Audio Scene Intelligence in domestic environment

Siddharth Rai, Kumar Vaibhav , K.Prabakar

SRM Institute of Science and Technology

sr6723@srmist.edu.in | kv3575@srmist.edu.in | ps3488@srmist.edu.in

Abstract

There have been many research and work on the topic of "classification of domestic data" to bring best possible model with minimum memory requirements for building a better architecture for speech recognition for the home automation devices that can automate the profiling of different person in different domestic condition.Applications like 'Siri' and 'Google assistant' are already in the market that are changing the human behavior ,Their have been many papers and research on our topic though there is minimum implementation of this at the practical level i.e to integrate this into Voice Assistant Devices. Audio Scene Detection, Speaker Verification and User Profiling in the domestic environment can help in developing many new use cases in Voice Assistant solutions.This is our approach and solution for the same.

Introduction

Acoustic scene classification is a task of widespread interest in the general topic of environmental audio analysis, and refers to the specific case of classifying environments based on their general acoustic characteristics . Other and many closely related and popular directions of research include classification of individual sound events ,area of their occurrence in the environment, sound event detection, localization and tagging. Specific applications for acoustic scene classification include services and devices that can benefit of context awareness , services or applications for indexing audio content , documentary and archival of everyday experience , wearable technology, navigation systems for robotics, etc .Recent years have seen a boom in deep-learning based solutions for various classification problems, also obvious in the last few editions of the DCASE Challenge[1]. While in 2016 just 22 of the 48 submissions used neural networks , in 2019 only five of the 146 systems submitted to the acoustic scene classification subtasks did not include a deep learning component . Generally, deep learning algorithms require large amounts of data for best performance, and the effort to produce more data for the task has resulted in gradual extension of the problem, from a classical textbook example to domain adaptation due to mismatched devices , and open-set classification .

Prior Work

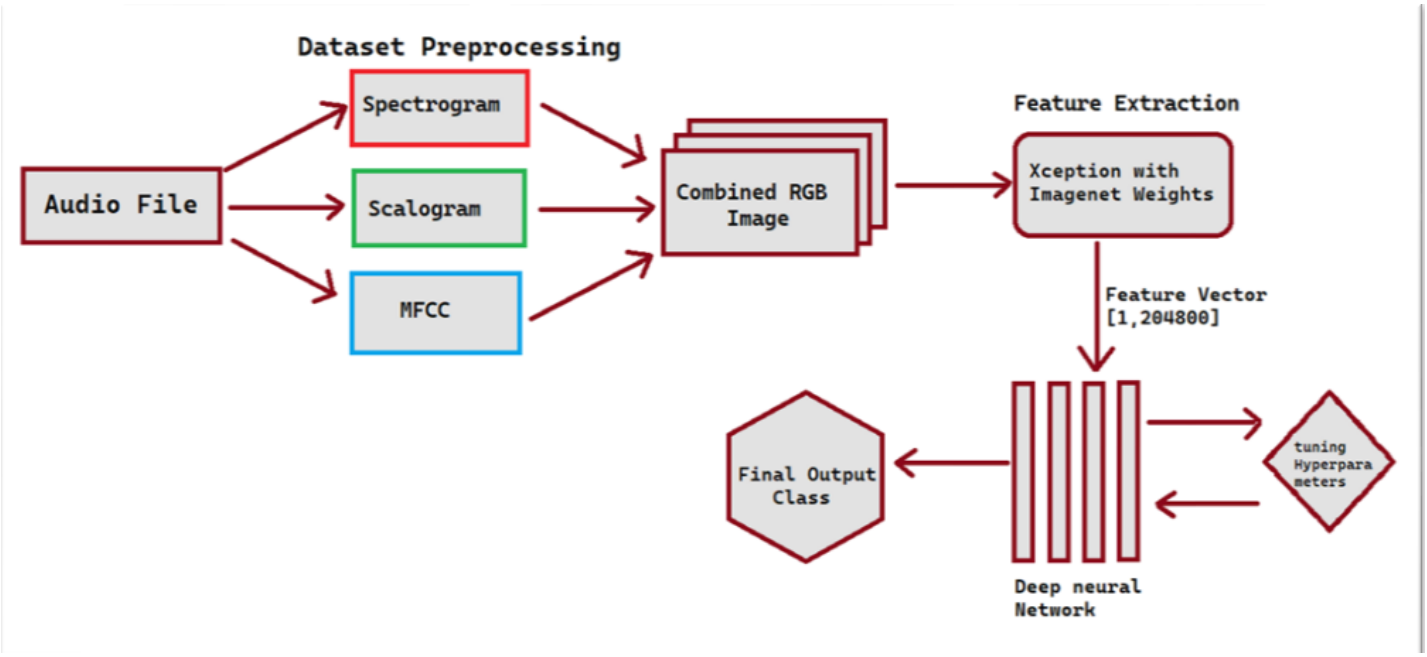
(THE RECEPTIVE FIELD AS A REGULARIZATION IN DEEP CONVOLUTIONAL NEURAL NETWORK FOR ACOUSTIC SCENE CLASSIFICATION)
This paper investigated the relation between CNNs' RFs over the input spectrograms and their generalization on unseen samples, for the acoustic scene classification task[3]. It showed that a large RF especially over the frequency dimension pushes CNNs to overfit, while a smaller than necessary RF forces a CNN to underfit the data and prevents it from learning decisive features. Although many factors contribute to a CNN's tendency to generalize, we show that for a specific training setup and network architecture, tuning the RF of the model is a crucial factor for its performance.

