AST: Audio Spectrogram Transforemr

Gong, Yuan, Yu-An Chung, and James Glass. MIT Computer Science and Artificial Intelligence Laboratory, Cambridge, MA 02139, USA*arXiv preprint arXiv:2104.01778* (2021)

경영과학연구실 이태헌
2023.05.03

Vision Transformer (VIT)

- **The Transformer architecture in NLP applied to the Vision domain**
- **Using the Transformer's Encoder architecture to divide a single image into patches and use them similar to words**

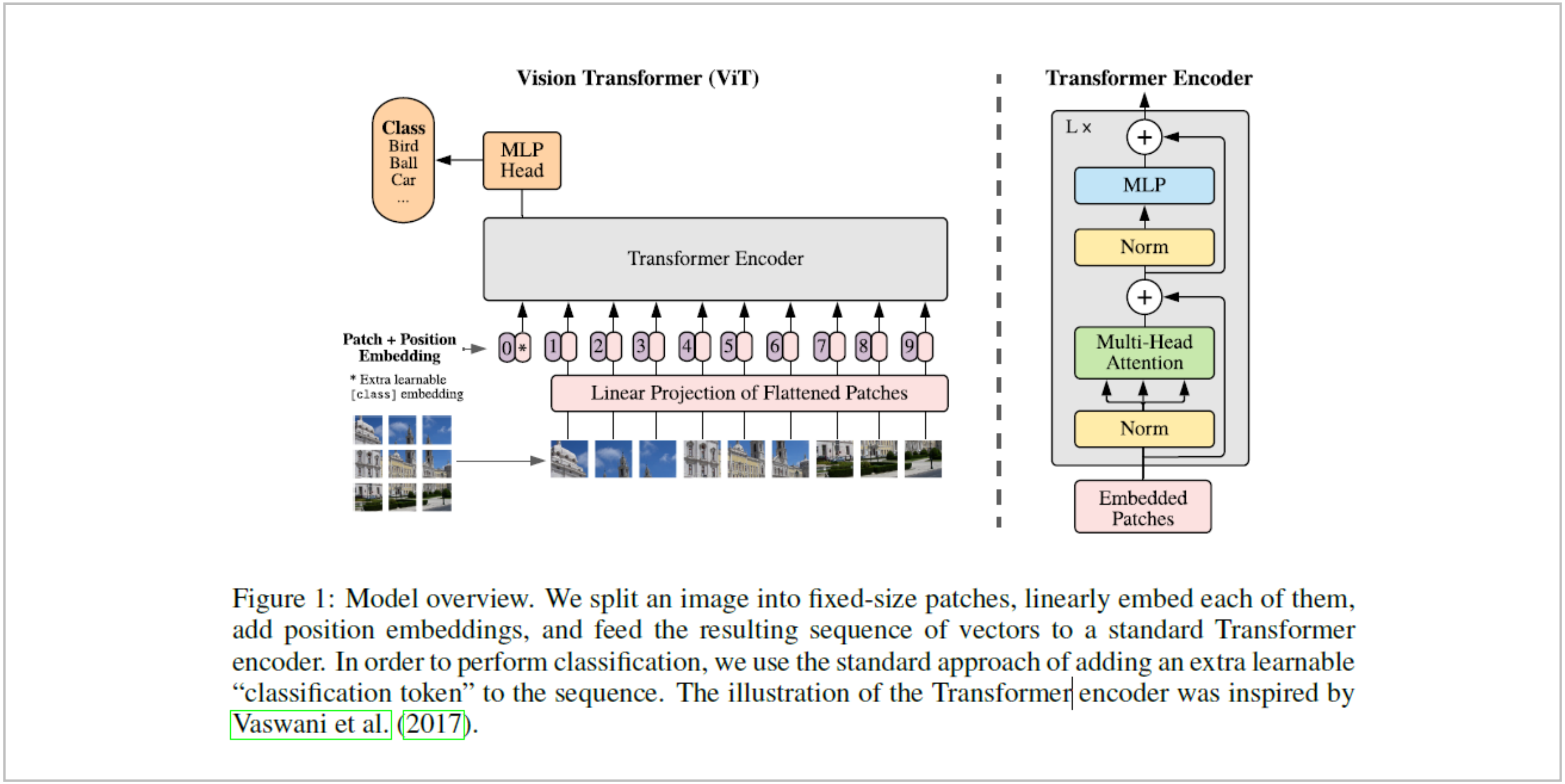


Figure 1: Model overview. We split an image into fixed-size patches, linearly embed each of them, add position embeddings, and feed the resulting sequence of vectors to a standard Transformer encoder. In order to perform classification, we use the standard approach of adding an extra learnable "classification token" to the sequence. The illustration of the Transformer encoder was inspired by Vaswani et al. (2017).

Low accuracy in music emotion classification

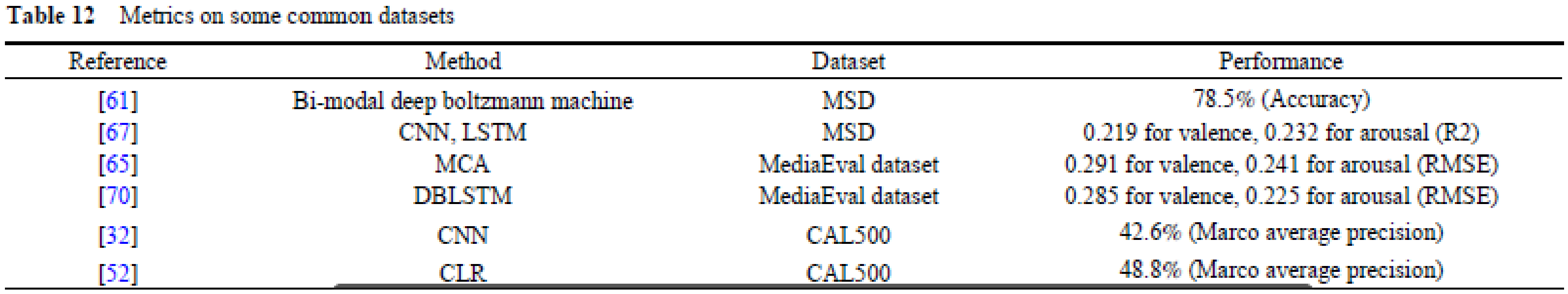


Table 12 Metrics on some common datasets

Reference	Method	Dataset	Performance
[61]	Bi-modal deep boltzmann machine	MSD	78.5% (Accuracy)
[67]	CNN, LSTM	MSD	0.219 for valence, 0.232 for arousal (R2)
[65]	MCA	MediaEval dataset	0.291 for valence, 0.241 for arousal (RMSE)
[70]	DBLSTM	MediaEval dataset	0.285 for valence, 0.225 for arousal (RMSE)
[32]	CNN	CAL500	42.6% (Marco average precision)
[52]	CLR	CAL500	48.8% (Marco average precision)

