SeCoST: SEQUENTIAL CO-SUPERVISION FOR LARGE SCALE WEAKLY LABELED AUDIO EVENT DETECTION

Anurag Kumar, Vamsi Krishna Ithapu
Facebook Reality Labs
Redmond, WA USA
anuragkr@fb.com, vamsi.ithapu@oculus.com

ABSTRACT

Weakly supervised learning algorithms are critical for scaling audio event detection to several hundreds of sound categories. Such learning models should not only disambiguate sound events efficiently with minimal class-specific annotation but also be robust to label noise, which is more apparent with weak labels instead of strong annotations. In this work, we propose a new framework for designing learning models with weak supervision by bridging ideas from sequential learning and knowledge distillation. We refer to the proposed methodology as SeCoST (pronounced Sequest) — Sequential Co-supervision for training generations of Students. SeCoST incrementally builds a cascade of student-teacher pairs via a novel knowledge transfer method. Our evaluations on Audioset (the largest weakly labeled dataset available) show that SeCoST achieves a mean average precision of 0.383 while outperforming prior state of the art by a considerable margin.

Index Terms— Audio Event Detection, Teacher Student Models, Weakly Labeled, Sequential Learning

1. INTRODUCTION

In the past decade, supervised learning has been extensively studied for audio event recognition and detection (AED) [1]. While several classical machine learning and deep learning methods have been developed for AED using strong labels [2, 1, 3, 4], much of the recent progress has focused on efficiently leveraging weakly labeled data [5]. In such a weak labeling paradigm, audio recordings are only tagged for the presence or absence of sound events; unlike the strong labeling alternative where explicit time stamps of sound events are required. Hence, the annotation efforts are substantially lower thereby giving the ability to scale AED to large datasets, e.g., Audioset [6]. It has also now become an important part of the annual DCASE challenge on sound events and scenes.

Several authors have shown promising results on weakly labeled AED [7, 8, 9, 10]. A significant fraction of these works use deep convolutional neural networks in one form or other. Some are driven by attention mechanisms in neural networks [11], so as to efficiently characterize the temporal occurrences of events in the audio recordings [8, 12, 13]. Other approaches have incorporated recurrent neural networks to model the temporal attributes of sound events [13, 14].

However, large scale AED using weakly labeled data remains an open problem. When the timestamps of event occurrences are not provided, one cannot use explicit example clips of sounds for training. This clearly makes it harder to learn the necessary features and characteristics that disambiguate different sound events. Additionally, noise and the presence of other irrelevant sounds complicates the learning, in particular with long recordings. Lastly, one can observe that noisy labels are also of concern in weak label learning [15, 16].

In this work, we address some of the above issues by presenting a novel learning framework for sounds. The proposal derives ideas from two distinct, but partly related, learning paradigms — sequential learning [17] and knowledge sharing [18]. In the seminal work on Sequence of Teaching Selves [19], the authors hypothesize that the human learning goes through different stages of development, where each such stage is “guided” by previous stages. This is, in principle, similar to lifelong learning where new knowledge is accumulated while retaining previous (learned) experiences [18, 20].

The central theme of these works is sequential learning, or, learning over time. Alternatively, there is recent interest in knowledge distillation (KD) through teacher - student frameworks [21, 22, 23], where the main motivation is model compression i.e., constructing a smaller, low-capacity, student model that emulates a high-capacity high-performance teacher. These student networks are optimized based on some carefully designed divergence measures [21, 24].

We tackle weakly labeled learning by constructing a sequence of reasonably well trained neural networks (on the weak labels), where each network in the series is designed to be better than the previous one. However, unlike the classical KD where “one-shot” distillation is done from a single teacher to a student with the goal of compressing the model, the student here aims to match the performance of the teacher while also correcting for teacher’s errors. This entails constructing a cascade of student-teacher pairs and allowing the student to learn from teacher’s mistakes over multiple generations.

We do this by controlling the amount of transferable knowledge between consecutive generations. This helps in correcting the implicit noise associated with weak labels, while distilling the necessary knowledge needed for generalization. We refer to the proposed framework as SeCoST — Sequential Co-Supervision for training generations of Students from Teachers.