(ACOUSTIC SCENE CLASSIFICATION IN DCASE 2020 CHALLENGE: GENERALIZATION ACROSS DEVICES AND LOW COMPLEXITY SOLUTIONS**)**
This paper presented an analysis of the solutions submitted to DCASE 2020 Challenge Task 1 Acoustic Scene Classification. The two different subtasks tackle the research problem from the point of view of real-world applications, in one case robustness and generalization to multiple devices, and in the other case requiring a low-complexity solution. The baseline system for the task implements a convolutional neural network (CNN) based approach, similar to the DCASE 2019 Task. 1 baseline. It uses 40 log mel-band energies, calculated with

an analysis frame of 40 ms and 50% hop size, to create an input shape of 40 × 500 for each 10 second audio file. The neural network consists of two CNN layers and one fully connected layer, followed by the softmax output layer.

(CP-JKU SUBMISSIONS TO DCASE'20: LOW-COMPLEXITY CROSS-DEVICE ACOUSTIC SCENE CLASSIFICATION WITH RF-REGULARIZED CNNS**)**
It showed that by adding a further limitation on the effective receptive field in the form of frequency-damping, we improve the accuracy of our RF Regularized baseline ResNet. Additionally, we investigated several approaches to reduce the number of parameters in our models

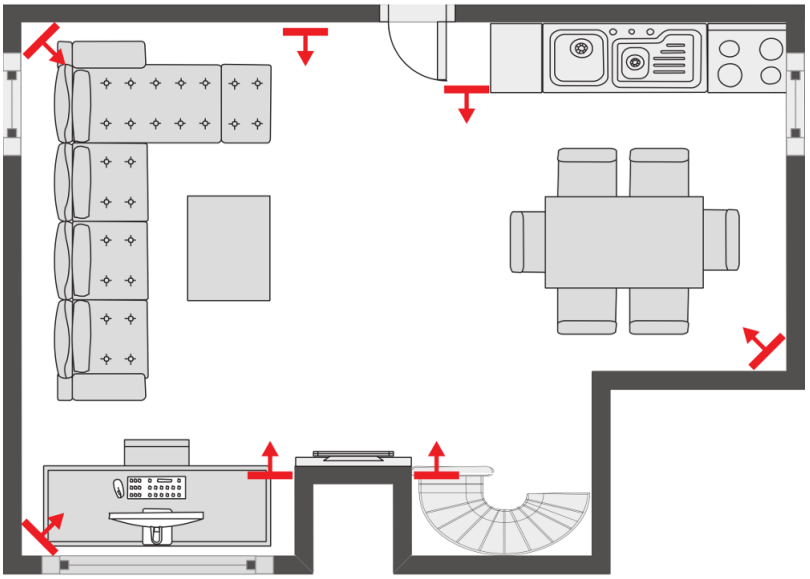
Our Approach:



Audio file is first converted into respective Spectrogram, Scalogram and MFCC ,later all the three are combined to form RGB image which are send through a Xception layer that consist of weights of Imagenet layer ,we extract a feature vector of [1,204800] form the process ,these features are used as training data for the neural network ,training hyperparameters are constantly updated through feedback mechanism , brief about these steps are described below.

