2022 16th International Conference on Ubiquitous Information Management and Communication (IMCOM) | 978-1-6654-2678-7/22/$31.00 ©2022 IEEE | DOI: 10.1109/IMCOM53663.2022.9721724

Data Cleaning of Sound Data with Label Noise Using Self Organizing Map

Pildong Hwang
Department of Computer & Information Sciences
Towson University
Towson, United States
phwang1@students.towson.edu

Yanggon Kim
Department of Computer & Information Sciences
Towson University
Towson, United States
ykim@towson.edu

***Abstract*— The noise label of data is a problem that can cause low performance of deep learning. It is difficult to manually re-label due to huge amounts of data. In addition, there are much more problems due to the similarity of sounds that are difficult to manually distinguish and label sound data. We proposed a data cleaning method using SOM (Self-Organizing Map), one of the unsupervised learning methods. In order to extract compact features from audio, densely connected layer with log scaled Mel-spectrogram is used. Data selection is performed based on the Euclidean distance of each Best matching unit (BMU) derived through the SOM. We also experiment with various grid sizes for SOM to find an efficient grid size. In addition, an appropriate distance finding experiment is conducted. This method is evaluated in sound classification using a pre-trained DenseNet model.**

Keywords—data cleaning, sound data, label noise, self-organizing map

I. INTRODUCTION

Label noise in datasets that can affect deep learning results is an issue that needs to be addressed. Noisy labeled data is also known as polluted or corrupted data, and the most reliable strategy for troubleshooting is to manually relabel the data. However, large datasets are too difficult to check and correct labels manually. The proposed method to solve the noise label problem can be largely divided into Label noise-robust method, probabilistic label noise-tolerant method, model-based label noise-tolerant method, and data cleaning [1]. In the method category, data cleaning focuses on depolluting data themselves. In case of sound data, label noise may be more complicated to relabel manually. This is because that sound objects are difficult and ambiguous to be categorized only by hearing[9]. For instance, some of musical instruments have similar sound. So that, only hearing the sound is not enough to recognize what instrument is playing. Additionally, some kind of sound in circumstance such as fart and trumpet. Those cases are difficult to distinguish even small size of data. Large amount of data must be harder to label manually than small sized data.

Enormous size of sound data with inaccurate labels lead us to consider the way that the machine learns a feature regardless of the label in data. Unsupervised learning is to train the machine without labels. This method looks for specific patterns among the input data and derives a function that acts as a filter[10]. This kind of method can be used alone to solve a problems such as autoencoder[11]. The feature that the machine does not need label for data in train, is the point to be used in selecting mislabeled data that is similar to semi-supervised learning is used for the missing labels[12]. The self-organizing map is the one of unsupervised learning methods, has the similarity with k-means clustering. Due to the method projects input items in a grid, it is convenient to visualize the clustering result [6].

Since dataset we handle with is sound files, proper method to extract feature from audio clips was necessary. Various methods are applied and improved to achieve the better performance in object classification with variety of data. Especially, many different processes for sound data have been tried to extract features from the sound. Deep convolutional neural networks(CNNs) are the most widely used method. Specifically, the classification task with image data has been stood out[8]. CNNs are also efficacious to extract features from sound data because sound data can be treated similarly to image. CNNs can be divided in two part, one is feature extraction part, and the other one is classifier part. So that, the extraction part is used separately and made combination with any other classifiers such as SVM. The convolutional layers have been improved as the layers become deeper, to solve problems of losing information from original source of data. Among the derived technologies, densely connected networks (DenseNet) is employed.

In this study, we propose a method to handle noisy labeled sound data using SOM with features extracted through the DenseNet with log scaled Mel-spectrogram. The detail of techniques we employed and our works are illustrated through the section II and III. The performance of our work is showed in the section IV. Furthermore, performance depending on the size of SOM grid is experimented.

II. RELATED WORK

In order to find the uncertainty of data labels, it is necessary to adopt a method of learning features regardless of labels. SOM is employed as unsupervised learning to determine how different the self-configuration collection of data labels is from a given labels. The convolutional neural networks with log scaled Mel-spectrogram is used as the input feature with the reason described in section II.D. Furthermore, densely connected convolutional networks is employed depending on

978-1-6654-2678-7/22/$31.00 ©2022 IEEE

the prior experiment to select proper model for an environmental sound classification. According to the results of [3], DenseNet showed admirable performance compared to other convolutional neural networks models. However, since performance was the result of using image data, sound classification was tested using some models, including VGG16, Resnet18 and Densenet-121. The Densenet-121 shows the most impressive performance in the sound classification results, the Densenet-121 model is selected according to the results.

