of Acoustic Scenes and Events (DCASE) 2023 Challenge Task 4b provides a dataset called MAESTRO Real for the development of sound event detection systems using soft labels. The dataset consists of real-life recordings with a length of approximately 3 minutes each, captured in different acoustic scenes to ensure the diversity of the data. The audio was annotated using Amazon Mechanical Turk, which is a platform that enables crowd-sourced annotations from multiple annotators [2].

The dataset is provided with a 5-fold cross-validation setup in which approximately 70% of the data (per class) is used in training, and the rest is used for testing. Out of the 17 sound classes present in the provided dataset, only 15 have values exceeding 0.5, and among those, 4 are very infrequent. Therefore, during our model training experiments, we focused solely on the 11 most prevalent classes.

An intriguing characteristic of the dataset under consideration is its composition, sourced from five distinct scenarios, with each scenario featuring only a subset of the total 17 classes. Thus, a comprehensive understanding of the audio scenario at hand alone can yield valuable insights into the predictive outcomes.

The system evaluation uses the macro-average F1 score with optimum threshold per class (F1MO) metric, calculated in 1-second segments. The F1MO score is based on the best F1 score per class obtained with a class-specific threshold, providing a comprehensive system performance assessment.

3. RELATED WORK

Sound event classification is an active area of research with a variety of techniques proposed in recent years. One approach that has shown promise is the use of Convolutional Recurrent Neural Networks (CRNNs) with various augmentation techniques. For example, the Forward-Backward CRNN pseudo-labeling approach and Bidirectional CRNN have been used for sound event detection with success [3], but rely on sharply marked start and end annotations.

Other modifications to above techniques that have been proposed include Mean-Teacher Models [4], which use a teacher-student approach to improve model generalization, and SK-CRNN [5], which utilizes a residual connection to improve the learning process. Frequency dynamic convolution (FDY) [6] has been proposed to remove the problem of frequency shift invariance faced by standard convolution, however, comes with an increased computational cost for each basis kernel. Asymmetric focal loss [7] has also been proposed as a solution to the data imbalance problem in SED. In terms of model architectures, CRNNs, and pretrained Models (such as PANNs and SSAST) have been widely used.

The use of time-warping augmentation, such as SpecAugment [8] has been shown to be effective for increasing robustness to variations in the duration of sound events. Frameshift, time mask, fre-

*Equal Contribution

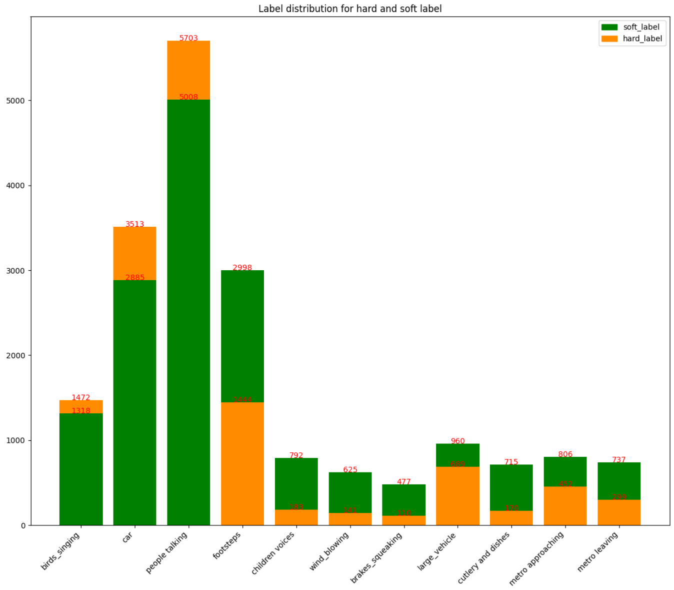

Figure 1: Distribution of classes in the dataset between hard labels and soft labels.

quency mask, gaussian noise, FilterAugment, ICT, and SCT are also commonly used augmentation techniques. In one study [5], a combination of mix-up, SpecAugment, and time-frequency shifting was found to improve performance over a standard EfficientNet-based mean-teacher model.

In addition to augmentation techniques, incorporating various types of background noise, such as Gaussian white noise, pure music, and other free sounds, has been found to be effective in improving the robustness of sound event detection models. The use of CRNNs and various augmentation techniques, in combination with carefully selected loss functions and model architectures, has shown promise for sound event classification tasks.