Data set

The dataset used in this task is a derivative of the **SINS dataset**. It contains a continuous recording of one person living in a vacation home over a period of one week. It was collected using a network of 13 microphone arrays distributed over the entire home. The microphone array consists of 4 linearly arranged microphones. For this task 7 microphone arrays in the combined living room and kitchen area are used.



The continuous recordings were split into audio segments of 10s. Segments containing more than one active class (e.g. a transition of two activities) were left out. This means that each segment represents one activity. Subsampling was then performed starting from the largest classes.These audio segments are provided as individual files along with the ground truth. Each audio segment contains 4 channels (e.g. the 4 microphone channels from a particular node). The dataset is passed on to extract Scalogram ,Spectrogram and MFCC to be passed through the Xception layer ,we proposed a combination of all of the leading extracting techniques so that all of the necessary features could be considered and could contribute to the identification of the audio data.

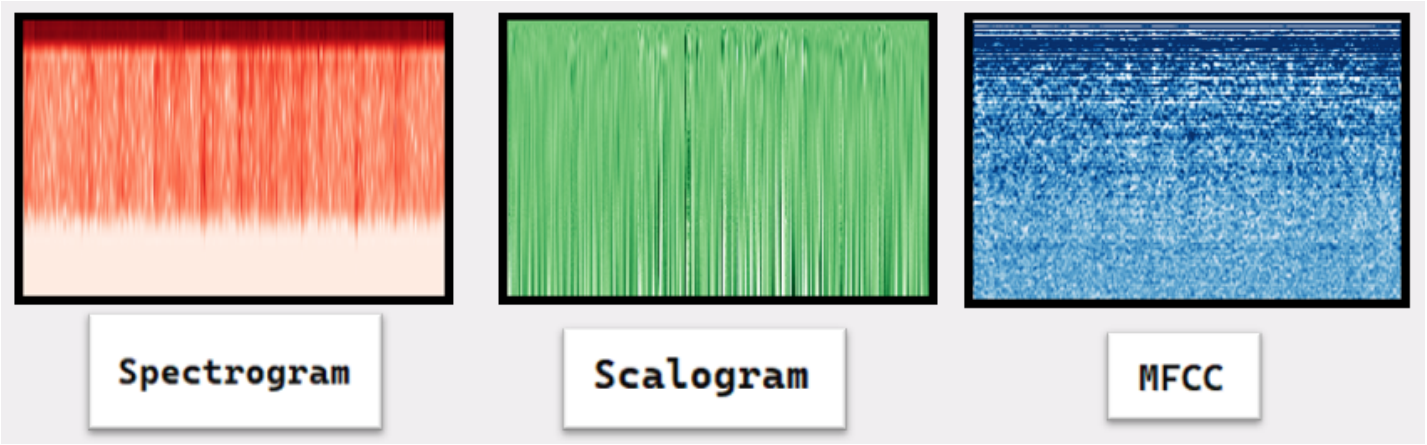
Data preprocessing

Converting sound to RGB picture

The features of all the layers form the spectrogram ,Scalogram and MFCC are used to develop the RGB image.Audio have many different ways to be represented, going from raw time series to time-frequency decompositions. The choice of the representation is crucial for the performance of your system. Initially, the DCASE task was many carried out based on handcrafted features or Mel-frequency cepstral coefficients (MFCCs).Additionally, they are adapted to the use of convolutional neural networks. Among time-frequency decompositions, Spectrograms have been proved to be a useful representation for DCASE task, and more generally for audio processing. They consist of 2D images representing sequences of Short Time Fourier Transform (STFT) with time and frequency as axes, and brightness representing the strength of a frequency component at each time frame. In such a way they appear a natural domain to apply CNN'S architectures for images directly to sound.

Another less commonly used representation is the Scalogram, 2 dimensional output of the wavelet transform with scale and time as axes. As wavelets can filter signals in a multiscale way, scalograms may outperform spectrograms for the DCASE task.

In this project, sound is represented as a RGB colour image, red being spectrogram, green being scalogram and blue MFCC and all of them are used to describe an audio file. The images are stored in a array format and they are resized to (299,299). For the training 10971 total examples are taken with approx 1000 examples per class.



