BEATs: Audio Pre-Training with Acoustic Tokenizers

Sanyuan Chen, Yu Wu, Chengyi Wang, Shujie Liu, Daniel Tompkins,

Zhuo Chen, Wanxiang Che, Xiangzhan Yu, Furu Wei

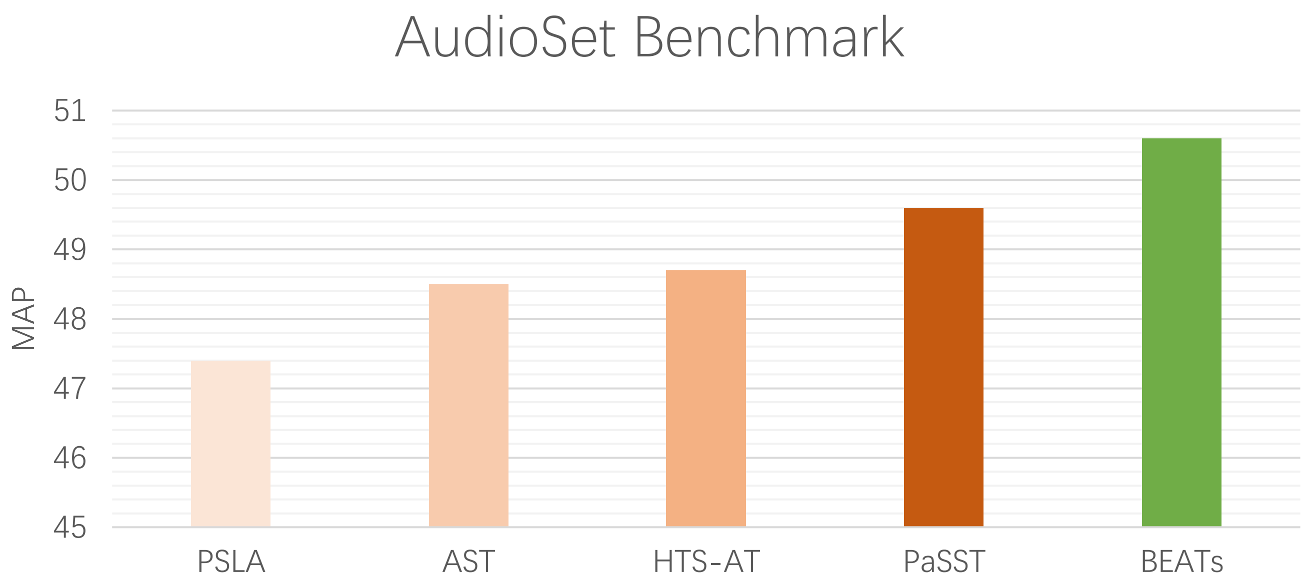
Presenter: Sanyuan Chen

Paper: https://icml.cc/virtual/2023/oral/25555

Codes and models: https://aka.ms/beats

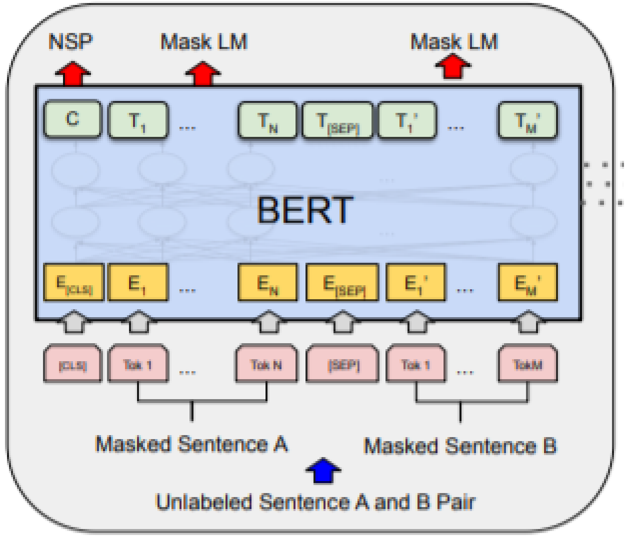
BEATs: Audio Pre-Training with Acoustic Tokenizers

Unlike the previous methods that employ continuous feature reconstruction loss for audio pre-training, we explore audio pre-training with discrete label prediction loss for **the first time** and outperform previous **state-of-the-art** models by **a large margin** with **much less training data** and **model parameters**.

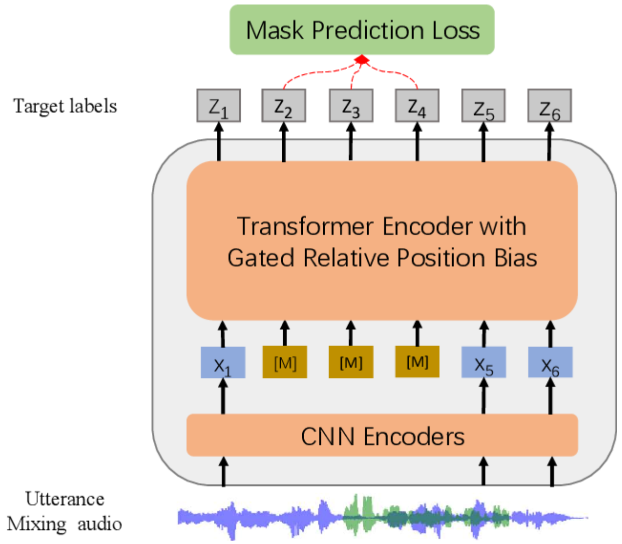


Background: SSL with discrete label prediction

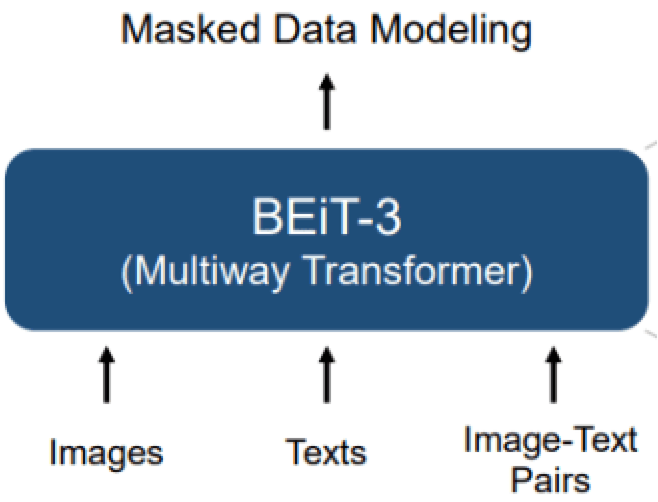
- Self-Supervised Learning (SSL) has achieved great success in language, vision, speech, and audio domains.
- SSL with **discrete label prediction loss** is widely adopted for **language**, **vision**, **speech** modalities, and shows better performance than **reconstruction loss**.