A. Dataset

Freesound Dataset Kaggle 2018(FSDK 2018) is a reduced subset of Freesound Dataset [2], which is an open audio dataset that is under development. The dataset that contains 41 classes of 11,073 audio clips and they are 44.1kHz mono sound with the duration from 30ms to longer than 30s. The training set consists of 9,473 clips, 3710 clips have manually verified labels, and 5,763 clip labels have not been verified. The test set contains 1600 clips, and all clips have verified labels.

B. DenseNet

As Convolutional Neural Networks (CNNs) become deeper, vanishing information of input data is the problem mainly discussed. Research recently published adopted the networks connected for bypassing information. In case of highway networks, the input is bypassed to the output according to the gating units [4]. In case of residual networks (ResNet), layers have the identity connection to the next layer for bypassing information [5]. The identity function makes gradients flow directly. Densely Connected Networks(DenseNet) consists of several dense blocks. The dense block is similar to the residual block from ResNet, but the connection for bypassing information is different. ResNet has the bypassing path from the earlier layer to the next one, but any layer of DenseNet has connection to all subsequent layers [3]. In this paper, the dense module of DenseNet is used for the feature extraction of sound data. The module is from pretrained densenet-121 model that the classification layer is removed.

C. Self Organizing Map

Self-organizing map (SOM) is the mapping that is projecting nonlinearly. SOM is similar to Vector Quantization(VQ) in perspective of using codebook vectors, but the models of codebook vectors are high-dimensional (usually 2D) grids[6], unlike conventional VQs using linear vectors. Simply say, input data selects the best matching model to input data, and the model is modified for better matching.

The original process of SOM is possible to be shown as following expression:

$$\boldsymbol{m_i}(t+1) = \boldsymbol{m_i}(t) + h_{ci}(t)[x(t) - m_i(t)]. \quad (1)$$

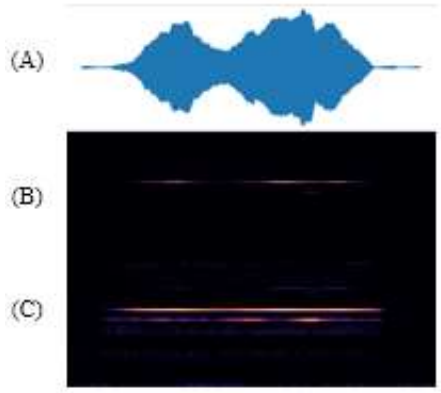

Fig.1. (A)Wave, (B)Mel-spectrogram, and (C)Log scaled Mel-spectrogram of "flute" sound

From the equation (1), $x(t)$ is the sequence of n-dimensional Euclidean vectors x, where t indicates the step in the sequence, and x also means the input data. $m_i(t)$ is the sequence of m_i which means the vector of the model computed, and i is the spatial index of model m_i's grid. The winner of the unit in the grid is selected through $h_{ci}(t)$ which is called neighborhood function, where c represents the index of the winner that is derived as follow:

$$\boldsymbol{c} = \boldsymbol{argmin}\{||\boldsymbol{x}(\boldsymbol{t}) - \boldsymbol{m\_i}(\boldsymbol{t})||\}. \quad (2)$$

Moreover, the neighborhood function for the modification rate is:

$$\boldsymbol{h_{ci}(t)} = \boldsymbol{\alpha(t)}\, \boldsymbol{exp}[-\boldsymbol{sqdist}(\boldsymbol{c},\boldsymbol{i})/\boldsymbol{2\delta^2(t)}], \quad (3)$$

Where α(t) is the learning rate that is decreasing, $sqdist(c,i)$ is the square of the distance between node c and i in the grid, and $\delta(t)$is another parameter that is reduced gradually along the steps of learning. The grid of SOM is trained through iterative sequence of fitting process that is mentioned above. At last, the winners, which is also called best matching units (BMUs), are projected in the grid where their values close to grid. The projection result of SOM is based to select data from our method.