Feature Extraction

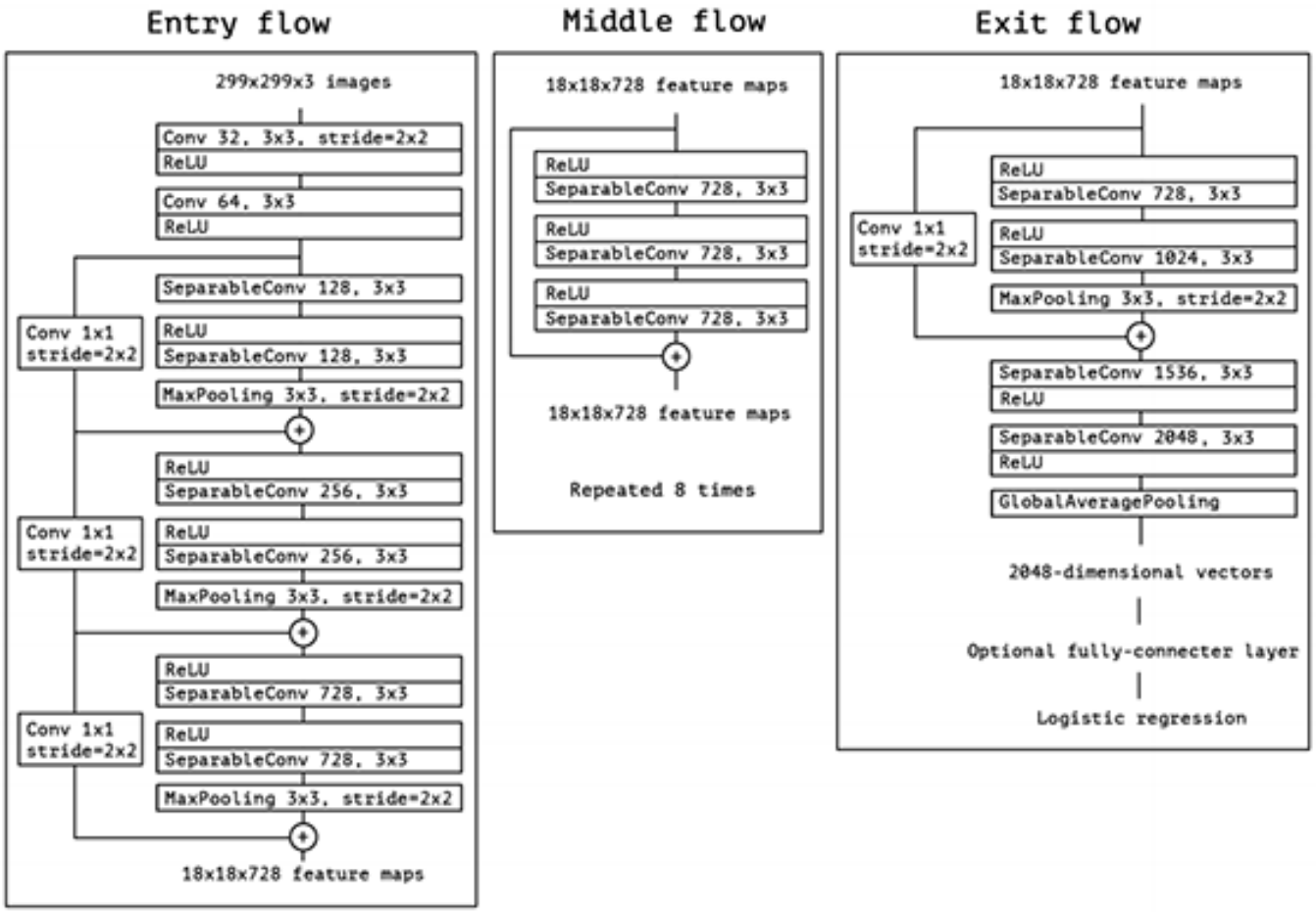
Classifying domestic sound using transfer learning

Domestic sound data generally suffers from limited labeled datasets. This makes it difficult to train deep models from scratch. On the contrary, In computer vision, the availability of large datasets such as ImageNet enables training very deep models for the task of image classification. Since 2012, state of the art computer vision networks have been trained on ImageNet, a dataset containing approximatively 1.2 million images separated into 1000 categories. Best performing architectures relied on convolutional neural networks and some of the pre-trained weights are available in the Keras library.

For our pretrained model, we chose Xception.Xception is an extension of Inception architecture with the use of depth wise separable convolutions. The pre-trained weights of these networks can be used as features extractors on our RGB representation of sound in a transfer learning approach.