The rest of the presentation is as follows. In Section 2, we first describe the baseline deep convolutional network that drives the work-flow of the overall framework. We then describe our proposed SeCoST framework in Section 3, followed by experimental evaluations and discussion in Section 4. We conclude the paper in Section 5.

2. DEEP CNN FOR WEAKLY LABELED AED

Notation: Let $\mathcal{S}$ be the set of audio recordings. $\mathcal{C}$ denotes the set of labeled sound classes in these recordings. Each recording is rep-
represented by logmel features (denoted by $X$). Let $y \in R^{[C]}$ be the (weak) label vector for the input $X$, and $y_i = 1$ corresponds to $i^{th}$ class being present (i.e., tagged).

We use a deep convolutional neural network as the prototypical learning model that drives our proposed framework [10, 25]. Given an input audio recording, the idea is to first produce segment level predictions. The segments are audio snippets of small duration (e.g., 1 second length). The resulting segment level outputs are then mapped to a recording level prediction. The appropriate loss is then calculated using this prediction and the recording level weak label. The mapping from segments to recording may be done via simple mean or max operation over segment level outputs or even by a neural network, if necessary.

Figure 1 and Table 1 summarize the network schematic and the specific architectural details. $B1$ to $B4$ blocks consist of two convolutional layers followed by a pooling layer. The size of segment level output at layer $L4$ will be $[C] \times K \times 1$, where $[C]$ is the number of classes and $K$ denotes the number of possible segments for a given input. The recording level output for any given class is obtained by taking the average of segment level outputs (layer $P$).

We will refer to the above network as WELS-Net – WEakly Labeled Sound Network. WELS-Net is flexible with respect to input size, thereby allowing processing of audio recordings of variable lengths. Moreover, it produces segment level outputs which can then be used for temporal localization of sound events in the audio recordings.

Note that, WELS-Net takes Logmel spectrograms as input. Specifically, we use 64 mel-filters and mel features are extracted using a 16 ms window moving by a stride of 10 ms. The sampling rate of all audio recordings is 16 kHz. Table 1 shows output sizes at each layer for a logmel input with 1024 frames (size $1024 \times 64$). For this input, we get $K = 30$ segments at layer $L4$. In the time domain, this network produces outputs for ~1 second segments (with a stride of ~0.33 seconds).

### 2.1. Network Training

Recall that the inputs are logmel features denoted by $X$. We assume a realistic scenario where $X$ may be tagged with multiple labels, i.e., multiple sound events may be present in a single recording. The goal is to train a neural network $N(\theta)$ which can generalize well on unseen data. The network is trained by minimizing a loss function which measures divergence between the network outputs $N(\theta, X)$ and the target $y$. Let $\mathcal{L}(N(\theta, X), y)$ denote the loss function. We use binary-cross entropy loss. Thereby, we have

$$l(p_i, y_i) = -y_i \log(p_i) - (1 - y_i) \log(1 - p_i).$$

$$\mathcal{L}(N(\theta, X), y) = \frac{1}{[C]} \sum_{i=1}^{[C]} l(y_i, p_i).$$  \hspace{1cm} (1)

$p_i$ is the output of the network for the $i^{th}$ class. $l(y_i, p_i)$ is the loss for this $i^{th}$ class, and $\mathcal{L}(N(\theta, X), y)$ computes the overall loss for the input $X$ and the corresponding target $y$.

### 3. SECoST FRAMEWORK

Using the network architecture presented in Section 2, we will now describe the proposed sequential co-supervision learning. As mentioned in Section 1, SeCoST follows the principle of sequence of teaching selves. A sequence of learners (neural networks here) are trained. At each stage, the learning of a new network is supervised by the already trained network(s) from previous stages. First recall that training a network with the loss function in Eq 1 corresponds to learning from available ground truth labels $y$. This is our initialization a.k.a. the first teacher. We denote this base model by $N^{(0)}$. Once it is learned, we propose that it can co-supervise training a new network (identical in architecture to the teacher) from scratch. In other
Algorithm 1 SeCoST:

