Curious Case of Transfer Learning in Audio Domain

Haozhe Huang
University of Toronto
marko.huang@mail.utoronto.ca

Jiabin He
University of Toronto
jiabin.he@mail.utoronto.ca

Yudong Xu
University of Toronto
wil.xu@mail.utoronto.ca

Abstract

Research has shown that machine learning models trained for a specific image classification task can be used, through transfer learning, to improve the performance of a different image classification task. This paper explores this concept in the audio classification domain. By re-purposing a music genre classification model for speaker recognition and vice versa, we show that transfer learning in the audio classification domain might not be as effective for image classification.

1 Introduction

In deep learning, extensive studies on image classification have led to many interesting observations. In particular, one can easily fine-tune models trained on a specific image classification task to solve a different image classification task. This phenomenon, coined transfer learning, dramatically reduces the training time required to produce high-performing image classification models. In this project, we explore the implications of this particular phenomenon in the signals/audio domain and aim to illustrate the similarities and dissimilarities between machine learning models in music genre classification and speaker recognition.

2 Related Works

With the ever-increasing collection of digitized music libraries, music genre classification has seen growing popularity among industries and researchers alike. In 2018, Defferrard *et al.* held the CrowdAI Music Genre Classification Challenge [1] on the Free Music Archive (FMA) dataset [2]. In this competition, Kim *et al.* successfully applied transfer learning with music metadata (i.e., artist/album information) [3] and achieved great success with classifying the various music waveforms. As a representative example, this made a strong case for transfer learning toward its application in the audio domain.

Speaker recognition systems are not new, and in fact, the first voice recognition algorithm made its debut as early as 1952 [4]. Before the deep learning era, speaker classification relied on hand-crafted audio features (i.e., spectrograms). For images, AlexNet [5] was able to dwarf its competitors in the ImageNet Large Scale Visual Recognition Challenge. In speaker recognition, however, even before the introduction of deep neural network (DNN) based speaker recognition models, Gaussian Mixture Models (GMM-UBM) [6] and Hidden Markov Models (HMM-UBM) [7] already performed decently well. This led to the natural development of HMM/DNN hybrid speaker recognition models, which are currently the highest performing models in the domain.

CSC413: Neural Networks and Deep Learning (Winter 2021) Course Project, University of Toronto, Canada.

3 Methods

In this project, we explore an end-to-end audio model (e.g., [8], [9]) that directly takes unprepared audio as input, as well as a spectrogram-based model (e.g., [10]) which takes pre-processed audio spectrograms as input for classification. First, section 3.1 will discuss the dataset used for this paper. Then, to explore potential connections existing between the two different audio classification tasks, we propose experimentation in two parts detailed in section 3.2 and section 3.3. The resulting model performances are then compared and analyzed in section 4.

3.1 Dataset

We used the FMA medium dataset and a subset of the Voice Cloning Toolkit (VCTK) English Multi-speaker dataset for our experiments. The FMA dataset contains 25,000 tracks of 30-second clips from 16 different genres, and the VCTK dataset contains 7,000 tracks of readings from 16 different speakers. In particular, we did a 90:10 split on each dataset for training and validation.

Table 1: RawNet2 architecture

Layer	Input:59,049 samples	Output shape

Sinc-conv	Sinc(251,1,128)
	MaxPool(3)
	BN
	LeakyReLU
		(19683, 128)

Res block	BN LeakyReLU Conv(3,1,128) BN Conv(3,1,128) MaxPool(3) ×2	(2187, 128)
Res block	BN LeakyReLU Conv(3,1,256) BN Conv(3,1,256) MaxPool(3) ×4	(27, 256)
GRU	GRU(1024)	(1024,)
Speaker embedding	Dense(1024)	(1024,)
Output	Dense(16)	(16,)

Table 2: ConvNet Architectures

Layer	Output shape
Input layer	(1, 128, 513)
BatchNorm	(1, 128, 513)
Conv2d(4, 513), ReLU	(256, 125, 1)

Padding(2, 1)
Conv2d(4, 1)
ReLU
×4
	(256, 125, 1)

MaxPooling2d (125, 1)	(512,)
AvgPooling2d (125, 1)	(512,)
Cat(MaxsPool2d, AvgPool)	(1024,)
Dropout (0.2)	(1024,)
Dense(480), ReLU	(480,)
Dropout (0.2)	(480,)
Dense(240), ReLU	(240,)
Dropout (0.2)	(240,)
Dense(16), Softmax	(16,)
Output layer	(16,)

3.2 Music Genre Classification Model

After a comprehensive survey of Machine Learning models used for music genre classification, the Convolutional Neural Network on Short Time Fourier Transformation Spectrograms Model (ConvNet) proposed by Chumbalov *et al.* [11] was selected as our music genre classification model due to its simplicity and high placement in the competitive music genre classification challenge hosted by CrowdAI. The model uses the spectrogram approach and transforms audio samples into 2-D spectrograms to leverage the well-studied frameworks of image classification. Since image classification tasks are generally effective under transfer learning, we deemed the model as a good fit for our objective.

