TRANSFER LEARNING FOR MUSIC EDUCATION

Pedro Ramoneda
Music Technology Group
pedro.ramoneda01@estudiant.upf.edu

ABSTRACT

Music education is an important topic that attracts a lot of attention. However, there have been few computational approaches to pursue educational outcomes, datasets are small, and models can not generalise. This final project aims to assess whether an instrument quality sound is good or bad. The few available data issue is managed with transfer learning techniques. Four transfer learning strategies have been carried out to conduct the sound quality assessment. All results and the code are public [1] .

1. INTRODUCTION

Advances in music information retrieval and computational musicology have benefited several areas related to classical music performance in recent years [14]. One example is music education, which is the study of how music is studied and taught. Some rather similar areas, such as music libraries and music editorials, may benefit from data-driven technologies.

In recent years, several computational technologies have been proposed to enhance music education and performance evaluation. However, in all these projects, the lack of data is notable. In these applications, automated musical output assessment is critical [3]. Besides, human assessment is often subjective, rendering the implementation of an automatic evaluation method challenging. Musical education evaluations may rely on a variety of aspects of music, including pitch and timing precision, musical vocabulary, and technique [6]. However, the parameters are often arbitrary, as mentioned by Wesolowski et al. [20].

Despite subjectivity, there is a general agreement on what constitutes a good and bad sound for a given instrument. With this in mind, Knight et al. [13] used a machine learning approach to identify sound characteristics in terms of consistency automatically. A similar technique is used in the Good-sounds project [2, 16], which attempts to grade isolated musical notes based on their output quality or sound goodness. Five sound attributes that influence the note goodness are described in collaboration with a community of music teachers, as is shown in Table 1. Afterwards, Giraldo et al. [9] focus his research in violin tone assessment based on a survey of experts. Giraldo et al. [9] stated in the conclusion that his system and Good-sounds failed to generalise.

Dynamic stability
Pitch stability
Timbre stability
Timbre richness
Attack clarity

Table 1. Sound attributes that influence the note goodness.

Transfer learning may be the solution to the lack of data in music education. As is stated on Jialin et al [15]: In such cases, knowledge transfer, if done successfully, would greatly improve the performance of learning by avoiding much expensive data-labeling efforts. Also, there are success stories in other domains of music technology with transfer learning in neural networks, in genre recognition [8] or in emotion recognition. Choi et al. [4] expose a transfer learning approach for multiple music classification and regression tasks for music purposes.

Machine learning has achieved a lot of success in data-intensive systems, but it has a lot of issues where the data collection is minimal. Few-Shot learning area has recently been suggested as a solution to this issue [19]. This new methods can easily generalize to new tasks involving just a few samples of supervised data using prior knowledge. This final project wants to be a first naive approach to few shot learning in music education.

Instrument	Sounds
clarinet	1688
violin	1383
flute	980
cello	948
trumpet	934
sax_alto	720
sax_tenor	680
piccolo	388
sax_soprano	334
sax_baritone	288
oboe	247
bass	159

Table 2. Instruments and number of sound recordings on Good-sounds dataset

[1] https://github.com/PRamoneda/transfer_learning_for_music_education

Attribution: Pedro Ramoneda. "Transfer Learning For Music Education", 22st International Society for Music Information Retrieval Conference, Online edition, 2021.

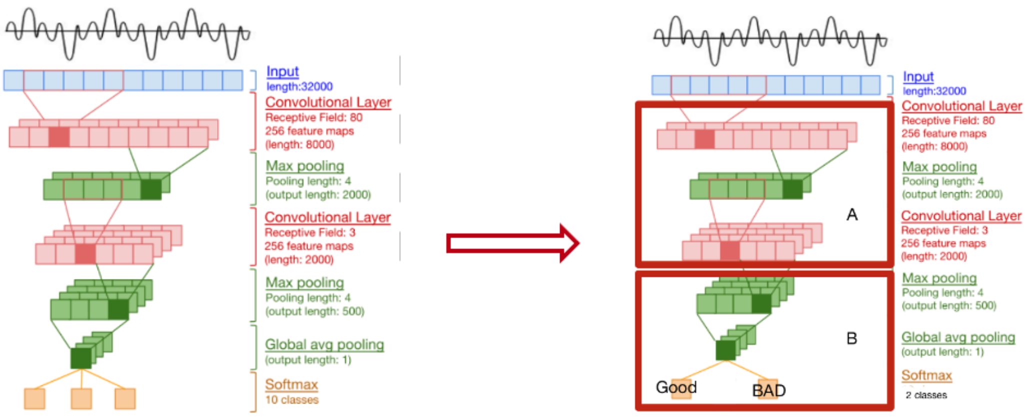

Figure 1. The first subplot is the M5 architecture. The second subplot is the M5 divided in two areas: A and B, for transfer learning purposes.