**Input:** Training data $\mathcal{D} = \{X^t, y^t\}$, Number of stages $S$, $\{\alpha^s\}$ for each stage $s = 1$ to $S$

**Output:** Trained Network after $S$ stages

1. Train base WELS-Net or Teacher-0 ($\mathcal{N}^T_0$) using $\mathcal{D} = \{X^t, y^t\}$
2. for $s = 1, 2, \ldots, S$
   3. Compute new target $\bar{y}^t$ for all training points $(X^t, y^t)$ from $y^t$, $\alpha^s$ and prediction of $N^{T_{s-1}}$ on $X^t$ using Eq 2
   4. Train new WELS-Net ($\mathcal{N}^{S_s}$) for current stage using $\mathcal{D} = \{X^t, \bar{y}^t\}$
   5. $\mathcal{N}^{T_s} = \mathcal{N}^{S_s}$ // Student becomes teacher for next stage
   6. end
7. return $\mathcal{N}^{T_S}$

words, the output of the teacher networks drive the supervision in future generations. We now formalize this process.

If the vector $\bar{y}$ denotes the output of the teacher with input $X$, then the new target denoted by $\bar{y}$, for the same input $X$, is given by

$$\bar{y} = \alpha y + (1 - \alpha)\bar{y}$$

where $\alpha$ is a hyper-parameter that controls the contribution of the teacher network’s supervision.

With a single teacher, the class-wise target for the student model is $\bar{y}_i = \alpha y_i + (1 - \alpha)\bar{y}_i$. The corresponding loss for the $i^{\text{th}}$ class now becomes

$$l(p_i, \bar{y}_i) = -\bar{y}_i \log(p_i) - (1 - \bar{y}_i) \log(1 - p_i)$$

$$= -\alpha y_i p_i - (1 - \alpha y_i) \log(1 - p_i) + (1 - \alpha) \bar{y}_i \log \frac{1 - p_i}{p_i}$$

$$l(p_i, \bar{y}_i) = l(p_i, \alpha y_i) + (1 - \alpha) \bar{y}_i \log \frac{1 - p_i}{p_i}$$

Using this new class-wise loss, the overall new loss function for the student network, denoted by $\mathcal{L}(\cdot; \bar{y})$, is

$$\mathcal{L}(\cdot; \bar{y}) = \frac{1}{|C|} \sum_{i=1}^{|C|} l((\alpha y_i, p_i) + (1 - \alpha) \bar{y}_i \log \frac{1 - p_i}{p_i})$$

$$= \mathcal{L}(\cdot; \alpha y) + (1 - \alpha) \frac{1}{|C|} \sum_{i=1}^{|C|} \bar{y}_i \log \frac{1 - p_i}{p_i}$$

which is a combination of loss w.r.t. ground truth $y$, although weighted by a factor of $\alpha < 1$; and a term representing supervision from the teacher. We can rewrite this as

$$\mathcal{L}(\cdot, \bar{y}) = \mathcal{L}(\cdot, y) + (1 - \alpha) \frac{1}{|C|} \sum_{i=1}^{|C|} (y_i - \bar{y}_i) \log \frac{1 - p_i}{p_i}$$

Here, the first term is same as computing the loss w.r.t. $y$. The second term now involves both the ground truth labels and the teacher’s predictions. Hence, the additional supervision for student is entirely determined by how much the teacher’s predictions ($\bar{y}_i$) differ from the true labels $y_i$, i.e., the student is trying to learn from the mistakes made by the teacher network. This additional information is hypothesized to improve the generalization capabilities.

This overall procedure can now be emulated to multiple stages $S$. Algorithm 1 summarizes this sequential procedure over $S$ stages. The output of this procedure is a network trained over $S$ generations with one or more teachers in each stage. Note that we expect the improvement in generalization to have diminishing returns after some stages, and we discuss more about this behaviour in Section 4.

Table 2. Comparing SeCoST with state of the art methods on Audioset.