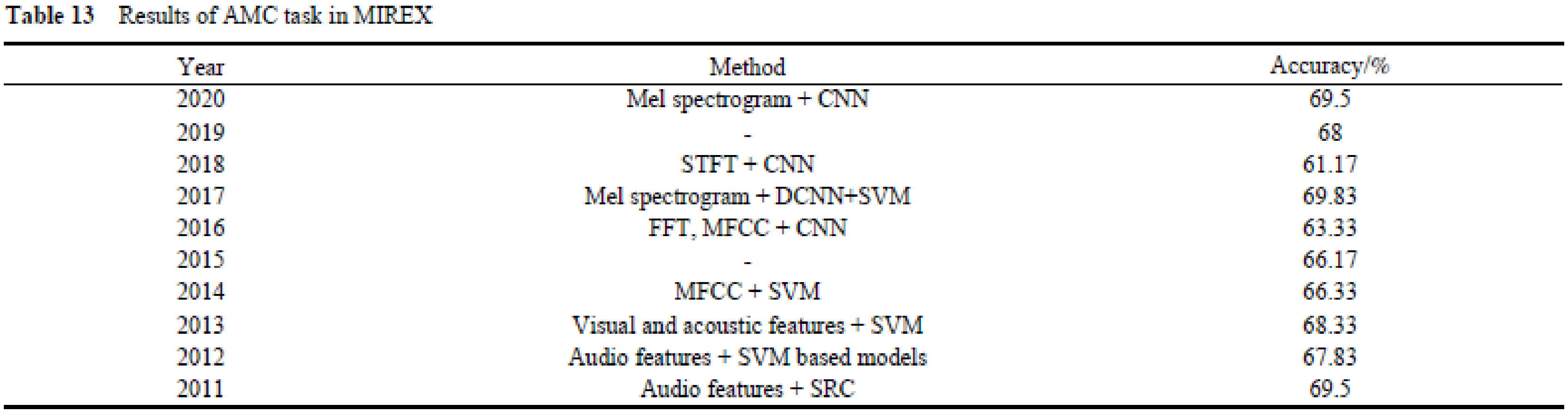


Table 13 Results of AMC task in MIREX

Year	Method	Accuracy/%
2020	Mel spectrogram + CNN	69.5
2019	-	68
2018	STFT + CNN	61.17
2017	Mel spectrogram + DCNN+SVM	69.83
2016	FFT, MFCC + CNN	63.33
2015	-	66.17
2014	MFCC + SVM	66.33
2013	Visual and acoustic features + SVM	68.33
2012	Audio features + SVM based models	67.83
2011	Audio features + SRC	69.5

Han, Donghong, et al. "A survey of music emotion recognition." *Frontiers of Computer Science* 16.6 (2022): 166335.

Related works

Transformer based models

- Vaswani, Ashish, et al. "Attention is all you need." *Advances in neural information processing systems* 30 (2017).
- Dosovitskiy, Alexey, et al. "An image is worth 16x16 words: Transformers for image recognition at scale." *arXiv preprint arXiv:2010.11929* (2020).
- *H. Touvron, M. Cord, M. Douze, F. Massa, A. Sablayrolles, andH. J ́egou, "Training data-efficient image transformers & distillation through attention," arXiv preprint arXiv:2012.12877, 2020.*

The importance of the number of data in Transformer

- *A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16x16 words: Transformers for image recognition at scale," in ICLR, 2021.)*

Problem statement

- **Using Transformer-based models to address the weaknesses of CNN-based models that were commonly used in existing image and audio spectrogram tasks. Through this, the goal is to solve audio classification problems**

- CNN is widely used in spectrogram tasks because it is believed that the inductive bias and translational equivalence inherent in CNN are helpful

- However, since audio spectrograms are long-term context images, it is difficult to extract features with only CNN. Therefore, self-attention mechanism is often added to better capture global context

- Based on the success of VIT in the image domain, there is an attempt to apply purely attention-based models to audio classification problems

Key idea

1. Applying the successful ViT in the image vision field to audio spectrogram tasks that are similar to images

2. Modifying the model structure to enable transfer learning of pre-trained ViT to AST

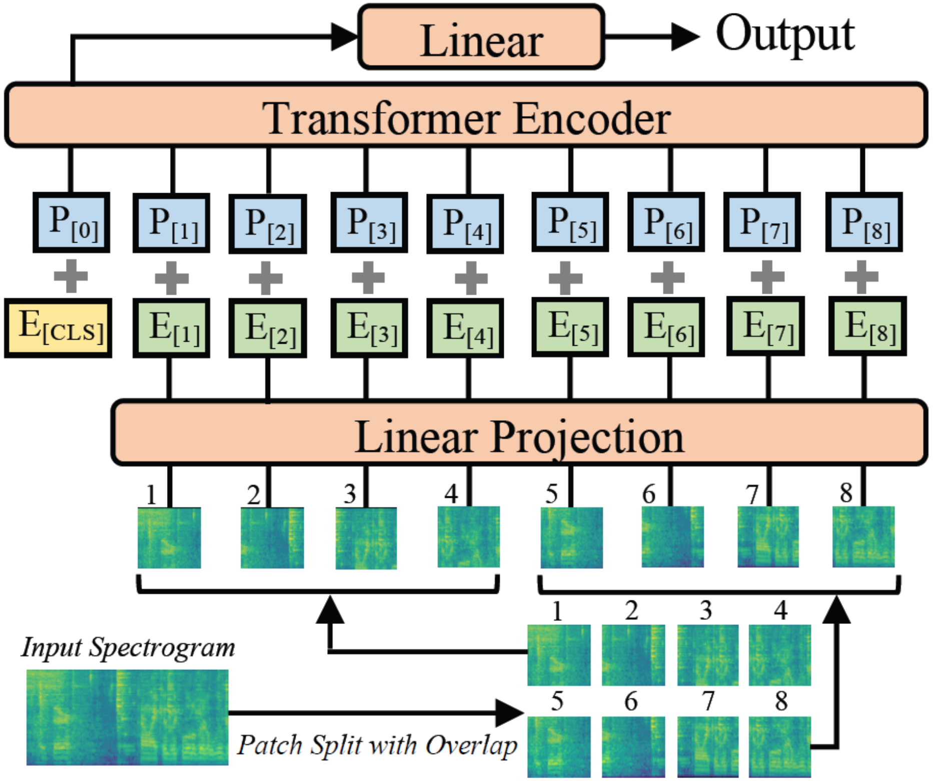
    - Resizing input images and using interpolation methods
    - Changing the weight of 3-channel image input to apply to 1-channel grayscale spectrogram
    - Modifying the structure of the output layer

