Improved student model training for acoustic event detection models

Anthea Cheung, Qingming Tang, Chieh-chi Kao, Ming Sun, and Chao Wang
Amazon Alexa

Introduction

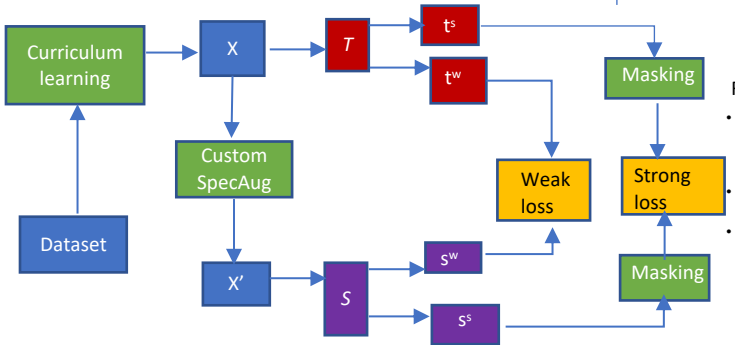
- Acoustic event detection (AED) is the task of predicting sound events and their time boundaries.
- State-of-the-art models are usually ensembles, comprised of multiple layers of convolutional layers, or contain custom architecture.
- This work: applied novel knowledge distillation techniques on DCASE 2019 task 4[1] dataset for acoustic event detection
  - Curriculum learning
  - Custom SpecAugment data augmentation
  - Loss masking
- After distillation, obtained strong event-based F1-score of 42.7%, compared to 34.7% when training with a generic knowledge distillation method. Performance matches top submission of challenge

Data

- Dataset: 2019 DCASE task 4 dataset (sound event detection)
- 10 different event types (speech, dog, cat, alarm/bell/ringing, dishes, frying, blender, running water, vacuum cleaner, electric shaver/toothbrush)
- Contains weakly labeled (event predictions only), strongly labeled (onset and offset times for events included), and unlabeled data
- Input: extracted mel spectrogram features: 64 (20) frequency bands x 500 time frames for teacher (student) model

Model training

- Teacher model: top performing submission for DCASE 2019 task 4 (Lin et al 2019)[2]
  - Ensemble of 6 CNN models
  - Includes custom architecture (“disentangled features”)
- Student model:
  - 3-layer LSTM followed by fully connected layer for frame predictions
  - Attention mechanism over frame predictions to obtain clip-level (weak) predictions



Curriculum learning

- Tagging predictions from teacher model t are generated for unlabeled dataset and used as targets for student model. Let t_c = teacher prediction for class c
- We use heuristic

  $\mu(t, s) = \max_c (|t_c - s_c|)$,

- i.e. the maximum difference in the teacher and student scores across all classes. This is a measure of how “difficult” the sample is.
- At each generation (5 epochs) we rank the unlabeled dataset by μ(t, s)
- Two schemes for adding samples from unlabeled dataset:
  - 1) Easier samples first: add bottom 20%, then bottom 40%, then all samples
  - 2) Harder samples first: add top 20%, then top 40%, then all samples

Custom SpecAugment method

- Motivation: for sound detection, the different classes have varying average durations. For example, vacuum cleaner have longer durations than durations
- We devised a custom SpecAugment method based on this observation, by using variable-length time masks
- For each clip, randomly pick top 2 events from the label. Apply time mask of 0.25*median duration of those events to spectrograms

Loss masking

- Observation: the strong labels (500 frames x 10 classes) are sparse.
- When learning from the strong labeled set (or strong pseudolabels generated by the teacher model), we want to focus more on the positive frames
- We address the sparsity issue with two types of masking in the loss function:
  - Event masking: only events that are present in the clip contribute to the strong loss
  - Segment masking: only frames containing sounds in the clip with 12ms buffer before/ after contribute to the strong loss

Results

- The results show that the best performance is attained by adding easier samples first, with a best event-based macro F1 score of 42.7%, on par with the best performing challenge submission. We compared our results to the top submission in the challenge (Lin ICT 3)
- Comparatively speaking, we find the custom SpecAug works best, followed by vanilla SpecAug and no data augmentation.
- The best results were achieved using event masking (EM), followed by segment masking (SM) and no masking (NM).

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, Oct. 2019. [Online]. Available: https://hal.inria.fr/hal-02160855
2 L. Lin and X. Wang,“Guided learning convolution system for dcase 2019 task 4,” Institute of Computing Technology, Chinese Academy of Sciences, Beijing, China, Tech. Rep., June 2019.

Experiment	Best val F1	Best test F1	Mean ± sd val F1	Mean ± sd test F1
Lin ICT 3	45.3	42.7	N/A	N/A
Easier first + event mask + custom SpecAug	41.6	42.7	40.7 ± 0.6	41.3 ± 0.9
Event masking + custom Specaug	41.3	42.5	40.3 ± 0.7	41.0 ± 1.1
Vanilla KD	34.1	34.7	33.1±0.7	33.5±0.9
Easier first + event masking+ standard Specaug	40.7	41.6	40.0± 0.4	40.9± 0.6
Easier first + event masking + no Specaug	40.2	41.2	39.5 ± 0.4	39.9 ± 0.7
Easier first + segment masking + custom Specaug	39.9	40.0	39.2± 0.4	39.1 ± 0.5
Easier first + no masking + custom specaug	35.7	34.5	33.9 ± 0.8	32.9 ± 0.8

Comparison	t-statistic	Statistically significant at :		
		α = 0.2	α = 0.05	α = 0.01
Easier first vs all	1.413	Y	N	N
Harder first vs all	0.143	N	N	N
Easier vs harder first	1.483	Y	N	N
Custom SpecAug vs standard	2.802	Y	Y	N
Custom SpecAug vs none	7.055	Y	Y	Y
Standard SpecAug vs none	2.989	Y	Y	Y
Event mask vs segment	6.800	Y	Y	Y
Segment mask vs none	19.875	Y	Y	Y
Event mask vs none	23.021	Y	Y	Y

Statistical significance and conclusion

- For each of the techniques, we perform a t-test on the validation F1 scores of ten trials of each .
- We find that adding easier samples first in the pseudolabeled dataset is statistically significant at the α = 0.2 level, while the other techniques are significant at the α = 0.05 level.
- Progressively applying pseudolabeled samples, using variable-length time masking in SpecAug augmentation, and applying event masking to the loss function all contribute to a single model with a 42.7% macro event-based F1- score on the test set, matching state of the art performance of 42.7%.