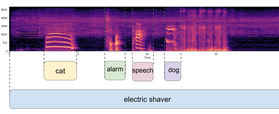
DCASE 2021 WORKSHOP
15-19 November 2021

Improving Sound Event Detection with Auxiliary Foreground-Background Classification and Domain Adaptation

Michel Olvera, Emmanuel Vincent, Gilles Gasso

Université de Lorraine, CNRS, Inria, Loria, LITIS Lab, Université & INSA Rouen Normandie
Emails: {michel.olvera, emmanuel.vincent}@inria.fr, gilles.gasso@insa-rouen.fr

Sound Event Detection



For a given audio recording, the goal of Sound Event Detection (SED) is to identify the class of the active sounds as well as their onset and offset time.

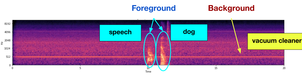
Motivations

- The **DCASE Challenge Task 4** encourages the development of SED methods through a baseline system.
- **Previous improvements to the DCASE Task 4 baseline system:**
  - Augmentation schemes to improve generalization.
  - Alternative time-frequency representations to log-Mel spectrograms.
  - Modifications to the deep network architecture.
  - Post-processing techniques to refine outputs.

- **Our proposed improvements:**
  - **Joint foreground-background classification to improve generalization.**
  - **An optimal transport-based domain adaptation strategy to reduce data mismatch.**

Foreground-Background Classification

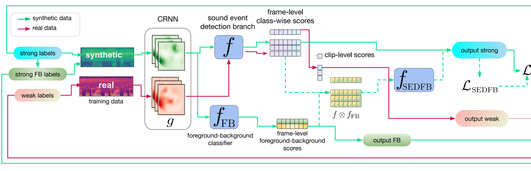
We propose to categorize domestic sounds according to their spectro temporal characteristics.



Foreground events:
Rapidly varying sounds

Background events:
Slowly varying sounds

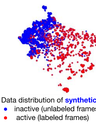
- It is customary to train a sound event detection system under the semi-supervised learning paradigm.
- We learn a foreground-background (FB) classifier jointly with the SED model.
- We investigate the combination of the FB classifier with the SED branch.



Domain Adaptation Strategy for Sound Event Detection

The semi-supervised learning paradigm to train SED systems aims to learn invariant representations for synthetic and real data.

However, there is still a gap in performance when testing systems on real environments due to mismatch between synthetic and real soundscapes.



Data distribution of **synthetic** data
- inactive (unlabeled frames)
- active (labeled frames)

Data distribution of **real** data
- inactive (unlabeled frames)
- active (labeled frames)

Deep Joint Distribution Optimal Transport

Damodaran, Bharath Bhushan, et al. "Deepjdot: Deep joint distribution optimal transport for unsupervised domain adaptation." *Proceedings of the European Conference on Computer Vision (ECCV)*. 2018.

Empirical distributions

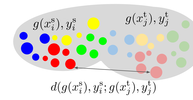
$$\mu_s = \sum_{i=1}^{n_s} a_i \delta_{g(x_i^s), y_i^s} \qquad \mu_t = \sum_{j=1}^{n_t} b_j \delta_{g(x_j^t), y_j^t}$$

Data points

$$(g(x_i^s), y_i^s), (g(x_j^t), y_j^t)$$

Weights

$$a_i \geq 0,\ b_j \geq 0 \qquad \sum_{i=1}^{n_s} a_i = \sum_{j=1}^{n_t} b_j = 1$$



$$d(g(x_i^s), y_i^s; g(x_j^t), y_j^t) = \alpha c(g(x_i^s), g(x_j^t)) + \beta \mathcal{L}(y_i^s, y_j^t)$$

Transportation coupling

$$\Gamma(\mu_s, \mu_t) = \{\gamma \in \mathbb{R}^{n_s \times n_t};\ \gamma \mathbf{1}_{n_t} = \mu_s,\ \gamma^T \mathbf{1}_{n_s} = \mu_t\}$$

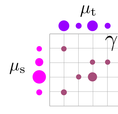
Optimization problem

$$\min_{\gamma \in \Gamma(\mu_s, \mu_t), g, f} \sum_{i,j} \gamma_{i,j} d(g(x_i^s), y_i^s; g(x_j^t), y_j^t))$$

Two-step process

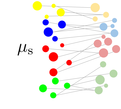
First: Compute optimal coupling matrix γ with fixed model parameters g and f

$$\min_{\gamma \in \Gamma(\mu_s, \mu_t)} \sum_{i,j} \gamma_{ij} (\alpha ||g(x_i^s) - g(x_j^t)||^2 + \beta \mathcal{L}(y_i^s, f(g(x_j^t))))$$