3. Patch overlap to preserve the characteristics of audio spectrograms

4. Applying ensemble techniques to Transformer models

Model Architecture

- **Audio spectrogram transformer (AST) architecture**



- **The Transformer is composed of multiple encoder and decoder layers. However, since AST is designed for classification tasks, only the encoder of the Transformer is used**

- **AST and ViT have similar structures**

1) The standard Transformer architecture is available in TensorFlow and PyTorch, and it is easy to implement and reproduce

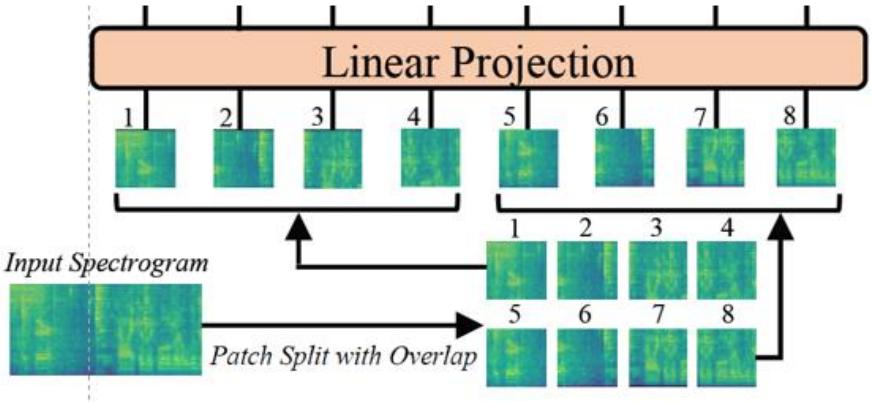
2) It is easy to transfer learning of pre-trained ViT on ImageNet to AST

Model Architecture

- **Audio spectrogram transformer (AST) architecture**

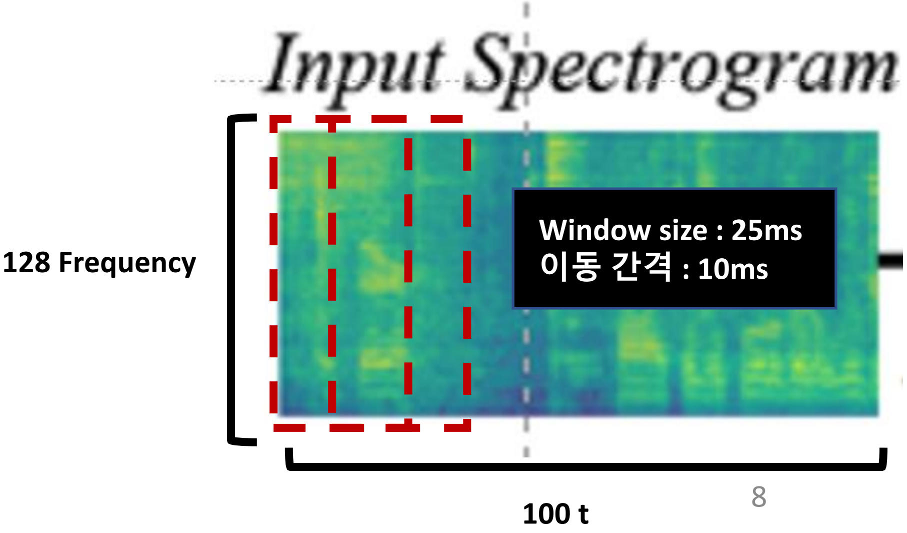
Window function

- In audio analysis, the main role of the window function is to divide the signal into multiple frames and reduce discontinuity and distortion that can occur at frame boundaries
- Using a window function can lead to better results in frequency analysis



Input spectrogram

- The input spectrogram for T seconds is calculated with a window size of 25ms every 10ms
- Transforming the waveform into a 128-dimensional space through Mel-filter bank conversion
- Generating a 128 x 100t spectrogram that is input to AST