BERT/GPT
for **language**



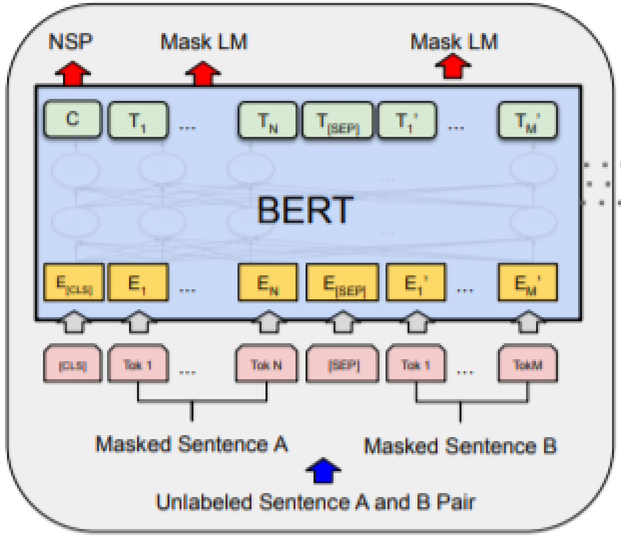
HuBERT/WavLM
for **speech**



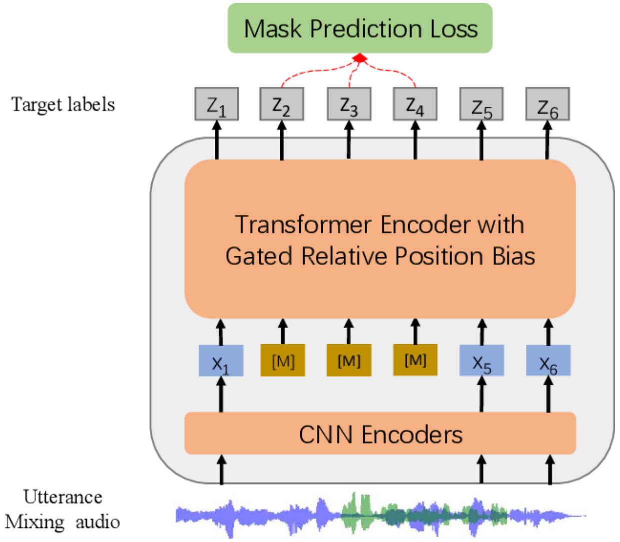
BEiT series
for **vision** and **vision-language**

Background: SSL with discrete label prediction

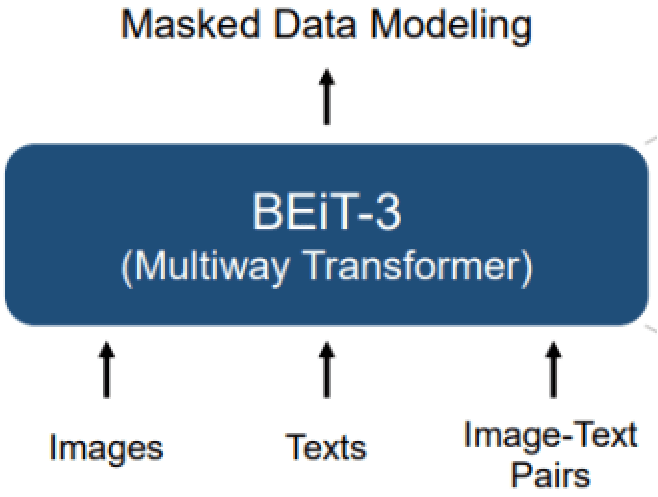
- Self-Supervised Learning (SSL) has achieved great success in language, vision, speech, and audio domains.
- SSL with **discrete label prediction loss** is widely adopted for **language**, **vision**, **speech** modalities, and shows better performance than **reconstruction loss**.
- Compared with reconstruction loss, semantic-rich discrete label prediction encourages the SSL model to **abstract the high-level semantics** and **discard the redundant details**.



BERT/GPT
for **language**



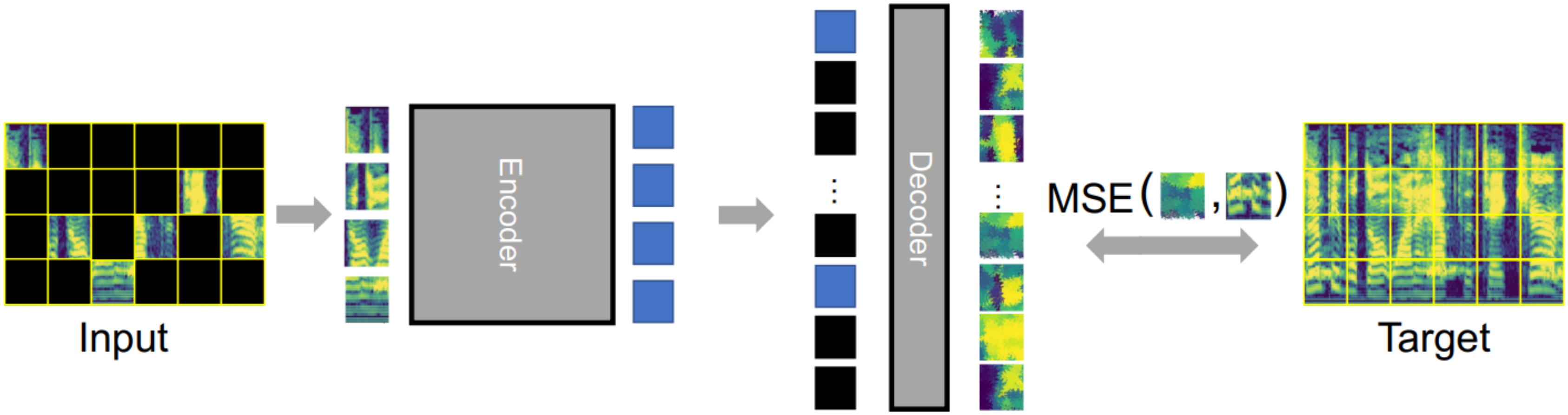
HuBERT/WavLM
for **speech**



BEiT series
for **vision** and **vision-language**

Background: SSL with discrete label prediction

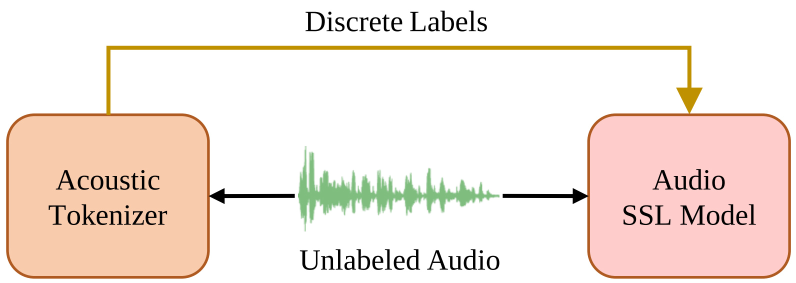
- Self-Supervised Learning (SSL) has achieved great success in language, vision, speech, and audio domains.
- SSL with **discrete label prediction loss** is widely adopted for **language, vision, speech** modalities, and shows better performance than **reconstruction loss**.
- The state-of-the-art **audio** SSL model still employ reconstruction loss for pre-training.