ConvNet uses five convolutional neural networks followed by three linear layers and implements batch normalization, max pooling, skip connections, and dropout. The specific model architecture is shown in Table 2. The model was trained using the FMA dataset.

To run the model, the audio samples must be first pre-processed following the steps detailed in [11]. Audio samples from both datasets were padded to exactly 30 seconds long, then split into 19 3-second Short-Time-Fourier-Transformed (STFT) frames to be taken as input by the model. After the data have been pre-processed, the model was first used to evaluate the FMA dataset. Then, the last linear layer of the model was replaced, and the model was retrained using the VCTK dataset with all weights except those on the last layer fixed.

3.3 Speaker Recognition Model

Similarly, the second part of our experimentations is to study Machine Learning models used for Speaker Recognition. We picked state-of-art speaker recognition model RawNet2 proposed by Jee-weon *et al.* [12] because it achieved the lowest EER of 2.57% in VoxCeleb1 dataset [13] in 2020. The model takes the raw waveform as input into a front-end embedding and back-end classification structure for speaker recognition.

RawNet2 uses a sinc-convolution layer of SincNet [14] followed by six residual blocks into a GRU layer and a speaker embedding fully connected layer with techniques such as feature map scaling. The specific model architecture can be found in Table 1.

We pre-processed the audio samples from the FMA and VCTK datasets into sequences of 59049 samples ($\approx$ 3 seconds) using Librosa [15]. For both datasets, we replaced the last layer with a new fully-connected layer followed by a softmax activation layer in the pre-trained RawNet2 model and trained our model until convergence.

4 Experiments & Discussions

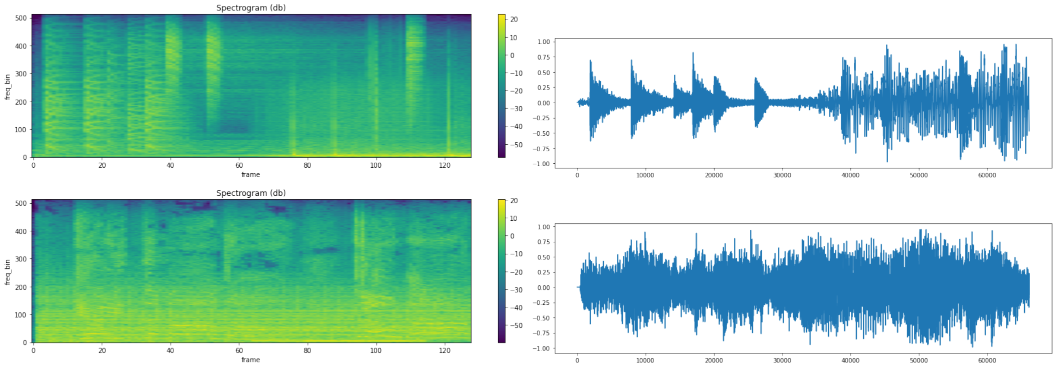

Figure 1: FMA music samples (Left: spectrogram. Right: raw audio waveform)
Electronic music (top) and Rock music (bottom)