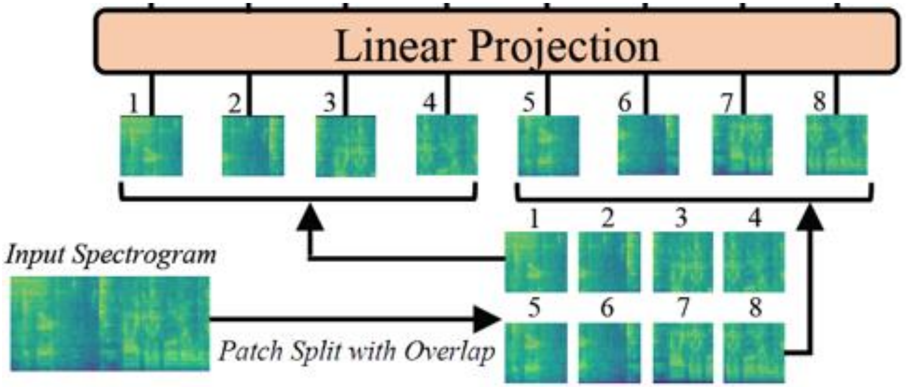
Model Architecture

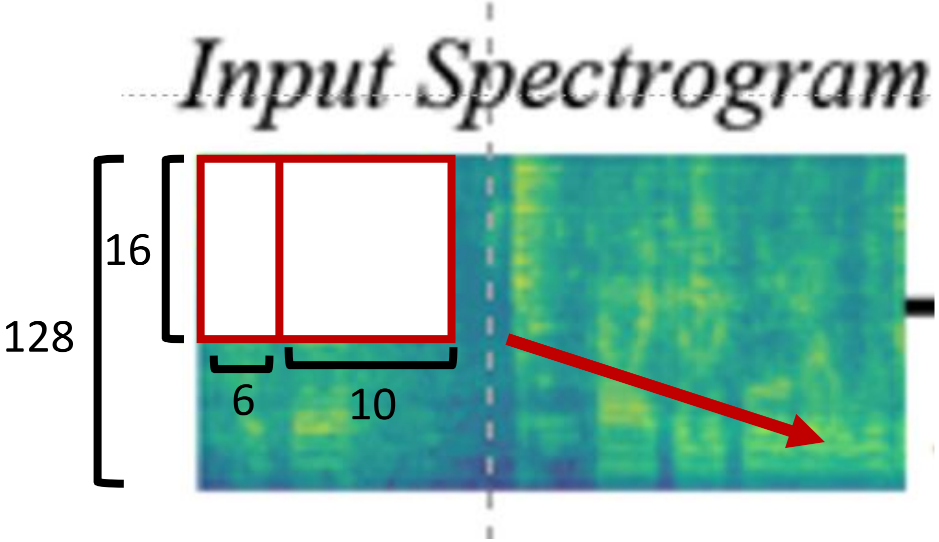
- **Audio spectrogram transformer (AST) architecture**

Dividing spectrograms into N 16 x 16 patches

- **Split the spectrogram into a sequence of N 16x16 patches with an an overlap of 6**
- **N is the number of patches, which is the actual input sequence length of AST**

- Creating 16x16 size patches starting from the top left of the spectrogram
- When creating the next patch, move horizontally and vertically to avoid 10 overlapping elements in both axes due to overlap consideration
- Repeat the above process until the bottom right of the spectrogram is reached

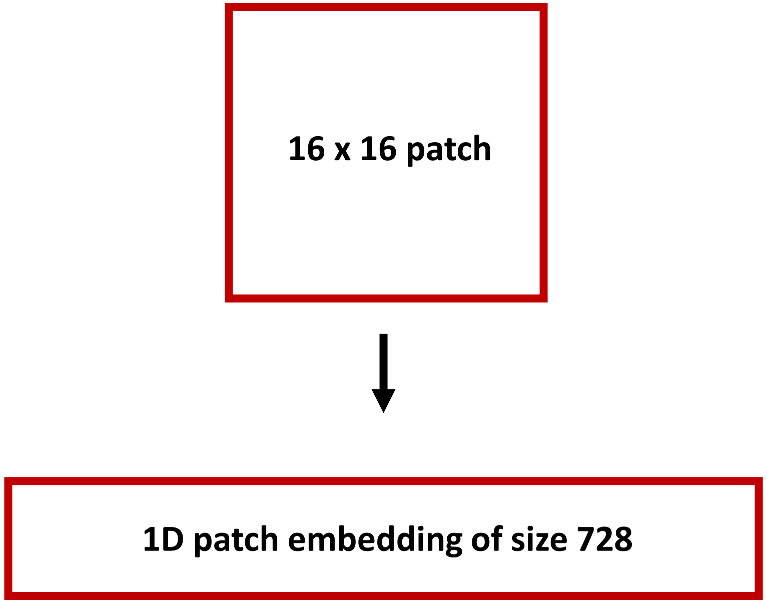




Model Architecture

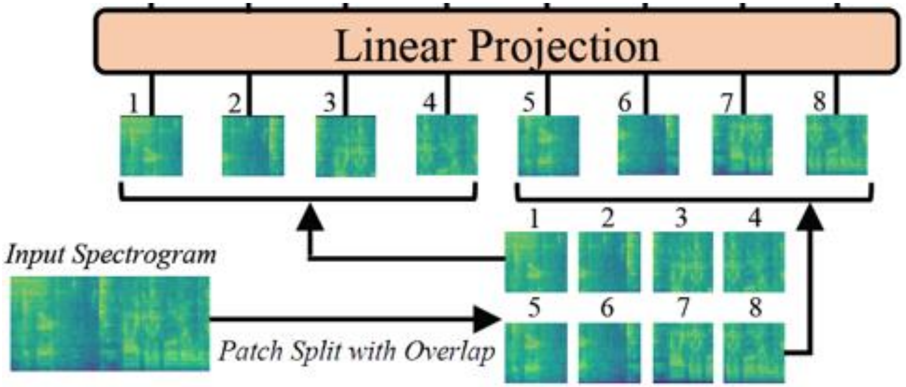
- **Audio spectrogram transformer (AST) architecture**

Linear Projection Layer



16 x 16 patch

1D patch embedding of size 728



- **Flattening each 16x16 patch into a 1D patch embedding of size 768 using a Linear projection layer**

Model Architecture

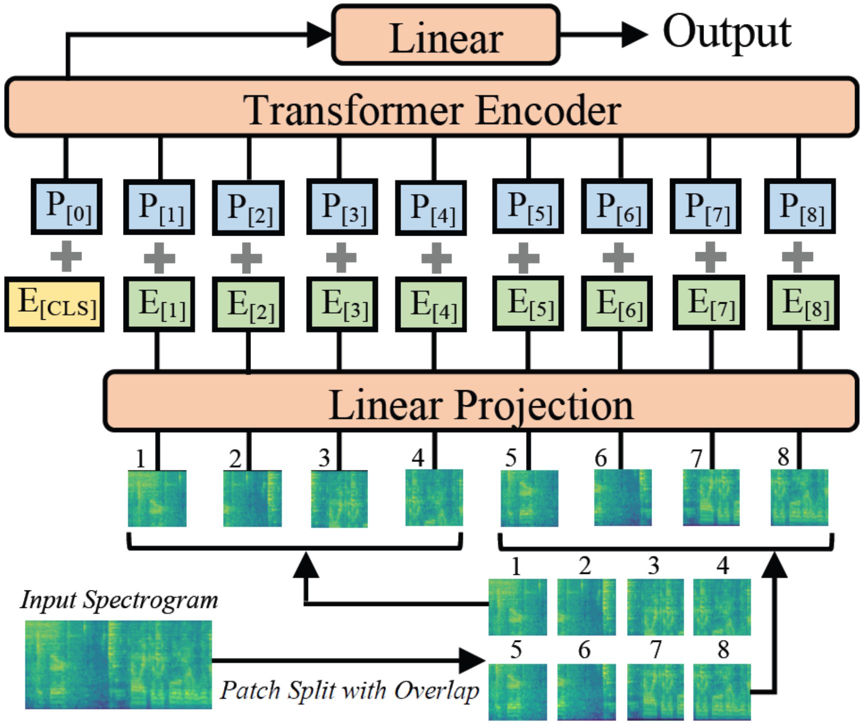
- **Audio spectrogram transformer (AST) architecture**

