DOMAIN-ADAPTED SOUND EVENT DETECTION SYSTEM WITH AUXILIARY FOREGROUND-BACKGROUND CLASSIFIER

Technical Report

Michel Olvera[1], *Emmanuel Vincent*[1], *Gilles Gasso*[2],

[1] Université de Lorraine, Inria, Loria, F-54000 Nancy, France, {michel.olvera, emmanuel.vincent}@inria.fr
[2] LITIS EA 4108, Université & INSA Rouen Normandie, 76800 Saint-Étienne du Rouvray, France,
gilles.gasso@insa-rouen.fr

ABSTRACT

In this technical report, we propose a sound event detection system for the DCASE 2021 task 4 challenge, which consists of a foreground-background classification branch that is jointly trained with the baseline architecture. Furthermore, to account for the mismatch between synthetic annotated data and real unlabeled data used for training, we also propose a frame-level domain adaptation scheme to improve detection performance over real soundscapes. We show that these improvements to the baseline method help in the generalization of the sound event detection task.

Index Terms— Sound Event Detection, Domain Adaptation, foreground-background classification

1. INTRODUCTION

Task 4 of the DCASE 2021 challenge [1] offers the opportunity to design systems for the analysis of ambient sound scenes of domestic environments. To this end, synthetic and real-world recordings are provided to come up with solutions that overcome commonly found problems in sound event detection (SED). In this task, although it is customary to generate synthetic soundscapes that try to match as much as possible the acoustic conditions found in real-life recordings, there still exists a shift between the simulated (source domain) and the actual generating process of real environments (target domain). This difference in domain distributions motivates the use of adaptation strategies to reduce data mismatch. Thus, we particularly tackle the feature distribution shift between synthetic and real data used to train data-driven SED systems. To account for any differences between synthetic and real soundscapes, we propose a domain adaptation strategy based on the DeepJDOT method [2] relying on optimal transport that aims to improve performance on real test data. Also to help the SED system to learn relevant sound representations, we investigate adding foreground-background detection as an auxiliary classification task. Altogether domain adaptation and the auxiliary task lead to enhanced SED performances.

2. MODEL ARCHITECTURE

The selected model architecture is the same as the baseline system. It consists of a mean-teacher model in which both the student and the teacher model have the same network architecture. Our proposed domain adaptation strategy chooses only the student model to undergo adaptation.

The network architecture is the same as the baseline system to perform sound event detection. It comprises a convolutional-recurrent neural network (CRNN). The CNN part is composed of 7 layers, each layer having [16, 32, 64, 128, 28, 128, 128] filters, respectively. A kernel of size 3x3 was used and the max-pooling for each layer are [[2, 2], [2, 2], [1, 2], [1, 2], [1, 2], [1, 2], [1, 2]], respectively. A gated linear unit activation is applied to the convolution operations.

The RNN part is composed of 2 layers of 128 bidirectional gated recurrent units. The output of the CRNN is followed by a dense layer with sigmoid activation function to produce frame-level (strong) predictions. Clip-level (weak) predictions are obtained by multiplying the aforementioned linear layer with a dense layer with softmax activation function.

The foreground-background classification branch consists of a dense layer with sigmoid activations, which acts upon the outputs of the RNN block.

In the training stage, the model is trained with Adam optimizer, and a dropout value of 0.5, with a gradually increasing learning rate with max value of $1e-3$ as in [3]. During the adaptation stage, learning rate is fixed to $1e-4$.

3. DATASET

We used the DCASE 2020 available data to trained our systems. The dataset is composed of 2,045 synthetic audio clips generated by Scaper [4], 1,578 real soundscapes with clip-level annotations and 14,412 unlabeled real recordings. All this available data for training was used to trained our proposed system with foreground-background classifier, while only the synthetic and weakly labeled datasets were used in our proposed domain adaptation strategy. We did not rely on the new synthetic dataset for this 2021's challenge.

