TEACHER-STUDENT TRAINING FOR ACOUSTIC EVENT DETECTION USING AUDIOSET

Ruibo Shi[†] *Raymond W. M. Ng*[†] *Pawel Swietojanski*[*]

[†]Emotech Labs, London, UK
[*] School of Computer Science and Engineering, The University of New South Wales, Sydney, Australia

ABSTRACT

This paper studies Acoustic Event Detection (AED) systems and the problem of their rapid and easy customisation to arbitrary deployment scenarios. Due to inherent challenges related to annotation processes of AED data (time-consuming and error-prone due to often unclear time-stamping), most of the available large-scale datasets for AED are released with weak clip-level labels, which also affects how one should design weakly-supervised training procedures. In this paper, we investigate a teacher-student training approach of learning low-complexity student models, using large teachers. We first show that state-of-the-art performance can be achieved by a Convolutional Neural Network (CNN) model with appropriate attention mechanism. Then we describe a framework that enables learning arbitrary small-footprint, generic or domain-expert, AED systems from generic teachers. We carry experiments on *Audioset* - a large-scale weakly labelled dataset of acoustic events.

Index Terms— Acoustic Event Detection, Weakly-supervised training, Teacher-Student Training, Attention

1. INTRODUCTION

Acoustic Event Detection (AED) or Sound Event Detection systems aim to understand the content of acoustic signals by automatically predicting single or multiple *simultaneous* events. It has attracted much interest due to its wide range of potential applications in activity monitoring, multimedia inspection and public surveillance, *etc.*. The application of AED, however, has been limited due to the insufficient amounts of training data and their relatively small coverage of event classes. Most relevant datasets used to be domain-specific [1,2]. Recently, Google released a large-scale corpora *Audioset* [3], in which events are annotated at the clip level (up to 10 seconds long each). Precise timing information is however not available, requiring weakly supervised learning techniques that are capable of recovering higher temporal resolution for runtime deployments.

Weak-label problem is typically approached by either applying a strong label assumption (SLA) where the annotations are assumed to be valid across the entire clip [4], or as a Multiple Instance Learning (MIL) problem, in which case the audio clip is treated as a bag of units with smaller time duration (*i.e. segments*) [5]. In this paper, we apply an attention mechanism similar to the system proposed in [6] that learns how to attend to relevant parts of audio clips given the class, and we demonstrate state-of-the-art performance on *Audioset*.

Large-scale AED systems are expected to operate accurately for different subsets of detected events and across various acoustic conditions, often requiring large and complex underlying models. This becomes a challenge when the goal is an embedded deployment on a device with limited compute and memory footprint. In this paper, our goal is to build small-footprint AED systems that can be quickly tailored to the task of interest (*i.e.* customisable decision space, and flexibility in temporal resolution), and can run with low latency. To achieve this goal, in contrast to other published systems such as [6] that are based on embeddings produced by a pre-trained neural feature extractor, we build end-to-end systems using standard acoustic features as input. End-to-end systems allow us to conduct Teacher-Student (TS) training to learn a small model that significantly reduces the overall resource footprint while maintaining a good performance.

The primary contributions of this paper include 1) building an end-to-end AED system with attention mechanism that achieves state-of-the-art performance; 2) proposition of a teacher-student training framework for learning generic or domain-expert small-footprint AED systems.

2. RELATED WORKS

While AED has been historically addressed using Gaussian mixture models [7], hidden Markov models [8] or support vector machines [9], typically trained on mel-frequency cepstral coefficients acoustic features, recent progress in deep neural networks (DNN) has inspired many DNN-based AED systems, including variants of both feed-forward (convolutional) (CNN) and recurrent neural networks (RNN) [4, 5, 10, 11]. Hershey et al. [4] compares several AED-tuned CNN architectures that were proved to work well in computer vision tasks, *i.e.* AlexNet [12], VGG [13], Inception-V3 [14] and ResNet-50 [15]. CNNs were found to outperform fully connected non-recurrent models, and the state of the art performance on *Audioset* (with 485 classes) was obtained with embeddings produced by the ResNet-50 architecture. Above

works rely on the SLA approach assuming the presence of reliable labels in all segments in a clip, which particularly degrades the performance for temporally short events (*i.e.* gunshots, glass breaking, doors). Further analyses of the effect of SLA can be found in [16] where the authors find that model performance tends to degrade proportionally to how short an event in a clip is (under the SLA assumption). This issue can be mitigated with the attention mechanism [6, 17], where one or more attention modules parallel to the classification layers are co-learned to infer the unobserved latent variables representing the relative importance of each segment. While our attention model is similar to the system in [6], our models were trained directly on raw acoustic features rather than embeddings.

