A LIGHTWEIGHT APPROACH FOR SEMI-SUPERVISED SOUND EVENT DETECTION WITH UNSUPERVISED DATA AUGMENTATION

Heinrich Dinkel, Xinyu Cai, Zhiyong Yan, Yongqing Wang, Junbo Zhang, Yujun Wang

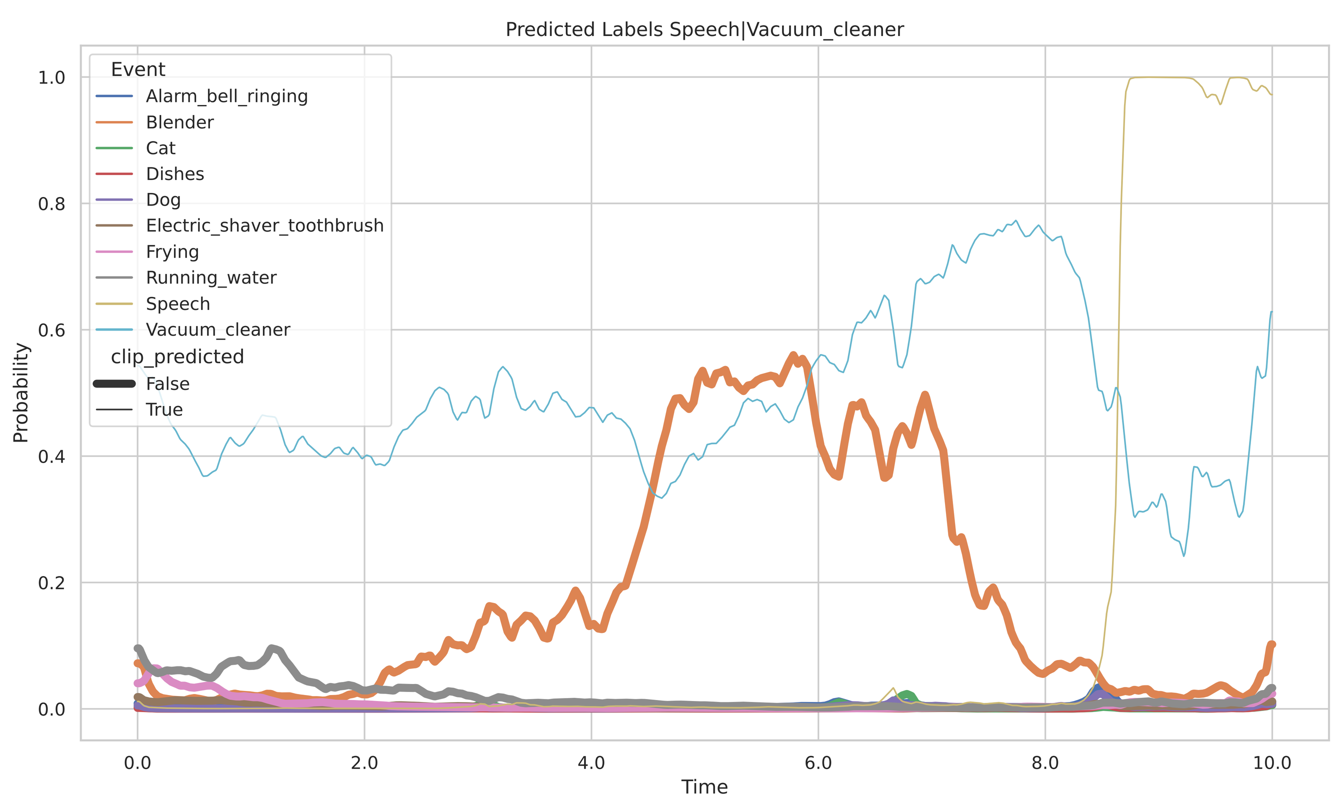
Xiaomi Corporation

Highlights

- Propose a lightweight (parameters) approach for semi-supervised sound event detection.
- A simple learnable clip-smoothing approach to enhance **consistency**.
- First time using unsupervised data augmentation (**UDA**) in SED.
- **7**th best approach in the challenge.
- **2**nd best approach from a single model.

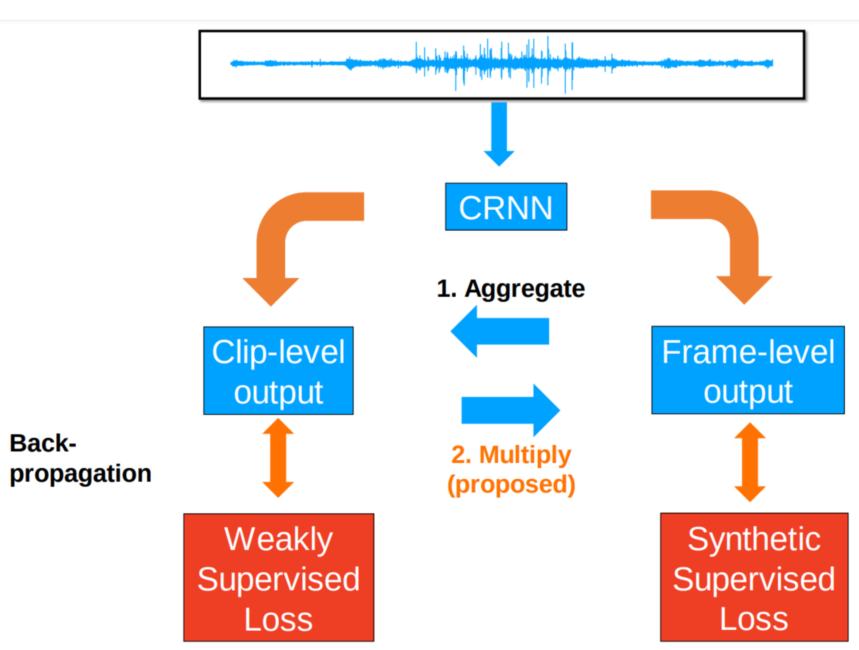
Problem statement

- Clip-level output is a (generally non-linear) combination of frame-level outputs.
- This leads to conflicting predictions between frame-level and clip-level outputs.



