A Spectrogram is Worth 16 x 16 Words: Vision Transformers for Spoken Language Understanding

Pooja Sethi
Stanford University
pjasethi@stanford.edu

Jared Weissberg
Stanford University
jared1@stanford.edu

Kaushal Alate
Stanford University
kalate@stanford.edu

Abstract

The Audio Spectrogram Transformer (AST), a vision transformer adapted for audio, showed that CNNs are not a strictly necessary building block for end-to-end audio classification. AST achieves SOTA results for environmental sound classification with a simpler architecture and faster training convergence compared to traditional convolution-based audio encoders. We investigate AST's effectiveness in spoken language processing. We fine-tune AST for speech emotion recognition (SER), genre classification for music with singing, and automatic speech recognition (ASR). AST outperforms SOTA baselines for music genre classification as measured on the GTZAN dataset. Results on SER were mixed, with AST beating competitive baselines on the RAVDESS dataset but slightly under-performing on IEMOCAP. Finally, we introduce a novel CNN-free ASR architecture, ASTForCTC. In an ultra-low-resource setting using just 10 hours of training data, we achieve a train WER of 0.059 on Google FLEURS en_US. With 100 hours of training data, we achieve a test WER of 0.738 on Librispeech. As most SOTA ASR models are pre-trained on significantly more data (50K-1M hours), we expect test WER to improve given more training data. Our code can be found here.

1 Introduction

Convolutional neural networks (CNNs) have been widely used to learn representations of spectrograms for end-to-end audio and speech modeling tasks. Recently, the Audio Spectrogram Transformer (AST) (Gong et al., 2021), a vision transformer (ViT) (Dosovitskiy et al., 2021) for audio classification, has been shown to achieve superior performance on certain tasks while having a simpler architecture than CNN-based models. The simpler architecture makes it faster to train, which is advantageous for large-scale machine learning applications. Gong et al. (2021) show that AST achieves state-of-the-art results on classification tasks such as AudioSet (Gemmeke et al., 2017) and ESC-50 (Piczak), which primarily focus on discriminating across environmental sounds.

AST also achieves SOTA results on Speech Commands (Warden, 2018), which is a simple classification task of 35 words. Beyond this, there has been little additional investigation into AST's performance on tasks involving human speech. The aim of this work is to explore AST's effectiveness for spoken human language.

We assess AST's performance on three tasks. First, we fine-tune and test AST on recognizing the emotion in a track of spoken language. Emotion recognition in speech is important for applications such as consumer sentiment analysis and human-computer interaction, where understanding the speaker's emotional state can change the voice agent's responses and enhance user experience. Second, we examine AST's performance in determining the genre of a piece of music that includes singing. This is a complex task, as the best models will use both musical and lyrical features when making the classification. Finally, we devise a new architecture that adds a connectionist temporal classification (CTC) head to AST for automatic speech recognition, and we evaluate the performance of this architecture by calculating Word Error Rates (WERs).

2 Related Work

We compare AST against competitive baselines for the tasks selected, all of which use convolution-based architectures.

Speech Emotion Recognition Since we fine-tuned separately on two datasets (explained in detail in Section 4), we selected separate SER baselines for each dataset. For the RAVDESS dataset, the SOTA is Att-Net (Mustaqeem and Kwon, 2021), a self-attention model where a dilated CNN uses channel and spatial attention for the extraction of cues from the input tensors. The SOTA model for

the IEMOCAP dataset is DS-CNN (Huang et al., 2014), based on a depthwise separable convolution operation (a depthwise convolution followed by a pointwise convolution).

Music Genre Classification The state-of-the-art model for music genre classification is MUSER (MUsic SEquence Representation) (Chen et al., 2022). MUSER uses a tri-modal contrastive learning framework with audio, spectrum, and text inputs. A CNN (ResNet-50) image encoder serves as a music spectrum encoder, and ESResNeXt is used as an audio encoder. Music metadata is converted into text form and is used to train a text encoder (a transformer network).