Teacher-Student (TS) training for knowledge distillation has been proposed in [18–20] to learn a model with fewer parameters to approximate the function transformation learned by the large model. TS training has been applied successfully in automatic speech recognition [19] and speaker recognition [21]. To the best of our knowledge, our work is the first to study knowledge distillation for weakly-supervised AED systems with attention using *Audioset* labels, including in-place domain specialisation of student models to arbitrary subsets of the original 527 *Audioset* classes.

3. APPROACH

RNNs are capable of modelling long-term sequential patterns [22] and bidirectional RNNs were in particular proven to be effective for AED [23, 24], but they require the entire temporal sequence to be scanned, incurring runtime latency. Furthermore, while many models have been proposed for AED under the weakly-labelled premise, little effort has been put in place to study lightweight models under constrained runtime latency regime. In this work, we propose a weakly-supervised TS training framework for AED, which, depending on the design choice, transforms a large model to small models that receive inputs of an arbitrarily short duration independent of the clip length used in training, thereby providing more accurate temporal information and the student can be easily adapted to work with only a subset of classes produced by the large-scale teacher model.

3.1. CNN Model with Attention

Our proposed teacher CNN model with attention is depicted in Figure 1. Given the m-th audio clip, represented as a matrix of log-mel spectrogram features $X^{(m)} \in \mathbb{R}^{W \times F}$, where W and F are the number of log-mel frames and mel-bins extracted from the clip, we divide $X^{(m)}$ into a group of segments each as $X_n^{(m)} \in \mathbb{R}^{T \times F}$, $n = 1, ..., B$ with $B = W/T$. $X_n^{(m)}$ forms an input acoustic feature to a CNN.

In this work, we use the same VGGish CNN that was used to generate the published *Audioset* embeddings [4]. To recap, block C1 to C4 each is followed by a 2×2 max pooling layer and ReLU activation with 3×3 filters used throughout. Specifically, the number of filters in each block is {C1: 64; C2: 128; C3: 256, 256; C4: 512, 512}. The output of C4 is flattened and fed into block E that implements three feed-forward layers of 4096, 4096 and 128 dimensions respectively and all with ReLU activation, resulting in an 128-dimensional embedding.

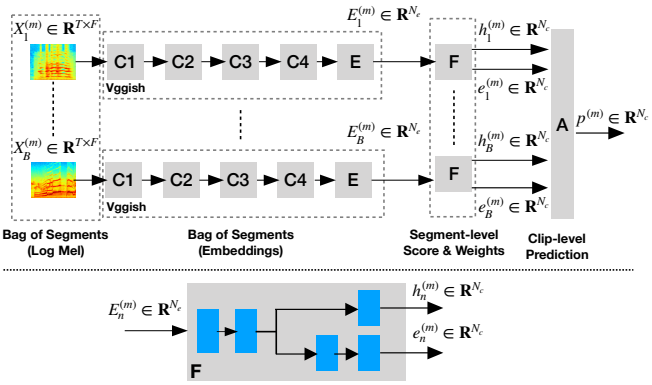

Fig. 1. Top: Attention CNN teacher model architecture; Bottom: Connection of a feed-forward layers in block F.