4. MODEL TRAINING

For the sound event detection task we have access to a synthetic dataset with strong labels $\mathcal{D}^S = \{\mathbf{x}_i^s, y_i^s\}_{i=1}^{n^s}$, and two datasets with real recordings: one with weakly labels $\mathcal{D}^W = \{\mathbf{x}_i^w, y_i^w\}_{i=1}^{n^w}$, and the other without labels of any sort $\mathcal{D}^U = \{\mathbf{x}_i^u, y_i^u\}_{i=1}^{n^u}$.

We use CRNN from student model as embedding function $g : \mathbf{x} \rightarrow \mathbf{z}$, where the log-mel representations are mapped to a latent space $\mathcal{Z}$. The sound event detection (SED) branch is represented by the function $f : \mathbf{z} \rightarrow \mathbf{y}$, that maps the latent space to the sound events class label space, and the proposed foreground-background

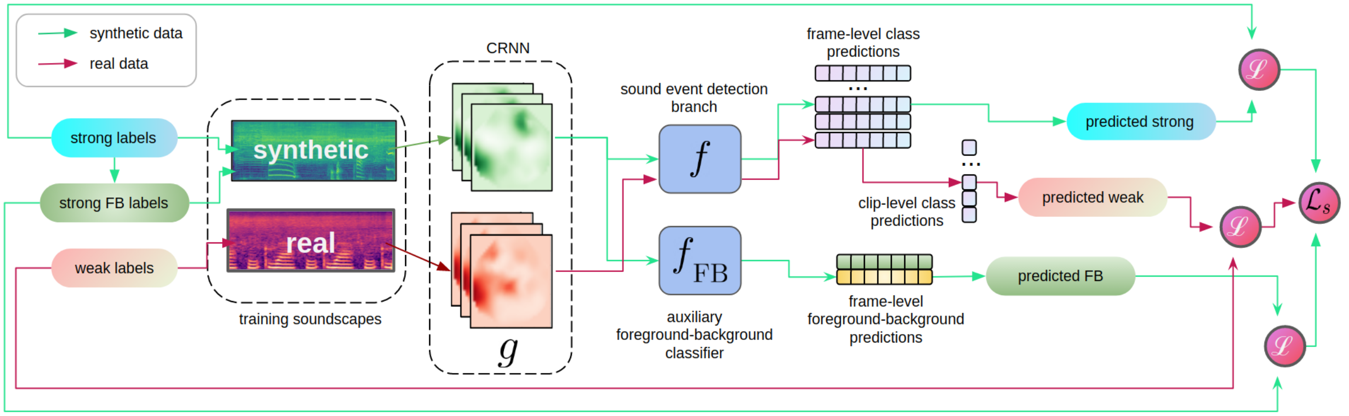

Figure 1: Proposed model with added foreground-background auxiliary classifier to the baseline model with color coded data flow for synthetic and real soundscapes. For simplicity, the diagram depicts only the student model and the associated classification costs for training.

(FB) auxiliary classifier $f_{\text{FB}} : \mathbf{z} \rightarrow \mathbf{y}^{fb}$ maps the latent space to the foreground-background label space. Analogously, for the teacher model we denote by g', f' and f'_{FB} the CRNN embedding function, SED branch and FB auxiliary classifier, respectively. Figure 1 show a general depiction of the proposed system.

Motivated by the fact that discriminating the spectro-temporal characteristics of domestic sounds in foreground (e.g., speech, cat, dog) and background (e.g., vacuum cleaner, blender, electric razor) is possible in source separation [5], we incorporate a classifier that aims to categorize the domestic sound events in these two broad categories. We hypothesize that learning it jointly with the baseline system will help the network improve generalization on the sound event detection task.