Learnable Clip Smoothing

Smoothing frame-level prediction by the clip-level probabilities. Learning how to smooth via the synthetic supervised loss.



Unsupervised data augmentation

Calculate a consistency loss for an unlabaled sample between its original and augmented variants. Augmentation is done on wave-level.

$$\begin{aligned} x^{\dagger} &= \text{Aug}(x), \\ \mathcal{M}(x) &\mapsto (\hat{y}, \hat{y}_t), \\ \mathcal{M}(x^{\dagger}) &\mapsto (\hat{y}^{\dagger}, \hat{y}_t^{\dagger}), \\ \mathcal{L}_{\text{UDA}}(x) &= \mathcal{L}_{\text{consistency}}(\hat{y}^{\dagger}, \hat{y}) + \mathcal{L}_{\text{consistency}}(\hat{y}_t^{\dagger}, \hat{y}_t). \end{aligned}$$

Dataset and Training

Aggregation function is Linear softmax:

$$\hat{y} = \frac{\sum_t \hat{y}_t^2}{\sum_t \hat{y}_t}$$

Training dataset consists of a weakly supervised dataset (weak), a strongly supervised synthetic (syn) dataset and an unlabaled (un) dataset.

$$\begin{aligned} \mathcal{D}_{weak} &= \{(x_1, y_2), (x_2, y_2), \ldots, (x_N, y_N)\}, \\ \mathcal{D}_{syn} &= \{(x_1, y_2), (x_2, y_2), \ldots, (x_M, y_M)\}, \\ \mathcal{D}_{un} &= \{x_1, \ldots, x_P\}. \end{aligned}$$

Our approach optimizes the following loss functions:

$$\begin{aligned} \mathcal{L}_{\text{sup}} &= \text{BCE}(\hat{y}, y), \{y, \hat{y}\} \in \mathcal{D}_{weak}, \\ \mathcal{L}_{\text{syn}} &= \text{BCE}(\hat{y}_t, y_t), \{y_t, \hat{y}_t\} \in \mathcal{D}_{syn}, \\ \mathcal{L}_{\text{unsup}} &= \mathcal{L}_{\text{UDA}}(x) = \text{BCE}(\hat{y}^{\dagger}, \hat{y}) + \text{BCE}(\hat{y}_t, \hat{y}_t), x \in \mathcal{D}_{un}. \end{aligned}$$

Development dataset results

Baseline results

Data	d'	E-F1	I-F1	**PSDS-1**	**PSDS-2**
Weak	2.28	22.71	49.06	15.17	33.47
+ Syn	2.23	30.39	49.63	19.01	28.12
++ Unlabel	2.47	32.11	52.14	26.87	42.19

With learn-able clip smoothing

Data	d'	E-F1	I-F1	**PSDS-1**	**PSDS-2**
Weak	2.27	22.99	49.14	19.98	46.57
+ Syn	2.21	35.31	54.84	29.85	47.34
++ Unlabel	2.50	37.21	57.12	34.41	54.90

Challenge results against competition

We achieved the 7th place in the DCASE2021 Task4 Challenge, **without** post-processing.

Model	**PSDS-1**	**PSDS-2**	**PSDS-Avg**	Post
Baseline	31.5	54.7	43.1	Median
1st	**45.2**	**74.6**	**59.9**	
2nd	44.2	67.4	55.8	
3rd	39.9	71.5	55.7	
3rd	41.9	68.6	55.2	Median
4th	41.6	63.7	52.6	
5th	41.3	58.6	49.9	
6th	37.0	62.6	49.8	
S1	36.1	58.4	47.2	
S2	37.3	58.5	47.9	-
S3	37.0	59.6	48.3	
S4 (Single)	33.9	50.4	42.1	
Ours (best)	**37.3**	**59.6**	**48.4**	-

Post-challenge results

Reevaluation of our results on the evaluation set **with median post-processing**.

Model	#Param (M)	**PSDS-1**	**PSDS-2**	**Score**	Single?
1st	14.3	**45.2**	**74.6**	**1.40**	N
2nd	20.2	44.2	67.4	1.32	Y
3rd	79.2	33.9	71.5	1.29	N
3rd	50.0	41.9	68.6	1.29	N
4th	119.8	41.6	63.7	1.24	N
S3	**3.4**	38.2	65.4	1.20	Y
S2	**2.7**	37.9	64.3	1.19	Y
5th	8.5	41.3	58.6	1.19	Y
S1	**2.0**	36.1	64.3	1.16	Y
6th	6.7	37.0	62.6	1.16	Y

Conclusion

- Learn-able clip-smoothing largely improves performance for PSDS-1 and 2.
- Successfully deployed UDA for SED.
- Most lightweight top-scoring model in the DCASE 2021 Task4 Challenge.