Environmental Sound classification

Tong Lin
tol009@ucsd.edu

Yuchen Zhang
yuz095@ucsd.edu

Shiqi Liu
shl352@ucsd.edu

***Abstract*—This project is about environmental sound classification which is essential to research and real-life applications. It is based on dataset for environment sound recognition (ESC-50), we used Mel Frequency Cepstral Coefficient as the feature, used models including MLP, CNN, random forest, SVM and KNN for classification and did detailed analysis.**

I. Introduction

With voice assistants gaining more and more popularity, there are many tasks around speech recognition based on machine learning or deep learning. While another critical issue apart form that in the sudio domain is environment sound. Human beings rely on both visual and auditory information to recognize surrounding conditions and events happening in the environment, and those are just same with intelligent agents or automated robots. Environment sound classification is essential to applications for disabled people, automated analysis of urban sounds as well as the newly developed technologies, including autonomous vehicles. Environment sound classification could be more challenging since those sounds are less structed than music or speech and more like noise thus hard to detected and classified.

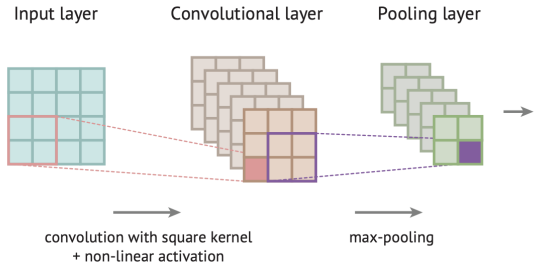
Thanks to the great datasets like ESC-50 which includes environment sound classes in everyday life (eg. “baby crying”, “street music”, “children playing”, “snoring”, “siren”), it is easier for us now to deal with this comparably young field. Classification of environmental sounds ,which excluding speech or music, sounds in everyday life that more diverse and chaotic-structured is mainly based on manually extracted features such as mel-frequency cepstral coefficients and with comparably basic classifiers including support vector machines. We are managing to solve and improve this problems with more models.

II. Related Work

Though environmental sound clasiification is a comparably young field, there are certain progress made after the dataset as ESC-50 and UrbanSound8K (US8K).

In Environmental sound classification with convolutional neural networks by Karol J.Piczak [1], a deep model concisting of 2 convolutional layers with max-pooling and 2 fully connected layers (shown as Fig. 1) was trained on a low level representation of audio data with deltas. Three datasets: ESC-10, ESC-50 and UrbanSound8K were used for evaluation of the models. The model was evaluated in 5-fold (ESC-10 and ESC-50) and 10-fold (UrbanSound8K) cross-validation regime, and with a single training fold as an intermittent validation set. Their experiments indicated that convolutional neural network seems to concentrate on distinct aspects of sound events, thus less complex models may work efficiently as well in the future.

Fig. 1. A typical convolution-pooling operation for the input data



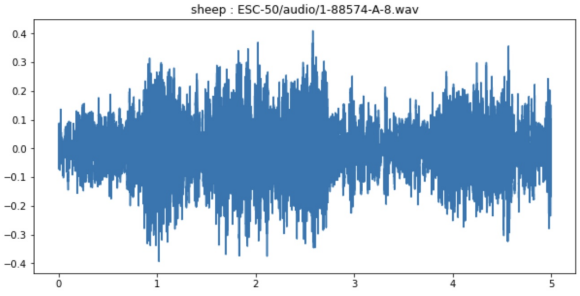
In Environment sound classification using multiple fearture channels and attention based deep convolutional neural network [2], they proposed a model consists ofmultiple feature channels given as input to a DCNN with attention mechanism. Multiple feature channels including Mel-Frequency Cepstral Coefficient (MFCC), Gammatone Frequency Cepstral Coefficients (GFCC), Constant Q-transform (CQT) and Chromagram for audio processing. Also, their DCNN model consists of spatially separable convolutions separately working for time and feature domains. Attention modulea that performs both for channel and spatial attention was utilized. Evaluated with datasets (ESC-10, ESC-50 and UrbanSound8K), their model achieved high accuracy and thus worth referring.

III. Dataset and Features

The dataset we used is called ESC-50, which can be obtained in Kaggle and github.[3] ESC-50 is a labeled collection of 2000 environmental audio recordings. All audio clips in the dataset have been extracted from public field recordings. 2000 audios are organized in 50 classes, each class has 40 clips. These 50 classes can be further loosely arranged into 5 major categories, which are animals, water sounds, human sounds, interior sounds and urban sounds. Every audio clip in the dataset is 5 seconds long, sample rate is 22050 Hz.

From the figure of category classification, it can be seen that although 50 categories can be arranged into 5 major categories, all classes in one major category don’t share lots of similarity and have lots of difference. For example, sound from dog and cat is very different but they are arranged into animal category. Less similarity in one class may cause bad result and increase difficulty when training model. Considering this situation, we

Fig. 2. one of 5 seconds long audio clip in ESC-50 dataset



decided to classify 50 categories instead of classifying 5 major categories in our project. From the given dataset, 50 categories cover a wide range of sound in reality.

