Transfer Learning in Voice Recognition based on Pre-trained Image model

Yayu Mo
ECE 7365 Adaptive Algorithm for Machine Learning
2023 Fall
49336397
yayum@mail.smu.edu

***Abstract*—The potential of transfer learning has brought benefits for cross-modal tasks to build a more efficient structure and novel training strategies. In our paper, we explore the relations between voice signals and general images using this specific learning technology and generate amplitude and MEL spectrogram. And we also proposed a structure embedded with 5 pre-trained models separately and applied a fine-tuning strategy during training. In our experiments, we achieved the best training accuracy of 93.53% and testing accuracy of 86.54%, which indicates the effectiveness of our method.**

***Index Terms*—Transfer Learning, voice recognition, pre-trained, image model**

I. Introduction

With the rapid development of machine learning technology, the innovative concept of applying transfer learning has significantly affected cross-modal research. The learning methodology takes advantage of the models that have already been trained on large datasets to apply to different learning tasks, to understand and interpret different kinds of sensory data. In our method, we applied several models that have been trained thoroughly on a large image dataset, to a specific task of gender recognition in voice signals. Our method aims to explore the connection between the general image data and spectrum of voice signals and utilize the potential of transfer learning to build a resource-efficient model and its influence on training and validating.

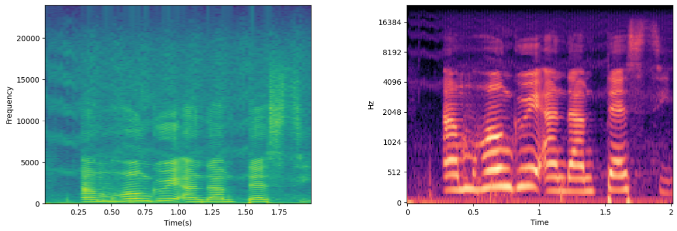
II. Literature Survey

Several works have been done to verify the effectiveness of applying the capability of mature pre-trained models in images to the voice recognition problem. In 2010, Muda et al.[1] proposed a feature extraction model based on DFT and Mel filter Bank to generate the MEL spectrum and get the MFCC vector, and compared the similarities between training and testing data. Their work proves the effectiveness of identifying the feature of voice signals through the spectrum operation, however, they just took the MFCC output vector, and compared it with reference patterns without considering applying the spectrograms as images, which could consider the image processing method for further feature extractions. In 2018, Lech et al.[2] proposed an automatic speech emotion recognition (SER) model that took advantage of pre-trained AlexNet[3] and different types of spectrogram. They also proposed a method utilizing the different emotion significance in separate RGB color channels. Le et al.[4] also worked on a baby cries classification method using the pre-trained ResNet50[5] and SVM on ImageNet, and achieved accuracies of more than 90%, which strongly supports the effectiveness of transfer learning based on pre-trained image models. In 2021, Gunawan et al.[6] proposed a model based on pre-trained EfficientNet[7] for owl sound recognition task, applied both MEL spectrogram and MFCC vector, and achieved 99.27% mAP on testing data. The studies that have been introduced above show the capability of transfer learning based on pre-trained image models, which could also achieve a high level of accuracy and performance.

III. Approach

A. Spectrogram Generation

In the data pre-processing, we generate 2 types of spectrogram, the general amplitude spectrogram and MEL spectrogram[8]. The general amplitude spectrogram is generated through a Short-time Fourier transform, and applying Hamming window. The MEL spectrogram is also generated through a Short-time Fourier transform but uses a MEL scale to present the spectrogram. For convenience, to generate the amplitude spectrogram, we applied the embedded function in ”matplotlib” API in python, and also applied the ”librosa” API to generate the MEL spectrogram. The output of the transform from .wav audio file to 2 types of spectrogram is shown as Fig. 1.



(a) General amplitude spectrogram

(b) MEL spectrogram

Fig. 1: Spectrogram example

B. Model

In our method, we first applied data augmentations such as rotation of 180 degrees and mirror flip to expand the datasets. The model contains a pre-trained model, a global average pooling layer to down-sampling the tensor output by the pre-trained model layers, and a dropout layer in the parameter of 0.2 to avoid overfitting, and a dense layer for the output, as shown in Fig. 2. Due to the reason that our task is a binary classification question, we applied Binary Cross-Entropy as our cost function

$$Cost = -\frac{1}{N}\sum_{i=1}^{N}[y_i log(\hat{y_i}) + (1 - y_i) log(1 - \hat{y_i})] \quad (1)$$

