EECS 545 Project Final Report
Spoken Language Identification using Audio Spectrogram Transformer

Mingliang Duanmu, Yucheng Gu, Haoyu Huang, Jing Peng, Minheng Xiao *
University of Michigan
{duanmuml,ycgu,haoyuh,jingpeng,minhengx}@umich.edu

March 3, 2023

Abstract

In this paper, we investigate the performances of previously proposed Audio Spectrogram Transformer (AST) model on Spoken Language Identification tasks. We apply the model to classify audio of eight languages and examine the performances of the model with various hyper-parameters. We achieve around 91.4% average accuracy. We also tested and enhance the robustness of the model on audio with accents in different languages. Furthermore, to address the low-resolution issue of transformers, we introduce multi-patch AST and achieved around 94% accuracy. We further trained a Pyramid Vision Transformer using transfer learning techniques and achieved around 98% accuracy.

1 Introduction

Spoken Language Identification (LID) is the process to classify the language of human speech from audio files. It is typically the first step of many human speech recognition problems and has found various applications including voice assistant such as Google Assistant and Siri, multilingual customer services and automatic spoken language translation. In the past decade, many work and experiments for this task are researching into the use of deep learning techniques. However, questions remain on the necessaity of neutral networks.
Inspired by the success of transformer in vision domain, we intend to transfer the knowledge of vision transformers to LID tasks. We first extend the work of Audio Spectrogram Transformer (AST) proposed in Gong et al. [1]. It is a convolution-free model that can be implemented directly with audio spectrogram as input. It is advantageous in using simpler structure with significantly fewer parameters and converging faster in training process. The original publication

*Authors have equal contribution and are listed in alphabetical order.

shows that the model achieved new state-of-art results when tested on data from AudioSet, ESC-50, and Speech Commands V2, but to our best knowledge, the model has not been tested on LID tasks before. In this project, we test the performances of AST on classifying the languages in audio from VoxForge [2], evaluate the robustness of this model on identifying languages that are spoken with accents from Speech Acent Archive [3]. Furthermore, we transfer the knowledge of a Pyramid Vision Transformer (PVT) to modify the AST model in an effort to address the limitations observed in vision transformer.

2 Related Works

In recent years, neural networks have been widely used in audio tasks such as audio identification, recognition, representation etc. Among them, convolutional neural networks (CNNs) and recurrent neural networks (RNNs) have shown dramatic successes. CNNs based on Mel-Scale Spectrum are able to capture higher-level features that are invariant to local spectral and temporal variations. By leveraging local filtering and max-pooling in frequency domain, it could normalize speaker variance to achieve higher speech recognition performance [4]. Meanwhile, RNNs and Long-Short Term Memory networks (LSTM), a special model of RNNs that uses a gating mechanism to model long-term dependencies in the data, are also widely used in LID tasks. They could process long sequential data such as text or sound and are powerful in learning long-term context in the audio signals [5].

Since the use of Mel-Scale spectrum has proved to be useful to capture higher-level features, it successfully transfers the recognition problem from audio to computer vision domain. Among many computer vision methods, one of the most popular model is the Vision Transformer (ViT). Transformers found their first application in natural language processing (NLP) tasks. A well-known example is BERT, which shows great success in capturing a wide range of characteristics such as fluency, suprasegmental pronunciation, syntactic and semantic characteristics [6]. As a special type of transformer, ViT applies a transformer architecture on image classification. It is achieved by splitting images into tokens a series of patches with a fixed length, embedding the patches, and including positional embedding as an input to the transformer encoder, which learns relations between these tokens [7]. However, there are some disadvantages of ViT. For example, ViT has much less image-specific inductive bias than CNNs and it needs more training dataset than CNNs [7].

Despite the success of ViT in computer vision, it is seldom used in audio recognition until recently when AST, inspired by ViT, was introduced as the first convolution-free, purely attention-based model for audio classification [1]. However, there are yet no applications in LID. Moreover, it fails to improve on the aforementioned disadvantages of ViT.

Therefore, we aim to extend the work of AST in the following two categories. First, realizing the transfer learning of ViT on LID through AST and investigating performances of the model with various patch selection. Second, modifying the

structure of current AST in efforts to address the shortcomings as observed in ViT.

3 Methods

3.1 Data

We use data from the VoxForge dataset [2]. This dataset is a well-known speech dataset for training LID models. We mainly worked with the European subset of the dataset, including eight classes: English, French, German, Italian, Spanish, Russian, Dutch and Portuguese. We split out about 30% data as the test set, and use the rest as the train set. This gives a train set of about 750 samples per language. Then each sample is cut to 5.12s to fit the transformer.
In addition, we use the Speech Accent Archive [3] as a complement dataset for accent-robust training and testing. This dataset contains about $2,100$ English audios read by non-native speakers. Furthermore, we develope our own data scrapers in Python to collect data from YouTube. These collections are planned to be used as augmentation set for accent robustness in languages other than English.