Automatic Speech Recognition ASR models typically fall under two categories: 1) Encoder-only models, such as Conformer (Gulati et al., 2020) and XLSR (Conneau et al., 2020) with a CTC head on top and 2) Encoder-decoder models. Our ASTForCTC approach falls under the first category. However, unlike all other established methods, which use a CNN for feature extraction from the waveform, our method is purely vision transformer-based.

The Conformer (Gulati et al., 2020) is one example of an encoder-only ASR model. Each Conformer block contains a feed-forward module followed by a multi-head self-attention module, a convolution model, another feed-forward module, and then layernorm. When it was released, the Conformer was a great improvement over the then state-of-the-art CNN and transformer-based models since it achieved a WER of 4.3% on the Librispeech test set without using a language model. However, the transformer-only model that Conformer was compared to was TransformerTransducer (Zhang et al., 2020), which in contrast to AST, takes in audio sequences rather than images (spectrograms) as input.

It is also important to note the amount of data that industry-leading ASR models are trained on. The original Conformer (Gulati et al., 2020) is trained using the full 970-hour Librispeech dataset. Since the introduction of the Conformer, Wav2Vec 2.0 (Baevski et al., 2020), XLS-R (Babu et al., 2021), and Whisper (Radford et al., 2022) have also proven to be leading models for ASR. Wave2Vec 2.0 is pre-trained on 53K hours of unlabeled speech. Whisper is pre-trained on 680K hours of multilingual and multitask supervised data.

The top-performing model on the Librispeech test set achieves a test WER of 1.34 and combines FAdam optimization with a 600M parameter Conformer model pre-trained on a 60K hour corpus (Hwang, 2024). In contrast, AST is 86M parameters. AST was pre-trained on AudioSet, (Gemmeke et al., 2017), a collection of over 2 million 10-second audio clips excised from YouTube videos, which totals to only 5.5k hours of speech.

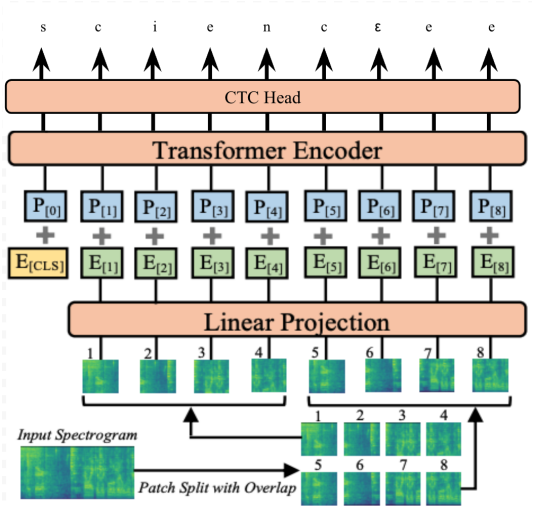

Figure 1: The `ASTForCTC` architecture only takes in image patches from the Mel-spectrogram as input. Characters may be repeated, and the model may predict blank tokens (ϵ). Figure is adapted from (Gong et al., 2021).

3 Approach

The AST model is a 12-layer transformer encoder with a hidden size of 768 and 86M parameters. The input to AST is prepared as follows: an audio waveform of t seconds is converted into a sequence of 128-dimensional log Mel filterbank (fbank) features. This results in a 128×100t spectrogram as input. The spectrogram is then split into a sequence of N 16x16 patches. Each 16×16 patch is flattened to an embedding of size 768 using a linear projection layer.

Classification For classification tasks, we use the `ASTForClassification` implementation already available on HuggingFace. We fine-tuned the model with a batch size of 16, the AdamW optimizer, and cross-entropy loss. We used a learning rate of 1e-5 with a weight decay of 0.01 and fine-tuned the model for either 5 or 20 epochs. We also used a linear learning rate scheduler.