Xception

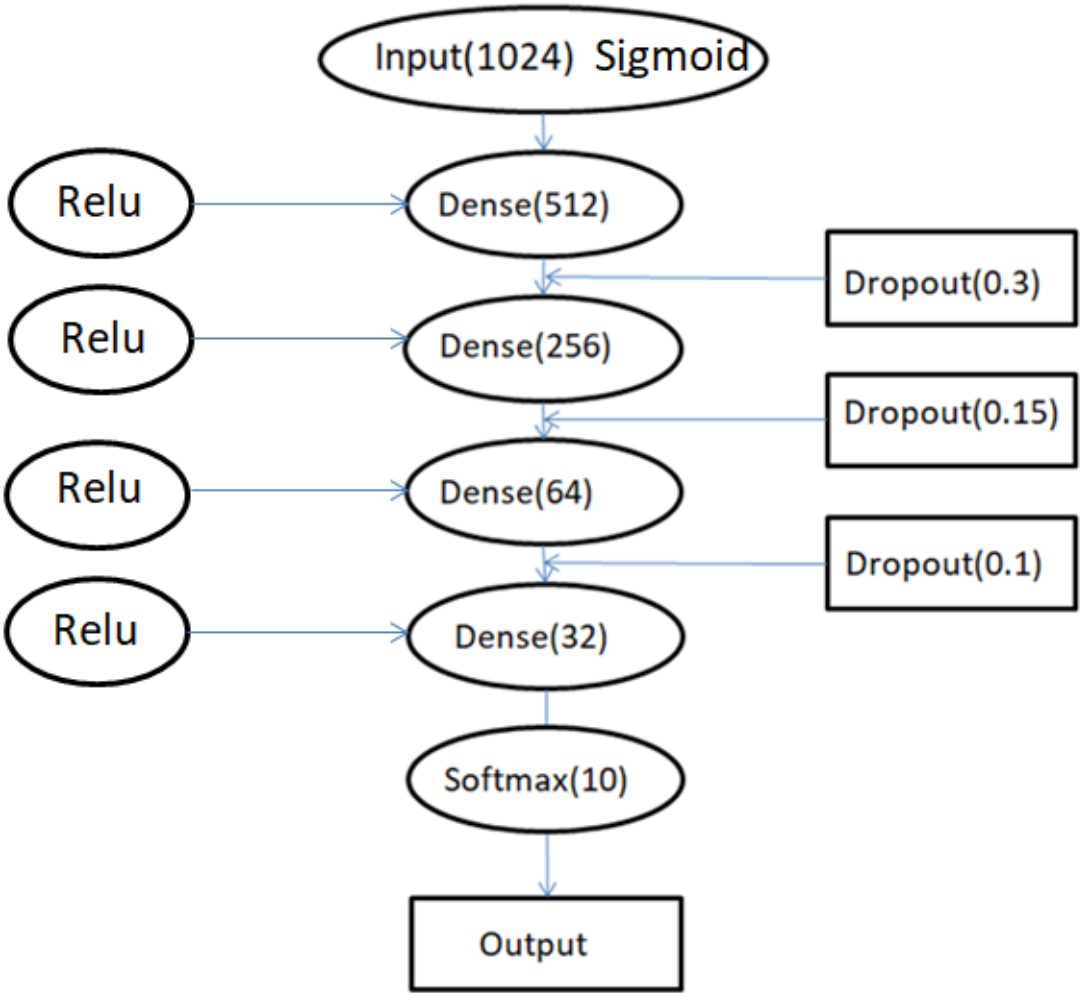
A complete description of the specifications of the network is given in figure below. The Xception architecture has 36 convolutional layers forming the feature extraction base of the network. In our experimental evaluation we will exclusively investigate image classification and therefore our convolutional base will be followed by a logistic regression layer. The 36 convolutional layers are structured into 14 modules, all of which have linear residual connections around them, except for the first and last modules[6]. In short, the Xception architecture is a linear stack of depthwise separable convolution layers with residual connections. This makes the architecture very easy to define and modify; it takes only 30 to 40 lines of code using a high level library such as Keras or TensorFlow-Slim , not unlike an architecture such as VGG-16 , but rather un- like architectures such as Inception V2 or V3 which are far more complex to define. An open-source implementation of Xception using Keras and TensorFlow is provided as part of the Keras Applications module2 , under the MIT license.



