SSAST: Self-Supervised Audio Spectrogram Transformer

Yuan Gong, Cheng-I Jeff Lai, Yu-An Chung, James Glass *(MIT CSAIL)*

Code and Pretrained Models at: github.com/yuangongnd/ssast Contact: yuangong@mit.edu

AAAI-22

Introduction

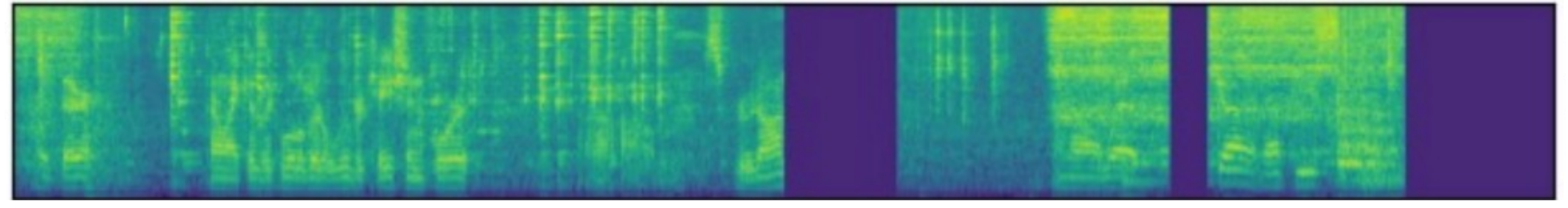
Audio Spectrogram Transformer (AST) is the first ***convolution-Free***, ***purely*** attention-based model for audio classification and achieves SOTA performance.

Problem: Original AST needs ***more labeled*** data to train, previous ImageNet supervised pretraining constrains AST to use ***16*16 square patch***.

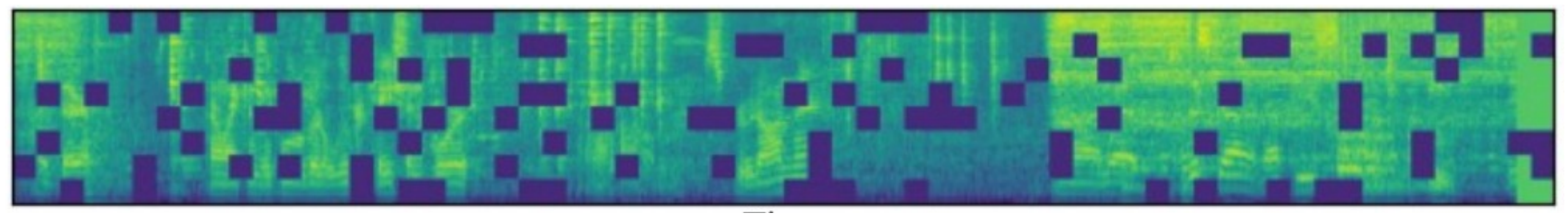
Proposed: A self-supervised pretraining framework that ***matches or outperforms*** previous supervised pretraining methods and supports ***arbitrary*** patch size and shape.

Key Novelty and Contribution

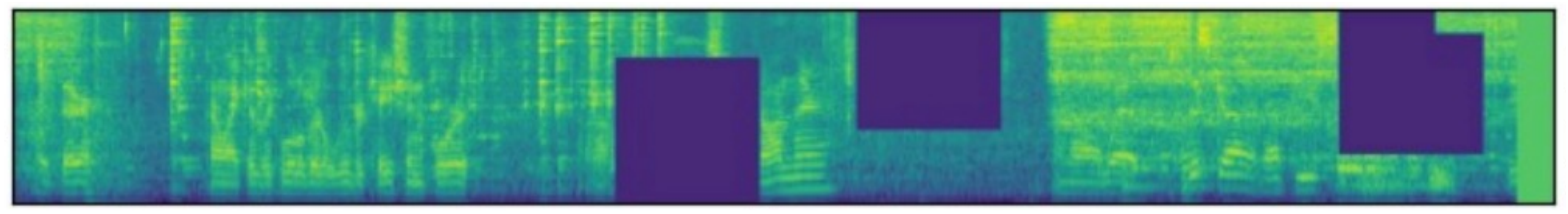
- ❑ First self-supervised pretraining framework for purely attention-based audio classification models.
- ❑ First patch masking based self-supervised pretraining framework in the audio & speech field.