ASR To our knowledge, there is no prior AST-based implementation or pre-trained AST model for ASR. We built a novel `ASTForCTC` model and

data processor for ASR. To create `ASTForCTC`, we added a linear layer to the base AST model. For each final hidden state returned by the AST encoder, the model predicts a corresponding probability distribution across all characters in the vocabulary. We minimize the CTC loss between the output and ground truth sequence. The architecture is illustrated in Figure 1. To prepare the data for training, we also built a custom `ASTProcessor` that combines the `ASTFeatureExtractor` for audio processing with a `Wav2Vec2CTCTokenizer` for label processing. Aside from initializing our `ASTForCTC` model from the AudioSet weights, we essentially trained for `ASTForCTC` model for ASR from scratch. We used the same learning rate and scheduler as for classification.

4 Experiments

For all of our tasks, we initialize from an AST model released on HuggingFace that was pre-trained with AudioSet (Gemmeke et al., 2017), which itself was initialized from a ViT-Base model pre-trained on ImageNet. The AudioSet task is multi-label classification across a set of 527 labels, such as music, speech, vehicle, etc.

We ran all of our experiments on one Google Colab A100 GPU using 1000 compute credits. For the classification tasks, fine-tuning time took an average of 10-15 minutes for 5 epochs and 40-60 minutes for 20 epochs.

For ASR, we fine-tuned for 150 epochs on the Google FLEURS dataset and 25 epochs for the Librispeech 100 hour train set. Fine-tuning time was on the order of 12-24 hours for each dataset.

4.1 Datasets

SER We fine-tuned AST on two different SER datasets. The IEMOCAP (Busso et al., 2008) dataset consists of 12 hours of audiovisual data, segmented into 1-3 second audio clips across 5 actors and 5 emotions (angry, excited, fear, sad, surprised, frustrated, happy, disappointed, and neutral). The RAVDESS (Livingstone and Russo, 2018) dataset contains 1440 files (60 trials per actor x 24 actors) of audio with calm, happy, sad, angry, fearful, surprise, and disgust speech emotions.

Music Genre Classification The GTZAN (Tzanetakis et al., 2001) dataset covers 10 genres of music: blues, classical, country, disco, hip-hop, jazz, metal, pop, reggae, and rock. For each genre, it contains 100 audio tracks, each of which is 30 seconds long. The audio samples were collected in 2000-2001 from a range of sources, such as personal CDs, radio, and microphone recordings, in order to capture a variety of sound conditions.

ASR We used the Google FLEURS en_US dataset (Conneau et al., 2022) for ASR fine-tuning, which contains 3,640 examples (approximately 12 hours) of spoken Wikipedia articles. We also used the 100-hour split of the Librispeech (Panayotov et al., 2015) train set and the full test set. Though we also tried to use additional splits of the Librispeech dataset (350 hours of clean data and 500 hours of "other" speech), we ran into disk and memory usage limitations.

4.2 Results

To evaluate the AST's performance on classification tasks, we measured accuracy. For ASR, we measured the Word Error Rate (WER).

Emotion and Music Genre Our results for the classification tasks are summarized in Table 1. The AST model achieved SOTA performance on music genre classification. Figure 2 illustrates there are clear differences in what AST attends to when comparing tracks of two different genres. The results of the SER tasks were mixed. On the RAVDESS dataset, AST outperformed the current SOTA baseline, but on the IEMOCAP dataset, it underperformed. A possible explanation for the performance difference is AudioSet pre-training. AudioSet contains over one million music segments from YouTube, but may not have a wide variety of emotional speech.

ASR The WERs of `ASTForCTC` on FLEURS and Librispeech are shown in Table 2. On FLEURs, `ASTForCTC` performed surprisingly well on the training set, despite being trained in an extremely low-resource setting (10 hours of data). We obtained a WER of 0.059 on the FLEURS train set and 1.083 on the test set. Sample predictions are shown in Figure 3. Despite overfitting to the train set, we did find that the AST model did a good job of learning to detect the number of words in an utterance – the ratio of the number of words predicted to the number of actual words was 1.00 in the train set and 1.05 in the test set for FLEURS. Training with more data helped to combat the overfitting. Using the 100-hour split of Librispeech for training, the WER for train and test was around 0.7. This indicates that more data can help reduce WER and combat overfitting.