Thus, to train the FB classifier in a supervised way, we generated foreground-background ground-truth annotations y_s^{fb} from the strong labels of synthetic data y_s by combining the sound event labels in two categories:

- Foreground: *alarm - bell ringing, speech, cat, dog* and *dishes*
- Background: *blender, vacuum cleaner, frying, electric shaver - toothbrush* and *running water*

The mean-teacher model is optimized by the combination of three classification-consistency cost pairs:

$$\mathcal{L} = L(\mathbf{y}_i^s, f(g(\mathbf{x}_i^s))) + \lambda L_{\text{strong}}(f(g(\mathbf{x}_i)), f'(g'(\mathbf{x}_i))) + \quad (1)$$

$$L(\mathbf{y}_i^w, f(g(\mathbf{x}_i^w))) + \lambda L_{\text{weak}}(f(g(\mathbf{x}_i)), f'(g'(\mathbf{x}_i))) + \quad (2)$$

$$L(\mathbf{y}_i^{fb}, f_{\text{FB}}(g(\mathbf{x}_i^s))) + \lambda L_{\text{strong}}(f_{\text{FB}}(g(\mathbf{x}_i)), f'_{\text{FB}}(g'(\mathbf{x}_i))) \quad (3)$$

where $L(\cdot,\cdot)$ is a binary cross-entropy classification loss, $L_{\text{strong}}(\cdot,\cdot)$ and $L_{\text{weak}}(\cdot,\cdot)$ are mean-square error consistency costs on strong (frame-level) and weak (clip-level) predictions, respectively. The consistency weight λ is tied to all consistency costs and increases gradually as training progresses.

5. DOMAIN ADAPTATION FOR SOUND EVENT DETECTION

The main idea of the proposed adaption method for sound event detection relies on joint distribution optimal transport (JDOT) of feature embeddings [2], which seeks to align the joint distribution of embedded feature representations and labels of two shifted domains.

Let μ_s and μ_t are measures on the product space $\mathcal{X} \times \mathcal{Y}$. The cost associated to this space can be expressed as a weighted combination of costs in the feature and label spaces as follows,

$$d(\mathbf{x}_i^s, \mathbf{y}_i^s; \mathbf{x}_j^s, \mathbf{y}_j^s) = \alpha c(\mathbf{x}_i^s, \mathbf{x}_j^t) + \beta t \mathcal{L}(\mathbf{y}_i^s, \mathbf{y}_j^t) \quad (4)$$

for the i-th source and j-th target element, and where $c(\cdot,\cdot)$ is chosen as a ℓ_2^2 distance and $\mathcal{L}(\cdot,\cdot)$ is a classification loss. Parameters α and β are two scalar values weighting those two terms. Since no labels are available in target domain, $\mathbf{y}_j^t$, they are replaced with a class prediction $f(\mathbf{x}_j^t)$ from a classifier $f : \mathcal{X} \rightarrow \mathcal{Y}$. Accounting for the classification loss, leads to the following minimization problem:

$$\min_{f,\boldsymbol{\gamma}\in\Gamma(\mu_s,\mu_t)} < \boldsymbol{\gamma}, \mathbf{C}_f >_F, \quad (5)$$

where $\mathbf{C}_f$ depends on f and comprises all the pairwise costs d.

For the adaptation task, we regard the synthetic dataset with strong labels as the source domain $\mathcal{S} = \mathcal{D}^S$, and the combination of real recordings from the weakly and unlabeled dataset as the target domain $\mathcal{T} = \mathcal{D}^W \cup \mathcal{D}^U$.

We propose a two-step frame-level domain adaptation method based on the joint distribution matching of the learned semantic audio embeddings using optimal transport (OT). In the first step, with CRNN fixed parameters (evaluation mode) we compute the optimal coupling matrix

$$\min_{\boldsymbol{\gamma}\in\Gamma(\mu_s,\mu_t)} \sum_{i,j=1}^{m} \gamma_{ij}(\alpha||g(\mathbf{x}_i^s) - g(\mathbf{x}_j^t)||^2 + \beta\mathcal{L}(\mathbf{y}_i^s, f(g(\mathbf{x}_j^t)))) \quad (6)$$

In the second step, with fixed $\boldsymbol{\gamma}$, we update model parameters g and f using the following objective

$$\frac{1}{m}\sum_{i=1}^{m} \mathcal{L}_s + \sum_{i,j=1}^{m} \gamma_{ij}(\alpha||g(\mathbf{x}_i^s) - g(\mathbf{x}_j^t)||^2 + \beta\mathcal{L}(\mathbf{y}_i^s, f(g(\mathbf{x}_j^t)))) \quad (7)$$

where $\mathcal{L}_s$ correspond to the classification cost on source domain to avoid forgetting the distribution of synthetic data.