Second: With fixed γ, update model parameters g and f

$$\min_{g,f} \mathcal{L}_s + \sum_{i,j} \gamma_{ij} (\alpha ||g(x_i^s) - g(x_j^t)||^2 + \beta \mathcal{L}(y_i^s, f(g(x_j^t))))$$



Domain Adaptation Strategy for Sound Event Detection

After an initial pretraining stage:
- We sample **active** learned feature representations $\tilde{z}^s$ and $\tilde{z}^t$
- We sample **inactive** learned feature representations $\tilde{z}^s$ and $\tilde{z}^t$

We construct the following objective function to account for the mismatch between these empirical distributions.

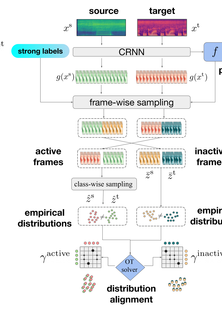
$$\mathcal{L}_s + \mathcal{L}_{\text{active}} + \mathcal{L}_{\text{inactive}}$$

where,

$$\mathcal{L}_{\text{active}} = \frac{1}{|\mathcal{C}_{\text{active}}|} \sum_{i,j}^{N_{\text{active}}} \gamma_{ij}^{\text{active}} (\alpha ||\tilde{z}_i^s - \tilde{z}_i^t||^2 + \beta \mathcal{L}(\tilde{y}_i^s, \tilde{y}_j^t))$$

and

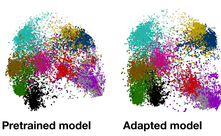
$$\mathcal{L}_{\text{inactive}} = \sum_{i=1}^{N_{\text{inactive}}} \gamma_{ij}^{\text{inactive}} (\alpha ||\tilde{z}_i^s - \tilde{z}_i^t||^2)$$



Domain Adaptation Strategy for Sound Event Detection

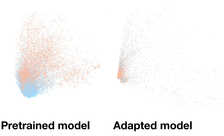
PCA sound embeddings

Distribution alignment of **active** frames
- Synthetic data
- Real data



Pretrained model **Adapted model**

Distribution alignment of **inactive** frames
- Synthetic data
- Real data



Pretrained model **Adapted model**

Main Results

We performed experiments on the Domestic Environment Sound Event Detection (DESED) dataset to test our proposed improvements to the sound event detection task. Results are in terms of the event-based macro F1-score.

Method	F1 score [%] validation		F1 score [%] evaluation	
	+ MF	+ HMMs	+ MF	+ HMMs
Baseline	34.8		38.1	
+ DA	42.41	43.89	44.8	47.12
+ FB	**43.12**	**45.42**	**46.06**	**49.38**
+ FB + DA	45.68	47.77	50.79	53.10
+ SEDFB	46.15	46.20	48.40	49.79
+ SEDFB + DA	47.61	47.75	52.12	53.30
DCASE 1	45.13	48.07	50.58	53.35
DCASE 2	45.15	47.08	50.28	52.23

1. Foreground-background classification improves generalization.

Method	F1 score [%] validation		F1 score [%] evaluation	
	+ MF	+ HMMs	+ MF	+ HMMs
Baseline	34.8		38.1	
+ DA	42.41	43.89	44.8	47.12
+ FB	43.12	45.42	46.06	49.38
+ FB + DA	45.68	47.77	50.79	53.10
+ SEDFB	**46.15**	**46.20**	**48.40**	**49.79**
+ SEDFB + DA	47.61	47.75	52.12	53.30
DCASE 1	45.13	48.07	50.58	53.35
DCASE 2	45.15	47.08	50.28	52.23

2. Combining the foreground-background classifier with the sound event detection branch leads to higher detection scores.

Method	F1 score [%] validation		F1 score [%] evaluation	
	+ MF	+ HMMs	+ MF	+ HMMs
Baseline	34.8		38.1	
+ DA	**42.41**	**43.89**	**44.8**	**47.12**
+ FB	43.12	45.42	46.06	49.38
+ FB + DA	**45.68**	**47.77**	**50.79**	**53.10**
+ SEDFB	46.15	46.20	48.40	49.79
+ SEDFB + DA	**47.61**	**47.75**	**52.12**	**53.30**
DCASE 1	45.13	48.07	50.58	53.35
DCASE 2	45.15	47.08	50.28	52.23

3. The proposed domain adaptation further reduces data mismatch between synthetic and real soundscapes.