The remainder of this paper is structured as follows. Section 2 details the data source of this study. Section 3 addresses the transfer learning tasks used to model the sound quality assessment problem. Section 4 expose the evaluation system and Section 5 details the results. Finally, Section 6 draws the conclusions of the study and states areas for future work.

2. GOOD-SOUNDS DATASET

The Good-sounds dataset is born of the collaboration between the Music Technology Group and Korg. Good-sounds [2, 16] is carried out recording a training dataset of single note excerpts including six classes of sounds per studied instrument. Twelve different instruments are recorded, as is shown in Table 2. For each instrument, the complete range of playable semitones is captured several times with various tonal characteristics. There are two classes: Good and Bad sounds. Bad sounds are divided into five sub-classes, one for each musical dimension stated by the expert musicians, shown in Table 1. Bad sounds are composed by examples of note recordings that are intentionally badly played. The last class includes examples of note recordings that are considered to be well played. The instrument sounds are recorded up to four different microphones and performed by up to two different musicians. The different subsets of labels are very unbalanced, bringing another challenge.

3. METHODOLOGY

This section exposes the several transfer learning tasks designed for expanding the knowledge of sound quality assessment from other domain fields. The goal of all tasks is to classify between two quality classes: good and bad sounds.

3.1 Task 0

This strategy can be understood as a benchmark task to compare the rest of the tasks with. The model is built from the M5, a neural network proposed in class and explained carefully in Dai et al. [5]. M5 takes raw waveform data as input and has several convolutional and a classifier layer at the end, as shown in the first subplot of Figure 1.

3.2 Task 1

In the M5 architecture, filters or convolutional layers are trained on the processed raw audio, and it is not possible to transfer learning from very different sounds. Therefore, the strategy followed in this task is to use other string family instruments to train the sound quality assessment on the target. For example, training cello and violin and transfer learning to bass. For this purpose, M5 is trained for the rest of the string family. Then, the layers marked as a in Figure x are frozen and classifier B is trained with the target's sounds.

3.3 Task 2

Image recognition is a machine learning area with several publications and data. The image recognition models are trained on much larger datasets than the music education systems [10]. The strategy is to convert the audio into a Mel-spectrogram and process it as a grey-scale image. The image recognition neural networks are very good at recognising patterns in their first layers. The models are going to keep the first layers, the feature extractors. Finally, the last layer of the classifier is trained on the good sounds data. In this work, we transfer learning from three image recognition state-of the-art models, resnet18 [10], VGG19 [18] and densenet161 [12], to sound quality assessment.

3.4 Task 3

Sound detection and classifications is a field that has grown a lot in recent years. Audioset's publication [7], a collection of more than 2 million sounds, has considerably increased the interest in the field. VGGish is a variant of the VGG architecture but adapted for audio with a Mel-spectrogram as input [11]. It is trained with Audioset data, the most extensive dataset in audio [7]. As in Task 2, transfer learning is carried out by retraining only the last classification layer, as in previous proposals [17].

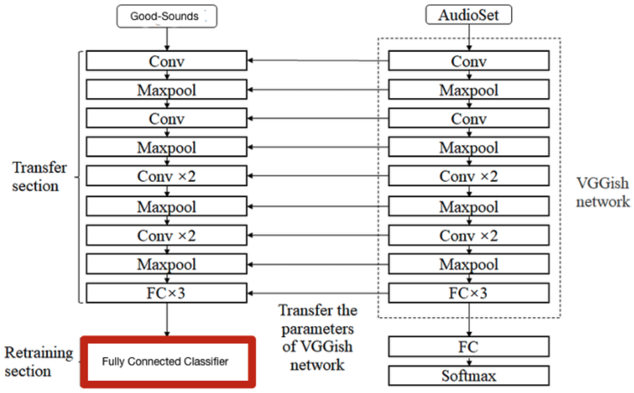

Figure 2. Transfer learning from Audioset VGGish to Good-sounds.

