Pseudo strong labels for large scale weakly supervised audio tagging

Heinrich Dinkel, Zhiyong Yan, Yongqing Wang, Junbo Zhang, Yujun Wang

Xiaomi Corporation

Highlights

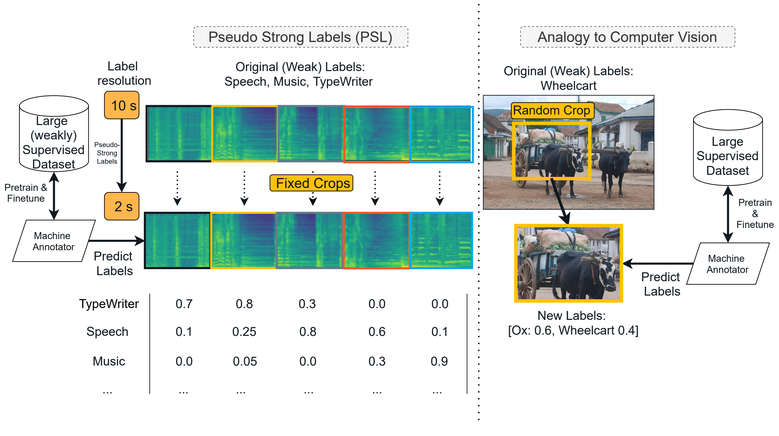
- **Simple** MobileNetV2 approach: Fast training and inference.
- **State-of-the-art performance** on the 60h long Audioset balanced subset.
- Achieves **35.95** mAP on Audioset without Augmentation.
- Obtained **87%** of the performance using **1%** of the data.

Problem statement

- Audioset contains 5200 h hours of training data, with 527 **ambiguous** labels.
- Most labels in Audioset are **missing** i.e., “Liquid” is present, but “Water” is not.
- 10 s of audio contains **too many labels**.

Proposed approach: Pseudo Strong Labels (PSL)

1. First train a machine annotator (MA).
2. Predict **soft** targets on a finer scale (5s, 2s) using MA.
3. Train a model on these new soft labels.



Dataset and Training

The largest Audio tagging dataset, Audioset is used. Three training datasets and one evaluation.

Dataset	Purpose	# Clips	Duration (h)
Balanced		21,155	58
Aud-300h	Train	109,295	300
Full		1,904,746	5244
Eval	Evaluation	18,229	50

Training objective, between the (soft) label $\mathbf{y} \in [0,1]^C$ and the model ($\mathcal{F}$) prediction $\hat{\mathbf{y}} \in [0,1]^C$, is the binary cross entropy:

$$\mathcal{L}_{\text{BCE}}(\mathbf{x}, \mathbf{y}) = \mathbf{y} \log \hat{\mathbf{y}} + (\mathbf{1} - \mathbf{y}) \log(\mathbf{1} - \hat{\mathbf{y}}),$$
$$\hat{\mathbf{y}} = \mathcal{F}(\mathbf{x}),$$

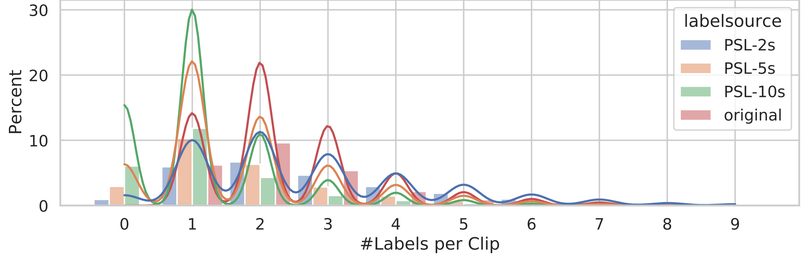
Main results

Results for models trained on the balanced subset of Audioset

Method	Label	mAP	d'
Baseline (Weak)	$\mathbf{y}^{10}_{\text{weak}}$	17.69	1.994
PSL-10s (Proposed)	$\hat{\mathbf{y}}^{10}_{\text{PSL}}$	31.13	2.454
PSL-5s (Proposed)	$\hat{\mathbf{y}}^{5}_{\text{PSL}}$	34.11	2.549
PSL-2s (Proposed)	$\hat{\mathbf{y}}^{2}_{\text{PSL}}$	**35.48**	**2.588**
CNN14 [**Kong2020d**]	$\mathbf{y}^{10}_{\text{weak}}$	27.80	1.850
EfficientNet-B0 [**gong2021psla**]		33.50	-
EfficientNet-B2 [**gong2021psla**]		34.06	-
ResNet-50 [**gong2021psla**]		31.80	-
AST [**gong21b_interspeech**]		34.70	-

PSL Label count distribution

- The naïve PSL-10s mainly predicts single labels (1).
- 18% of the naïve PSL-10s does not predict a single label.
- PSL-5s/2s perform uniformly better than 10s, due to finer time resolution.
- PSL-2s outperforms the **original labels for > 4 labels**.
- PSL is capable of predicting missing labels.



Transfer Learning

Transferring weights to 3 different Audiotagging datasets using the MA and our proposed PSL-2s.

Dataset	Metric	MA	PSL-2s	Imp.
FSD50k	mAP	44.41	**54.23**	+9.82
FSD2018	mAP@3	87.31	**89.21**	+1.90
FSD2019-Curated	*lw*lwrap	68.84	**71.86**	+3.02
FSD2019-Noisy	*lw*lwrap	53.57	**54.49**	+0.92

Conclusion

- PSL mitigates missing labels for temporally weakly-supervisec methods.
- A time-window of 2s seems to be a reasonable choice.
- Transfer learning experiments show that the improvement ir label-quality also transfers to other tasks.

Reevaluation of our results on the evaluation set **with median post-processing**.

Model	#Param (M)	**PSDS-1**	**PSDS-2**	**Score**	Single?
1st	14.3	**45.2**	**74.6**	**1.40**	N
2nd	20.2	44.2	67.4	1.32	Y
3rd	79.2	33.9	71.5	1.29	N
3rd	50.0	41.9	68.6	1.29	N
4th	119.8	41.6	63.7	1.24	N
S3	**3.4**	38.2	65.4	1.20	Y
S2	**2.7**	37.9	64.3	1.19	Y
5th	8.5	41.3	58.6	1.19	Y
S1	**2.0**	36.1	64.3	1.16	Y
6th	6.7	37.0	62.6	1.16	Y

Code Available

github.com/RicherMans/PSL