A HYBRID SYSTEM OF SOUND EVENT DETECTION TRANSFORMER AND FRAME-WISE MODEL FOR DCASE 2022 TASK 4

Yiming Li[1, 2], Zhifang Guo[1, 2], Zhirong Ye[1, 2], Xiangdong Wang[1], Hong Liu[1], Yueliang Qian[1], Rui Tao[3], Long Yan[3] , Kazushige Ouchi[3]

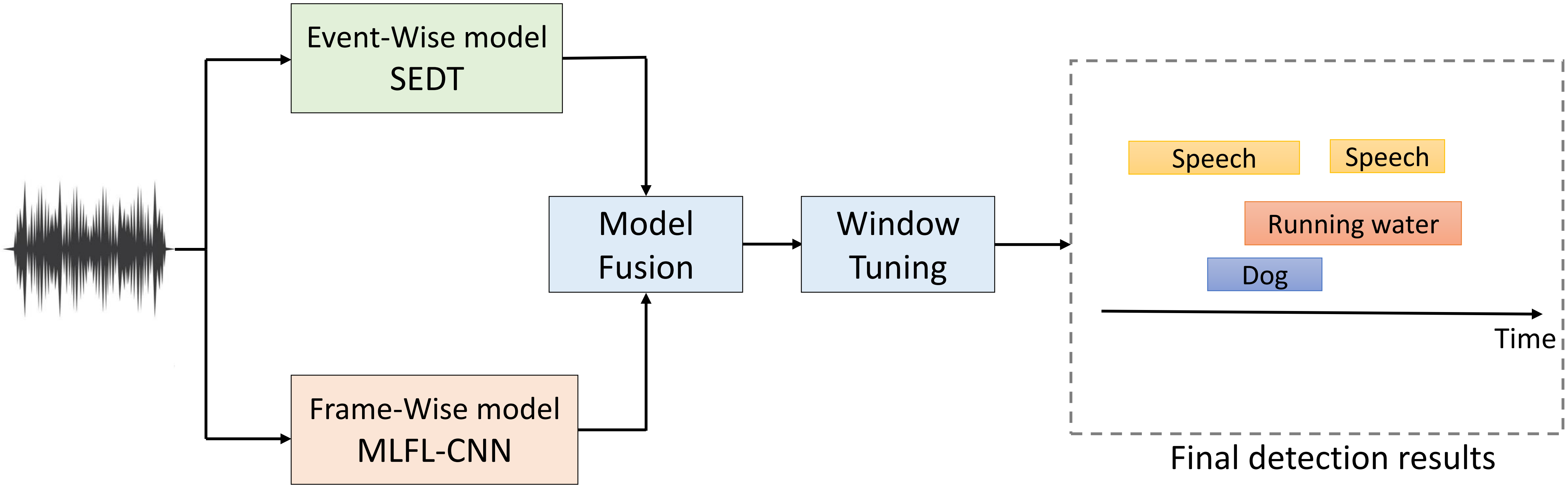
[1] Institute of Computing Technology, Chinese Academy of Sciences
eamon.y.li@gmail.com, {guozhifang21s, yezhirong19s, xdwang, hliu, ylqian}@ict.ac.cn
[2] University of Chinese Academy of Sciences
[3] Toshiba China R&D Center
{taorui, yanlong}@toshiba.com.cn, kazushige.ouchi@toshiba.co.jp