Audio-MAE: SOTA audio SSL model

Background: SSL with discrete label prediction

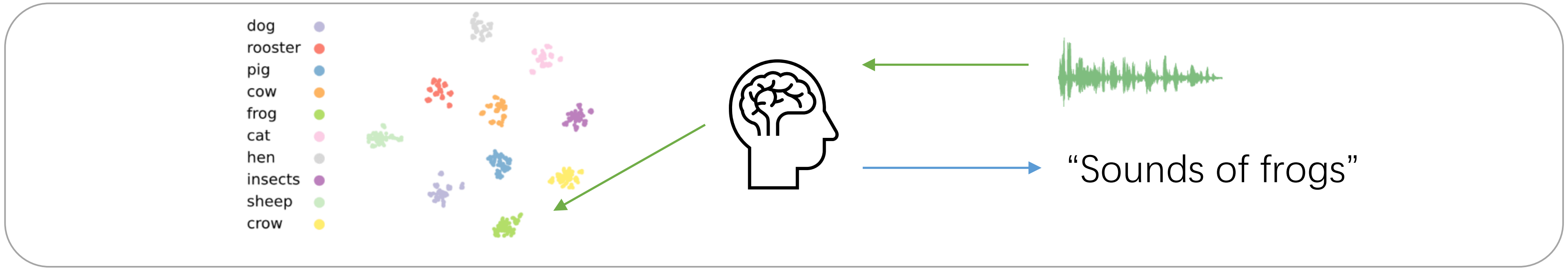
- Self-Supervised Learning (SSL) has achieved great success in language, vision, speech, and audio domains.
- SSL with **discrete label prediction loss** is widely adopted for **language**, **vision**, **speech** modalities, and shows better performance than **reconstruction loss**.
- The state-of-the-art **audio** SSL model still employ reconstruction loss for pre-training.
- **Questions:**
    1. Would **discrete label prediction** be a better choice for audio pre-training?
    2. How to **design the acoustic tokenizer** for semantic-rich discrete label generation?



Would discrete label prediction be a better choice for audio pre-training?

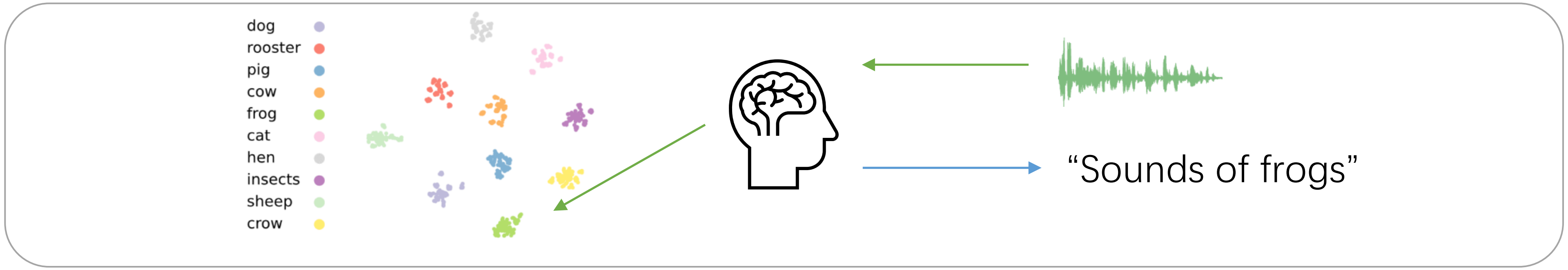
Would discrete label prediction be a better choice for audio pre-training?

1. **Humans understand audio** by extracting and clustering the high-level semantics instead of focusing on the low-level time-frequency details.



Would discrete label prediction be a better choice for audio pre-training?

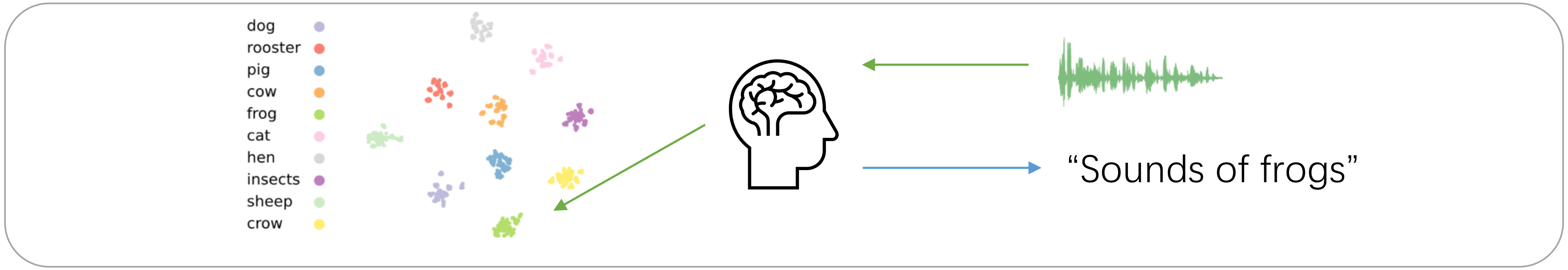
1. **Humans understand audio** by extracting and clustering the high-level semantics instead of focusing on the low-level time-frequency details.



2. **Better audio modeling efficiency** by encouraging the model to focus on the high-level semantics and discard the redundant details.

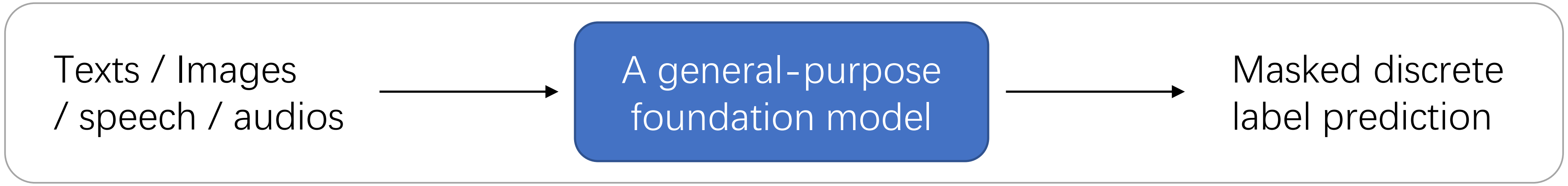
Would discrete label prediction be a better choice for audio pre-training?