Task	Baseline	AST (5 Epochs)	AST (20 Epochs)
Music Genre Classification (GTZAN)	0.825	0.865	**0.900**
SER (IEMOCAP)	**0.788**	0.640	0.7
SER (RAVDESS)	0.800	0.760	**0.816**

Table 1: Accuracy results for different classification tasks. The best results are in bold.

Task	Train	Test
FLEURS, en_us (150 epochs)	0.059	1.083
Librispeech, (25 epochs)	0.719	0.738

Table 2: Word error rates (WERs) of `ASTForCTC`.

On both FLEURS and Librispeech, we observed a few hallucinations toward the end of sentences. For example, on FLEURs, `ASTForCTC` tended to add extra "s"'s to the end of the sentence. On Librispeech, it sometimes hallucinated extra vowels.

5 Conclusion

The goal of this work was to evaluate the Audio Spectrogram Transformer (AST), a novel vision transformer and convolution-free model that achieves SOTA results on various audio classification tasks in the context of spoken language understanding. We found that AST performed competitively on emotion recognition and music genre classification. AST achieved SOTA results compared to attention (Att-Net) and hybrid attention-CNN (MUSER) architectures. Though AST underperformed the CNN (DS-CNN) architecture for emotion recognition, it did not lag far behind. We believe our results indicate that AST shows promise as a foundation model for human speech classification tasks.

In addition, we contribute a novel `ASTForCTC` architecture for ASR and demonstrate promising early results when training on only 10 hours and 100 hours of labeled data, respectively. A limitation of our work was the amount of compute we had available. To achieve performance potentially more comparable to SOTA ASR models like Conformer, we recommend fine-tuning `ASTForCTC` with at least 970 hours of labeled Librispeech data and possibly further pre-training AST with more unlabeled speech. Alternatively, to ablate the effect of different scales of pre-training and fine-tuning data, one could train AST and Conformer, for example, without any initialization from pre-trained model weights.

Finally, we recommend evaluating the performance of AST on multilingual classification tasks, such as language identification. In addition, we did not benchmark the training time and inference speed of AST against CNN-based architectures and leave that to future work.

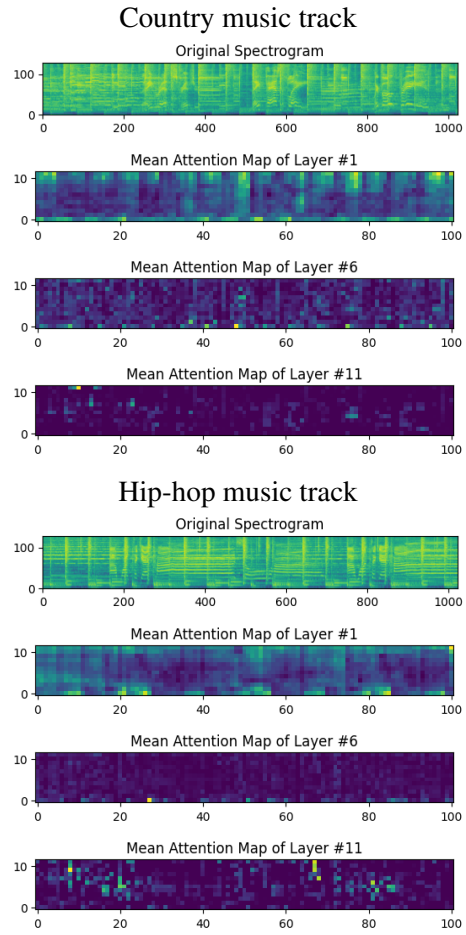

Figure 2: Attention maps for two distinct music genres. The differences illustrate how the AST model learns to attend to different regions of the spectrogram for classifying different genres of music.