Conventional frame-masking based SSL that only learns *temporal* spectrogram structure



Proposed patch-masking based SSL that learns *both temporal and frequency* structure



The model is forced to learn more *local* spectrogram structure with *smaller* masked patch size and more *global* spectrogram structure with *larger* masked patch size

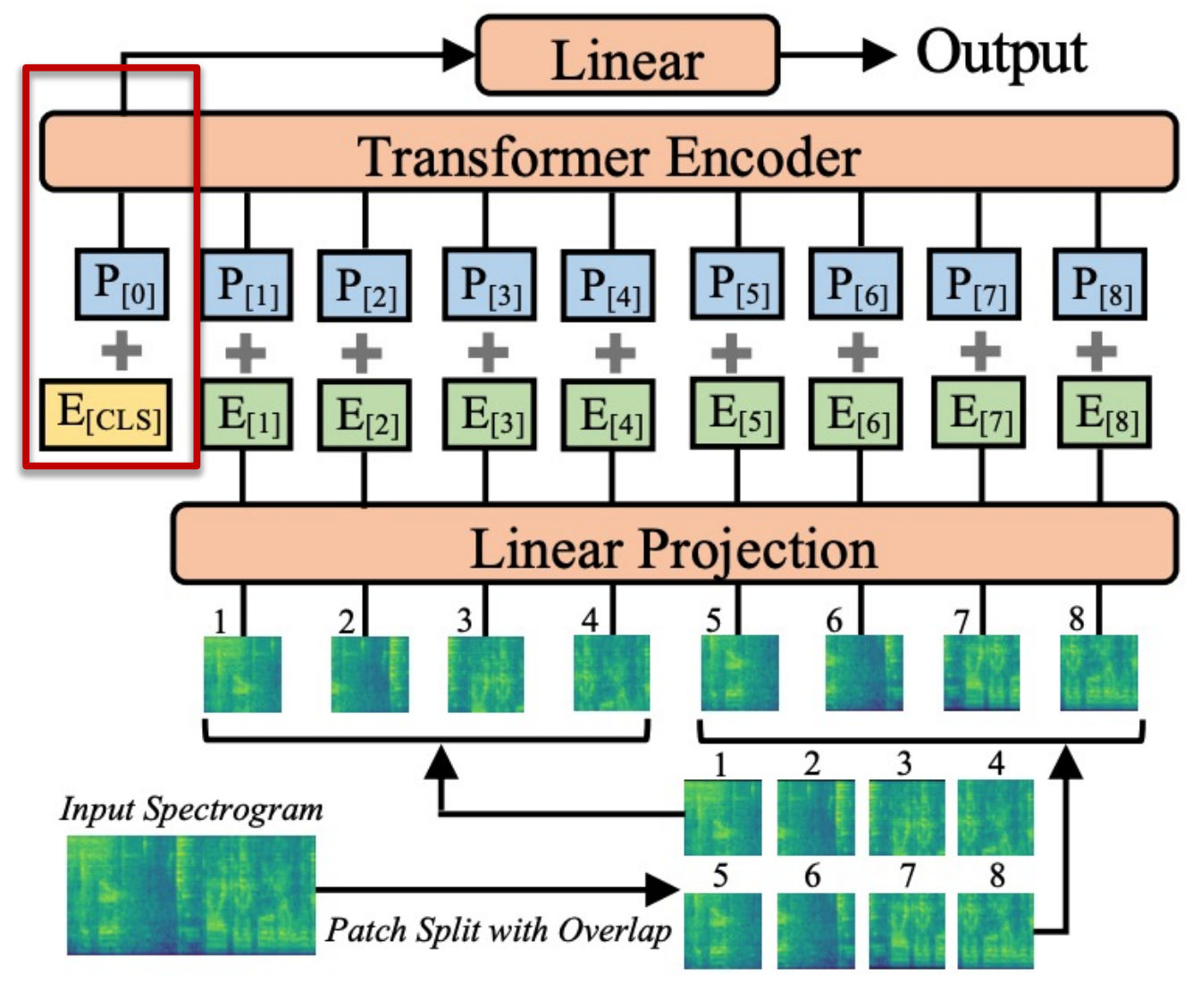
- ❑ Unlike ImageNet pretraining, SSAST supports arbitrary patch size and shape.