3.5 Task 4

The fourth task strategy is to transfer learning from a model that has already been transferred. For this purpose, a model trained to recognise voice/instrumental music audios proposed on Essentia TensorFlow [1] is used as a starting point. This model is already been transferred from the VGGish model trained in Audioset [11]. Therefore the layers trained by Audioset and an embedding layer trained in Essentia TensorFlow [1] are used, and only the fully connected classifier is trained on top of all these layers.

4. EVALUATION

All the experiments conducted from task 0 to 4 have been achieved with the string sound instruments from Good sounds: violin, cello and bass. The data have been divided in two subsets, train and test, stratified by the good sound subclases. Classes are very unbalanced so accuracy metric (A) is not the best option:

$$A = \frac{\text{tp} + \text{tn}}{\text{tp} + \text{fp} + \text{tn} + \text{fn}} \tag{1}$$

In this project, it is computed by educational purposes. A more suitable metric for this problem is balanced accuracy (BA):

$$BA = \frac{\text{atp} + \text{atn}}{2} \tag{2}$$

BA is the mean of the averaged true positives (atp) and the averaged true negatives (atn). Since the classification is binary, even though it is unbalanced, it is possible to compute f-score (F), precision (P) and recall (R) metrics:

$$P = \frac{\text{tp}}{\text{tp} + \text{fp}} \quad , \quad R = \frac{\text{tp}}{\text{tp} + \text{fn}} \quad , \quad F = \frac{2PR}{P + R} \tag{3}$$

5. RESULTS

All the results have been computed for the five tasks described in the Section 4. The results range between 80 and 90% Accuracy and between 50% and 70% balanced Accuracy. The 50% accuracy occurs when the recall is 100%, this value is reached by several transferred models due to the lack of data.

In task 0, the baseline accuracy is approximately 86% for the distinct instruments: violin, cello and double bass. After transferring the family knowledge to the specific instrument, the cello and violin reach the same accuracy as task 0. It means that the feature extractor is similarly trained with stringed instruments. So in the creation of a new dataset, it would only have to annotate one instrument massively. The double bass has worse transfer learning results (50% of accuracy). This issue may be due to its tessitura or because, as shown in Table 2, it has very few samples.

A	BA	F	P	R
88,44	71,33	93,36	90,72	96,15

Table 3. Quality sound assessment Transfer learning from Resnet18 to violin.

In task 2, similar results to the other tasks are obtained in the transfer learning process from the Densenet161 model and the VGG19. However, with Resnet18, the best results are obtained in the evaluation; for example, the violin results can be seen in Table 3. Similarly to the rest of the tasks, the recall is always higher than the precision.

In task 3 and 4, the results are not better than the other tasks. Moreover, the classifier layer hardly varies in the epochs iteration, although the loss is gradually decreasing. The VGGish embeddings may not distinguish the two classes at all, and most of them are classified as bad class unequivocally.

6. DISCUSSION AND FUTURE RESEARCH LINES

Transfer learning results have not improved the baseline task 0 substantially. The higher recall than precision indicates that to ensure that the larger class is more accurate (bad class) the smaller class has few hits (good class). The feature extractor may cause this issue. The feature extractor is not trained to separate the two classes, and maybe the features extracted are not enough to classify correctly. Possibly, training in the last epochs of the feature extractor could improve the transfer learning tasks. However, the training process is faster in time. Furthermore, it can be seen that metrics are very important in this type of unbalanced audio dataset. Assessing the audios by hand, we have discovered that it is crucial to consider the whole audio, not just the attack or the first three seconds. Fully CNN architectures can be used for this purpose.

Finally, in the future, we want to try other few-shot learning techniques to train neural networks by exploiting the fact that two or three distinct musicians are playing each instrument. The data needed to model systems capable of evaluating humans for music education purposes is very expensive to annotate. Therefore, in the future, it is expected that researchers will use domain knowledge plus the few-shot learning methodologies to achieve significant results in this field.

7. REFERENCES

[1] Pablo Alonso-Jiménez, Dmitry Bogdanov, Jordi Pons, and Xavier Serra. Tensorflow audio models in essentia. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 266–270. IEEE, 2020.