Final Prediction

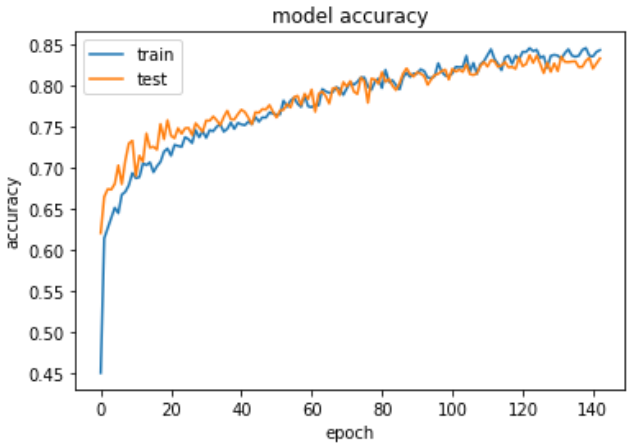
Deep Neural Network Architecture

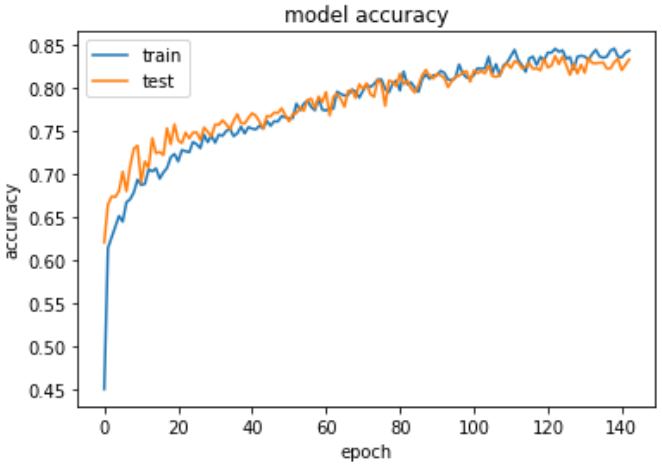
The output of Xception layer is fed into Our neural Network that proposes a sequential model that consist of 5 Dense Layer and an output layer. The dataset is divided into test and train splits of 70% and 30% and with a random state of 200.
The input for the model is a vector of shape [1,204800]. `Adam` is used as a model optimiser. We found that a learning rate of 0.001 gives a better accuracy , and prevents model form overfitting. For overfitting we introduced a decay rate of 1e-5 with 2 callbacks which are early stopping with patience of 20 with monitor of validation loss and model checkpoint which saves best weights on the basis of low validation loss, we then trained the model for 150 epochs with a batch size of 128.



Conclusion:

We achieved a accuracy of 83.78 percentage with a validation loss of 0.4716 with the evaluation dataset after training it for 143 epochs after which model stopped training due to early stopping callback. The model accuracy can be increased by increasing the dataset size and therefore decrease the chance of overfitting. The examples for the classes which have similar pre-processed images should be increased to decrease intra class similarity.
We chose a single learner for our model prediction but if the resources are available we can go for ensemble approach which will give better result by combining multiple weak learners.





REFERENCES

[1] Monitoring of Domestic activities based on multi- channel acoustics. http://dcase.community/challenge2018/task-monitoring-domestic-activities

[2] Kosmider, M. (2019, June), "Calibrating neural networks for secondary recording devices," in Proceedings of the Detection and Classification of Acoustic Scenes and Events Workshop (DCASE), New York, NY, USA (pp. 25-26).

[3] H. Zhang, M. Cisse, Y. N. Dauphin, and D. LopezPaz, "Mixup: Beyond empirical risk minimization," in International Conference on Learning Representations, 2018.

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

[5] Koutini, K., Eghbal-zadeh, H., and Widmer, G. (2019), "CPJKU submissions to DCASE'19: Acoustic Scene Classification and Audio Tagging with Receptive-FieldRegularized CNNs," in Proceedings of the Detection and Classification of Acoustic Scenes and Events 2019 Workshop (DCASE2019)

[6] Xception: Deep Learning with Depthwise Separable Convolutions Franc¸ois Chollet Google, Inc. https://arxiv.org/pdf/1610.02357.pdf

[7] RGB representation of sound for environmental sound classification: https://github.com/vbelz/audio_classification