Outline

- Overview
- Method
- Experiments
- Conclusions

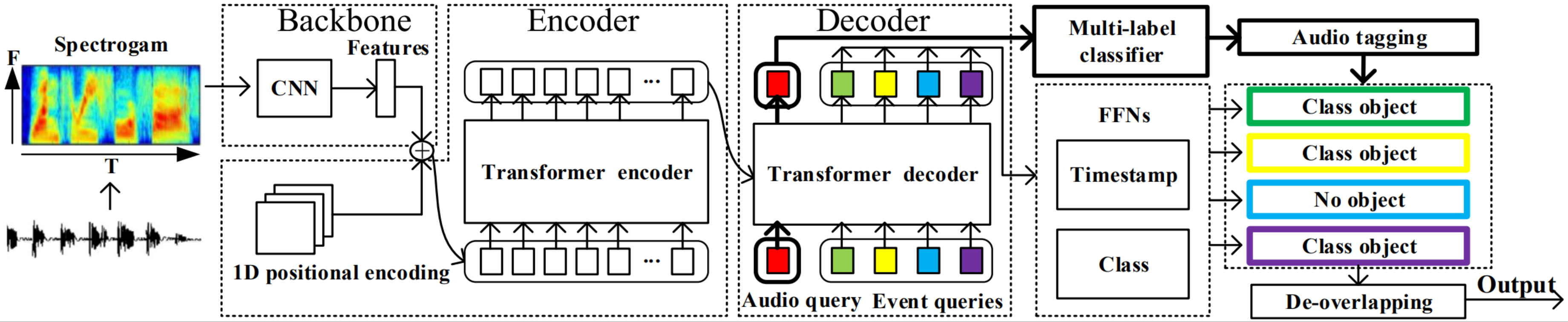
Overview

Pipelines of the Hybrid System



Method

Sound Event Detection Transformer (SEDT)



- Aims

  To learn mapping function from input spectrogram to event boundaries directly

- Key Components

  - Backbone and Encoder

    Extract primary features from input spectrograms

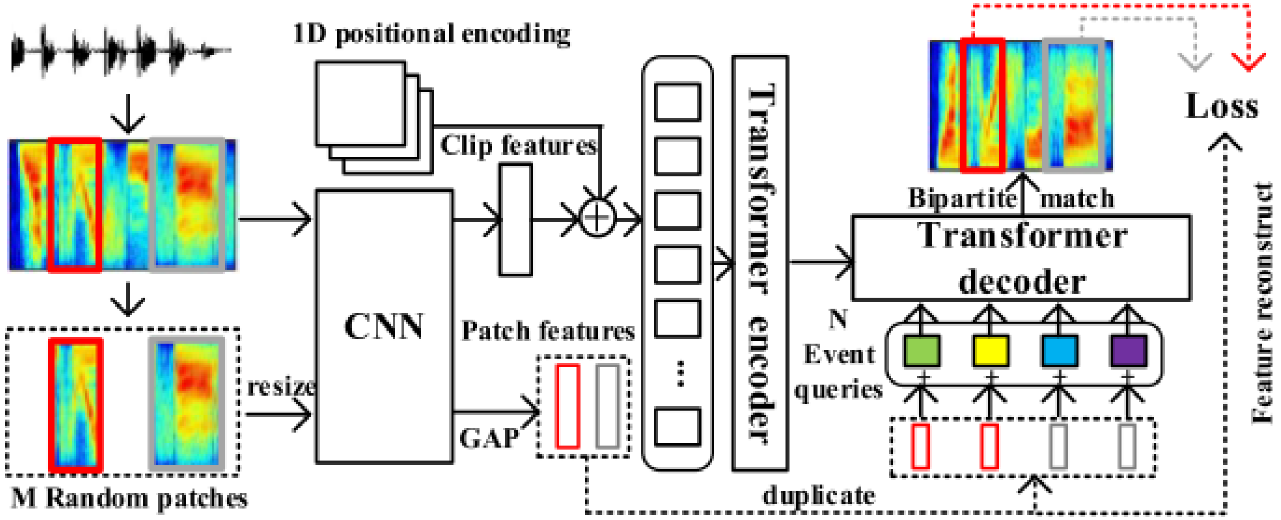
  - Audio / Event Queries and Decoder

    Gather information from the encoder outputs via encoder-decoder cross-attention mechanism to generate event-level representations (event queries) and clip-level representations (audio queries)