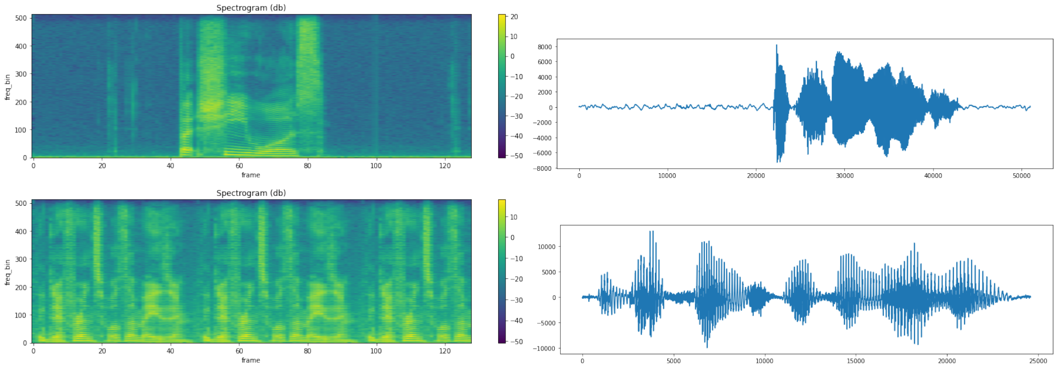

Figure 2: VCTK audio samples (Left: spectrogram. Right: raw audio waveform)
'But she was' (top) and 'The difference was amazing' (bottom)

All experiments reported in this paper were conducted using PyTorch [16], and the code is available at https://github.com/markohuang/csc413-final-project.

Table 3 reports the validation accuracy of ConvNet and RawNet on both the FMA dataset and the VCTK dataset. Surprisingly, RawNet not only outclasses ConvNet on speaker classification but also

Table 3: Validation Accuracy

Dataset Model	FMA	VCTK
ConvNet	45.14%	24.39%
RawNet 2	59.99%	93.55%

on music genre classification. While there may be many other factors at play, we present two primary reasons that may have contributed to this outcome.

Firstly, ConvNet seems to overfit the FMA dataset on the architectural level - the model is fed only selected frequencies from STFT, has peculiar placements of BatchNorm layers, and makes the questionable choice of using the AdaDelta optimizer instead of the Adam optimizer. The severely imbalanced dataset also emphasizes the overspecificity of ConvNet. We see that while ConvNet completely failed to recognize any Blues, Classical, or Country music, the model has a whopping 95% recall for Rock music and 72% recall for Electronic music, which collectively accounts for 53% of the entire FMA dataset. Contrastingly, RawNet directly takes in the raw audio waveform as input, uses a much richer hypothesis space to capture temporal information, and does not make any assumptions of the audio's frequency spectrum.

Secondly, it is speculated that in this instance, the music genre classification task is inherently easier than speaker recognition. For example, Rock music often has a very distinctive drum pattern, and electronic music a distinctive baseline. Which conveniently shows up as patterns on the spectrogram in Figure 1. However, speaker recognition can be much more nuanced. As we are dealing with text from newspapers, rainbow passages, and elicitation paragraphs, the utterances wildly differ across samples resulting in somewhat arbitrary spectrograms/waveforms in Figure 2. Effectively, to perform classification, the phonetic content of the audio samples needs to be separated from the speaker's pitch, tone, and identity. As is apparent with the disparity in model performance, the ConvNet model, which was trained on the music recognition dataset and more tailored towards a pattern recognition approach, is less likely to capture the hierarchy of information within the audio samples. Therefore, it could be hypothesized that the learned features carried over from RawNet are more expressive and may contain more useful information for audio processing tasks in general.

In the larger context of transfer learning, most successful transfer learning cases involve tasks more similar in nature than music genre classification and speaker recognition. For example, a digit recognition model offers the best performance when applying transfer learning for character recognition [17], but using transfer learning from ImageNet to medical imaging tasks yields little to no performance benefits [18]. Therefore, although the two specific models we explored in this paper did not perform well in each other's domain, it is possible that some other models used for audio classification tasks that are more similar in nature will have better success using transfer learning. Finally, for the scope of this paper, we only implemented the method of fine-tuning pre-trained models, which is one specific method of transfer learning [19]. Potentially, other transfer learning methods could lead to more promising results. However, for the two audio classification tasks that we have investigated for this paper, we conclude that transfer learning by fine-tuning pre-trained models is not effective.

5 Summary

This paper showed that transfer learning techniques that work well in the image classification domain might not perform as well in the audio classification domain. Specifically, we fine-tuned Speaker Recognition models for Music Genre Classification tasks and vice versa, showing that the resulting validation accuracy was significantly worse. However, since the resulting accuracy was still better than a random model, we can conclude that although the direct fine-tuning method of transfer learning does not apply well in audio classification tasks, other forms of transfer learning may find success. Therefore it is a direction worth pursuing.

Contribution

All team members contributed equally to this project.

References

[1] M. Defferrard, S. P. Mohanty, S. F. Carroll, and M. Salathé, “Learning to recognize musical genre from audio,” *Companion of the The Web Conference 2018 on The Web Conference 2018 - WWW ’18*, Mar 2018.