<table>
<thead>
<tr>
<th>Method</th>
<th>mAP</th>
<th>mAUC</th>
<th>Method</th>
<th>mAP</th>
<th>mAUC</th>
</tr>
</thead>
<tbody>
<tr>
<td>TALNet [13] - Pooling</td>
<td>0.369</td>
<td>0.899</td>
<td>TALNet [13] - Attention</td>
<td>0.361</td>
<td>0.969</td>
</tr>
<tr>
<td>DFN [14] - Attention</td>
<td>0.702</td>
<td>0.990</td>
<td>WELS-Net [15], [16]</td>
<td>0.352</td>
<td>0.926</td>
</tr>
<tr>
<td>Net-Ont (Fixed $\alpha = 0.0$)</td>
<td>0.379</td>
<td>0.759</td>
<td>SeCoST (Variable $\alpha$)</td>
<td>0.385</td>
<td>0.971</td>
</tr>
</tbody>
</table>

**Multiple Teachers per Stage:** Observe that Eqs 2 and 5 are parameterizing learning from a single teacher at a given stage. The above procedure can also be extended to incorporate multiple teachers (denoted by $T$) per stage. The new target is given by a convex combination of all available supervision, from all the $T$ teachers, as shown below.

$$\bar{y} = \sum_{k=0}^T \alpha_k \bar{y}_k$$

$$\text{and } \sum_{k=0}^T \alpha_k = 1$$

where $\bar{y}_k = y$ represents the ground truth labels, and $\alpha$, $k = 0$ to $T$, parameterize the contribution of the ground truth and the $T$ teacher networks respectively. The class-wise loss $l(p_i, \bar{y}_i)$ becomes

$$l(p_i, \bar{y}_i) = l(p_i, \alpha y_i) + \sum_{k=1}^T \alpha_k \bar{y}_k \log \frac{1 - p_i}{p_i}$$

4. EXPERIMENTS AND RESULTS

We evaluate SeCoST using Audioset (the largest available dataset for sound events) [6]. It has weakly labeled examples for 527 sound events, with approximately 2 million training examples and 20k evaluation recordings. We use 25k samples from the training set for validation. Each recording is 10 seconds long (and we resample them at 16kHz). Audioset is multi-label in nature with ~2.7 labels per recording. PyTorch is used for implementing the networks [26]. Training utilizes Adam [27] where hyperparameters (like learning rates) are tuned using the validation set. Similar to existing AED works [6, 10, 7], Average precision (AP) and area under ROC (AUC) are used to measure the performance [28, 29]. Further, mean average precision (mAP) and mean AUC (mAUC) over all classes summarizes the overall performances. In this work, we use a single teacher at each stage (Alg 1), leaving evaluation of multiple teacher per stage for future work.

**Performance Comparison:** Table 2 compares SeCoST with existing state-of-the-art methods on Audioset. Note that the authors of [7] use embeddings from a network trained on a very large database of audio recordings (YouTube-70M) [30], thereby the resulting feature representations already lead to enhanced performance. However, we work directly with audio recordings and use their logmel representations. Our base WELS-Net model ($\mathcal{N}^T_0$) trained on the ground truth labels gives an mAP of 0.352 over all the 527 events. SeCoST gives an mAP of 0.383, improving WELS-Net by 8.8%. Notably, it is also 3.8% better than the best reported performance in literature. Our best performance of 0.383 mAP is obtained by increasing the contribution of teacher ($1 - \alpha$) as the sequence progresses. Using a fixed $\alpha$ of 0.3, we improve the base WELS-Net by 7.7% (from 0.352 to 0.379).

**Effect of $\alpha$:** Figure 2 (Left) analyzes the influence of teacher’s contribution (parameterized by $1 - \alpha$) in SeCoST (see Alg 1). Here a single stage of SeCoST is done with base WELS-Net as the teacher. $1 - \alpha = 0$ represents no contribution from the teacher, i.e training only on ground truth labels. We can see that the performance improves as teacher’s contribution increases, but only up to a certain point. This occurs at $1 - \alpha = 0.3$, with the corresponding mAP of
from

ments, although with diminishing returns after each stage. After the

to

vision remains same for all 4 stages with

SeCoST (S = sequential training aspect of SeCoST. We do

SeCoST Stage-wise performance: Fig 2 (Right) evaluates the se-

quential training aspect of SeCoST. We do 4 stages of training in

SeCoST (S = 4 in Alg 1). The contribution of teacher in the super-

vision remains same for all 4 stages with 1 − αs = 0.7, for s = 1 to 4.