  - Multi-label classifier and FFNs

    Transform the event-level and clip-level representations into event detection and audio tagging results

Self-supervised SEDT (SP-SEDT)



Patch Localization and **Feature Reconstruction** as pre-training task

Randomly crop spectrogram along the time axis to obtain several patches, and then pre-train the model to predict the corresponding locations of the patches as well as reconstruct the features

Semi-supervised SEDT (SS-SEDT)

Require: $\mathcal{B}_L$ = labeled batch, $\mathcal{B}_U$ = unlabeled batch
Require: $S_\theta(x)$ = student model, $T_{\theta'}(x)$ = teacher model
Require: $A_w(x)$ = weak augmentation function
Require: $A_s(x)$ = strong augmentation function
Require: α = learning rate, γ = EMA ratio
Require: $\mathcal{L}$ = loss function
Ensure : θ, θ'

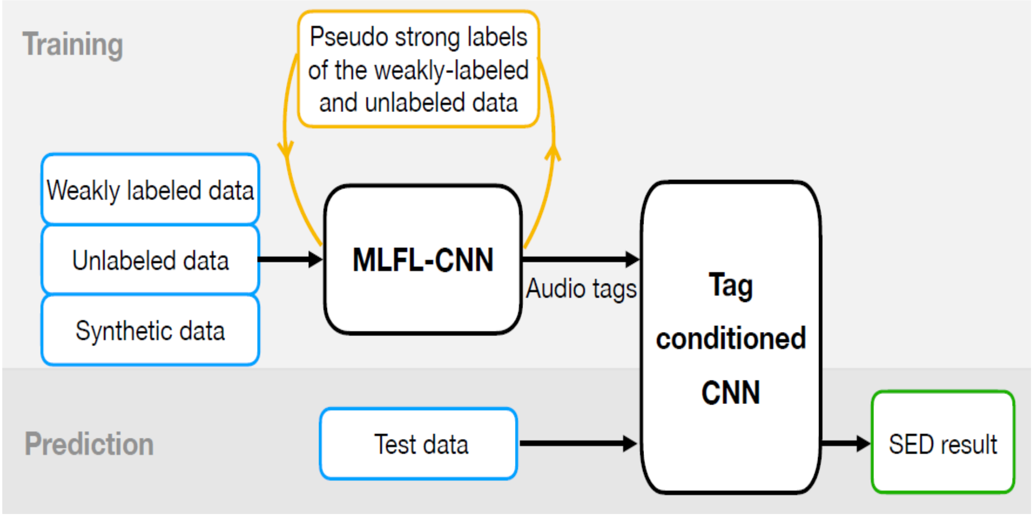
1 **for** $i \rightarrow 1$ **to** *max_epochs* **do**
2   **foreach** $\mathcal{B}_L \cup \mathcal{B}_U \in \mathcal{B}$ **do**
3         $\mathcal{J}_{\text{sup}} \leftarrow \frac{1}{|\mathcal{B}_L|} \sum_{(x_i, y_i) \in \mathcal{B}_L} \mathcal{L}\left(S_\theta\left(A_w(x_i)\right), y_i\right);$
4   **foreach** $x_i \in \mathcal{B}_U$ **do** $y_i \leftarrow T_{\theta'}(A_w(x_i));$
5         $\hat{\mathcal{B}} \leftarrow \mathrm{M}ixup(\mathcal{B}_L, \mathcal{B}_U);$
6         $\mathcal{J}_{\text{unsup}} \leftarrow \frac{1}{|\hat{\mathcal{B}}|} \sum_{(\hat{x}_i, \hat{y}_i) \in \hat{\mathcal{B}}} \mathcal{L}\left(S_\theta\left(A_s(\hat{x}_i)\right), \hat{y}_i\right);$
7         $\theta \leftarrow \theta - \alpha\left(\frac{\partial \mathcal{J}_{\text{sup}}}{\partial \theta} + \frac{\partial \mathcal{J}_{\text{unsup}}}{\partial \theta}\right);$
8         $\theta' \leftarrow \gamma\theta' + (1-\gamma)\theta;$
9   **end**
10 **end**