4. METHOD

4.1. Data Processing

The steps involved include resampling the data to a standard sampling rate of 16000 Hz and converting stereo audio to mono. The audio signal is then broken down into small time frames, and a Mel Spectrogram is calculated using a filter bank to obtain a visual representation of the frequency content of the signal. The Mel Spectrogram can be computed with different hop sizes to control the amount of overlap between time frames. Finally, the pre-trained wav2vec2 [9] model is used to extract embedding features from each 1-second segment of audio data. These embeddings capture higher-level characteristics of the audio signal that can be used as input features for machine learning models.

4.2. Backbone model

Convolutional Recurrent Neural Network (CRNN) combines the spatial feature extraction capabilities of Convolutional Neural Networks (CNNs) and the temporal modeling capabilities of Recurrent Neural Networks (RNNs). It is widely used for tasks such as video classification, object detection, speech recognition, and audio classification. We rely on the baseline model [10] of the organizers to improve the accuracy.

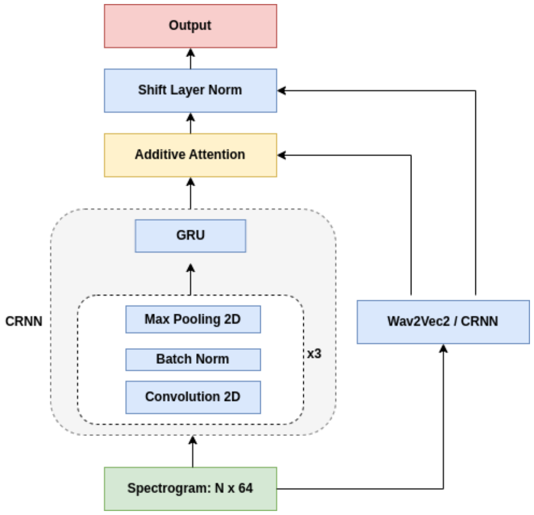

Figure 2: Architecture Overview for Global Attention CRNN.

The model takes in an audio chunk that includes the primary target segment that needs to be predicted (Figure 2) and surrounding audio segments. The output of the model is the probability of each event occurring in the main segment.

4.3. Self-attention for global scene information

The model pipeline involves using a Global CRNN model to extract embeddings that represent the entire audio context. These embeddings act as the key in the attention mechanism (**??**). Then, the local feature of the predicted segment is extracted using a Wav2Vec2 model or another CRNN, and this is used to form a query. The attention score is then computed by multiplying each audio segment by the query through additive attention. Based on the resulting attention distribution, the attention output is a weighted sum of all the context audio segments. In essence, this approach allows the model to focus its attention on the relevant audio segments while taking into account the surrounding context.

$$Q = CRNN/Wav2vec2(audiotarget) \tag{1}$$

$$K = CRNN(audiocontext) \tag{2}$$

$$\text{Attention Output} = \text{softmax}\left(QW_q + KW_k\right)V \tag{3}$$

The attention output is combined with the feature of the segment that requires event prediction either by concatenation or by a shift layer norm. This combined feature is then passed through fully connected layers to predict the event of the 1-second segment.

4.4. Augmentation

SpecAugment is an augmentation technique commonly used in sound event classification tasks. It applies two types of masking to the spectrogram: frequency masking and time masking. The frequency masking randomly masks out a continuous frequency band