Positional Embedding

- Positional Embedding plays a role in conveying the spatial position information of each patch in the image to the Transformer model
- Since the Transformer cannot encode information about order or position on its own, positional embeddings are necessary

Append a [CLS] token

- Summarizing the information of the entire input image
- The [CLS] token receives information about each patch embedding and is used to understand the overall meaning of the input image
- Adding the [CLS] token at the beginning of the sequence
- The resulting sequence is then input to the Transformer

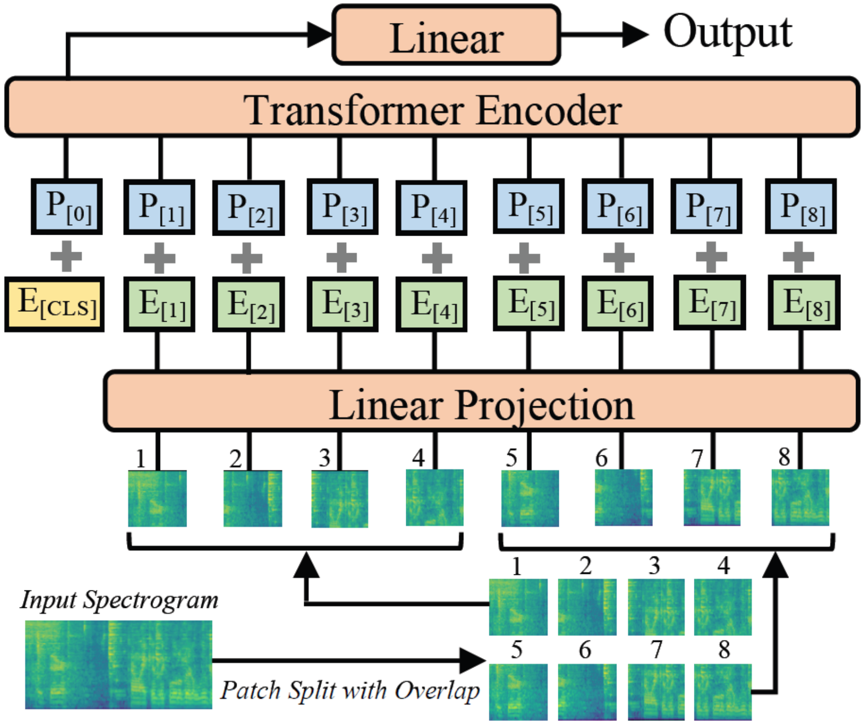


Model Architecture

- **Audio spectrogram transformer (AST) architecture**

Linear layer

- A linear layer with sigmoid activation maps the audio spectrogram representation to labels for classification



ImageNet Pretraining

- One disadvantage of the Transformer compared with CNNs is that the Transformer needs more data to train

- According to A. Dosovitskiy, et al. "An image is worth 16x16 words: Transformers for image recognition at scale", ICLR, 2021, Transformers outperform CNNs in image classification tasks only when the data volume is more than 14 million

- However, audio datasets generally do not have large amounts of data

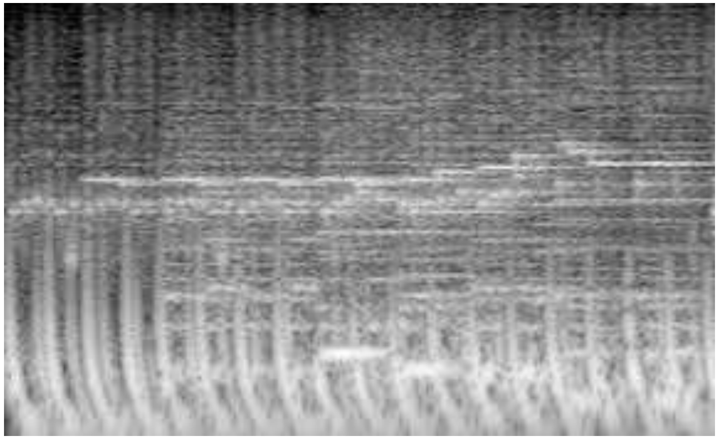
- Since images and audio spectrograms have similar formats, this motivates the idea of applying cross-modality transfer learning to AST

ImageNet Pretraining

- ViT and AST have similar architectures, but they are not identical
- The weights corresponding to the three input channels of the ViT patch embedding layer are averaged and used as the weights of the AST patch embedding layer

1. The input of ViT is a 3-channel image, while the input of AST is a single-channel spectrogram

3-channel RGB image



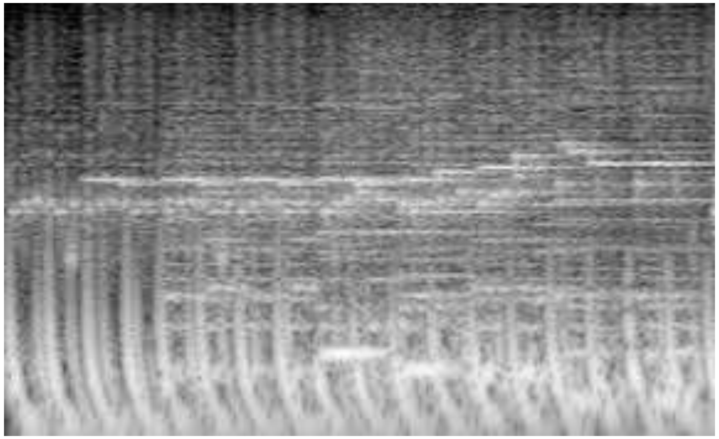
1-channel gray-scale spectrogram

ImageNet Pretraining

- **ViT and AST have similar architectures, but they are not identical**
- **The input size of ViT is fixed at 224 x 224 or 384 x 384, but the length of audio spectrograms is variable**