D. Environmental sound classification

The environmental sound classification is to classify sounds that we can hear around generally [9]. Category of sound can be specific such as animals sound [13,14] or musical instruments [15], or can be broad [16]. The Log-scaled Mel-spectrogram with CNNs is well proven feature for the sound classification [17]. It leads to the improvement of feature extraction part of classification model using convolutional layer such as convolutional recurrent neural networks (CRNNs) [18], ResNet[19], and DenseNet[20]. The environmental sound classification result is used for the evaluation of our method.

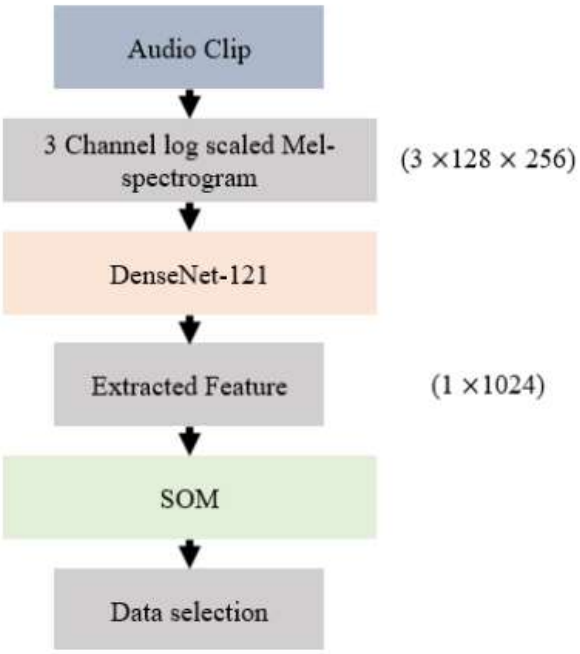

Fig. 2. The process flow

III. METHOD

The proposed method is briefly illustrated in Fig. 2. The whole process is divided in big two part, the one is the part of extraction of feature from audio clips, and the other one is the data selection using SOM with extracted feature from the prior process. The extraction part is to process input data fit into the SOM architecture. The log scaled Mel-spectrogram extracted from each sound clip has shape of 128 × 256. In the final step of feature extraction, the shape is transformed to suit the SOM in the shape of 1 × 1024 through DenseNet module. The data selection part is based on the cluster of best matching units(BMUs) in the grid that are derived through the SOM. The detail of each part is described following sections A and B. In addition to processing of data, evaluation of the proposed method is described in the section C.

A. Feature Extraction

First task of the feature extraction part is transforming raw wave data to the log scaled Mel-spectrogram. The wave form of the sound is extracted and transformed into the Mel scale spectrogram. The size of the frame is 256 and the size of the Mel bin is 128. In addition, 256 of the frame size is about 2.5 seconds at an audio sampling rate of 44100. Therefore, the sample with a length exceeding 2.5 seconds is cropped by allowing the largest energy to come from the center of the cut shape. Energy is calculated as the sum of Mel bins located in the same time frame of the sequence. If the audio is less than 2.5 seconds, zero is filled at the beginning and end of the feature.

The extracted features go through one more simple process before being inserted into the SOM model. This process is to convert the input shape from a single channel to three channels similar to image data. Since the Densenet model pre-trained with 3-channel image data is used, the input shape must be reconstructed accordingly. Therefore, the shape of input data is converted from 1 × 128 × 256 to 3 × 128 × 256. Thereafter, the compressed feature is extracted through the dense module of the Densenet-121 model from which the classifier portion is removed.

B. Data selection

The data selection part begins with training the grid of the SOM. Unlike general classification, dataset that trains the SOM grid and is projected onto it are the same. This is because the purpose of constructing the map is not to classify it, but to obtain data projected on the grid. The Data are projected onto each BMU of the grid, and the BMUs create clusters. The data projection results can be visualized as shown in Fig, 3, and Fig, 4 is a grid that data are projected only for one class with the same grid as Fig. 3. The selection is made according to clusters of data that are projected for each class. Data projected on the BMU below the threshold is selected.

Some classes show more than one cluster. To this end, it is assumed that there may be diversity in some classes. For instance, the dataset we used has a class called "telephone". If you look at the audio clips of the "telephone" class, some are classical ringtones and some are digital sounds. Therefore, data selection is performed in consideration of this. The data are selected according to the Euclidean distance from the item to another item. At first, a random item is selected from BMUs. Items within the threshold distance from the selected item are then collected into the queue and removed from BMUs. The following item is selected and removed from the queue, and adjacent items of the selected items are collected in the same way as the previous items. This sequence is then repeated until the queue is empty. At the end of the sequence, repeat the entire previous sequence until BMUs does not have items to select. If the item selected by the BMU does not have an adjacent item within the threshold distance, the item is deleted from the dataset.