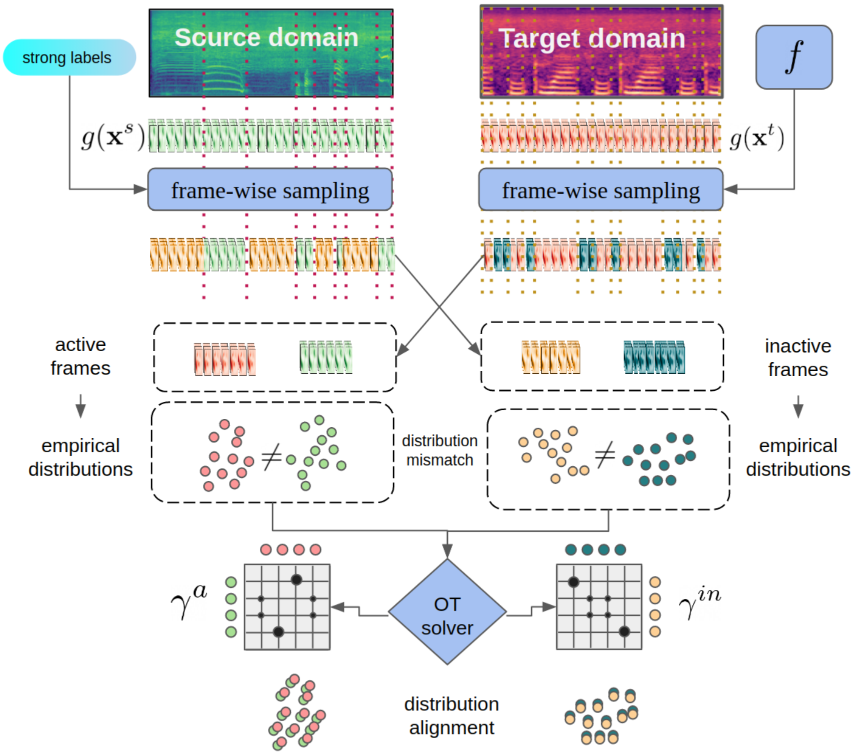

Figure 2: Proposed optimal transport-based method to correct domain mismatch between sound embeddings with active and inactive sound events.

5.1. Sampling strategy for domain adaptation

The sampling strategy for domain adaptation corresponds to selecting active (labeled) and inactive (unlabeled) frames from the semantic representation $g : \mathbf{x} \rightarrow \mathbf{z}$ learned by the CRNN model.

For each data batch we sample N_a^s active frames from synthetic as indicated by the strong labels (oracle sampling), as well as N_a^t active and inactive frames from real data as indicated by the frame-level pseudo-labels assigned by the model to the real data. For both types of data we keep only frames with no sound event overlap, i.e., only one sound event class is active per sampled frame. The amount of sampled active frames per class can vary considerably for synthetic and real data and in some cases classes might be missing in one type of data. To partially account for this class data imbalance we keep only active frames from all classes appearing in both synthetic and real data, and from that, we re-sample in a class-wise manner $N_c = \min(N_c^s, N_c^t)$, $c = 1, \ldots, C_a$ samples, where C_a is the total number of sound classes active in the batch, leading to $N_a = \sum_{c=1}^{C_a} N_c$ frames sampled for each type of data.

Similarly, the amount of inactive frames in synthetic and real data varies from each batch. So, after sampling N_{in}^s and N_{in}^t inactive frames as indicated by the strong labels and pseudo-labels from synthetic and real data, respectively, we keep only $N_{in} = \min(N_{in}^s, N_{in}^t)$.