2. 고정된 ViT 입력 사이즈와 가변적인 AST 입력 사이즈

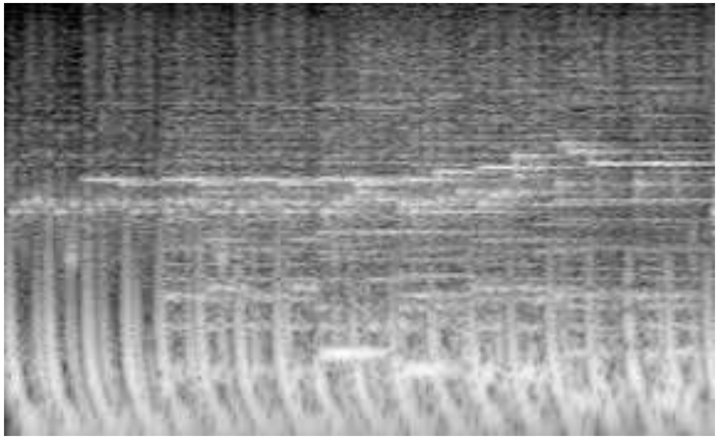
고정된 ViT 입력 사이즈



가변적인 Audio spectrogram 사이즈

ImageNet Pretraining

- **CUT and bi-linear interpolated methods are used for adaptive positional embedding**



- If divided into patch sizes of 16 x 16, for ViT, the number of patches and corresponding positional embeddings would be 24 x 24 = 576
- ViT divides patches without overlapping

- AST, which uses a 10-second audio input, has 12 x 100 patches, and each patch requires a positional embedding

Even in cases where the input size is different, pre-trained ViT can transfer 2D spatial knowledge to AST

Dataset

- The experiments were primarily focused on the Audioset dataset
- In addition, experiments were conducted on the ESC-50 and Speech Commands v2 datasets

1. Audioset dataset

- Audioset is a collection of over 2 million 10-second audio clips excised from Youtube videos and labeled with the sounds that the clip contains from a set of 527 labels

2. Esc-50 dataset

- ESC-50 dataset consists of 2,000 5-second environmental audio recordings organized into 50 classes

3. Speech commands v2 dataset

- Speech commands v2 is a dataset consists of 105,829 1-second recordings of 35 common-speech commands

Audioset Results

Two subsets of Audioset: the Balanced set and the Full set

- Audioset consists of two subsets: the Balanced set and the Full set
- Balanced set : The Balanced set includes an equal number of samples from each class to minimize the data imbalance problem
- Full set : Audioset dataset includes all samples. The data distribution between classes may be imbalanced

Two experimental settings for the ensemble method: Ensemble-S and Ensemble-M

- Ensemble-S : the experiment three times with the exact same setting, but with a different random
- Ensemble-M : Involves ensembling models trained with different settings. Specifically, three models trained with the same patch division strategy are ensembled with three models trained with a different patch division strategy

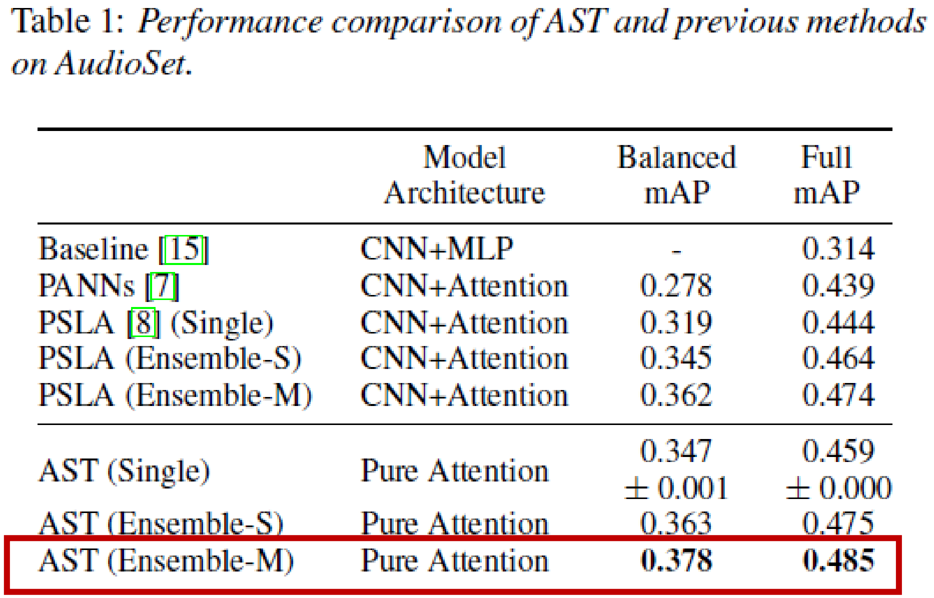


Table 1: *Performance comparison of AST and previous methods on AudioSet.*

	Model Architecture	Balanced mAP	Full mAP
Baseline [15]	CNN+MLP	-	0.314
PANNs [7]	CNN+Attention	0.278	0.439
PSLA [8] (Single)	CNN+Attention	0.319	0.444
PSLA (Ensemble-S)	CNN+Attention	0.345	0.464
PSLA (Ensemble-M)	CNN+Attention	0.362	0.474
AST (Single)	Pure Attention	0.347 ± 0.001	0.459 ± 0.000
AST (Ensemble-S)	Pure Attention	0.363	0.475
AST (Ensemble-M)	Pure Attention	**0.378**	**0.485**

mAP (mean Average precision)

- It is a useful method for evaluating the performance of a model on multiple classes in classification problems by comprehensively evaluating the relationship between the model's precision and recall
- after calculating the Precision and Recall for each class, the average performance across all classes is obtained by taking the mean of these values
- This approach is useful for evaluating the performance of a model for multiple classes in a classification task

Audioset Results

- Comparison between ImageNet pre-trained AST and randomly initialized AST
- Table 2 shows that ImageNet pre-trained AST performs better in both Balanced Audioset and Full Audioset compared to randomly initialized AST
- Table3 shows the results of a study on the impact of pre-trained weights. The performance of AST models initialized with pre-trained weights of Vit-Base, Vit-Large, and Deit models are compared