3.2 Transfer Learning of Audio Spectrum Transformer

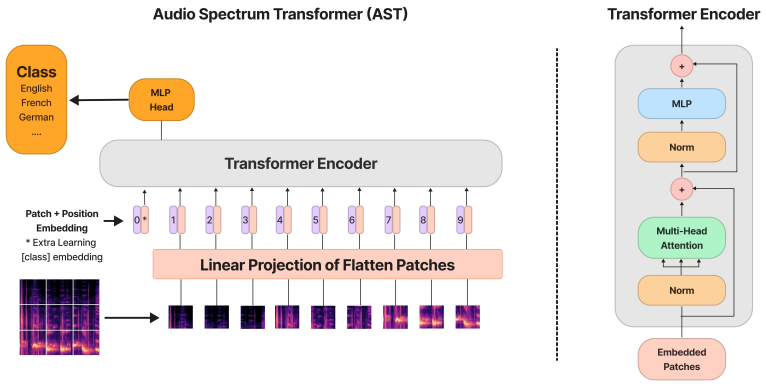


Figure 1: Architecture of Audio Spectrum Transformer

We first refer to [1] for the implementation of AST. The structure of the model is shown in Fig. 1. As Gong et al. experimented in their paper, we start to train an AST model based on transfer learning techniques from ImageNet pretrain parameters. Since the model structure is slightly different from ViT, we load the parameters as follows.
First, while ViT is designed for inputs with certain fixed shape (usually 224×224 or 384×384), we set the audio input length to be 5120 ms, which is equivalent to

a 512×128 spectrogram. Additionally, the input of ViT is a three-channel image, while the spectrogram itself is sufficient on one channel. Therefore, we adopt the cut and bi-linear interpolate method that Gong et al. proposed. In our case, we start from the 384×384 pretrained ViT model with a patch size of 16×16. This results in $24 \times 24 = 576$ positional embeddings with each embedding for a patch without overlap.
We use several different AST models for experiments. As the results are similar, we demonstrate with one of them in which we use 12×56 patches. In this case, we cut at the first dimension of the 24×24 positional embeddings, and interpolate at the second dimension. This gives a 12×56 positional embedding. Additionally, the weights of ViT over 3 channels are averaged, so that it works on the single-channel input. This allows us to train AST on our language dataset from the ImageNet pretrained parameters.

3.3 Model Modification

Although ViT is proved to be successful in image classification, it is difficult to adapt it directly to pixel-level dense predictions such as object detection and segmentation, specially for the mel-spectrum. It is mainly because the feature map of ViT is single-scale and low-resolution, and its computational and memory costs are relatively high [8]. Therefore, in this section, we introduce some methods to overcome the problems.

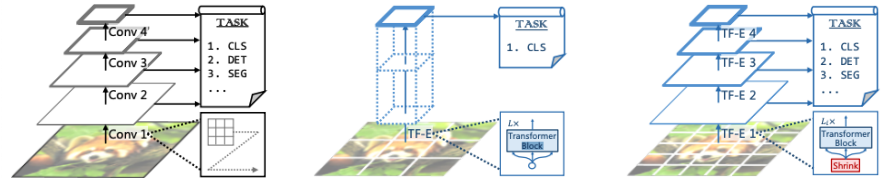
3.3.1 Multi-Patch Audio Spectrum Transformer

To handle the low-resolution problem of ViT, we introduce the multi-patch AST. The normal AST splits the image into several patches with certain time and frequency stride. In multi-patch model, we split the image into two groups of patches with different time and frequency strides, respectively. This method enables the transformer to capture the patches from different aspects, which better detects the features between pixels. Therefore, we perform some experiments on different selections of stride combination and compare the performances.

3.3.2 Transfer Learning of Pyramid Vision Transformer

We also refer to Pyramid Vision Transformer (PVT), which is a hybrid combination of CNNs and ViT, which is illustrated in Fig. 2. It is mainly implemented by introducing a progressive shrinking pyramid to control the scale of feature maps by patching embedding layers and reducing the sequence length as the network deepens, which reduces the computational cost. It uses four stages that generate features maps of different scales. Furthermore, it leverages a spatial-reduction attention (SRA) layer to further reduce the resource consumption when learning high-resolution features [8]. Therefore, we conduct transfer learning on the PVT model. Since PVT follows the rules of ResNet, i.e. as the network depth increases, the hidden dimension increases gradually, and the output resolution

shrinks progressively [8], we experiment on different choices of network depth and compare their performances.