and in the output layer, we applied sigmoid as our activation function

$$sigmoid = \frac{1}{1 + e^{-x}} \quad (2)$$

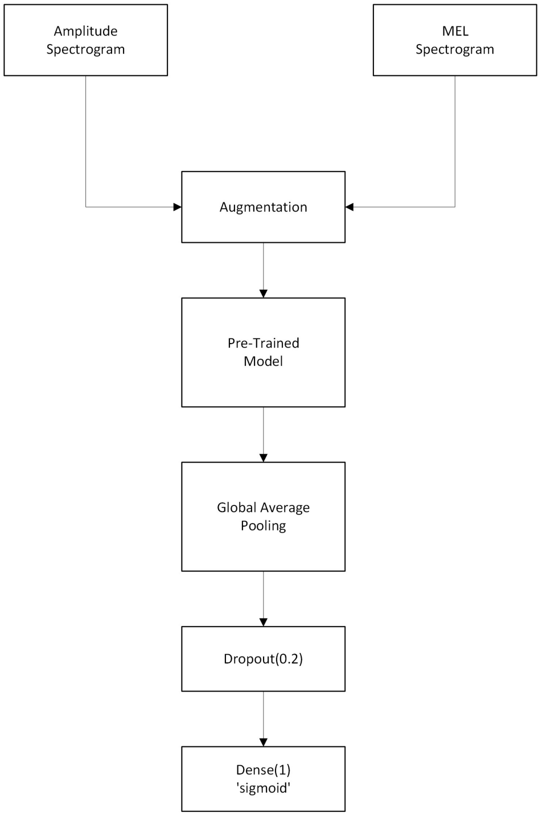


Fig. 2: Model structure

C. Fine-tuning

In our training, we applied a fine-tuning strategy that left the weights of the pre-trained model unchanged and only updated the weights of the classifier part in the first half of iterations during training. During the last half of iterations, we started to update the weights of the entire model with lower learning rates and different optimizers. For example, the detailed properties of fine-tuning in the VGG19-based model are shown as TABLE. I.

TABLE I: Fine-tuning properties of VGG19-based model

Training type	**Learning rate**	**optimizer**
freeze training	1e-5	Adam
unfreeze training	1e-6	RMSprop

D. Dataset

The BVC Voice biodata[9] for this experiment is obtained from kaggle, and the dataset contains the collected voice signal from 526 individuals of different ages, genders and ethnicities. Due to the reason that we choose the specific subject of gender recognition, which includes 2149 male and 1815 female voice utterances in the dataset. The detailed information related to gender in the dataset are listed as TABLE. II.

TABLE II: Dataset properties in the experiment

voice type	**male counts**	**female counts**	**total counts**
1 sentence	501	170	671
multi sentence	710	1068	1778
s2multi sentence	938	577	1515

After generating the amplitude spectrogram and MEL spectrogram datasets, we split the data into the ratio of 8:2, where 80% are the train set, and 20% are the validation set. In addition, we also take 20% of the validation set as the test set to predict the model's accuracy.