[2] M. Defferrard, K. Benzi, P. Vandergheynst, and X. Bresson, “Fma: A dataset for music analysis,” 2017.

[3] J. Kim, M. Won, X. Serra, and C. C. S. Liem, “Transfer learning of artist group factors to musical genre classification,” *Companion of the The Web Conference 2018 on The Web Conference 2018 - WWW ’18*, 2018. [Online]. Available: http://dx.doi.org/10.1145/3184558.3191823

[4] K. H. Davis, R. Biddulph, and S. Balashek, “Automatic recognition of spoken digits,” *The Journal of the Acoustical Society of America*, vol. 24, no. 6, pp. 637–642, 1952. [Online]. Available: https://doi.org/10.1121/1.1906946

[5] A. Krizhevsky, I. Sutskever, and G. E. Hinton, “Imagenet classification with deep convolutional neural networks,” vol. 25, 2012. [Online]. Available: https://proceedings.neurips.cc/paper/2012/file/c399862d3b9d6b76c8436e924a68c45b-Paper.pdf

[6] C. Miyajima, Y. Hattori, K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitamura, “Speaker identification using gaussian mixture models based on multi-space probability distribution,” in *2001 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings (Cat. No.01CH37221)*, vol. 1, 2001, pp. 433–436 vol.1.

[7] P. Bansal, A. Kant, S. Kumar, A. Sharda, and S. Gupta, “Improved hybrid model of hmm/gmm for speech recognition,” *Intell. Technol. Appl*, 01 2008.

[8] Y. Shao, Y. Wang, D. Povey, and S. Khudanpur, “Pychain: A fully parallelized pytorch implementation of lf-mmi for end-to-end asr,” 2020.

[9] J. Rohdin, A. Silnova, M. Diez, O. Plchot, P. Matějka, L. Burget, and O. Glembek, “End-to-end dnn based text-independent speaker recognition for long and short utterances,” *Computer Speech & Language*, vol. 59, pp. 22–35, 2020. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0885230818303632

[10] Y. Zeng, H. Mao, D. Peng, and Z. Yi, “Spectrogram based multi-task audio classification,” *Multimed. Tools Appl.*, vol. 78, no. 3, pp. 3705–3722, Feb. 2019.

[11] C. Daniyar and P. Philipp, “pushnyakov/wwwmusicalgenrerecognitionchallenge.” [Online]. Available: https://github.com/pushnyakov/WWWMusicalGenreRecognitionChallenge

[12] H. Tak, J. Patino, M. Todisco, A. Nautsch, N. Evans, and A. Larcher, “End-to-end anti-spoofing with RawNet2,” Nov. 2020.

[13] A. Nagrani, J. S. Chung, and A. Zisserman, “VoxCeleb: A large-scale speaker identification dataset,” in *Interspeech 2017*. ISCA, Aug. 2017. [Online]. Available: https://doi.org/10.21437/interspeech.2017-950

[14] M. Ravanelli and Y. Bengio, “Speaker recognition from raw waveform with sincnet,” 2019.

[15] B. McFee, C. Raffel, D. Liang, D. Ellis, M. McVicar, E. Battenberg, and O. Nieto, “librosa: Audio and music signal analysis in python,” in *Proceedings of the 14th Python in Science Conference*. SciPy, 2015. [Online]. Available: https://doi.org/10.25080/majora-7b98e3ed-003

[16] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Kopf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, and S. Chintala, "Pytorch: An imperative style, high-performance deep learning library," in *Advances in Neural Information Processing Systems*, H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, Eds., vol. 32. Curran Associates, Inc., 2019. [Online]. Available: https://proceedings.neurips.cc/paper/2019/file/bdbca288fee7f92f2bfa9f7012727740-Paper.pdf

[17] R. Saha, "Transfer learning - a comparative analysis," 2018. [Online]. Available: http://rgdoi.net/10.13140/RG.2.2.31127.39848

[18] M. Raghu, C. Zhang, J. Kleinberg, and S. Bengio, "Transfusion: Understanding transfer learning for medical imaging," in *Advances in Neural Information Processing Systems*, vol. 32. Curran Associates, Inc., 2019.

[19] S. J. Pan and Q. Yang, "A survey on transfer learning," *IEEE Transactions on Knowledge and Data Engineering*, vol. 22, no. 10, pp. 1345–1359, 2010.