After the VGGish CNN, all segment-level embeddings from the m-th clip, $E_n^{(m)}$, $n = 1, ..., B$, are fed into the block F, implemented as two hidden feed-forward layers with ReLU activations (1000 units each) $g_n^{(m)} = f(E_n^{(m)})$ followed by two parallel output layers. The first output produces sigmoid normalised scores $h_n^{(m)} = \gamma(g_n^{(m)})$, while the second attention weights $e_n^{(m)} = \omega(g_n^{(m)})$. The attention module uses one 1000 unit ReLU hidden layer followed by a softmax output layer. Finally, the clip-level posteriors are obtained by combining the segment-level outputs in the block A as follows:

$$A(X^{(m)}) = \frac{1}{\sum_{n=1}^{B} e_n^{(m)}} \sum_{n=1}^{B} e_n^{(m)} \cdot h_n^{(m)} = \sum_{n=1}^{B} \alpha_n^{(m)} \cdot h_n^{(m)} \quad (1)$$

In a weakly supervised training the ground truth labels are given at the clip level and multiple events can be simultaneously present in a clip. Therefore, we use a multi-label cross-entropy objective for training after the clip-level posteriors are obtained. Eq. 2 below shows the binary cross-entropy loss for the class k.

$$l(p_k^{(m)}, y_k^{(m)}) = -y_k^{(m)} \cdot \log(p_k^{(m)}) - (1 - y_k^{(m)}) \cdot \log(1 - p_k^{(m)}) \quad (2)$$

where $p_k \in [0, 1]$ is the k^{th} element of the posterior output $A(X^{(m)})$ from Eq. (1) and y_k is the hard label that is either 1 if k^{th} event is present and 0 otherwise. The multi-class objective is obtained by taking the mean across all considered classes:

$$L(X^{(m)}, y^{(m)}) = \frac{1}{N_c} \sum_{k=1}^{N_c} l(p_k^{(m)}, y_k^{(m)}), \quad (3)$$

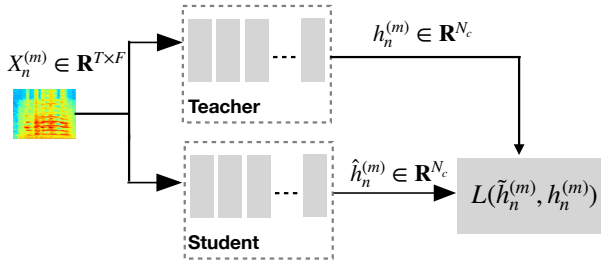

Fig. 2. Teacher-Student training framework: the same segment-level input mel-log spectrograms $\hat{X}$ is fed to both the teacher and student models and the student learns to approximate h generated by the teacher.

where $N_c = |\mathbb{S}|$ denotes the cardinality of the target label set.

Such a CNN architecture has a large modelling capacity, which comes at a high memory and computation cost of approximately 7.5×10^7 parameters.

3.2. Teacher-Student Training

For many practical AED applications, one typically seeks for a model characterised by a small runtime footprint, and tailored to the given deployment domain (i.e. only few classes are required). Those two objectives can be achieved together by the TS training framework. Specifically, a teacher model may be trained and used to produce soft targets on a transfer data set. A small student model then learns to mimic predictions from the teacher.

In a student model, we abandon the two-level hierarchy with "clips" against "segments". In training, data is presented as independent segments to the model. We treat the n^{th} segment in clip $X^{(m)}$ independently, i.e. $B = 1$. Thus, Eq.(1) boils down to

$$A(X_n^{(m)}) = \alpha_n^{(m)} \cdot h_n^{(m)} = h_n^{(m)} \tag{4}$$

as $\alpha_n^{(m)} = e_n^{(m)} / \sum_{n=1}^{B} e_n^{(m)} = 1$. Given Eq.(4), Figure 2 shows our proposed TS training framework where both the trained teacher and the to-be-trained student model are conditioned on the same segment input. The student is then tasked to approximate the soft targets generated by the teacher model by minimising the Kullback-Leibler Divergetnce (KLD) between its output $\hat{h}_n^{(m)}$ and the soft target $h_n^{(m)}$.

With the proposed TS framework, one can quickly train the student models to work on a sub-set of classes of interest, i.e. $\hat{\mathbb{S}} \subseteq \mathbb{S}$ by extracting only the interested classes from the generated soft targets. Furthermore, since neither weak nor strong labels are required for this TS training, one may also make use of the huge amount of unlabelled data in the wild by leveraging the generalisation power of the teacher model. Finally, during offline testing, a clip can be easily batched along segments with overlapping shifts before presenting to the model to further reduce latency.

