Self-trained Audio Tagging and Sound Event Detection in Domestic Environments

Janek Ebbers, Reinhold Haeb-Umbach

Paderborn University, Germany

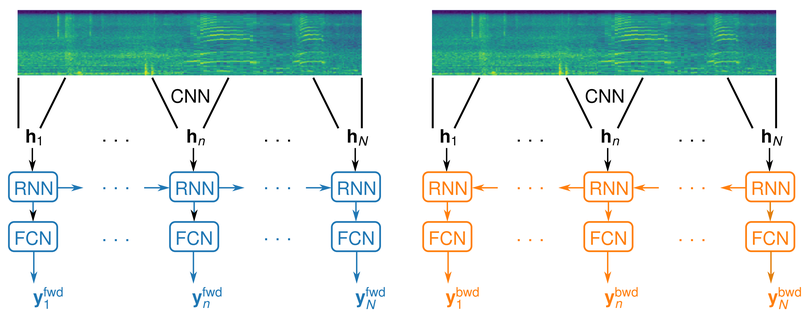
Overview

- Presentation of our system for DCASE 2021 Challenge Task 4: Sound Event Detection and Separation in Domestic Environments
- Our focus: Weakly labeled semi-supervised Sound Event Detection (SED)
- Advancement of our DCASE 2020 Challenge system:
  - Forward-Backward Convolutional Recurrent Neural Network (FBCRNN) for tagging (and strong pseudo labeling during training)
  - Followed by tag-conditioned SED (TCSED)
- Novelties:
  - Strong label loss (sll) in FBCRNN training to better exploit strongly labeled synthetic data
  - Multiple iterations of self-training (ST) (training → pseudo labeling → re-training → ...)
  - Exploration of convolutional, recurrent and transformer architectures for TCSED
  - Non-linear score transformation for smooth polyphonic sound detection (PSD)-ROC
- Results:
  - Fourth rank in terms of PSD scores (PSDSs)
  - Best performance in terms of collar-based F_1-score

Forward-Backward CRNN (FBCRNN)

Forward Tagging:

Backward Tagging:

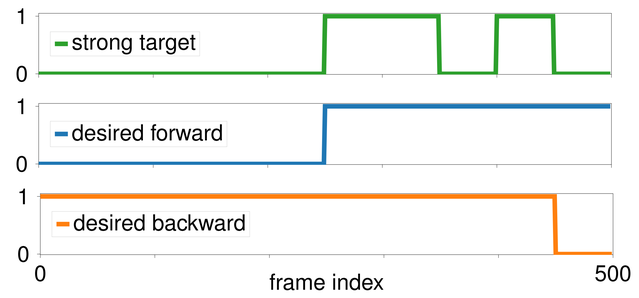


Shared CNN with two separate recurrent classifiers:

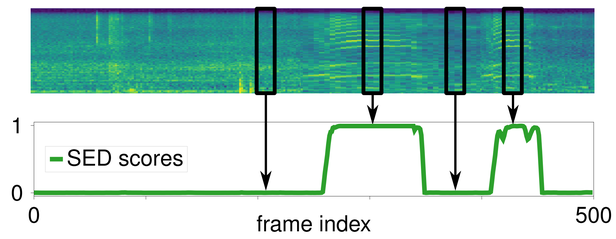
- One processing audio in forward direction (blue)
- Other processing audio in backward direction (orange)

FBCRNN Objectives

- Tag events as soon as it is seen in the input signal

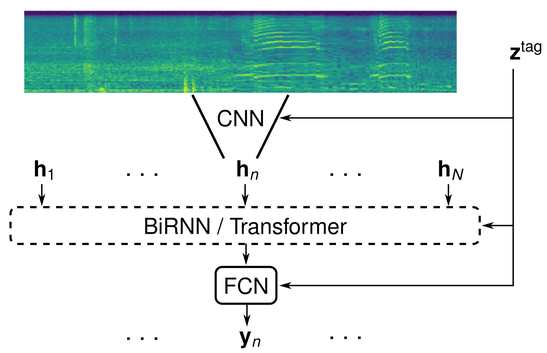


- If strong labels given: Compute separate losses for fwd & bwd classifiers w.r.t. above desired signals
- Else: use previously proposed weak label loss
- Perform SED by applying FBCRNN to small windows around each frame



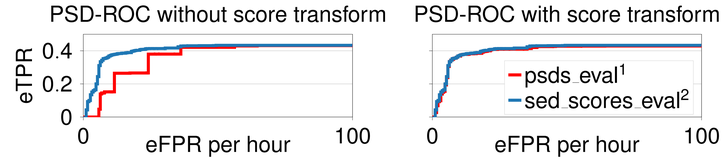
Tag-conditioned SED (TCSED)

- Use FBCRNN for tagging and strong pseudo labeling of weakly and unlabeled data
- Train TCSED models in strongly supervised manner



Non-linear Score Transformation

- PSDS metric: Normalized area under PSD-ROC
- In Challenge: PSD-ROC approximated using psds_eval[1] package with 50 decision thresholds linearly spaced between 0.01 and 0.99 (red curves)
- Recently published sed_scores_eval[2] package computes true PSD-ROC (blue curves)



- Without transform PSD-ROC is underestimated

[1] https://github.com/audioanalytic/psds_eval [2] https://github.com/fgnt/sed_scores_eval

Results

FBCRNN:

ST	It.	PSDS1	PSDS2	$F_1^{(collar)}$	$F_1^{(tag)}$
0		31.6±0.6	67.3±1.7	44.1±1.1	83.8±0.8
	w/o sll	29.0±2.1	67.2±3.0	41.2±1.9	83.3±0.6
1		36.4±0.5	68.0±1.0	49.1±1.4	84.6±0.3
	w/o psll	33.2±0.7	68.9±1.3	47.4±0.6	85.1±0.7
2		**38.2**±0.9	68.9±1.3	**50.9**±1.0	85.1±0.4
3		37.9±1.4	**70.2**±1.2	50.7±1.2	**85.6**±0.6

TCSED:

ST	It.	Model	PSDS1	PSDS2	$F_1^{(collar)}$
0		CNN	38.2±2.7	64.4±0.4	54.4±0.1
		CRNN	39.7±0.8	**66.7**±0.8	54.5±0.1
		CTNN	**40.9**±1.5	66.2±0.6	**55.7**±0.5
1		CNN	39.6±1.2	64.3±0.6	54.4±0.3
		CRNN	39.8±0.6	**67.0**±1.0	**56.6**±0.1
		CTNN	**40.8**±1.6	66.3±0.4	56.5±0.6

- Strong label loss improves temporal event localization
- Self-training improves FBCRNN performance
- RNN / Transformer layers improve TCSED performance
- No significant improvement due to TCSED self-training

Challenge (8 FBCRNNs followed by 6 TCSED models):

	eval-public			eval-2021		
Model	PSDS1	PSDS2	$F_1^{(collar)}$	PSDS1	PSDS2	$F_1^{(collar)}$
Baseline	35.9	59.6	40.8	31.5	54.7	37.3
Winner	**51.7**	**77.8**	57.4	**45.2**	**74.6**	52.3
FBCRNN	40.6	70.7	52.4	-	-	-
TCSED	45.5	68.4	**59.6**	41.6	63.7	**56.7**

Our system ...

- significantly outperforms baseline w.r.t. to all metrics
- is outperformed by winner system w.r.t. PSDSs
- achieves best performance w.r.t. collar-based F_1-score

Funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) - Project 282835863