Impact of ImageNet Pretraining

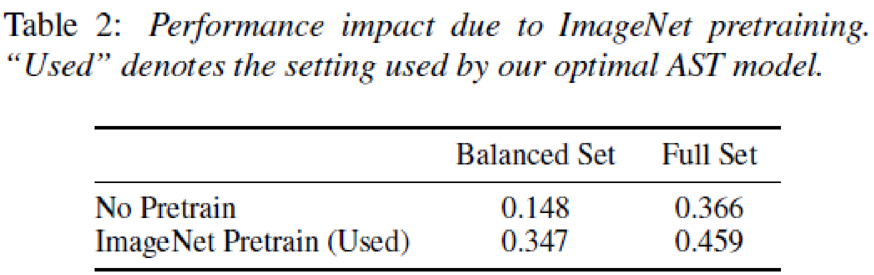


Table 2: *Performance impact due to ImageNet pretraining. "Used" denotes the setting used by our optimal AST model.*

	Balanced Set	Full Set
No Pretrain	0.148	0.366
ImageNet Pretrain (Used)	0.347	0.459

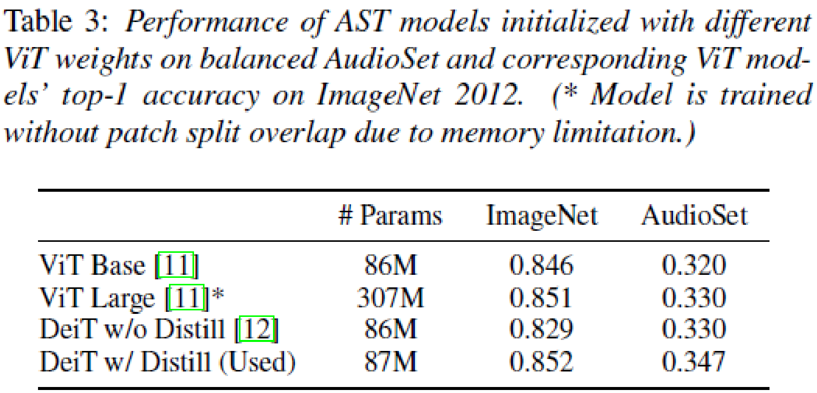


Table 3: *Performance of AST models initialized with different ViT weights on balanced AudioSet and corresponding ViT models' top-1 accuracy on ImageNet 2012. (* Model is trained without patch split overlap due to memory limitation.)*

	# Params	ImageNet	AudioSet
ViT Base [11]	86M	0.846	0.320
ViT Large [11]*	307M	0.851	0.330
DeiT w/o Distill [12]	86M	0.829	0.330
DeiT w/ Distill (Used)	87M	0.852	0.347

DeiT (Data-efficient Image Transformer)

- A model that uses the Knowledge Distillation technique, which is a method of transferring knowledge from a Teacher Model to a Student Model during training
- DeiT w/o Distill : The DeiT model without undergoing the process of knowledge distillation.
- DeiT w Distill : DeiT model with Knowledge distillation process

Audioset Results

- **In AST, CUT and bi-linear interpolation methods are used to adapt the positional embeddings when receiving knowledge from ViT**
- **Table4은 demonstrates the importance of transferring spatial knowledge. Bi-linear interpolation and nearest-neighbor interpolation**

Impact of Positional Embedding Adaptation

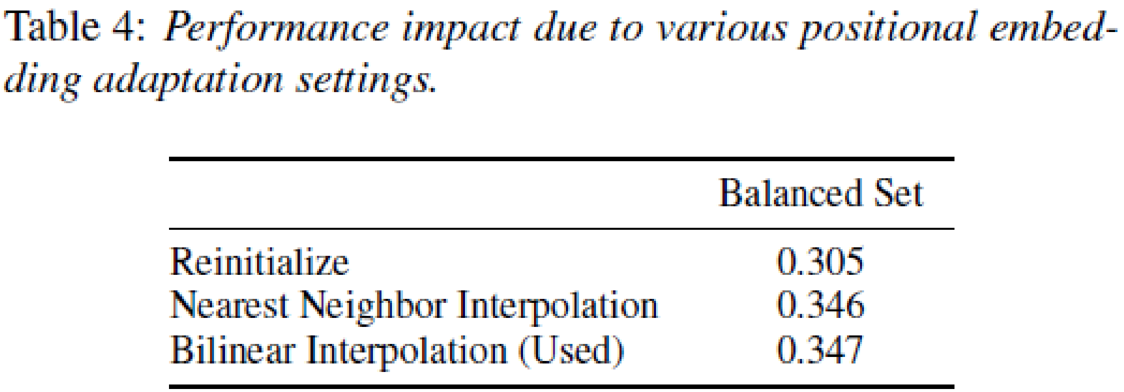


Table 4: *Performance impact due to various positional embedding adaptation settings.*

	Balanced Set
Reinitialize	0.305
Nearest Neighbor Interpolation	0.346
Bilinear Interpolation (Used)	0.347

Audioset Results

- Compare the performance of models trained with different patch split overlap
- Table 5 shows that performance improves as the overlap size increases for both Balanced set and Full set