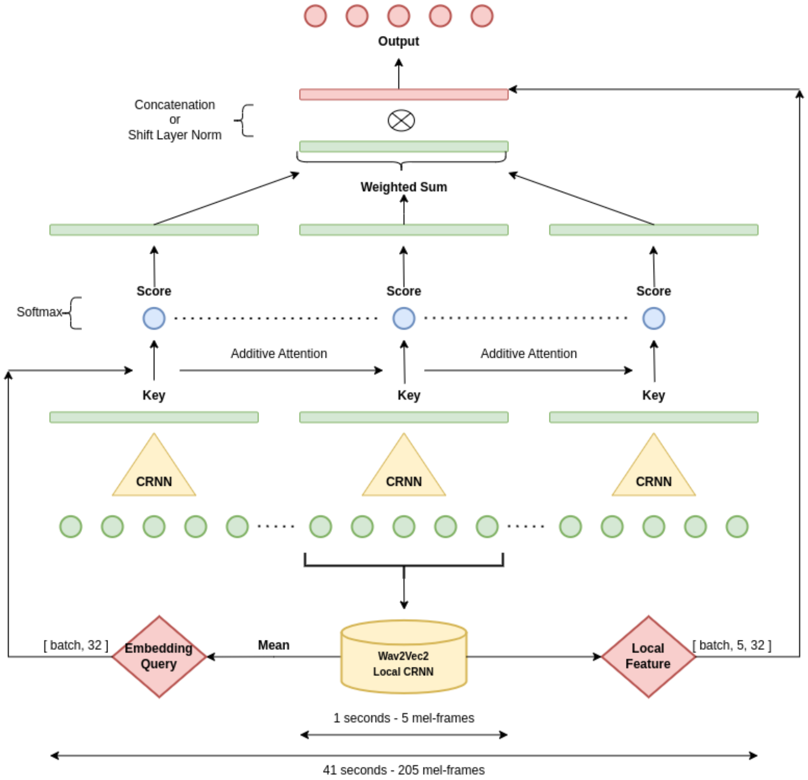

Figure 3: CRNN model with self-attention mechanism using wav2vec2 feature as query.

by setting the corresponding spectrogram coefficients to zero. The time masking randomly masks out a continuous time segment by setting the corresponding spectrogram coefficients to zero. In our experiments, the number of masks applied is set to 2. The hyperparameters freq masking and time masking are set to 0.15 and 0.20, respectively, which means that on average, 15% of the frequency band and 20% of the time segment are masked out.

Audio waveform augmentation is also used for the purpose of adding variety to the training data and improving the robustness of the model. One of the methods used is pitch augmentation, which alters the pitch of the audio by scaling the frequency axis. Another method is clipping augmentation, which simulates audio clipping by truncating the waveform. The third method is reverb augmentation, which adds simulated reverberation to the audio waveform.

We also use a balanced sampler[1] for the data loader to achieve a balanced distribution by randomly undersampling the majority class and oversampling the minority class.

4.5. Focal Loss

Focal Loss [11] is a modified version of cross-entropy loss that is designed to address the class imbalance in classification tasks. It assigns higher weights to misclassified examples of the minority class, thereby focusing the learning process on hard-to-classify examples. The formula for Focal Loss is:

$$FL(p_t) = -a(1-p_t)^{\gamma} log(p_t) \tag{4}$$

where p is the predicted probability of the correct class, and γ is a user-defined parameter that adjusts the degree of focusing.

To address the issue of imbalanced data, we employed the use of focal loss in our approach. This allowed us to effectively calculate error predictions against soft event label targets. Additionally, we utilized cross-entropy to account for scene loss for local features.

$$\begin{aligned} Loss_{total} = {} & FL(prediction_{event}, target_{event}) \\ & + CE(prediction_{scene}, target_{scene}) \end{aligned}$$

5. EXPERIMENTAL RESULTS

5.1. Self-Attention CRNN Architecture

We tested the proposed methods on the development dataset and evaluated them using the F1 Macro Optimized metric. The results presented in the table indicate that the length of the input chunk has an impact on the accuracy of the model. Specifically, increasing the input size leads to higher accuracy.

Based on the experimental results, we find that using wav2Vec2 used as query feature extraction is less effective than using a simple CRNN model, this pre-trained audio needs to be fine-tuned with even more event audio data.

5.2. Augmentation and Balanced Sampler

By leveraging augmentation techniques for both waveform and spectrogram data, the accuracy of the model can be improved significantly by approximately 2%. Additionally, utilizing a balanced al-

[1] https://github.com/khornlund/pytorch-balanced-sampler

Mean Prediction	Hop Size	Chunk Size	Augmentation	Balanced Sampler	F1MO dev
.	.	.	-	-	Baseline = 44.13%
✓	800	1s	✓	-	38.0%
✓	3200	1s	-	-	39.2%
-	3200	1s	-	-	41.1%
-	3200	5s	-	-	43.6%
-	3200	41s	-	-	44.2%
-	3200	41s	✓	-	46.2%
-	3200	41s	✓	0.5	**46.7%**

Table 1: Experimental results for the proposed methods on the development Set.