1. **Humans understand audio** by extracting and clustering the high-level semantics instead of focusing on the low-level time-frequency details.



2. **Better audio modeling efficiency** by encouraging the model to focus on the high-level semantics and discard the redundant details.

3. Advances the **unification of language, vision, speech, and audio pre-training**, which enables the possibility of building a foundation model across modalities with a single pre-training task.



How to design the semantic-rich acoustic tokenizer?

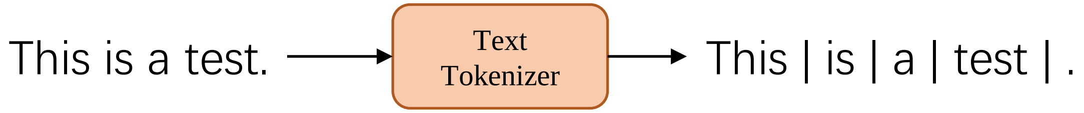
How to design the semantic-rich acoustic tokenizer?

- **Audio property:**
  1. **Continuous** signals.
  2. **Wide variations of environmental events** (human voices, nature sounds, musical beats)
  3. Each environmental events might have **various durations** in different occasions.

How to design the semantic-rich acoustic tokenizer?

- **Audio property:**
  1. **Continuous** signals.
  2. **Wide variations of environmental events** (human voices, nature sounds, musical beats)
  3. Each environmental events might have **various durations** in different occasions.

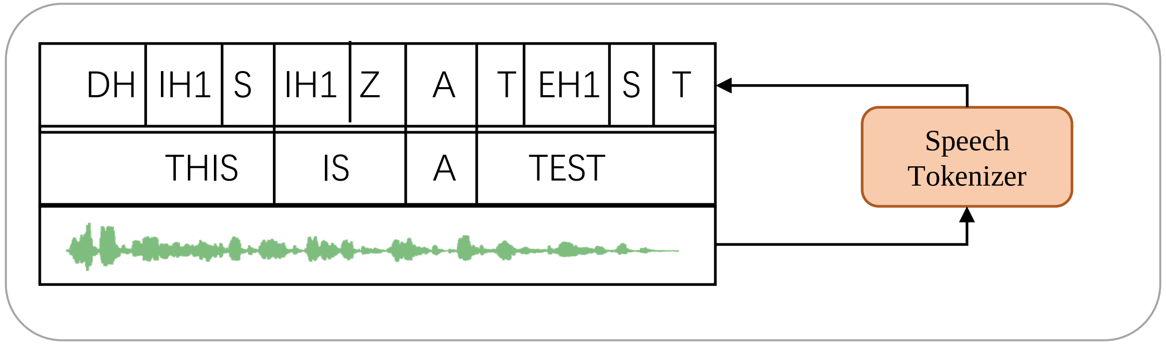
- Can we use the **text tokenizer**?



How to design the semantic-rich acoustic tokenizer?

- **Audio property:**
  1. **Continuous** signals.
  2. **Wide variations of environmental events** (human voices, nature sounds, musical beats)
  3. Each environmental events might have **various durations** in different occasions.

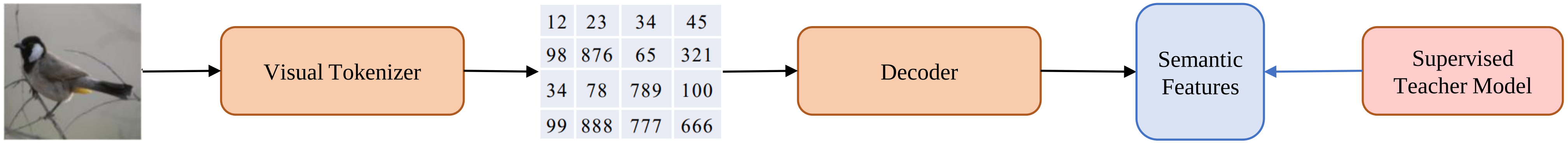
- Can we use the **speech tokenizer**?



How to design the semantic-rich acoustic tokenizer?

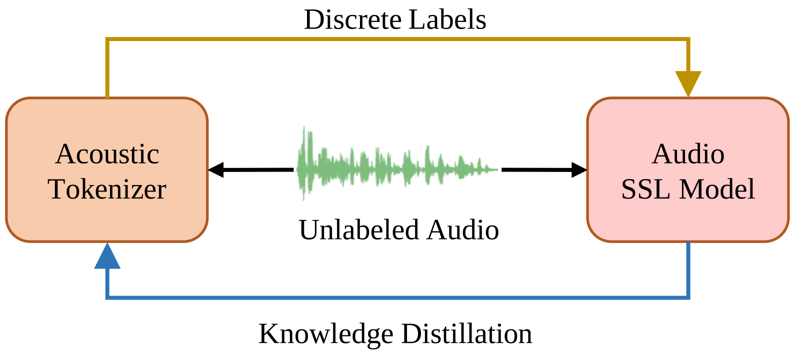
- **Audio property:**
    1. **Continuous** signals.
    2. **Wide variations of environmental events** (human voices, nature sounds, musical beats)
    3. Each environmental events might have **various durations** in different occasions.

- Can we use the **visual tokenizer**?



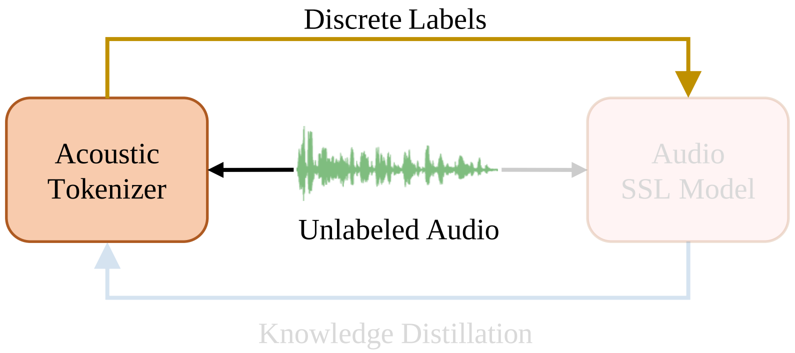
BEATs: an iterative audio pre-training framework

- An **acoustic tokenizer** and an **audio SSL model** are optimized by iterations.