5.2. Pseudo label refinement

To improve the reliability of the pseudo-labels assigned to real data, we leverage the provided annotations of the weakly labeled set. The refinement process consists of fusing the frame-level predictions of the model on the real data with their clip-level annotations by an element-wise multiplication,

$$\hat{\mathbf{y}}^t = f(g(\mathbf{x}_j^t)) \odot \mathbf{y}_j^w, \; j = 1, \ldots, n^w \tag{8}$$

This operation constrains the output prediction labels to contain at most the same classes present in the weakly labeled soundscapes. Filtering out all extra classes helps reduce false positive predictions and consequently more reliable pseudo-labels for domain adaptation are obtained.

5.3. Frame-level domain adaptation

After frame-level sampling and pseudo-label refinement, adaptation is performed by the DeepJDOT method by aligning the distributions of class-sampled active and inactive frames. To this end, two separate cost functions are jointly optimized to account for the mismatch between synthetic and real soundscapes. The system is adapted by minimizing the overall training objective

$$\mathcal{L}_s + \mathcal{L}_a + \mathcal{L}_{in} \tag{9}$$

where $\mathcal{L}_s$ corresponds to the first and third classification cost terms of the training classification cost,

$$\mathcal{L}_s = \frac{1}{n^s}\sum_{i=1}^{n^s} L(\mathbf{y}_i^s, f(g(\mathbf{x}_i^s))) + \frac{1}{n^s}\sum_{i=1}^{n^s} L(\mathbf{y}_i^{fb}, f_{\text{FB}}(g(\mathbf{x}_i^s))). \tag{10}$$

Table 1: Performance on development and evaluation sets.

Method	F1 score		F1 score	
	val	val +HMMs	eval	eval +HMMs
Baseline	34.8			
Baseline^{++}	40.1			
Baseline + FB	43.12	45.42	46.06	49.38
Baseline + FB + DA	**45.68**	47.77	**50.79**	53.10
Ensemble 1	45.13	**48.07**	50.58	**53.35**
Ensemble 2	45.15	47.08	50.28	52.23

Note that only the student model is undergoing adaptation, therefore, no consistency losses are included in the above objective to train the source domain classifier.

Costs function $\mathcal{L}_a$ corresponds to the distribution alignment loss of active frames,

$$\mathcal{L}_a = \frac{1}{C_a}\sum_{i,j}^{N_a} \gamma_{ij}^{a}(\alpha||g(\mathbf{x}_i^s) - g(\mathbf{x}_j^t)||^2 + \beta\mathcal{L}(\mathbf{y}_i^s, \hat{\mathbf{y}}^t)). \tag{11}$$

Note that $\mathcal{L}_a$ is averaged by the number of active classes C_a in the batch. The second term in the loss enforces regularity of the target classifier with the available source data.

Finally, cost function $\mathcal{L}_{in}$ corresponds to simply the distribution alignment loss of inactive representation embeddings, and no consistency loss is required.

$$\mathcal{L}_{in} = \sum_{i=1}^{N_{in}} \gamma_{ij}^{in}(\alpha||g(\mathbf{x}_i^s) - g(\mathbf{x}_j^t)||^2) \tag{12}$$

Experiments with optimal transport were performed using the Python Optimal Transport package [6]. We used cost weights $\alpha = 0.02$ and $\beta = 5.0$. Also, the contribution of the source classifier cost $\mathcal{L}_s$ was increased by 100 during the adaptation stage. Figure 2 depicts the proposed frame-level domain adaptation strategy.

5.4. Post-processing of final predictions

Rather than using median filtering to post-process predictions, we used Hidden-Markov-Model (HMM) decoding. Following the same procedure as in [7], a two-state HMM was employed for each sound class, while the silence self-loop transition probability was tied to be the same for all HMMs. The class-wise transition probabilities and silence were tuned using 50% of the validation set by using Random Forest and maximizing the event based F1- macro-average score of the trained model. The optimal computed values for the HMMs transition probabilities were used as prediction refinement by running Viterbi decoding on the model's emission probabilities for each class.