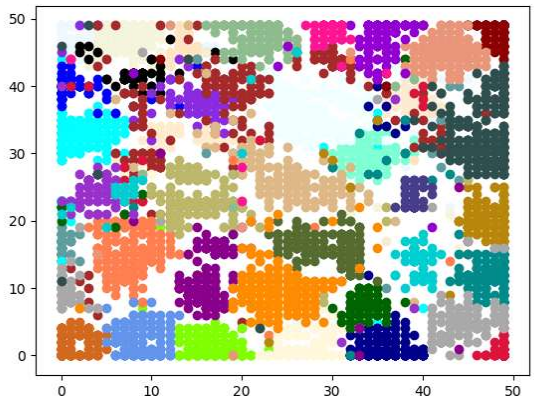

Fig. 3. Projection of 41 classes' BMUs in SOM grid of size 50×50

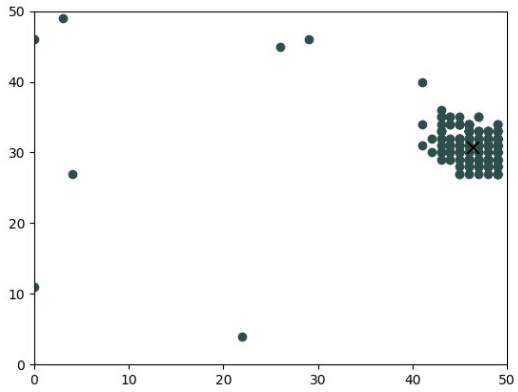

Fig. 4. Projected BMUs of "Fart" in the grid of SOM. X is the centroid of the cluster

TABLE I

SOUND CLASSIFICATION RESULTS OF THE TOP 5 CLASSES OUT OF 41 CLASSES

Label	***Original***	***Our Method***
Tearing	88.77%	92.47%
Computer keyboard	86.79%	90.64%
Scissors	71.20%	76.27%
Drawer	75.40%	79.77%
Squeak	47.70%	53.33%

C. Evaluation

This method is evaluated in acoustic classification using DenseNet-121, which has pre-trained parameters. Even after the 10 epochs, the classification results do not increase impressively, so they train until the 10th epoch. Also, the training set is not augmented. Because of the deviation of the test, the classification is conducted five times, and the average value of the results is used as the evaluation index. The average precision @ k (AP @ k) is used as an evaluation measure for each class. K is 3, which means that the AP calculation depends on the top three predictions of the classification.

IV. RESULT

At first, experiments are conducted with SOM grids of various sizes. The shape of the grid is m × n, where m = n and increases by 10 from 30 to 80. The initial α and δ values of the quicksom library[7]. The initial value of α is $m \times n /$ (number of input data) , and δ is $\sqrt{m \times n}/2$. Iteration number of SOM training is 100, and the batch size is 32. DenseNet-121 pre-trained model is used for the evaluation of results from proposed methods. In addition, the threshold distance to select neighbors of BMUs is 3. Tests for evaluation are performed five times each with the original data and the data which is selected through our method. The mean value of all AP@3 for each class(mAP@3) is used for the comparison by size of the grid. As a result, we obtained 89.19% with the minimum mAP@3 value and showed a maximum of 89.56%. This is the result of training with selected data through the 30×30 grid and the 50×50 grid of SOM, respectively, and the result obtained using the original data is 89.04%. The remaining results according to the size of the grid are shown in Figure 5.

The grid size for the maximum results and the largest grid in the experiment are selected to find an appropriate threshold distance for neighbor selection. The threshold distance for selecting proximity items increases one by one from 2, and the remaining parameters are the same as in the previous experiment. The reason the threshold increment starts at 2 is that some BMUs are not close enough to each other. Some of these are within a distance of 2, not 1. More specifically, it's within $\sqrt{2}$. Starting with a 50×50 shape grid with threshold 2, the test ends with an 80×80 shape grid with threshold 8. Each test is performed five times as in the initial test, and the average mAP@3 of each five tests is derived. The results are shown in Fig 6, and Judging from it, the results of the 50×50 grid size show the most appropriate results at the threshold of 3. However, the test was stopped because two attempts after thresholds 4 and 5 showed a decrease. In the case of 80×80, the threshold value 7 shows a result of 89.64%, indicating that it is suitable among the attempts.