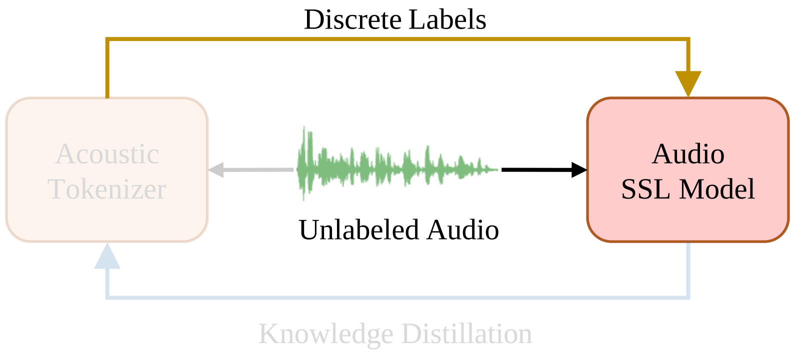
BEATs: an iterative audio pre-training framework

- An **acoustic tokenizer** and an **audio SSL model** are optimized by iterations.
- Each iteration:
  1. Generate **discrete labels** with the acoustic tokenizer



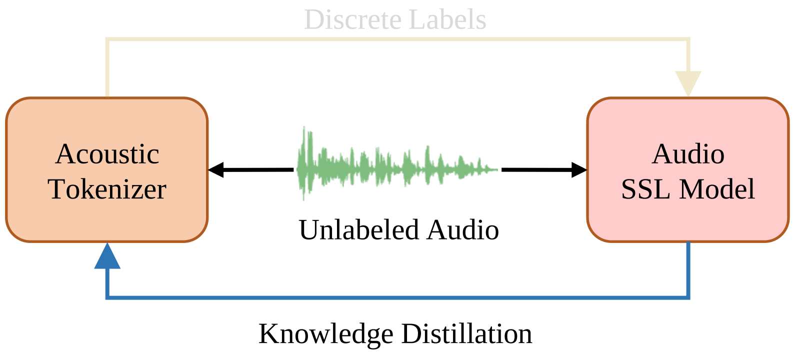
BEATs: an iterative audio pre-training framework

- An **acoustic tokenizer** and an **audio SSL model** are optimized by iterations.
- Each iteration:
  1. Generate **discrete labels** with the acoustic tokenizer
  2. Optimize the **audio SSL model** with mask and discrete label prediction loss



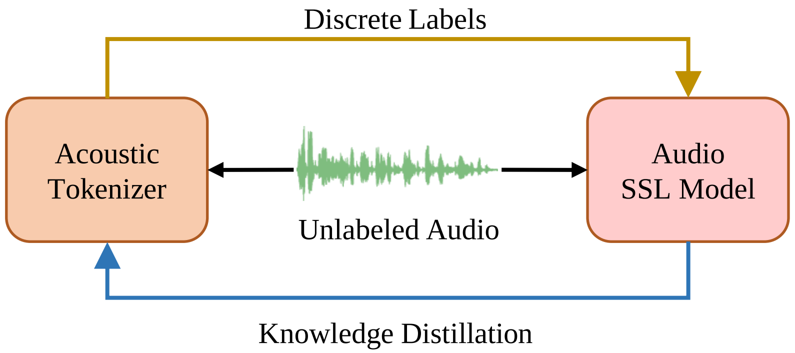
BEATs : an iterative audio pre-training framework

- An **acoustic tokenizer** and an **audio SSL model** are optimized by iterations.
- Each iteration:
    1. Generate **discrete labels** with the acoustic tokenizer
    2. Optimize the **audio SSL model** with mask and discrete label prediction loss
    3. Update the **acoustic tokenizer** with audio semantics distilled from the pre-trained or fine-tuned audio SSL model



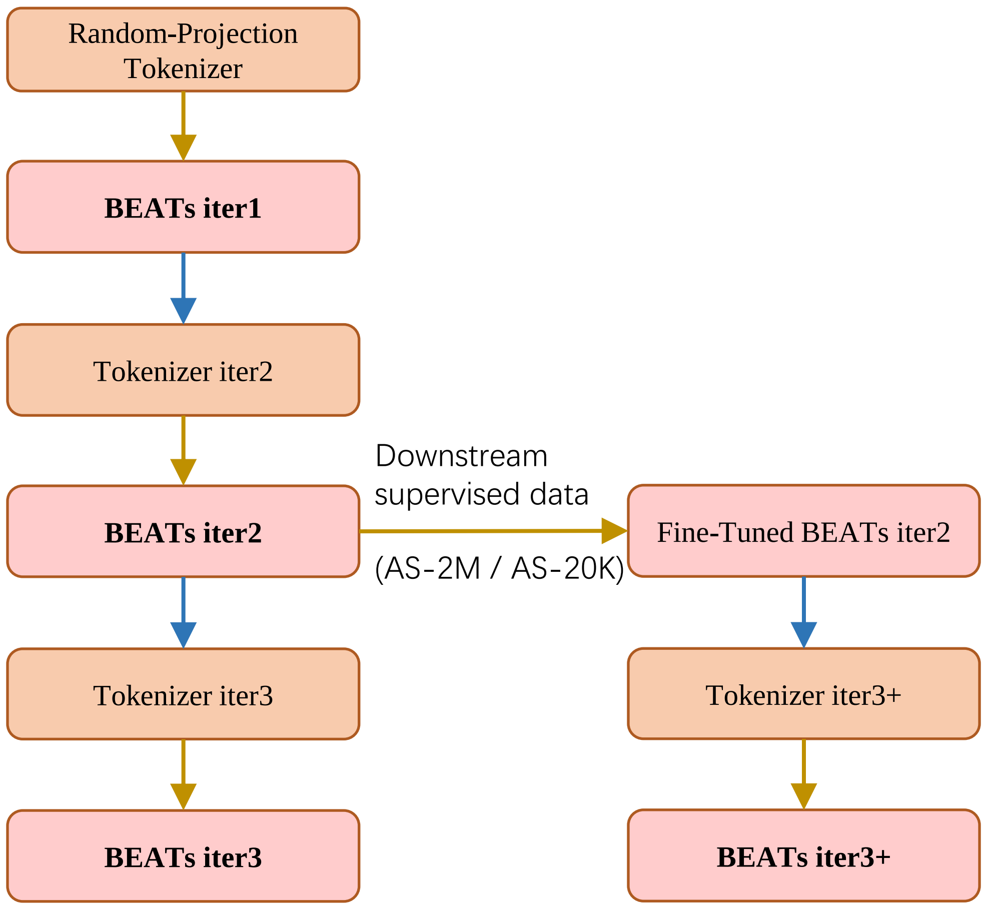
BEATs : an iterative audio pre-training framework

- An **acoustic tokenizer** and an **audio SSL model** are optimized by iterations.
- Each iteration:
  1. Generate **discrete labels** with the acoustic tokenizer
  2. Optimize the **audio SSL model** with mask and discrete label prediction loss
  3. Update the **acoustic tokenizer** with audio semantics distilled from the pre-trained or fine-tuned audio SSL model