4. EXPERIMENTS AND RESULTS

4.1. Datasets

We carry our experiments on a large-scale weakly labelled *Audioset* [3] corpora. It has three partitions: (1) balanced training set (approx. 20K clips), (2) an evaluation set (approx. 20K clips) and (3) an unbalanced training set (approx. 2M clips). Hereafter, we will refer to those as *Audioset-bal*, *Audioset-eval* and *Audioset-unbal*, respectively. In addition, we sampled a balanced set of 15K clips from *Audioset-unbal* and made *Audioset-val* for validation purposes. In total, *Audioset* contains approximately 2 million audio clips with 527 classes following a hierarchical ontology. Since the raw waveforms were not readily available (only embeddings from a pretrained VGGish model are officialy released), we crawled the corresponding raw audios from the Internet. Some clips were not available at the time of download, thus our target dataset is approximately 8% smaller when compared to the original.

To investigate domain-expert scenarios where the model is tasked to detect only a small number of classes, we manually selected 8 representative 10-class subsets. For each subset, we extracted in-domain samples from *Audioset-bal*, *Audioset-val* and *Audioset-eval* and formed *domain-train*$_s$, *domain-val*$_s$ and *domain-eval*$_s$ where $s = 1, .., 8$. Whereas the first 5 subsets are characterised by highly distinctive classes that differ in annotation quality as assessed by Google, the other 3 subsets contain similar sound events corresponding to musical instruments, bells and human voices.

4.2. Experiments

Log-mel features are extracted with 25ms window and 10ms hop size using 64 mel-filters per frame, and each segment consists of 96 such frames, i.e. $T = 96$ and $F = 64$ (Section 3.1), as a result our model updates detection scores every 0.96-second.

We trained two end-to-end AED teacher models. The first is a CNN model with attention as described in Section 3.1, and the second is trained with SLA, which resembles the same system architecture but with the attention module (block F in Figure 1) removed. To speed up training, the 5-block VGGish feature extractor is bootstrapped from a checkpoint pretrained on *YouTube-8M* [4] and fine-tuning is enabled during the entire training process. Specifically, we trained these two models using both *Audioset-bal* and *Audioset-unbal* for 3 epochs and the results are reported on *Audioset-eval*.

To evaluate the proposed TS training framework described in Section 3.2, we compare three small-footprint models of the same architecture trained using weak labels by (1) enforcing SLA, (2) distilling from the teacher trained with SLA and (3) distilling from the attention model respectively. These experiments are repeated for both full generic *Audioset* (527-class) and the 8 domain-expert scenarios (10 classes

each). The model consists of 3 blocks of CNN followed by 4 fully connected feed-forward layers. Given that 3×3 filters and ReLU activation are applied throughout the CNN blocks, the number of filters in each block is {C1: 16, 32, 64; C2: 64; C3: 128}. C1 and C2 are each followed by a 2×2 max pooling layer and the C3 is finished by a global max pooling, resulting in a 128-dimensional output. The activations are then connected to a 4-layer DNN with 64, 256, 256, and $\hat{N}_c$ units respectively, where $\hat{N}_c$ is the number of the output classes. All hidden layers have ReLU non-linearity, except for the output layer that uses Sigmoid activation. Depending on $\hat{N}_c$, the number of parameters of the student model is approximately 3×10^5. We train the generic / domain-expert small models using *Audioset-bal* / *domain-train*$_s$ until convergence on *Audioset-val* / *domain-val*$_s$ with a patience of 15 epochs and report results on *Audioset-eval* / *domain-eval*$_s$.

Model performance is evaluated by the balanced average Area Under Curve (AUC) of the Receiver Operating Curve (ROC) and the balanced mean Average Precision (mAP) [4]. While the teacher model with attention uses the segment-to-clip combination strategies described in Section 3.1 and Figure 1, for the teacher model trained with SLA and student models we derive clip-level output by averaging the predicted segment-level scores across a clip. Reported domain-expert student model results are an average over the eight representative 10-class subsets.

4.3. Results

The results of teacher models are shown in Table 1. The attention model t-CNN-ATT outperforms the model trained with SLA, i.e. t-CNN-SLA, in terms of both AUC and mAP.