V. CONCLUSION

In this study, a method of cleaning sound data containing label noise using SOM was proposed. The Freesound dataset Kaggle 2018 dataset is used for experimentation and evaluation of the method. Since it was necessary to transform features to fit the SOM model, DenseNet modules were used to extract compact functions. Prior to the process of extracting compressed features through the DenseNet module, the raw wave form of the audio clip is converted into a log-scale Mel-spectrogram with a size of $3 \times 128 \times 128$. SOM constructed a map to project the data via unsupervised learning, and selected the data according to the results. The selection method follows several rules for selecting neighbors for each BMU within a threshold distance. The initial experiment was performed with a threshold of 3 and various SOM grid sizes. According to the results of this experiment, the 50×50 shaped grid was 89.56% at mAP@3, and the results of training with the original data were 89.04%.

The following experiment was conducted to find a threshold suitable for data selection. In this experiment, grids of sizes 50×50 and 80×80 were selected. As a result, we obtained the graph of Fig. 6. Given the overall mAP@3 result alone, it seems that we have not derived such a meaningful result. However, we confirmed that our method was not meaningless when comparing the original dataset with class-specific results. We confirmed that some classes have increased in our way. However, due to the data cleansed, the results of the classes that were not removed are also reduced, or some classes that should not have been processed are excessively cleaned up, reducing the overall mAP@3. Complementing this will lead to higher results. Furethermore, we can study how Mahalanobis distance affects outcomes instead of Euclidean distance for selection. Also, according to

Fig. 3, we can consider re-labeling rather than removing data using this.

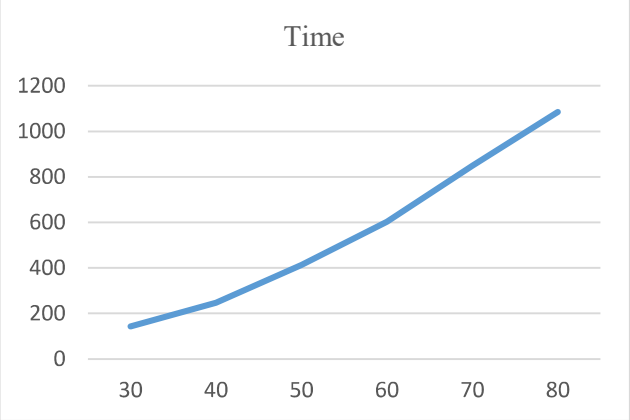

Fig. 5. Time spent depending on the grid size (seconds)