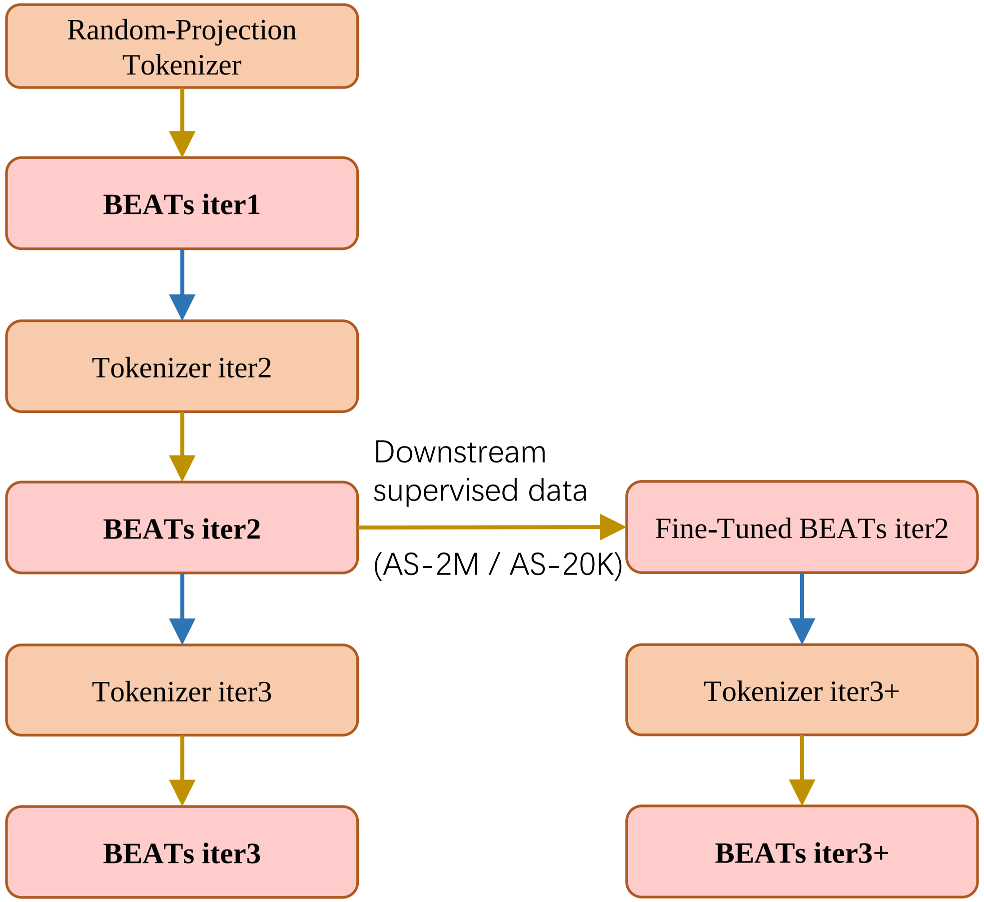
- Cold start:
  - we use **random projection** as the acoustic tokenizer in the first iteration.

BEATs : an iterative audio pre-training framework



Comparing with the SOTA Single Models

- We gray-out the models and results with external datasets.



Model	# Param	Data	Audio			Speech		
			AS-2M	AS-20K	ESC-50	KS1	KS2	ER
No Pre-Training								
PANN [Kong et al., 2020]	81M	-	43.1	27.8	83.3	-	61.8	-
ERANN [Verbitskiy et al., 2022]	55M	-	45.0	-	89.2	-	-	-
Out-of-domain Supervised Pre-Training								
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	95.5	98.1	56.0
MBT [Nagrani et al., 2021]	86M	IN-21K	44.3	31.3	-	-	-	-
PaSST [Koutini et al., 2021]	86M	IN	47.1	-	-	-	-	-
HTS-AT [Chen et al., 2022a]	31M	IN	47.1	-	-	-	98.0	-
Wav2CLIP [Wu et al., 2022]	74M	TI+AS	-	-	86.0	-	-	-
AudioCLIP [Guzhov et al., 2022]	93M	TI+AS	25.9	-	96.7	-	-	-
In-domain Supervised Pre-Training								
PANN [Kong et al., 2020]	81M	AS	-	-	94.7	-	-	-
ERANN [Verbitskiy et al., 2022]	55M	AS	-	-	96.1	-	-	-
AST [Gong et al., 2021a]	86M	IN+AS	45.9	-	95.6	-	97.9	-
PaSST [Koutini et al., 2021]	86M	IN+AS	47.1	-	96.8	-	-	-
HTS-AT [Chen et al., 2022a]	31M	IN+AS	47.1	-	97.0	-	-	-
CLAP [Elizalde et al., 2022]	190.8M	TA	-	-	96.7	-	96.8	-
Audio-MAE [Xu et al., 2022]	86M	AS	-	-	97.4	-	-	-
Self-Supervised Pre-Training								
Wav2vec [Schneider et al., 2019]	33M	LS	-	-	-	96.2	-	59.8
Wav2vec 2.0 [Baevski et al., 2020]	95M	LS	-	-	-	96.2*	-	63.4*
SS-AST [Gong et al., 2022a]	89M	AS+LS	-	31.0	88.8	96.0	98.0	59.6
MSM-MAE [Niizumi et al., 2022]	86M	AS	-	-	85.6	-	87.3	-
MaskSpec [Chong et al., 2022]	86M	AS	47.1	32.3	89.6	-	97.7	-
MAE-AST [Baade et al., 2022]	86M	AS+LS	-	30.6	90.0	95.8	97.9	59.8
Audio-MAE [Xu et al., 2022]	86M	AS	47.3	37.1	94.1	96.9	**98.3**	-
data2vec [Baevski et al., 2022]	94M	AS	-	34.5	-	-	-	-
Audio-MAE Large [Xu et al., 2022]	304M	AS	47.4	37.6	-	-	-	-
CAV-MAE [Gong et al., 2022b]	86M	AS+IN	44.9	34.2	-	-	-	-
Ours								
$\text{BEATs}_{\text{iter1}}$	90M	AS	47.9	36.0	94.0	**98.0**	**98.3**	65.9
$\text{BEATs}_{\text{iter2}}$	90M	AS	48.1	38.3	95.1	97.7	**98.3**	**66.1**
$\text{BEATs}_{\text{iter3}}$	90M	AS	48.0	38.3	**95.6**	97.7	**98.3**	64.5
$\text{BEATs}_{\text{iter3+}}$	90M	AS	**48.6**	**38.9**	98.1	98.1	98.1	65.0