Impact of Patch Split Overlap

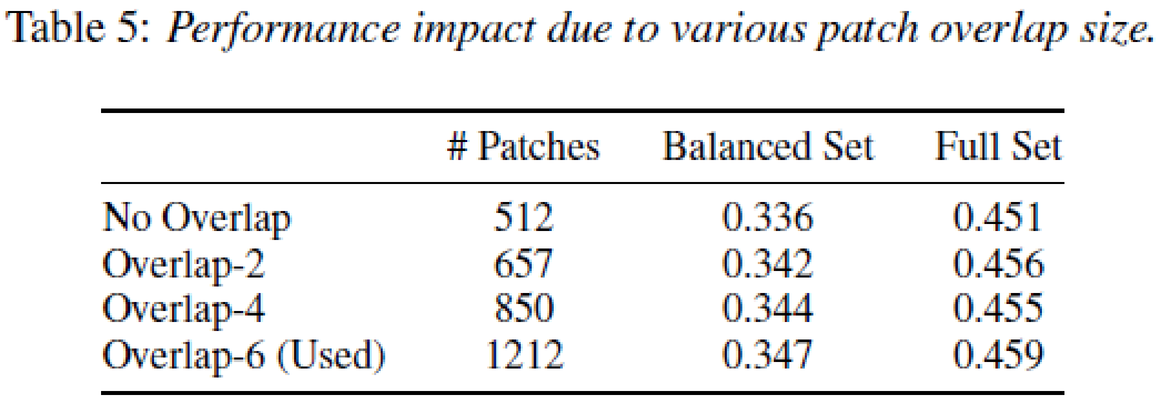
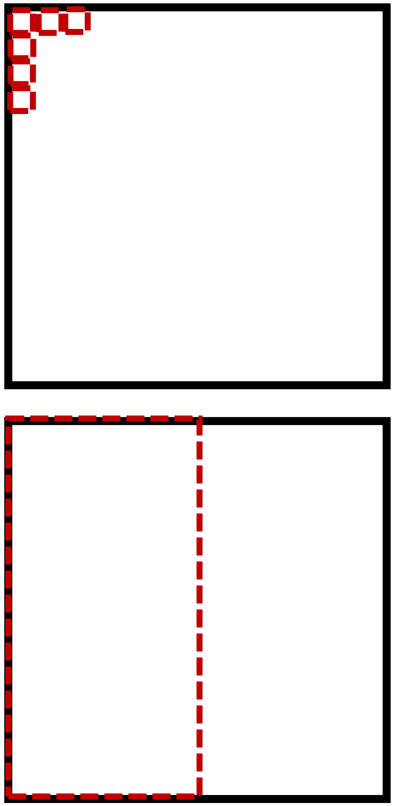


Table 5: *Performance impact due to various patch overlap size.*

	# Patches	Balanced Set	Full Set
No Overlap	512	0.336	0.451
Overlap-2	657	0.342	0.456
Overlap-4	850	0.344	0.455
Overlap-6 (Used)	1212	0.347	0.459

Audioset Results

- Split the audio spectrogram into 16 x 16 square patches, so the input sequence to the Transformer cannot be in temporal order
- Comparison of an alternative patching strategy, which involves cutting the audio spectrogram into rectangular patches in time order
- Table 6 shows that using 128x2 rectangular patches performs better than using 16x16 square patches
- However, since there is no ImageNet pre-trained model based on 128x2 patch, 16x16 patch is used



1. 16 x 16

256 x 256

2. 128 x 2

Impact of Patch shape and size

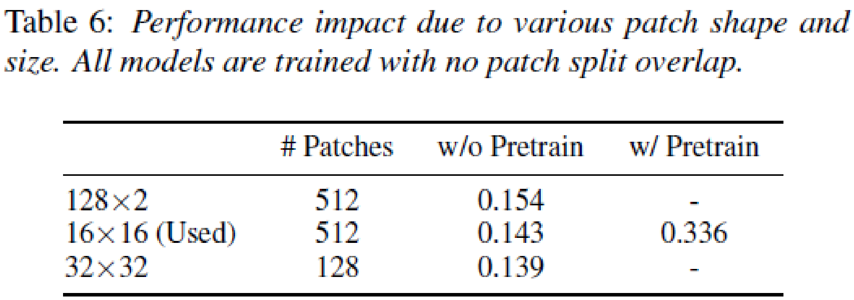


Table 6: *Performance impact due to various patch shape and size. All models are trained with no patch split overlap.*

	# Patches	w/o Pretrain	w/ Pretrain
128×2	512	0.154	-
16×16 (Used)	512	0.143	0.336
32×32	128	0.139	-

Results on ESC-50, Speech Commands

- SOTA-S : When the model is trained from no pretraining
- SOTA-P : When Audioset pretraining is used
- Comparing SOTA and AST models in the two different settings mentioned above
- The models are AST-S, which is trained only with ImageNet pretraining, and AST-P, which is trained with both ImageNet and Audioset pretraining

Comparison of S and P for SOTA and AST models

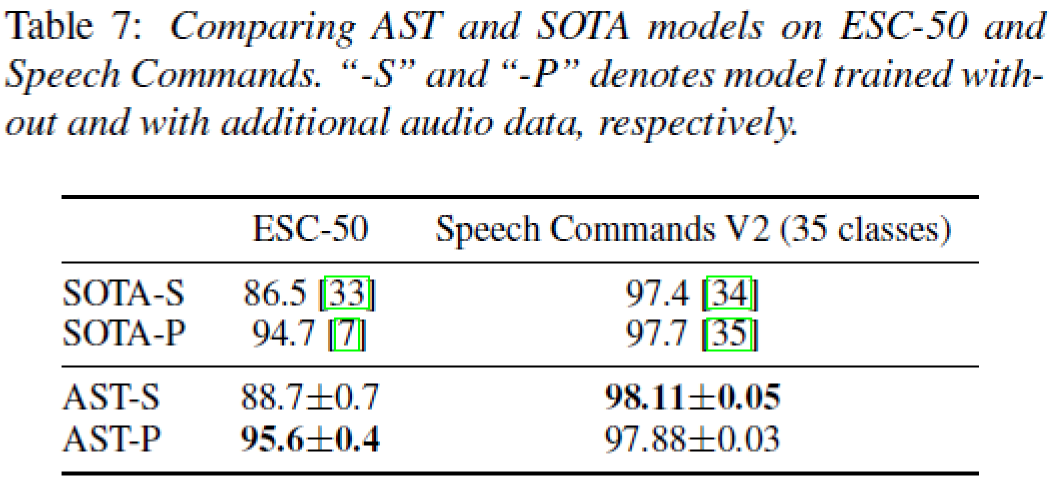


Table 7: *Comparing AST and SOTA models on ESC-50 and Speech Commands. "-S" and "-P" denotes model trained without and with additional audio data, respectively.*

	ESC-50	Speech Commands V2 (35 classes)
SOTA-S	86.5 [33]	97.4 [34]
SOTA-P	94.7 [7]	97.7 [35]
AST-S	88.7±0.7	**98.11±0.05**
AST-P	**95.6±0.4**	97.88±0.03

Conclusion & Further work

Conclusion

- **For a long time, CNN has been the most common and effective model in audio classification**
- **In this work, Audio Spectrogram Transformer (AST), a convolution-free, purely attention-based model for audio classification which features a simple architecture and superior performance**

Q & A