(a) CNN: ResNet[9] (b) ViT [7] (c) PVT [8]

Figure 2: Structures of CNNs, ViT and PVT

4 Experiments

4.1 Results of Transfer Learning from Pretrained Model

For the AST model, we trained it for 25 epochs with or without the pretrained parameters. For both settings, we use learning rate 10^{-5}, batch size 32 with frequency stride 12 and time stride 8 (see section 4.2 for experiments on these parameters). Fig. 3 shows the average accuracy of all classes. We can see that

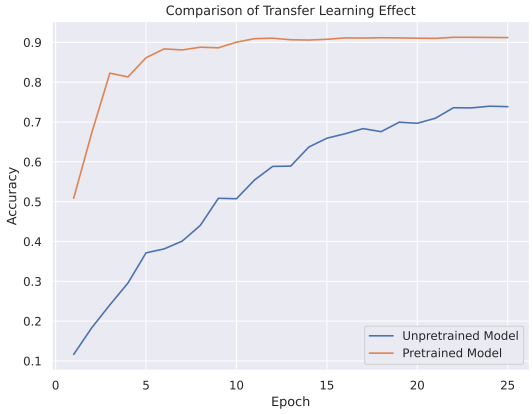


Figure 3: Comparison of Transfer Learning Effect

when pretrained parameters are used, the model converges much faster, and has significantly improved accuracy. For the model without pretrained parameters, the average accuracy stablizes around 0.80, which suggests that the model has stuck in a local minimum.

This result shows that the pretrained parameters of ViT on ImageNet is capable

of boosting the LID task. We will therefore experiment on other ViT models with the transfer learning technique in later experiments.

4.2 Results of Tuning the Stride Parameters

We also perform experiments on the strides of the patches in AST. Since the patches in AST could overlap by design, a larger stride means smaller overlap of patches. Note that we have two different kinds of stride in the model: the frequency stride (fstride) which is vertical, and the time stride (tstride), which is horizontal. We experiment on both fstride and tstride and train for 25 epochs, with a learning rate of 10^{-5}, batch size of 32. To reduce the computational burden, we use a reduced training/testing set (5 classes: English, French, German, Italian and Spanish) instead of 8 classes here.

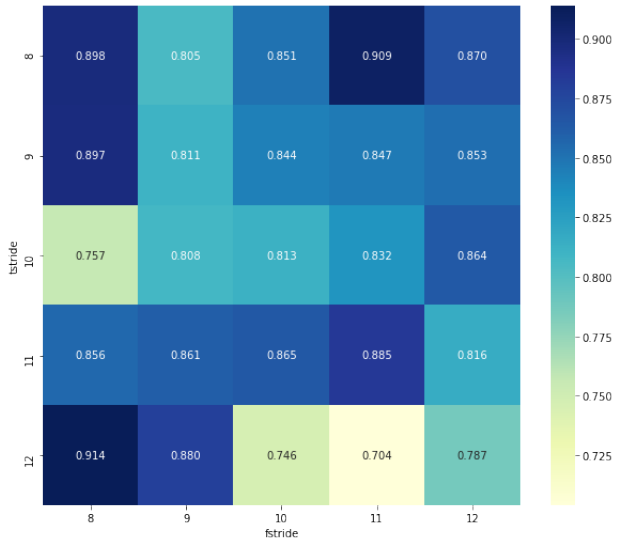


Figure 4: Effect of Different Strides

Fig. 4 shows the result. We can see that the AST model performs well when either fstride or tstride is large while the other is small. This indicates that the language audio spectrograms have latent features that extend on both frequency and time axis. We also observe that when fstride=8 and tstride=12, the model has the best performance. These parameters are therefore used in later experiments.

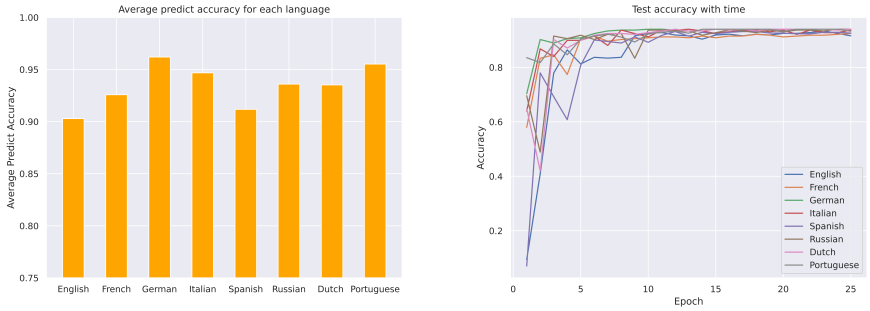
4.3 Results of Multi-Patch Audio Spectrum Transformer

As discussed in previous section, we find that fstride of 12 and fstride of 8 yield best results. Therefore, in the multi-patch AST, we fix one patch with time stride 12 and frequency stride 8 and adjust the size of the other patch. The results are

shown in Table 1. Fig. 5 shows the convergence and prediction accuracy of each class. It can be seen that the multi-patch model further improves the accuracy based on the single-patch AST with single patch size.