Square Patch Based AST

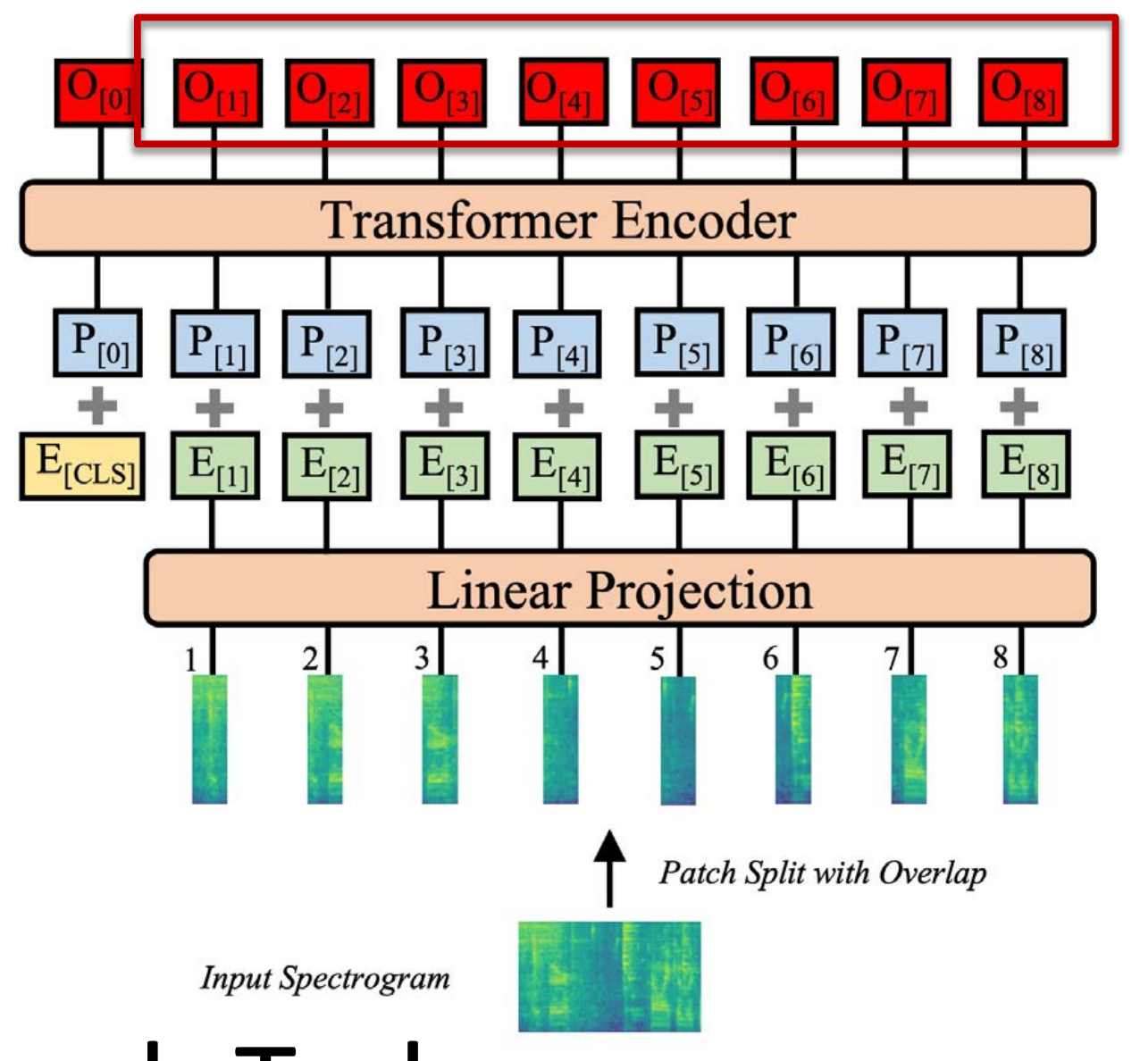
Cannot be used for frame-level feature extraction. Only designed for classification