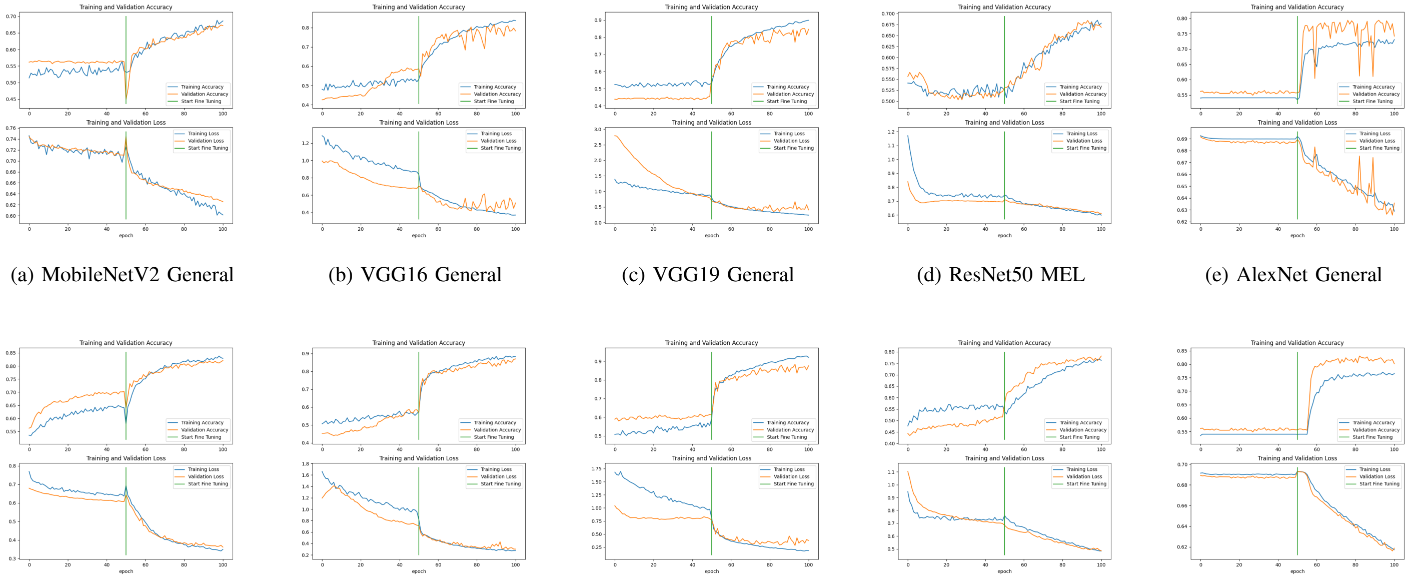
IV. Results

In our experiment, we applied 5 pre-trained models on 2 type of pre-processed spectrogram datasets, and set several specific parameters such as batch size, learning rate, and input image size. After times of trail, we determine the parameters with the best performance in our theory, as listed in TABLE. III.

TABLE III: Experiment Setting

Model	**Dataset**	**Batch size**	**Learning rate**	**Input size**
MobileNetV2[10]	General	4	1e-5	224x224
	MEL	4	1e-5	224x224
VGG16[11]	General	4	1e-5	224x224
	MEL	4	1e-5	224x224
VGG19[11]	General	4	1e-5	224x224
	MEL	4	1e-5	224x224
ResNet50[5]	General	8	1e-6	224x224
	MEL	8	1e-6	224x224
AlexNet[3]	General	4	1e-4	227x227
	MEL	4	1e-4	227x227

The training performances are shown in a series of loss curves and accuracy curves in Fig. 3, which indicates that applying VGG19[11] as base model achieves the best performance of 93.53% train accuracy and 87.75% validation accuracy. After training, we also predict the images in test set to verify the effectiveness of the model. As shown in Fig. 4, the accuracies are distributed in the range of 60%-90%, and VGG19 achieves the best accuracy of 86.54% in MEL dataset.



(a) MobileNetV2 General (b) VGG16 General (c) VGG19 General (d) ResNet50 MEL (e) AlexNet General

(f) MobileNetV2 MEL (g) VGG16 MEL (h) VGG19 MEL (i) ResNet50 MEL (j) AlexNet MEL

Fig. 3: Loss curves and accuracy curves

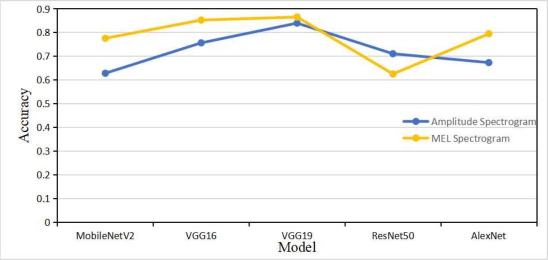


Fig. 4: Test accuracy of models

V. Summary

In this paper, we explore the significant potential of transfer learning by applying the pre-trained model from image data to a voice spectrogram. We generated 2 different kinds of spectrogram datasets, built our method with 5 pre-trained models, trained them with a specific fine-tuning strategy, and achieved the best accuracy of 86.54% in VGG19-based model. For future studies, we would continue exploring the performance on different pre-trained models, and instead of classification models, we would also take some of the NLP models pre-trained on language data and object detection models such as YOLO series into consideration.

References

[1] Lindasalwa Muda, Mumtaj Begam, and Irraivan Elamvazuthi. Voice recognition algorithms using mel frequency cepstral coefficient (mfcc) and dynamic time warping (dtw) techniques. *arXiv preprint arXiv:1003.4083*, 2010.

[2] Margaret Lech, Melissa Stolar, Robert Bolia, and Michael Skinner. Amplitude-frequency analysis of emotional speech using transfer learning and classification of spectrogram images. *Adv. Sci. Technol. Eng. Syst. J*, 3(4):363–371, 2018.

[3] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25, 2012.

[4] Lillian Le, Abu Nadim MH Kabir, Chunyan Ji, Sunitha Basodi, and Yi Pan. Using transfer learning, svm, and ensemble classification to classify baby cries based on their spectrogram images. In *2019 IEEE 16th International Conference on Mobile Ad Hoc and Sensor Systems Workshops (MASSW)*, pages 106–110. IEEE, 2019.

[5] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[6] A transfer learning strategy for owl sound classification by using image classification model with audio spectrogram. *International Journal on Electrical Engineering and Informatics*, 13(3):546–553, 2021.

[7] Mingxing Tan and Quoc Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International conference on machine learning*, pages 6105–6114. PMLR, 2019.

[8] Stanley Smith Stevens, John Volkmann, and Edwin Broomell Newman. A scale for the measurement of the psychological magnitude pitch. *The journal of the acoustical society of america*, 8(3):185–190, 1937.

[9] Ogechukwu Iloanusi and Samuel Ezichi. Gender age classification from voice data, 2020.

[10] Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. Mobilenetv2: Inverted residuals and linear bottlenecks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4510–4520, 2018.

[11] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.