Comparing with the SOTA Ensemble Models

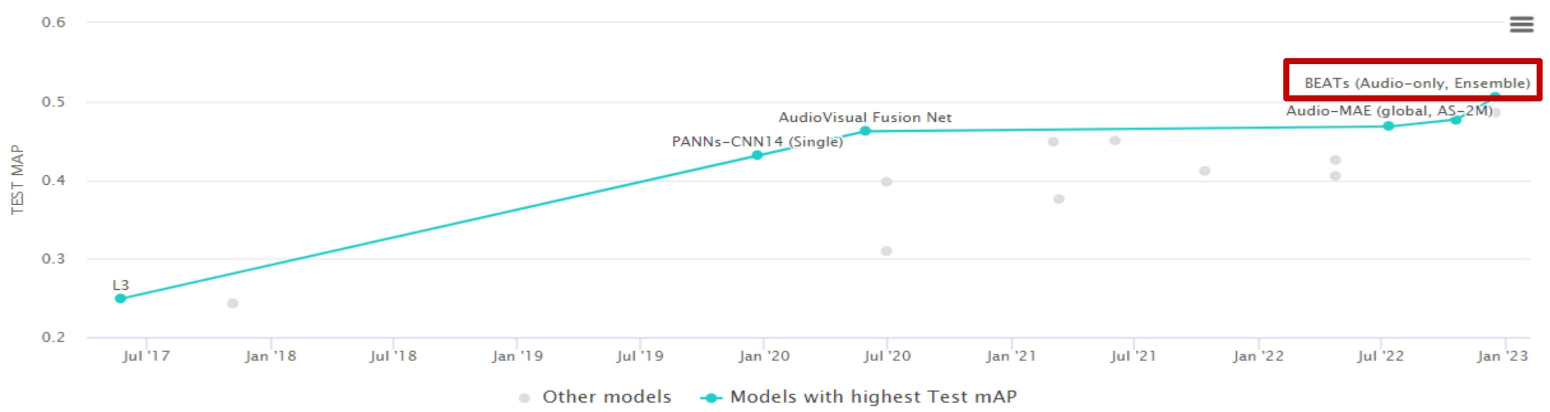
Model	SL Data	AS-2M
PSLA [Gong et al., 2021b]	IN+AS	47.4
AST [Gong et al., 2021a]	IN+AS	48.5
HTS-AT [Chen et al., 2022a]	IN+AS	48.7
PaSST [Koutini et al., 2021]	IN+AS	49.6
BEATs (5 models)	AS	50.4
BEATs (10 models)	AS	**50.6**

Audio Classification

Audio Classification on AudioSet

Leaderboard Dataset

View Test mAP by Date for Models not using extra training data

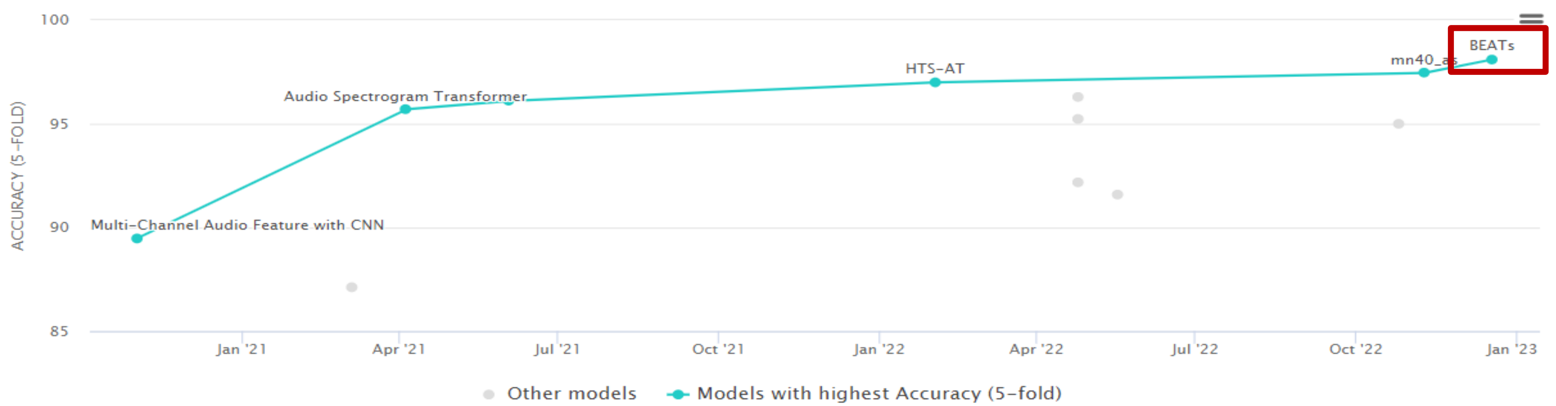


Audio Classification

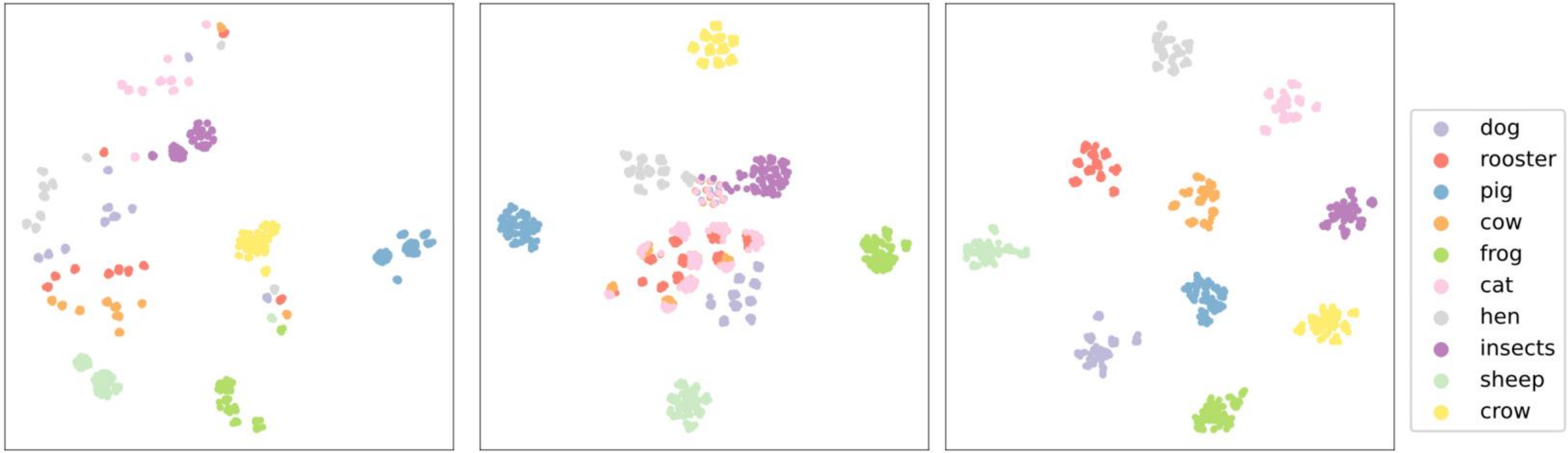
Audio Classification on ESC-50

Leaderboard Dataset

View Accuracy (5-fold) by Date for All models



Comparing Different Pre-Training Targets via Visualization



(a) Reconstruction (b) $\text{BEATs}_{\text{iter3}}$ (c) $\text{BEATs}_{\text{iter3+}}$

