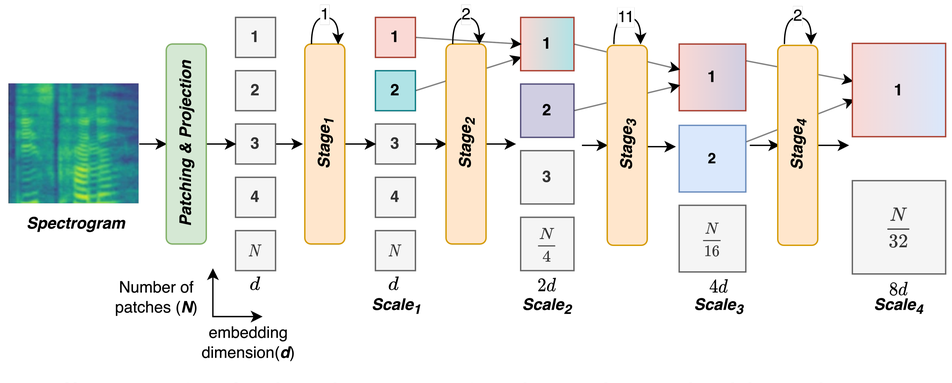
MAST: Multiscale Audio Spectrogram Transformers

Sreyan Ghosh[2*], Ashish Seth[1*], S.Umesh[1], Dinesh Manocha[2]
[1]Indian Institute of Technology, Madras, [2]University of Maryland, College Park

Motivation: Exploring multiscale hierarchical structures in audio

- We introduce MAST (Multiscale Audio Spectrogram Transformer), which builds on AST and modifies the AST architecture to incorporate the idea of multiscale feature hierarchies into it.
- We also present SS-MAST, a new SSL approach that helps MAST achieve higher performance in low-resource supervised learning settings.

Multi Head Pooling Attention (MHPA)



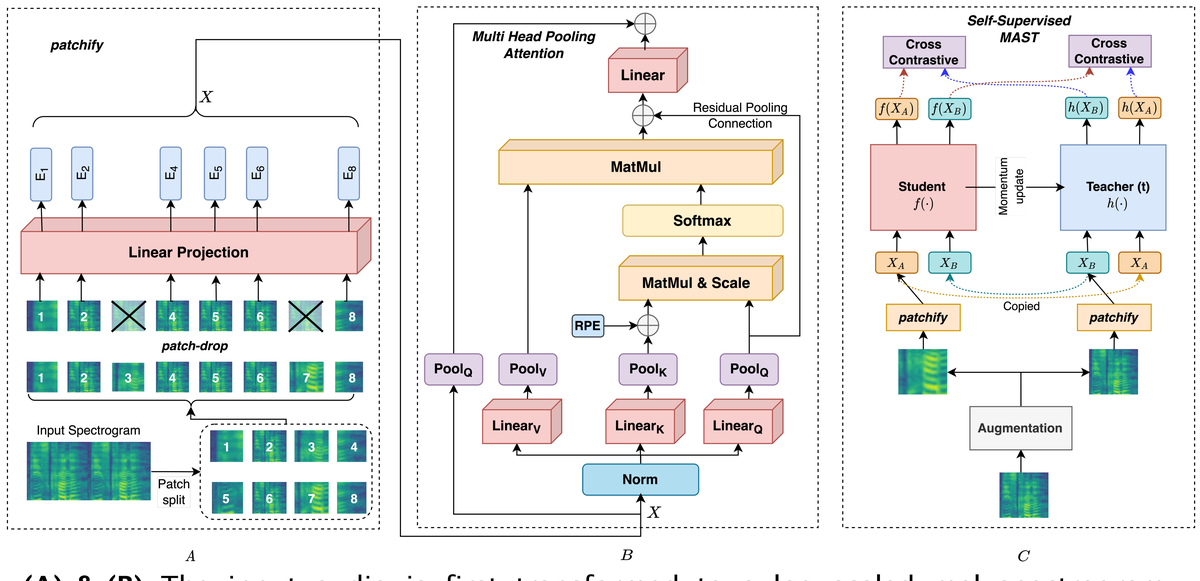
- Unlike vanilla MHA (multi-head attention), where the embedding dimension and the temporal resolution remain fixed, MHPA pools the sequence of latent tensors to reduce the sequence length along the time axis.
- MHPA first project input (**X**) to the key (**K**), value (**V**) and query (**Q**). Next, a pooling step is applied such that $K, Q, V \in \mathbb{R}^{\tilde{L} \times D}$ where $\tilde{L} = \lfloor (L + 2p - k)/s \rfloor + 1$

Results: MAST Vs AST

Models have been pre-trained on 10% of AudioSet and FSD50K and then linearly evaluated while updating the pre-trained weights on various downstream tasks

Model	Initialization	Speech Tasks						Non-Speech Tasks	
		SC-V1	SC-V2 (12)	SC-V2 (35)	VC	IC	VF	NS	US8K
AST	random	87.3	88.2	**92.7**	30.1	51.9	**72.3**	70.9	50.1
AST	IN weights	90.0	91.1	93.1	51.2	54.2	79.8	71.1	62.3
AST	IN+SSAST	95.5	94.2	94.4	53.3	55.2	81.8	74.3	68.3
AST	IN+SS-MAST	96.0	94.4	95.4	53.4	60.1	88.8	76.4	79.3
MAST	random	91.0	92.2	93.4	33.2	58.3	74.3	73.4	54.4
MAST	IN weights	92.0	93.1	94.2	54.4	61.0	87.3	75.4	64.4
MAST	IN+SS-MAST	97.0	96.8	96.4	56.7	64.0	89.2	80.6	84.0
MAST	IN+SS-MAST+pd	**97.4**	**96.8**	**96.6**	**57.3**	**64.4**	**90.0**	**81.2**	**84.8**

Proposed Architecture for MAST and SS-MAST



- **(A) & (B)** The input audio is first transformed to a log-scaled mel-spectrogram before it is patched and passed through multiple stages of MAST. We also introduce a patch-drop augmentation technique which randomly drops 20% of patches from the patched log-mel spectrogram and shows an additional improvement while pre-training MAST using SS-MAST
- **(C)** SS-MAST: For SSL pre-training of MAST, we make 2 copies of the randomly augmented log-mel-spectrogram and solve a cross-contrastive loss between the student and the momentum-teacher networks.

$$L_{InfoNCE}(f,h) = -\log\left(\frac{\exp\left(f\left(x_i^a\right)\cdot h\left(x_i^b\right)/\tau\right)}{\exp\left(f\left(x_i^a\right)\cdot h\left(x_i^b\right)/\tau\right) + \sum_{j=0}^{K}\exp\left(f\left(x_i^a\right)\cdot h\left(\tilde{x}_j\right)/\tau\right)}\right)$$

- Finally, cross-contrastive loss between the teacher and the student representations is calculated by: $L_{InfoNCE} = L_{InfoNCE}(f,h) + L_{InfoNCE}(h,f)$

Conclusion

- MAST outperforms AST across multiple pre-training settings: random, Image-Net (IN) weights, IN+SSAST, IN+SS-(MAST/AST). We don't implement SSAST on MAST due to it's pooling nature.
- Introduced patch-drop augmentation technique for pre-training MAST boots the performance of SS-MAST by **0.5%** averaged across all the downstream tasks.

Paper

Code