[2] Giuseppe Bandiera, Oriol Romani Picas, Hiroshi Tokuda, Wataru Hariya, Koji Oishi, and Xavier Serra. Good-sounds. org: A framework to explore goodness in instrumental sounds. In *Devaney J, Mandel MI, Turnbull D, Tzanetakis G, editors. ISMIR 2016. Proceedings of the 17th International Society for Music Information Retrieval Conference; 2016 Aug 7-11; New York City (NY).[Canada]: ISMIR; 2016. p. 414-9.* International Society for Music Information Retrieval (ISMIR), 2016.

[3] Baris Bozkurt, Sankalp Gulati, Oriol Romani Picas, and Xavier Serra. Musiccritic: a technological framework to support online music teaching for large audiences. In *Forrest D, editor. Proceedings of the International Society for Music Education. 33rd World Conference on Music Education (ISME); 2018 Jul 15-20; Baku, Azerbaijan. Malvern: International Society for Music Education; 2018.* International Society for Music Education, 2018.

[4] Keunwoo Choi, György Fazekas, Mark Sandler, and Kyunghyun Cho. Transfer learning for music classification and regression tasks. *arXiv preprint arXiv:1703.09179*, 2017.

[5] Wei Dai, Chia Dai, Shuhui Qu, Juncheng Li, and Samarjit Das. Very deep convolutional neural networks for raw waveforms. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 421–425. IEEE, 2017.

[6] Vsevolod Eremenko, Alia Morsi, Jyoti Narang, and Xavier Serra. Performance assessment technologies for the support of musical instrument learning. 2020.

[7] Jort F Gemmeke, Daniel PW Ellis, Dylan Freedman, Aren Jansen, Wade Lawrence, R Channing Moore, Manoj Plakal, and Marvin Ritter. 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, 2017.

[8] Deepanway Ghosal and Maheshkumar H Kolekar. Music genre recognition using deep neural networks and transfer learning. In *Interspeech*, pages 2087–2091, 2018.

[9] Sergio Giraldo, George Waddell, Ignasi Nou, Ariadna Ortega, Oscar Mayor, Alfonso Perez, Aaron Williamon, and Rafael Ramirez. Automatic assessment of tone quality in violin music performance. *Frontiers in psychology*, 10:334, 2019.

[10] 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.

[11] Shawn Hershey, Sourish Chaudhuri, Daniel PW Ellis, Jort F Gemmeke, Aren Jansen, R Channing Moore, Manoj Plakal, Devin Platt, Rif A Saurous, Bryan Seybold, et al. Cnn architectures for large-scale audio classification. In *2017 ieee international conference on acoustics, speech and signal processing (icassp)*, pages 131–135. IEEE, 2017.

[12] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4700–4708, 2017.

[13] Trevor Knight, Finn Upham, and Ichiro Fujinaga. The potential for automatic assessment of trumpet tone quality. In *ISMIR*, pages 573–578. Citeseer, 2011.

[14] Alexander Lerch, Claire Arthur, Ashis Pati, and Siddharth Gururani. An interdisciplinary review of music performance analysis. *Transactions of the International Society for Music Information Retrieval*, 3(1), 2020.

[15] Sinno Jialin Pan and Qiang Yang. A survey on transfer learning. *IEEE Transactions on knowledge and data engineering*, 22(10):1345–1359, 2009.

[16] oriol romani picas, hector parra rodriguez, dara dabiri, hiroshi tokuda, wataru hariya, koji oishi, and xavier serra. a real-time system for measuring sound goodness in instrumental sounds. *journal of the audio engineering society*, may 2015.

[17] Lukui Shi, Kang Du, Chaozong Zhang, Hongqi Ma, and Wenjie Yan. Lung sound recognition algorithm based on vggish-bigru. *IEEE Access*, 7:139438–139449, 2019.

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

[19] Yaqing Wang, Quanming Yao, James T Kwok, and Lionel M Ni. Generalizing from a few examples: A survey on few-shot learning. *ACM Computing Surveys (CSUR)*, 53(3):1–34, 2020.

[20] Brian C Wesolowski, Stefanie A Wind, and George Engelhard Jr. Examining rater precision in music performance assessment: An analysis of rating scale structure using the multifaceted rasch partial credit model. *Music Perception: An Interdisciplinary Journal*, 33(5):662–678, 2016.