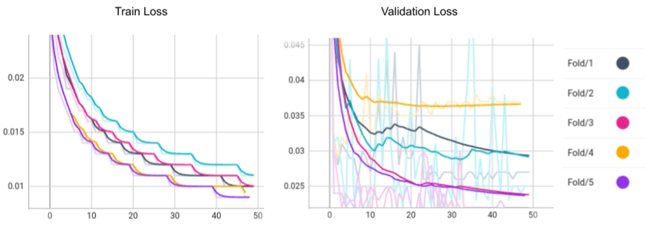

Figure 4: Visualization of the focal loss for training and validation set of 5 folds.

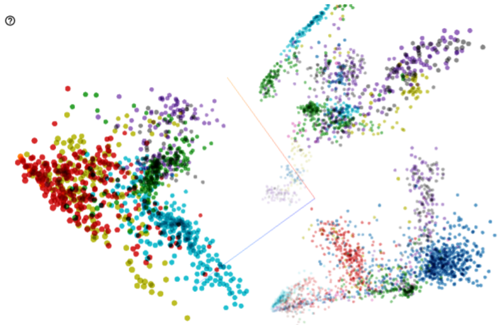

Figure 5: t-SNE visualization of sound event embeddings generated by our proposed method.

pha that is appropriate for the model's configuration can further enhance its performance. After experimenting with different balance weights, we found a ratio that improved F1MO score by around 0.5%. Striking a balance in the alpha value is crucial, as using too much or too little can negatively affect the model's accuracy.

5.3. TSNE Visualization and Loss plots

We extract event embeddings before the final linear layer of the model and subsequently visualize them using t-SNE. In Figure 5, it becomes evident that the majority of classes are densely concentrated in clusters, whereas the minority of classes are sparsely distributed.

Figure 4 shows that a decrease in the training loss does not correspond to a decrease in the validation loss, which fluctuates significantly. These observations suggest that the model is unstable and has not fully captured the distinctive features of sound events.

6. CONCLUSION

In conclusion, the proposed method for Sound Event Detection with Soft Labels builds upon a CRNN backbone model and utilizes data augmentation techniques to enhance model robustness. The introduction of self-attention mechanisms further improves the model's accuracy by capturing global context information. The experimental results on the development set indicate that incorporating soft labels and self-attention mechanisms leads to performance improvements compared to the baseline methods. Further experimentation is necessary to determine the optimal hyperparameters and configuration for this proposed method.

7. REFERENCES

[1] I. Martín-Morató, M. Harju, P. Ahokas, and A. Mesaros, "Training sound event detection with soft labels from crowdsourced annotations," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023, pp. 1–5.

[2] I. Martín-Morató and A. Mesaros, "Strong labeling of sound events using crowdsourced weak labels and annotator competence estimation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 31, pp. 902–914, 2023.

[3] J. Ebbers and R. Haeb-Umbach, "Pre-training and self-training for sound event detection in domestic environments," Paderborn University, Tech. Rep, Tech. Rep., 2022.

[4] K. He, X. Shu, S. Jia, and Y. He, "Semi-supervised sound event detection system for dcase 2022 task 4."

[5] J. W. Kim, G. W. Lee, H. K. Kim, Y. S. Seo, and I. H. Song, "Semi-supervised learning-based sound event detection using frequency-channel-wise selective kernel for dcase challenge 2022 task 4 technical report."

[6] H. Nam, S.-H. Kim, B.-Y. Ko, and Y.-H. Park, "Frequency dynamic convolution: Frequency-adaptive pattern recognition for sound event detection," *arXiv preprint arXiv:2203.15296*, 2022.

[7] S. Suh and D. Y. Lee, "Data engineering for noisy student model in sound event detection."

[8] D. S. Park, W. Chan, Y. Zhang, C.-C. Chiu, B. Zoph, E. D. Cubuk, and Q. V. Le, "Specaugment: A simple data augmentation method for automatic speech recognition," *arXiv preprint arXiv:1904.08779*, 2019.

[9] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," *Advances in neural information processing systems*, vol. 33, pp. 12 449–12 460, 2020.

[10] I. Martín-Morató, M. Harju, P. Ahokas, and A. Mesaros, "Strong labeling of sound events using crowdsourced weak labels and annotator competence estimation," in *Proc. IEEE Int. Conf. Acoustic., Speech and Signal Process. (ICASSP)*, 2023.

[11] 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.