Broader Impacts

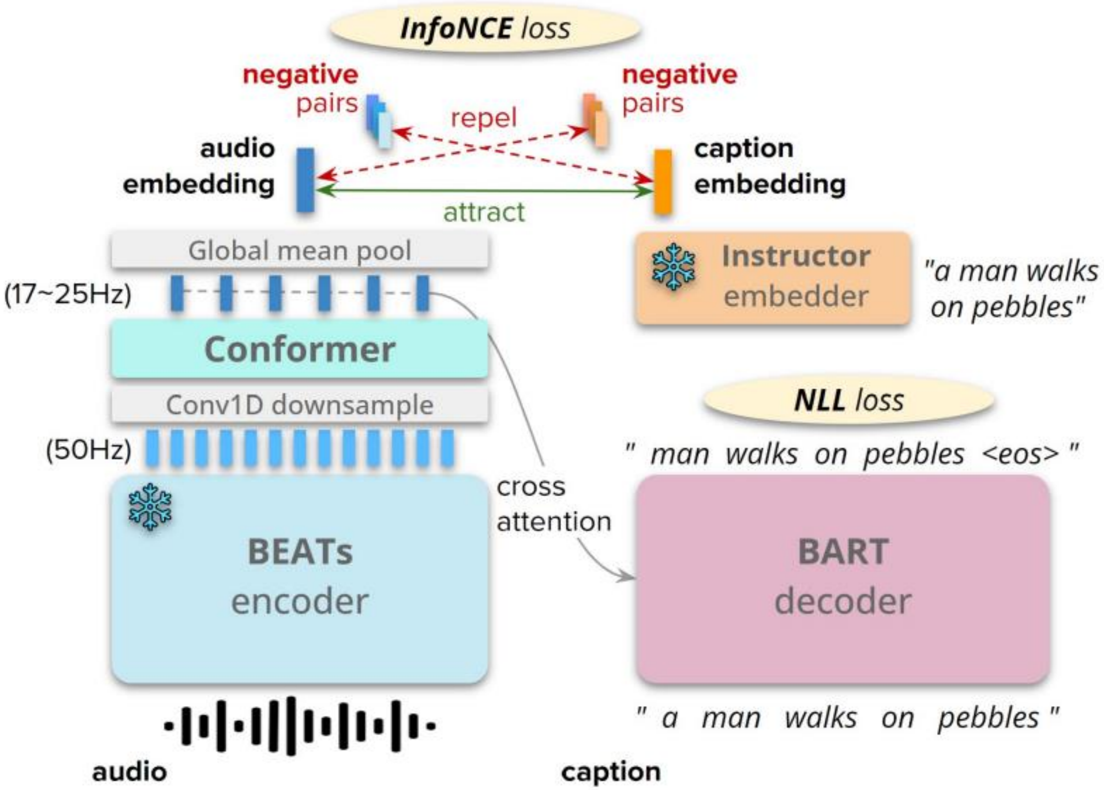
- Powers all the top 5 winning systems in DCASE 2023 Sound Event Detection Challenge

Systems with BEATs

Rank	Submission code (PSDS 1)	Submission code (PSDS 2)	Technical Report	Ranking score (Evaluation dataset)	PSDS 1 (Evaluation dataset)	PSDS 2 (Evaluation dataset)
1	Kim_GIST-HanwhaVision_task4a_2	Kim_GIST-HanwhaVision_task4a_3		1.68	0.591 (0.574 - 0.611)	0.835 (0.826 - 0.846)
2	Zhang_IOA_task4a_6	Zhang_IOA_task4a_7		1.63	0.562 (0.552 - 0.575)	0.830 (0.820 - 0.842)
3	Wenxin_TJU_task4a_6	Wenxin_TJU_task4a_6		1.61	0.546 (0.536 - 0.556)	0.831 (0.823 - 0.842)
4	Xiao_FMSG_task4a_4	Xiao_FMSG_task4a_4		1.60	0.551 (0.543 - 0.562)	0.813 (0.802 - 0.827)
4	Guan_HIT_task4a_3	Guan_HIT_task4a_4		1.60	0.526 (0.513 - 0.539)	0.855 (0.844 - 0.867)
5	Chen_CHT_task4a_2	Chen_CHT_task4a_2		1.58	0.563 (0.550 - 0.574)	0.779 (0.768 - 0.792)
6	Li_USTC_task4a_6	Li_USTC_task4a_6		1.56	0.546 (0.529 - 0.562)	0.783 (0.771 - 0.796)
7	Liu_NSYSU_task4a_7	Liu_NSYSU_task4a_7		1.55	0.521 (0.510 - 0.531)	0.813 (0.796 - 0.831)
8	Cheimariotis_DUTH_task4a_1	Cheimariotis_DUTH_task4a_1		1.53	0.516 (0.504 - 0.529)	0.796 (0.784 - 0.808)
9	Baseline_BEATS	Baseline_BEATS		1.52	0.510 (0.496 - 0.523)	0.798 (0.782 - 0.811)
10	Wang_XiaoRice_task4a_1	Wang_XiaoRice_task4a_1		1.50	0.494 (0.477 - 0.510)	0.801 (0.789 - 0.815)
11	Lee_CAUET_task4a_1	Lee_CAUET_task4a_2		1.28	0.425 (0.415 - 0.440)	0.674 (0.661 - 0.690)
12	Liu_SRCN_task4a_4	Liu_SRCN_task4a_4		1.25	0.412 (0.400 - 0.424)	0.663 (0.652 - 0.676)
13	Barahona_AUDIAS_task4a_2	Barahona_AUDIAS_task4a_4		1.21	0.380 (0.361 - 0.406)	0.673 (0.652 - 0.700)
14	Wu_NCUT_task4a_1	Wu_NCUT_task4a_1		1.15	0.391 (0.379 - 0.405)	0.596 (0.584 - 0.610)
15	Gan_NCUT_task4a_1	Gan_NCUT_task4a_1		1.12	0.365 (0.353 - 0.377)	0.603 (0.589 - 0.617)
16	Baseline	Baseline		1.00	0.327 (0.317 - 0.339)	0.538 (0.515 - 0.566)

Broader Impacts

- Powers all the top 5 winning systems in DCASE 2023 Sound Event Detection Challenge
- Powers the winning system in DCASE 2023 Automated Audio Captioning Challenge.



Wu et al. BEATs-based audio captioning model with instructor embedding supervision and ChatGPT mix-up

Conclusion

- We propose **BEATs**, an iterative audio pre-training framework, which opens the door to audio pre-training with a **discrete label prediction loss**.
- We provide **effective acoustic tokenizers** to quantize continuous audio features into semantic-rich discrete labels.
- We achieve **state-of-the-art results** on several audio understanding benchmarks.
- The pre-trained/fine-tuned models and codes are released at https://aka.ms/beats.