Proposed SSL techniques

- Line 5: **Mixup of Labeled and Unlabeled Data** to improve the model robustness to pseudo annotation noise
- Line 4, 6: **Asymmetric Augmentation** to regularize the consistency between student and teacher under data perturbations
- Line 6: **Focal Loss** to handle the unbalanced event categories in SED
- Line 8: **EMA model** to generate more reliable event-level pseudo labels

Metric Learning and Focal Loss CNN (MLFL-CNN)

MLFL-CNN is a traditional frame-wise model which obtains frame-level classification probabilities and then applies pooling mechanisms to acquire final event detection results



MLFL-CNN contains Three branches

- Trained in a **multi-branch** manner to exploit the heterogeneous dataset as well as narrow the gap between real and synthetic data
- A **tag-conditioned CNN** is utilized to generate final predictions according to audio tags provided by MLFL-CNN
- Traditional **mean-teacher** framework is adopted

Model Fusion and Window Tuning

Model Fusion

Derive Class-specific PSDS ⇨ Get ensemble weights for each model i and class c

$$\mu_{\mathrm{TP,c}} = r_{\mathrm{TP,c}} \quad \sigma_{\mathrm{TP,c}} = r_{\mathrm{TP,c}} - \mu_{\mathrm{TP,c}}$$

$$\mathrm{eTPR}_c: \quad r_c(e) \triangleq \mu_{\mathrm{TP,c}}(e) - \alpha_{ST} * \sigma_{\mathrm{TP,c}}(e)$$

$$\mathrm{PSDS}_c \triangleq \frac{1}{e_{\max}} \int_0^{e_{\max}} r_c(e)de$$

$$w_{i,c} = \frac{\mathrm{PSDS}_{i,c}}{\sum_{i=1}^{N} \mathrm{PSDS}_{i,c}}$$

$$\hat{p_c} = \sum_{i=1}^{N} w_{i,c} * p_{i,c}$$

Window Tuning

For a given event class, enumerate window length and find the optimal one to optimize PSDS

$$wl_c^* = \arg\max_{wl_c} \frac{\mathrm{PSDS}_c}{\mathrm{PSDS}}$$

Experiments

System Performance

Official Baseline

Optimize PSDS1

Optimize PSDS2

Table 1: The PSDS on the validation set

System	Extra data	PSDS1	PSDS2
Baseline 1		0.336	0.536
Baseline 2	✓	0.351	0.552
System 1	✓	0.449	0.645
System 2	✓	0.115	0.816
System 3		0.420	0.618
System 4		0.099	0.783

Our systems outperform the official baseline to a large extent, and finally ranked 6th/9th in the final challenge.

Ablation Study

Table 2: Ablation study on techniques in SS-SEDT

MU	FL	AA	EMA	PSDS1	PSDS2
	✓	✓	✓	0.372	0.570
✓		✓	✓	0.349	0.540
✓	✓		✓	0.369	0.566
✓	✓	✓		0.357	0.538
✓	✓	✓	✓	0.388	0.573

All proposed techniques can improve the performance of SS-SEDT

Table 3: Ablation study on window tuning and model fusion

Id	Model	MF	WT	PSDS1	PSDS2
1	Single SEDT			0.415	0.582
2	Ensemble SEDT			0.431	0.607
3	Single frame			0.349	0.668
4	Ensemble frame			0.392	0.673
5	Hybrid system	✓		0.437	0.740
6	Hybrid system	✓	✓	0.449	0.816

- Frame-wise model and SEDT can supplement each other while combining together
- Window Tuning and Model Fusion have beneficial effects

Conclusions

Summary

We developed a framework to fuse the detection results of the frame-wise model and event-wise model, which leads to improved PSDS scores on the DCASE2022 validation set

Thank You