EAT: Self-Supervised Pre-Training with Efficient Audio Transformer

Introduction

EAT is an audio SSL model with high effectiveness and efficiency in pre-training. Adopting the bootstrap paradigm, we propose the Utterance-Frame Objective (UFO) and adapt the inverse block masking on audio patches during its self-supervised training.
The paper has been released on arxiv.

Performance

Pre-training on AS-2M, EAT gain state-of-the-art (SOTA) performance on several audio and speech classification datasets like AS-20K, AS-2M, ESC-50 and SPC-2.

Model	#Param	Pre-training Data	AS-2M mAP(%)	AS-20K mAP(%)	ESC-50 Acc(%)	SPC-2 Acc(%)
Supervised Pre-Training						
PANN [Kong *et al.*, 2020]	81M	-	43.1	27.8	83.3	61.8
PSLA [Gong *et al.*, 2021b]	14M	IN	44.4	31.9	-	96.3
AST [Gong *et al.*, 2021a]	86M	IN	45.9	34.7	88.7	98.1
MBT [Nagrani *et al.*, 2021]	86M	IN-21K	44.3	31.3	-	-
PassT [Koutini *et al.*, 2021]	86M	IN	47.1	-	96.8	-
HTS-AT [Chen *et al.*, 2022a]	31M	IN	47.1	-	97.0	98.0
Wav2CLIP [Wu *et al.*, 2022]	74M	TI+AS	-	-	86.0	-
AudioCLIP [Guzhov *et al.*, 2022]	93M	TI+AS	25.9	-	96.7	-
Self-Supervised Pre-Training						
Conformer [Srivastava *et al.*, 2022]	88M	AS	41.1	-	88.0	-
SS-AST [Gong *et al.*, 2022]	89M	AS+LS	-	31.0	88.8	98.0
MAE-AST [Baade *et al.*, 2022]	86M	AS+LS	-	30.6	90.0	97.9
MaskSpec [Chong *et al.*, 2023]	86M	AS	47.1	32.3	89.6	97.7
MSM-MAE [Niizumi *et al.*, 2022]	86M	AS	-	-	85.6	87.3
data2vec [Baevski *et al.*, 2022]	94M	AS	-	34.5	-	-
Audio-MAE [Huang *et al.*, 2022]	86M	AS	47.3	37.1	94.1	**98.3**
BEATs$_{iter1}$ [Chen *et al.*, 2022c]	90M	AS	47.9	36.0	94.0	**98.3**
BEATs$_{iter2}$ [Chen *et al.*, 2022c]	90M	AS	48.1	38.3	95.1	**98.3**
BEATs$_{iter3}$ [Chen *et al.*, 2022c]	90M	AS	48.0	38.3	95.6	**98.3**
BEATs$_{iter3+}$ [Chen *et al.*, 2022c] *	90M	AS	48.6	38.9	98.1	98.1
Ours						
EAT	88M	AS	**48.6**	**40.2**	**95.9**	**98.3**

Table 1: **Model Comparison among existing methods in audio classification tasks.** Pre-training data sources include ImageNet (IN), AudioSet (AS), and LibriSpeech (LS), while CLIP utilizes 400M text-image pairs (TI). We gray-out the methods with additional supervised training on external datasets or additional pseudo-labels. *: Models employ knowledge distillation across iterations with extra pseudo-labels.

Efficiency

EAT achieves a total pre-training time reduction of ~15x compared to BEATs$_{iter3}$ and ~10x relative to Audio-MAE. It costs only 10 epochs during EAT's pre-training on AS-2M.

model	epoch	hour $\times$ GPU	speedup	mAP
BEATs$_{iter3}$	342	3600	1$\times$	38.3
Audio-MAE	32	2304	1.56$\times$	37.1
EAT	**10**	**230**	**15.65**$\times$	**40.2**

Table 2: **Comparison with BEATs$_{iter3}$ and Audio-MAE on pre-training cost.** We evaluate the pre-training wall-clock time of EAT on 4 RTX 3090 GPUs in Fairseq [Ott *et al.*, 2019] and it demands around 5.8 hours for each epoch. BEATs is pre-trained on 16 Tesla V100-SXM2-32GB GPUs for around 75 hours per iteration with 114 epochs while Audio-MAE on 64 V100 GPUs for approximately 36 hours in total. All models are uniformly fine-tuned on AS-20K.

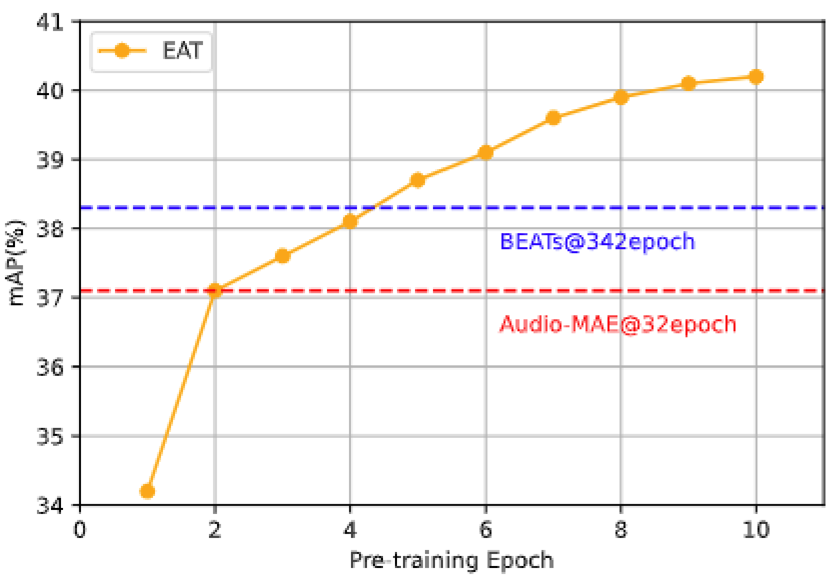

Figure 3: **Comparison with BEATs$_{iter3}$ and Audio-MAE on pre-training epoch during EAT's 10-epoch pre-training.** All models are uniformly fine-tuned on AS-20K and tested on the evaluation set.

Feature Extraction

TODO

Pre-training

TODO

Fine-tuning

TODO

TODO

- [ ] release the main pre-trained codes and pre-trained EAT model
- [ ] release the fine-tuned codes and fine-tuned EAT models (in AS tasks)
- [ ] release the inferrence codes

Citation

If you find our EAT code and paper useful, please kindly cite:

```
@article{chen2024eat,
  title={EAT: Self-Supervised Pre-Training with Efficient Audio Transformer},
  author={Chen, Wenxi and Liang, Yuzhe and Ma, Ziyang and Zheng, Zhisheng and Chen, Xie},
  journal={arXiv preprint arXiv:2401.03497},
  year={2024}
}
```