References

Arun Babu, Akshat Shrivastava, Armen Aghajanyan, Ahmed Aly, Angela Fan, and Marjan Ghazvininejad. 2021. Non-autoregressive semantic parsing for compositional task-oriented dialog. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics:*

Human Language Technologies, pages 2969–2978, Online. Association for Computational Linguistics.

Alexei Baevski, Henry Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations.

Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower Provost, Samuel Kim, Jeannette Chang, Sungbok Lee, and Shrikanth Narayanan. 2008. Iemocap: Interactive emotional dyadic motion capture database. *Language Resources and Evaluation*, 42:335–359.

Tianyu Chen, Yuan Xie, Shuai Zhang, Shaohan Huang, Haoyi Zhou, and Jianxin Li. 2022. Learning music sequence representation from text supervision. In *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4583–4587.

Alexis Conneau, Alexei Baevski, Ronan Collobert, Abdelrahman Mohamed, and Michael Auli. 2020. Unsupervised cross-lingual representation learning for speech recognition.

Alexis Conneau, Min Ma, Simran Khanuja, Yu Zhang, Vera Axelrod, Siddharth Dalmia, Jason Riesa, Clara Rivera, and Ankur Bapna. 2022. Fleurs: Few-shot learning evaluation of universal representations of speech. *arXiv preprint arXiv:2205.12446*.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. 2021. An image is worth 16x16 words: Transformers for image recognition at scale.

Jort F Gemmeke, Daniel PW Ellis, Dylan Freedman, Aren Jansen, Wade Lawrence, R Channing Moore, Manoj Plakal, and Marvin Ritter. 2017. Audio set: An ontology and human-labeled dataset for audio events. In *2017 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 776–780. IEEE.

Yuan Gong, Yu-An Chung, and James Glass. 2021. Ast: Audio spectrogram transformer.

Anmol Gulati, James Qin, Chung-Cheng Chiu, Niki Parmar, Yu Zhang, Jiahui Yu, Wei Han, Shibo Wang, Zhengdong Zhang, Yonghui Wu, and Ruoming Pang. 2020. Conformer: Convolution-augmented transformer for speech recognition.

Zhengwei Huang, Ming Dong, Qi rong Mao, and Yongzhao Zhan. 2014. Speech emotion recognition using cnn. *Proceedings of the 22nd ACM international conference on Multimedia*.

Dongseong Hwang. 2024. Fadam: Adam is a natural gradient optimizer using diagonal empirical fisher information.

Steven R. Livingstone and Frank A. Russo. 2018. The ryerson audio-visual database of emotional speech and song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions in north american english. *PLOS ONE*, 13(5):e0196391.

Mustaqeem and Soonil Kwon. 2021. Att-net: Enhanced emotion recognition system using lightweight self-attention module. *Appl. Soft Comput.*, 102:107101.

Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. 2015. Librispeech: An asr corpus based on public domain audio books. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5206–5210.

Karol J. Piczak. ESC: Dataset for Environmental Sound Classification. In *Proceedings of the 23rd Annual ACM Conference on Multimedia*, pages 1015–1018. ACM Press.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2022. Robust speech recognition via large-scale weak supervision.

George Tzanetakis, Georg Essl, and Perry Cook. 2001. Automatic musical genre classification of audio signals.

Pete Warden. 2018. Speech commands: A dataset for limited-vocabulary speech recognition.

Qian Zhang, Han Lu, Hasim Sak, Anshuman Tripathi, Erik McDermott, Stephen Koo, and Shankar Kumar. 2020. Transformer transducer: A streamable speech recognition model with transformer encoders and rnn-t loss.

A Appendix

See the following page for selected ASR predictions, spectrograms, and attention maps.

Prediction	Reference	pwc / rwc / ratio
some people believe that experiencing many artificially induced lucid dreams often enough can be very exhausting	some people believe that experiencing many artificially induced lucid dreams often enough can be very exhausting	16 / 16 / 1.0
permits must be reserved in advance you must have a permit to stay overnight at sirenasssssss	permits must be reserved in advance you must have a permit to stay overnight at sirena	16 / 16 / 1.0

(a) FLEURS ASR training predictions.