We can see that sequential co-supervision leads to improve-

ments, although with diminishing returns after each stage. After the

first stage of co-supervision (N2 as teacher), the mAP improves

from 35.2 to 37.4, a 6.3% improvement over the base WELS-Net.

Using this improved network as teacher in the second stage of SeC-

oST, we see a further improvement of 1.3% (overall a 7.6% improve-

ment from WELS-Net). The performance then saturates and we do

not see any additional improvements in future stages.

We saw above that, SeCoST in general works better with lower

α i.e., larger weight to teacher’s contributions. This suggests that as

newer generations of networks become teachers, it might be helpful

to increase their contribution in co-supervision. To evaluate this,

we run SeCoST for 5 stages with \{α1 = 0.3, α2 = 0.3, α3 =

0.2, α4 = 0.1, α5 = 0.05\}. The performance at different stages

is shown in Figure 3. As teacher’s contribution is increased from

0.7 to 0.8 at stage 3, we see an improvement in performance, unlike

the case where it was fixed at 0.7 and the performance remained

same at stage 3 (refer back to Figure 2 (left)). At stage 4, with 90% 

supervision from teacher, we get an improved mAP of 0.383 (overall

8.8% improvement on base WELS-Net).

Only Teacher Networks (α = 0) in SeCoST: In Figure 2 (left), we

showed that training the student using only the teacher’s output

(1 − α = 1) as the target already gives a better model. We get

0.366 mAP versus 0.352 when training only on ground truth labels.

A similar observations has been made in [21]. This may be because

the soft probability outputs from the teacher network provides richer

information compared to the ‘hard’ ground truth labels. However,

as shown in Figure 3 (right), this teacher-alone strategy does not

work well for future generations. Here we run SeCoST for 3 stages

using only teacher’s supervision. We see that the first stage leads to

an improvement over the base network, and then there is saturation.

This shows that co-supervision is necessary. Hence, the teacher’s

knowledge needs to be coupled with the ground truth labels while

training the sequence of students.

Some Class Specific Analysis: We observe that SeCoST improves

performance for > 85% percent of the sound classes in Audioset

(448 out of 527). Of these, for 110 classes we get > 20% improve-

ment in AP. Specifically, forCrushing, Harmonic and Mouse sounds

in Audioset vocabulary, we observe more than 100% improvement

using SeCoST. On the other hand, there are only 12 classes with

more than 10% drop in performance. Sound class Squish has max-

imum drop in performance, around 19%. Note that, these summaries

are based on the best SeCoST model with mAP of 0.383 and base

WELS-Net with 0.352 mAP. To further analyze class specific perfor-

mance, we try to see whether the improvements are coming for

classes where the base model already does well or if the classes

with low APs are actually improving. Figure 4 shows this relative

mAP improvement for classes with APs within a specified range. For

classes with the APs < 0.1, the mean AP improves by ∼ 28%. For

classes with APs ∈ [0.1, 0.2], the gain is 21%. This show that SeC-

oST leads to considerable improvement in classes which are harder

to learn for WELS-Net. For relatively easy classes (WELS-Net AP

> 0.5), we see ∼ 4.5% improvement.

5. CONCLUSION

We proposed a sequential co-supervision learning framework for au-

dio event detection. Our proposal, SeCoST, builds a generation of

networks by designing student models that learn to predict a con-

vex combination of teachers’ predictions instead of the given ground

truth. We showed that SeCoST gives a considerably better perfor-

mance on Audioset compared to baseline model and state of the art

performance. We note that the proposed framework is generally ap-

licable to learning from noisy, weak labels, and we intend to future

investigate the theoretical merits of the model in the future.

6. REFERENCES

[1] Tuomas Virtanen, Mark D Plumbley, and Dan Ellis, Computational analysis of