5.5. Results

In table 1 we compare results obtained on the validation and public evaluation sets in terms of the event-based macro F1 score by the proposed improvements to the baseline model. Models labeled as Baseline and Baseline^{++} correspond to the 2020 and 2021 challenge editions, respectively. For each set we also show performance after post-processing the final predictions with HMM smoothing.

We can see that adding a foreground-background auxiliary (FB) branch is beneficial to the sound event detection task as it improved results from Baseline and Baseline^{++} by 8.3% and 3%, respectively. Further improvement was achieved by refining predictions with HMM smoothing, as performance increased by 10.6% and 5.3%, respectively.

Performing adaptation by the proposed approach after training brought additional improvement, and combined with HMM smoothing the best score on the validation and public evaluation sets were obtained. Note that after adaptation the F1 score on the validation set only improved by around 2.5%, but a significant improvement was achieved on the public evaluation set, as the score improved by around 4.7%. This higher improvement might be due to the fact that the empirical distribution of the active and inactive frames of this set, resembles more that of the provided real training data from which adaptation was carried out. Ensemble 1 and Ensemble 2 are model ensembles comprising three and two Baseline + FB + DA systems from different training runs, respectively. Predictions from the ensembles are simply the average of their individual predictions. Only Ensemble 1 achieved an improvement in the validation set compared to a single system. These models correspond to the submissions made for the DCASE 2021 Challenge task 4 and are labeled *Olvera_INRIA_task4_SED_1* and *Olvera_INRIA_task4_SED_2*, respectively.

5.6. Conclusion

In this work we described our systems for the task 4 of the DCASE 2021 Challenge corresponding to the sound event detection task. Motivated by the categorization of the spectro-temporal characteristics of domestic sounds in foreground and background, we proposed the use of an auxiliary foreground-background classifier that is jointly trained with the baseline system to improve generalization in the detection of sound events. Furthermore, we proposed to incorporate an adaptation stage based on the joint distribution optimal transport of feature embeddings to account for the acoustic mismatch between the available synthetic and real data for training. We showed that the multi-task training approach together with the adaptation stage brought a substantial improvement to the performance of the baseline system.

5.7. Acknowledgment

This work was made with the support of the French National Research Agency, in the framework of the project LEAUDS "Learning to understand audio scenes" (ANR-18-CE23-0020). Experiments presented in this paper were carried out using the Grid'5000 testbed, supported by a scientific interest group hosted by Inria and including CNRS, RENATER and several Universities as well as other organizations (see `https://www.grid5000.fr`).

6. REFERENCES

[1] 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

[2] B. Bhushan Damodaran, B. Kellenberger, R. Flamary, D. Tuia, and N. Courty, "Deepjdot: Deep joint distribution optimal

transport for unsupervised domain adaptation," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 447–463.

[3] L. Delphin-Poulat and C. Plapous, "Mean teacher with data augmentation for dcase 2019 task 4," *Orange Labs Lannion, France, Tech. Rep*, 2019.

[4] J. Salamon, D. MacConnell, M. Cartwright, P. Li, and J. P. Bello, "Scaper: A library for soundscape synthesis and augmentation," in *2017 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*. IEEE, 2017, pp. 344–348.

[5] M. Olvera, E. Vincent, R. Serizel, and G. Gasso, "Foreground-background ambient sound scene separation," in *2020 28th European Signal Processing Conference (EUSIPCO)*. IEEE, 2021, pp. 281–285.

[6] R. Flamary and N. Courty, "Pot python optimal transport library," *GitHub: https://github. com/rflamary/POT*, p. 144, 2017.

[7] S. Cornell, M. Olvera, M. Pariente, G. Pepe, E. Principi, L. Gabrielli, and S. Squartini, "Domain-adversarial training and trainable parallel front-end for the dcase 2020 task 4 sound event detection challenge."