Prediction	Reference	pwc / rwc / ratio
a oy al al un ald wo hernm os an sth ldg hrndlsoni wn weshey iaodnighisssssssss	in some areas boiling water for a minute is enough in others several minutes are needed	16 / 16 / 1.0
aoas alsi toilnsd tans pan cailehtan s.iic tafnendan inirds...	oliver sacks in his paper the president's speech indicated how people...	28 / 28 / 1.0

(b) FLEURS ASR test predictions.

Prediction	Reference	pwc / rwc / ratio
my fathe was and still is wecivver generl at se he as aegret repputation their for voalty thanks tohich whe sable thein t secur o e e aa o e o e e e	my father was and still is receveur general at c he has a great reputation there for loyalty thanks to which he was able to find the security...	34 / 37 / 0.92
i cam to pars study vaw was calld to the baur and like omny other youung men but my deplo endmy poctet and lett myself drift o e e a oe o o e	i came to paris studied law was called to the bar and like many other young men put my diploma in my pocket and let myself drift...	34 / 34 / 1.0

(c) Librispeech ASR training predictions.

Prediction	Reference	pwc / rwc / ratio
conclored returned too hich placse am mits thetente	concord returned to its place amidst the tents	8 / 8 / 1.0
i am comvingced ive what i say said be count	i am convinced of what i say said the count	10 / 10 / 1.0
thus it is that the oner of the we is save acuntry ar laster's and nor hone	thus it is that the honor of three is saved our country's our master's and our own	17 / 17 / 1.0

(d) Librispeech ASR test predictions.

Figure 3: ASR predictions for the FLEURS and Librispeech training and test datasets. pwc is the predicted word count, and rwc is the reference word count. Hallucinations are shown in blue.

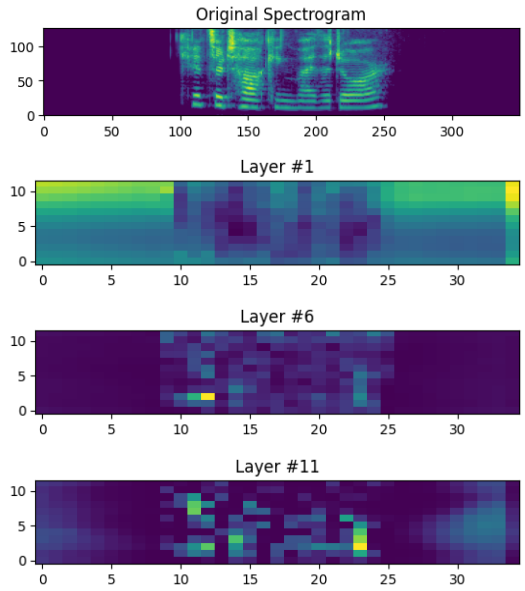

Figure 4: The Mel-spectrogram and selected attention maps for a "calm" sample audio track in SER.

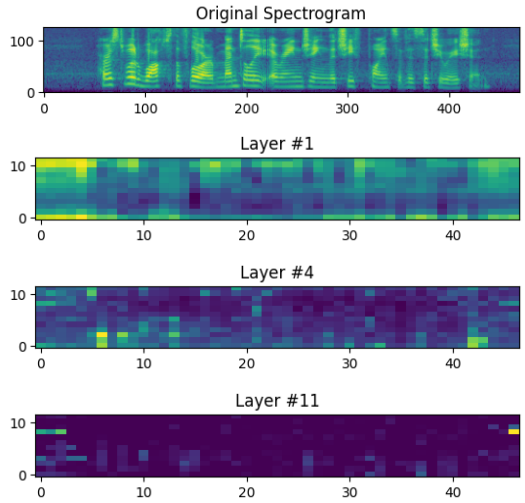

Figure 5: The Mel-spectrogram and selected attention maps for a sample from Librispeech. This track consists of a female voice saying, "Hurstwood walked the floor mentally arranging the chief points of his situation."