Frame Based AST

Ideal for both classification and frame-level audio representation extraction

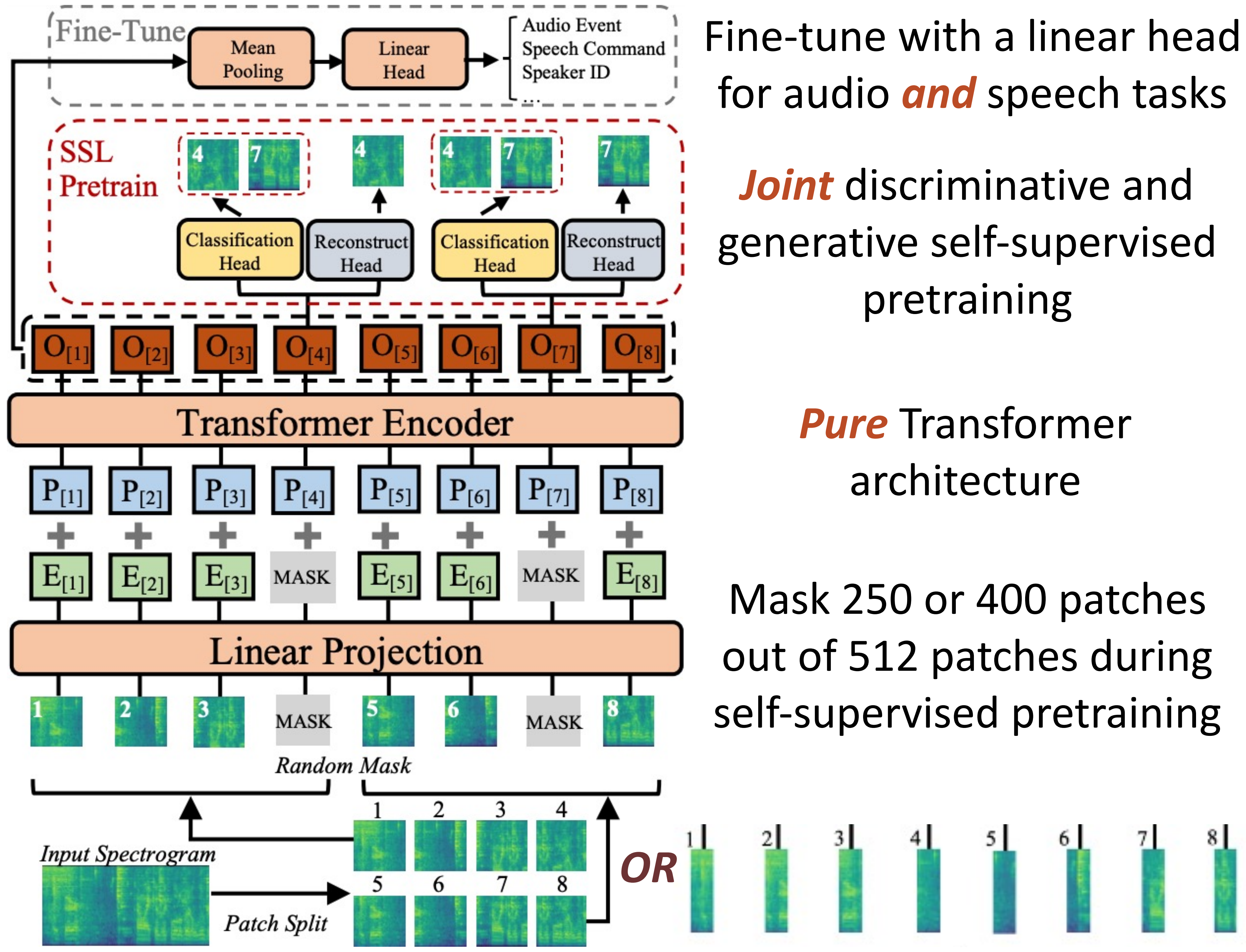


SSAST Support Both


O[0] O[1] O[2] O[3] O[4] O[5] O[6] O[7] O[8]
Transformer Encoder
P[0] P[1] P[2] P[3] P[4] P[5] P[6] P[7] P[8]
E[CLS] E[1] E[2] E[3] E[4] E[5] E[6] E[7] E[8]
Linear Projection
1 2 3 4 5 6 7 8
Patch Split with Overlap
Input Spectrogram