Table 1: Accuracy of Multi-Patch Audio Spectrum Transformer

T/F stride	8/11	8/12	9/11	9/12	10/11	10/12
12/8	0.87958	0.92597	0.93821	0.89094	0.85528	0.86335



(a) Average Prediction Accuracy

(b) Accuracy vs. Time

Figure 5: The convergence and accuracy of multi-patch AST

4.4 Robustness Towards Accents

We tested our model on a subset of Speech Accent Archive dataset. The dataset was cut to around 8,000 samples with 5.12s each, and we use about 800 samples as test set. Our result shows that the finetuned model with fstride=8 and tstride=12 can only achieve 67% accuracy on this accented dataset.
To encourage diversity and add robustness to the model, we inserted another 800 samples subset from the Speech Accent Archive dataset into the original training set, and retrained the model using the same parameters and epochs. Results show that while the model's accuracy on the accent test dataset rises to 82%, its accuracy on the original test dataset only drops by less than 2%. We believe this augmentation is necessary, and are still working on collection of a larger augmented dataset from YouTube and other open data sources.

4.5 Results of PVT

For PVT, we primarily change the parameters of the block size in the transformer block, the parameter setting and the result is shown in Table 6. All PVT models are trained using learning rate 5×10^{-4} using AdamW optimizer with $\epsilon = 10^{-8}$, batch size 192 and 0.1 dropout rate. We find that the use of PVT could improve

the performance of LID dramatically, with the highest accuracy being 98.83%. We also find that with the increasees in the block size, the performance is gradually improved until reaching the saturation at certain point.

Model Size	Block depth	#Params (M)	Acc
Tiny	16	14.0	93.45
Small	216	25.4	95.65
Medium	648	45.2	98.18
Large	1944	62.6	98.83

Table 2: Accuracy of PVT with different block depth

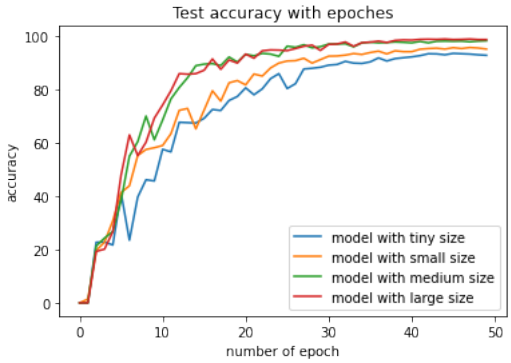


Figure 6: Accuracy vs. Time

5 Conclusion

Motivated by the success of transformer in vision domain, we intend to tackle the spoken language identification task by transferring knowledge from vision transformer. We start by implementing a audio spectrogram transformer and achieve an average accuracy of 91.4%. We also experiment on stride parameters to finetune the model. Furthermore, we introduce multi-patch AST to address the low-resolution issue and achieve higher accuracy. We also test the robustness of AST on audios with accents. We enhance robustenss by testing trainset augmentation with accents. Finally, we leveraged the transfer learning experience on a PVT model and significantly improve the accuracy to 98.8%, which is far above the 96.3% accuracy of the SOTA CNN models on the same dataset [10].

Code

The code for this work can be found at `https://github.com/Enoch2090/SpokenLanguageDetectionUsingTransformers`

References

[1] Y. Gong, Y.-A. Chung, and J. Glass, "Ast: Audio spectrogram transformer," *arXiv preprint arXiv:2104.01778*, 2021.

[2] VoxForge. `http://www.voxforge.org/`.

[3] S. Weinberger, "Speech accent archive," George Mason University, 2015.

[4] O. Abdel-Hamid, A.-r. Mohamed, H. Jiang, and G. Penn, "Applying convolutional neural networks concepts to hybrid nn-hmm model for speech recognition," in *2012 IEEE international conference on Acoustics, speech and signal processing (ICASSP)*, pp. 4277–4280, IEEE, 2012.

[5] A. Graves, A.-r. Mohamed, and G. Hinton, "Speech recognition with deep recurrent neural networks," in *2013 IEEE international conference on acoustics, speech and signal processing*, pp. 6645–6649, Ieee, 2013.

[6] J. Shah, Y. K. Singla, C. Chen, and R. R. Shah, "What all do audio transformer models hear? probing acoustic representations for language delivery and its structure," *arXiv preprint arXiv:2101.00387*, 2021.

[7] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.

[8] W. Wang, E. Xie, X. Li, D.-P. Fan, K. Song, D. Liang, T. Lu, P. Luo, and L. Shao, "Pyramid vision transformer: A versatile backbone for dense prediction without convolutions," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 568–578, 2021.

[9] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

[10] Sarthak, S. Shukla, and G. Mittal, "Spoken language identification using convnets," 2019.