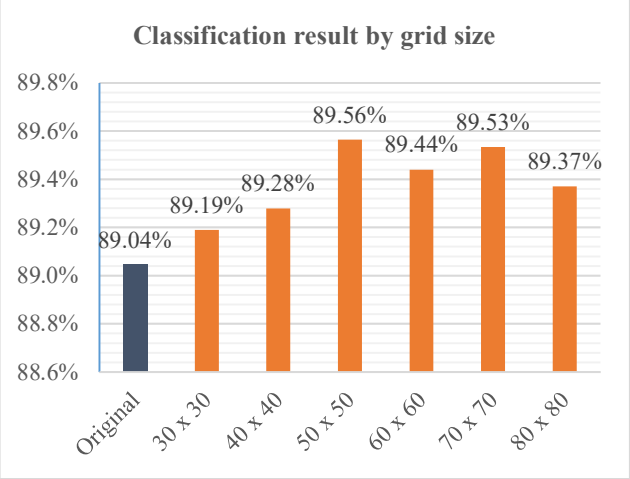

Fig. 6. Classification result by grid size in mAP@3

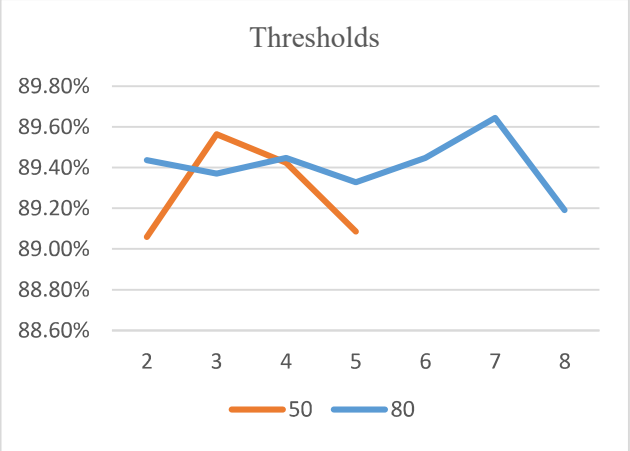

Fig. 7. Classification result according to the threshold for each grid size

REFERENCES

[1] Frénay, B., & Verleysen, M. (2013). Classification in the presence of label noise: a survey. *IEEE transactions on neural networks and learning systems*, *25*(5), 845-869.

[2] Fonseca, E., Plakal, M., Font, F., Ellis, D. P., Favory, X., Pons, J., & Serra, X. (2018). General-purpose tagging of freesound audio with audioset labels: Task description, dataset, and baseline. *arXiv preprint arXiv:1807.09902*.

[3] Huang, G., Liu, Z., Van Der Maaten, L., & Weinberger, K. Q. (2017). Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 4700-4708)..

[4] Srivastava, R. K., Greff, K., & Schmidhuber, J. (2015). Highway networks. *arXiv preprint arXiv:1505.00387*.

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

[6] Kohonen, T. (2013). Essentials of the self-organizing map. *Neural networks*, *37*, 52-65.

[7] Mallet, V., Nilges, M., & Bouvier, G. (2020). Quicksom: Self-Organizing maps on GPUs for clustering of molecular dynamics trajectories. *Bioinformatics*.

[8] Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. Advances in neural information processing systems, 25, 1097-1105.

[9] Houix, O., Lemaitre, G., Misdariis, N., Susini, P., & Urdapilleta, I. (2012). A lexical analysis of environmental sound categories. Journal of Experimental Psychology: Applied, 18(1), 52.

[10] Ghahramani, Z. (2003, February). Unsupervised learning. In Summer School on Machine Learning (pp. 72-112). Springer, Berlin, Heidelberg.

[11] Le, Q. V. (2013, May). Building high-level features using large scale unsupervised learning. In 2013 IEEE international conference on acoustics, speech and signal processing (pp. 8595-8598). IEEE.

[12] Zhou, X., & Belkin, M. (2014). Semi-supervised learning. In Academic Press Library in Signal Processing (Vol. 1, pp. 1239-1269). Elsevier.

[13] Kahl, S., Wilhelm-Stein, T., Hussein, H., Klinck, H., Kowerko, D., Ritter, M., & Eibl, M. (2017, September). Large-Scale Bird Sound Classification using Convolutional Neural Networks. In CLEF (Working Notes).

[14] Huang, C. J., Yang, Y. J., Yang, D. X., & Chen, Y. J. (2009). Frog classification using machine learning techniques. Expert Systems with Applications, 36(2), 3737-3743.

[15] Joder, C., Essid, S., & Richard, G. (2009). Temporal integration for audio classification with application to musical instrument classification. IEEE Transactions on Audio, Speech, and Language Processing, 17(1), 174-186.

[16] Piczak, K. J. (2015, September). Environmental sound classification with convolutional neural networks. In 2015 IEEE 25th International Workshop on Machine Learning for Signal Processing (MLSP) (pp. 1-6). IEEE.

[17] Salamon, J., & Bello, J. P. (2017). Deep convolutional neural networks and data augmentation for environmental sound classification. IEEE Signal Processing Letters, 24(3), 279-283.

[18] Sang, J., Park, S., & Lee, J. (2018, September). Convolutional recurrent neural networks for urban sound classification using raw waveforms. In 2018 26th European Signal Processing Conference (EUSIPCO) (pp. 2444-2448). IEEE.

[19] Naranjo-Alcazar, J., Perez-Castanos, S., Martin-Morato, I., Zuccarello, P., & Cobos, M. (2019). On the performance of residual block design alternatives in convolutional neural networks for end-to-end audio classification. arXiv preprint arXiv:1906.10891.

[20] Bian, W., Wang, J., Zhuang, B., Yang, J., Wang, S., & Xiao, J. (2019, August). Audio-based music classification with DenseNet and data augmentation. In Pacific Rim International Conference on Artificial Intelligence (pp. 56-65). Springer, Cham.