Model Name	AUC	mAP
t-CNN-SLA	0.961	0.332
t-CNN-ATT	0.965	0.349

Table 1. Comparison of teacher models.

Table 2 reports the performance of the three small-footprint models, where s-CNN-1 is directly trained from the clip-level weak labels by enforcing SLA across segments, s-CNN-2 and s-CNN-3 are distilled from t-CNN-SLA and t-CNN-ATT teacher models respectively.

4.4. Analysis

Our end-to-end CNN model with attention (t-CNN-ATT) achieves 0.965 / 0.349 of AUC / mAP, outperforming the Google baseline (with 485 classes) at 0.959 / 0.314 [4]. In addition, t-CNN-ATT directly trained on raw acoustic features gives a better mAP when compared to the attention model trained on VGGish embeddings (0.965 / 0.327) [6]. Considering also the competitive performance in the model trained with SLA (t-CNN-SLA), we believe the end-to-end training approach, i.e. fine-tuning the VGGish blocks as the entire model is learned, has contributed to the increase of model robustness.

Teacher Model	Model Name	AUC	mAP
N/A	s-CNN-1	0.917	0.139
	s-CNN-1(domain)	0.796	0.479
t-CNN-SLA	s-CNN-2	0.930	0.197
	s-CNN-2(domain)	0.879	0.572
t-CNN-ATT	s-CNN-3	0.937	0.211
	s-CNN-3(domain)	0.882	0.582

Table 2. Comparison of small-footprint models

In TS training for the 527-class student models, we utilised teacher targets from the t-CNN-SLA and t-CNN-ATT models. Compared to student model training from scratch with SLA, i.e. s-CNN-1, the student model with the same number of parameters (s-CNN-2) benefits from the soft targets provided by the teacher model, and the resultant AUC and mAP are 0.013 and 0.058 higher respectively. And distillation from the teacher model with attention has further improved AUC and mAP by 0.007 and 0.014. Provided that the student is less than 0.05% the size of the teacher models and only 1% of the original teacher training data is used for distillation, moderate degradation in student performance compared to the teachers is expected.

In the study of domain-expert student models, we focus on the ubiquitous application of a single, generic and unadapted teacher model. Domain-expert student models with knowledge distilled from t-CNN-SLA on average leads to an increase of 0.083 / 0.093 of AUC / mAP compared to a model trained from weak labels with SLA. More importantly, it is interesting to note that, s-CNN-3(domain) – the student model distilled from the unadapted teacher model t-CNN-ATT – shows 0.003 / 0.010 further increase of AUC / mAP compared to s-CNN-2(domain), demonstrating the effectiveness of attention mechanism in guiding the models to produce more robust soft targets for knowledge distillation, also at the segment-level.

5. CONCLUSION

In this paper we present a study on teacher-student training for AED using *Audioset*. An end-to-end CNN model with attention mechanism to remedy the weak-label issue was first introduced, with which we showed that state-of-the-art performance can be achieved. Furthermore, teacher-student training was explored to learn both generic and domain-expert small-footprint student models. It is found that students distilled from attention-equipped teacher model have higher robustness for both generic and domain-specific tasks. A study on improving the quality of knowledge distillation, potentially by a more effective use of the attention weights provided by the teacher model, along with an investigation on the performance of student models operating at the segment-level, will both be part of the future work.

6. REFERENCES

[1] Justin Salamon, Christopher Jacoby, and Juan Pablo Bello, "A dataset and taxonomy for urban sound research," in *ACM Multimedia*, 2014.

[2] Karol J. Piczak, "ESC: Dataset for environmental sound classification," in *ACM Multimedia*, 2015.

[3] Jort F. Gemmeke, Daniel P. W. Ellis, Dylan Freedman, Aren Jansen, Wade Lawrence, R. Channing Moore, Manoj Plakal, and Marvin Ritter, "Audio set: An ontology and human-labeled dataset for audio events," in *Proc. IEEE ICASSP 2017*, New Orleans, LA, 2017.