- ❑ Works for both audio and speech Tasks.

Method



Fine-tune with a linear head for audio ***and*** speech tasks

Joint discriminative and generative self-supervised pretraining

Pure Transformer architecture

Mask 250 or 400 patches out of 512 patches during self-supervised pretraining

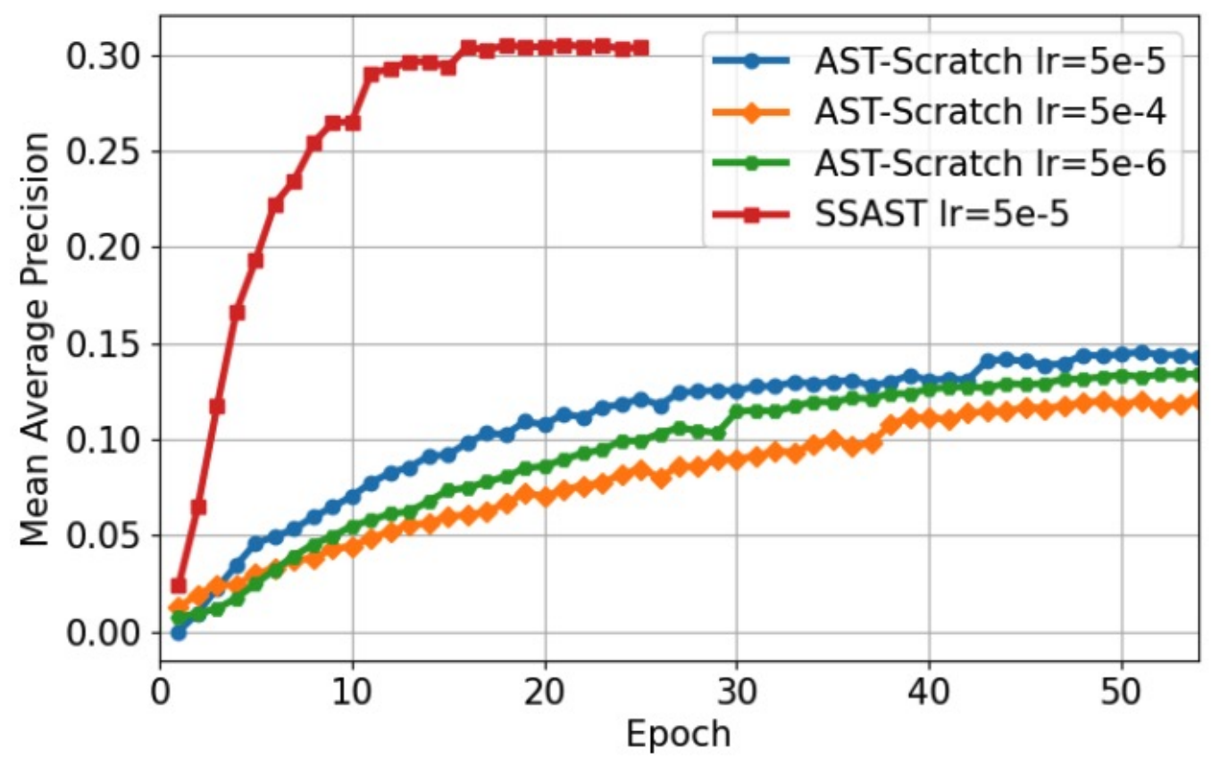
The spectrogram patch can be an ***arbitrary*** shape and size (e.g, a square patch or a conventional time frame).