[4] Shawn Hershey, Sourish Chaudhuri, Daniel P. W. Ellis, Jort F. Gemmeke, Aren Jansen, R. Channing Moore, Manoj Plakal, Devin Platt, Rif A. Saurous, Bryan Seybold, Malcolm Slaney, Ron J. Weiss, and Kevin W. Wilson, "CNN architectures for large-scale audio classification," in *Proc. IEEE ICASSP*, 2017, pp. 131–135.

[5] Anurag Kumar and Bhiksha Raj, "Audio event detection using weakly labeled data," in *Proc. ACM*, 2016, pp. 1038–1047.

[6] Qiuqiang Kong, Yong Xu, Wenwu Wang, and Mark D. Plumbley, "Audio set classification with attention model: A probabilistic perspective," in *Proc. IEEE ICASSP*, 2018, pp. 316–320.

[7] Xiaodan Zhuang, Xi Zhou, Mark Hasegawa-Johnson, and Thomas S. Huang, "Real-world acoustic event detection," *Pattern Recognition Letters*, vol. 31, pp. 1543–1551, 2010.

[8] Annamaria Mesaros, Toni Heittola, Antti J. Eronen, and Tuomas Virtanen, "Acoustic event detection in real life recordings," in *Proc. Eusipco*, 2010, pp. 1267–1271.

[9] Andrey Temko, Robert G. Malkin, Christian Zieger, Dusan Macho, Climent Nadeu, and Maurizio Omologo, "Clear evaluation of acoustic event detection and classification systems," in *Proc. CLEAR*, 2006.

[10] Keunwoo Choi, György Fazekas, and Mark B. Sandler, "Automatic tagging using deep convolutional neural networks," in *Proc. ISMIR*, 2016.

[11] Yong Xu, Qiuqiang Kong, Qiang Huang, Wenwu Wang, and Mark D. Plumbley, "Attention and localization based on a deep convolutional recurrent model for weakly supervised audio tagging," *CoRR*, vol. abs/1703.06052, 2017.

[12] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Proceedings of the 25th International Conference on Neural Information Processing Systems - Volume 1*, 2012, NIPS'12, pp. 1097–1105.

[13] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Proc. ICLR*, 2015.

[14] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jonathon Shlens, and Zbigniew Wojna, "Rethinking the inception architecture for computer vision," in *IEEE CVPR*, 2016, pp. 2818–2826.

[15] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, "Deep residual learning for image recognition," in *IEEE CVPR*, 2016, pp. 770–778.

[16] Ankit Shah, Anurag Kumar, Alexander G. Hauptmann, and Bhiksha Raj, "A closer look at weak label learning for audio events," *CoRR*, vol. abs/1804.09288, 2018.

[17] Changsong Yu, Karim Said Barsim, Qiuqiang Kong, and Bin Yang, "Multi-level attention model for weakly supervised audio classification," *CoRR*, vol. abs/1803.02353, 2018.

[18] Jimmy Ba and Rich Caruana, "Do deep nets really need to be deep?," in *Advances in Neural Information Processing Systems 27*, Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, Eds., pp. 2654–2662. Curran Associates, Inc., 2014.

[19] Jinyu Li, Rui Zhao, Jui-Ting Huang, and Yifan Gong, "Learning small-size DNN with output-distribution-based criteria," in *Proc. Interspeech*, 2014.

[20] Geoffrey Hinton, Oriol Vinyals, and Jeffrey Dean, "Distilling the knowledge in a neural network," in *NIPS Deep Learning and Representation Learning Workshop*, 2015.

[21] Raymond W. M. Ng, Xuechen Liu, and Pawel Swietojanski, "Teacher-student training for text-independent speaker recognition," in *Proc. IEEE SLT*, 2018.

[22] Hasim Sak, Andrew W. Senior, and Franoise Beaufays, "Long short-term memory recurrent neural network architectures for large scale acoustic modeling," in *Proc. Interspeech*, 2014.

[23] Yong Xu, Qiuqiang Kong, Wenwu Wang, and Mark D. Plumbley, "Surrey-CVSSP system for DCASE2017 challenge task4," Tech. Rep., DCASE2017 Challenge, September 2017.

[24] Jiakai Lu, "Mean teacher convolution system for dcase 2018 task 4," Tech. Rep., DCASE2018 Challenge, September 2018.