Main Results

All models are pretrained and *end-to-end* fine-tuned (i.e., without layer freezing) on downstream datasets
AS=AudioSet-20K; ESC=ESC-50; KS=SpeechCommands; SID=VoxCeleb1; ER=IEMOCAP

	Model	Task					
		AS	ESC	KS2	KS1	SID	ER
No Pretraining	AST-Scratch	14.8	41.9	92.6	87.2	30.1	51.9
	Supervised Pretraining Baselines						
ImageNet Pretrain	AST-IM + KD	**34.7**	88.7	98.1	95.5	41.1	56.0
AudioSet Pretrain	AST-AudioSet	28.6	86.8	96.2	91.6	35.2	51.9
	Proposed Self-Supervised AST						
SSL Pretrain with 250/400 Masked Patches	SSAST 250	30.4	86.7	**98.1**	**96.2**	**66.6**	57.1
	SSAST 400	31.0	**88.8**	98.0	96.0	64.2	**59.6**

- ❑ The proposed ***self-supervised*** pretrained models significantly outperforms from-scratch models with an average improvement of ***60.9%***, and can match or even outperform previous supervised pretrained models.
- ❑ During fine-tuning, SSAST model learns much faster and better. Using a different learning rate or increasing training epochs cannot improve the from-scratch model performance.



Ablation Study

Setting	Task						
	AS	ESC	KS2	KS1	SID	ER	
From Scratch	14.8	41.9	92.6	87.2	30.1	51.9	
# Masked Patches							
100	28.7	85.3	98.0	94.9	62.1	57.3	①
250	30.4	86.7	**98.1**	**96.2**	**66.6**	57.1	
400 (Default)	**31.0**	**88.8**	98.0	96.0	64.3	**59.6**	
Pretext Task							
Discriminative	30.6	85.6	98.0	94.2	61.4	57.5	②
Generative	16.1	74.2	96.6	93.3	40.1	54.3	
Joint (Default)	**31.0**	**88.8**	**98.0**	**96.0**	**64.3**	**59.6**	
Pretraining Data							
AudioSet-20K	25.7	82.2	97.6	93.8	43.8	55.4	③
AudioSet 2M	29.0	84.7	97.8	94.8	57.1	56.8	
AudioSet 2M Supervised	28.6	86.8	96.2	91.6	35.2	51.9	④
Librispeech	22.9	80.0	97.8	95.6	60.8	58.3	⑤
Joint (Default)	**31.0**	**88.8**	**98.0**	**96.0**	**64.3**	**59.6**	

1. ***More*** masked patches leads to better performance.
2. ***Joint*** pretraining objective helps.
3. The proposed SSL works with ***small*** data.
4. With ***same*** AudioSet 2M data, SSL pretraining generalizes better than supervised pretraining.
5. Joint AudioSet and Librispeech pretraining leads to best performance for all tasks.

Patch Based- vs Frame Based- AST

Compare models pretrained and fine-tuned with 16*16 square patches and 128 *2 time frames

Model	Task					
	Audio		Speech			
	AS	ESC	KS2	KS1	SID	ER
Frame-Scratch	**16.6**	**53.7**	**96.0**	**91.7**	**54.9**	51.2
Patch-Scratch	14.8	41.9	92.6	87.2	30.1	**51.9**
SSAST-Frame-250	27.1	84.0	98.0	**96.6**	**73.6**	**58.3**
SSAST-Patch-250	**30.4**	**86.7**	**98.1**	96.2	66.6	57.1
SSAST-Frame-400	29.2	85.9	**98.1**	**96.7**	**80.8**	**60.5**
SSAST-Patch-400	**31.0**	**88.8**	98.0	96.0	64.2	59.6
Frame-Improvement	12.6	32.2	2.1	5.0	25.9	**9.3**
Patch-Improvement	**16.2**	**46.9**	**5.4**	**8.8**	**34.1**	7.7

- ❑ Frame-based AST is better for ***speech*** tasks while patch-based AST is better for ***audio*** Tasks.
- ❑ SSL pretraining helps more for patch-based AST.

Acknowledgement

This work is partly supported by Signify.