Fig. 3. 50 categories and 5 major categories of ESC-50 dataset [4]

Animals	Natural soundscapes & water sounds	Human, non-speech sounds	Interior/domestic sounds	Exterior/urban noises
Dog	Rain	Crying baby	Door knock	Helicopter
Rooster	Sea waves	Sneezing	Mouse click	Chainsaw
Pig	Crackling fire	Clapping	Keyboard typing	Siren
Cow	Crickets	Breathing	Door, wood creaks	Car horn
Frog	Chirping birds	Coughing	Can opening	Engine
Cat	Water drops	Footsteps	Washing machine	Train
Hen	Wind	Laughing	Vacuum cleaner	Church bells
Insects (flying)	Pouring water	Brushing teeth	Clock alarm	Airplane
Sheep	Toilet flush	Snoring	Clock tick	Fireworks
Crow	Thunderstorm	Drinking, sipping	Glass breaking	Hand saw

In our project, we use all samples to train models. 2000 examples are divided into training set (80 percent of samples) and test set (20 percent of samples). The original data we get is a folder of all audio clips and a csv file which shows all audio files name and their relative categories. In order to label audio clips, in preprocessing part we build a dictionary, the key of the dictionary is filename, the value is audio category. In this way, when we load audio, we can easily get the label by searching in the dictionary. In order to train model easily, we also build another dictionary to link 50 category names and number 0 to 49, which can be used as data label.

In order to have a better performance when do the classification, original signal provides limit information, thus we do the feature extraction from the raw signal, grab more information without noise to make inferences. In machine learning problems, we want to extract features to be independent of others because it is easier to develop models and to train models with independent features. One popular audio feature extraction method is the Mel-Frequency cepstral coefficients (MFCC).[5]

In audio analysis, generating audio shape can tell where sounds come out, which can accurately produce the phoneme. MFCC can represent envelope of short time power spectrum of audio, which shows audio shape. The step to calculate MFCC is:

1. Frame signal into 20 40ms frames

2. Calculate the power spectrum of each frame (Mel spectrogram)

3. Take the logarithm of step 2 result (log amplitude of mel-spectrogram)

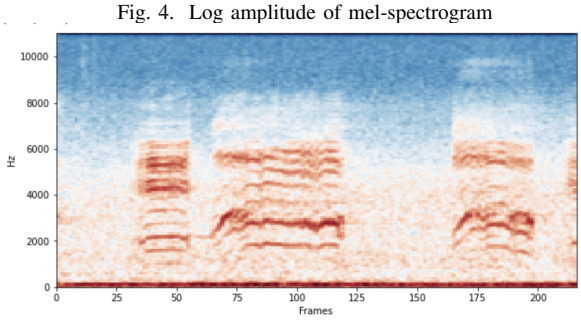
4. Compute the DCT of step 3 result

All steps of MFCC calculation can be done by package named librosa.

MFCC have 39 coefficients. In our project, we only take first 13 because first 13 coefficients have contained most of the information about the overall spectral shape. Even though higher order coefficients represent increasing levels of spectral details, selecting a large number of coefficients results in more complexity in the models and need more data. In our project, considering a 2000 sample audio dataset is not very large, which can't provide lots of data, thus we just capture first 13 coefficients of MFCC.

Mel Spectrogram is another feature we used. Mel Scale is the result of some non-linear transformation of the frequency scale. Mel spectrogram is a spectrogram with the Mel Scale as its y axis. We can acquire mel spectrogram by storing the result of step 2 in MFCC calculation.

Fig. 4. Log amplitude of mel-spectrogram



We also try mel-frequency cepstrum(MFC) as feature in our model training because MFCC is log scale of MFC. When we use some strong classifiers like Convolution Neural Network, it is often good to get the neural network to learn complex representations and not impose them. MFC can often performs better then MFCC in this situation.

After we get MFC and MFCC, we calculate mean and standard deviation of 13 coefficients in all frames as feature of data and start training.

IV. Methods

We applied various deep learning and machine learning models to train the environmental sound dataset. For each of the following models, we split the data into 80% training set and 20% test set.

A. Multilayer Perceptron (MLP)

Multilayer perceptron, or MLP, is a classical type of the feedforward neural network. A MLP model is composed of at least 3 fully connected layers of neurons: an input layer, an output layer and one or more hidden layers [6]. The training process involves the backpropagation algorithm which typically works by using gradient descent to adjust the weights of each layer and minimize the loss function. With enough

training data, MLP shows great advantages in speech and image recognition.

For this environmental sound classification problem, we constructed the MLP model with 3 hidden layers. Each layer contains 512 perceptrons and is followed by RELU activation function and 30% dropout. The output layer has 50 perceptrons corresponding to 50 environmental sound classes and then activated by softmax activation function.

B. Convolutional Neural Network (CNN)

Compared to the architecture of MLP, CNN models do not have fully connected layers in all of its hidden layers. Instead, hidden layers in CNN consist of combinations of convolutional layers, max-pooling layers, normalization layers and fully connected layers, as shown in Figure 5. In each hidden layer, only a small portion of neurons is connected to the next layer so that the dimensionality of the data can be reduced, and networks can have deeper structures with fewer parameters [8] [9].

Our CNN model has 3 convolutional layers, each with 64 feature maps and 3*3 filter banks followed by RELU activation function and then passed through a max pooling layer with size 2*2. Between each convolutional and maxpooling layer, we apply 10% dropout to reduce overfitting. Then we flatten the generated features and connect to a fully connected layers with 128 perceptrons with RELU activation function. Lastly connect to the output layer with 50 perceptrons and activated by the softmax activation function.

C. Random Forest

Random forest builds the ensemble of decision trees from sub-samples randomly selected from the training data, and then collect and average votes from these decision trees to make the classification decision [11].

After several attempts of hyper-parameter tuning, we find the random forest model with 1000 trees and minimum samples split of 2 provides the best result.

D. Support Vector Machine

The main objective of support vector machine is to find a hyper-plane in a N-dimensional feature space that can best separate the different categories [12]. This is achieved by finding the maximum distance of data points in different classes.

Here we use the SVC function under sklearn.svm package with regularization parameter $C = 0.02$ and the linear kernel.

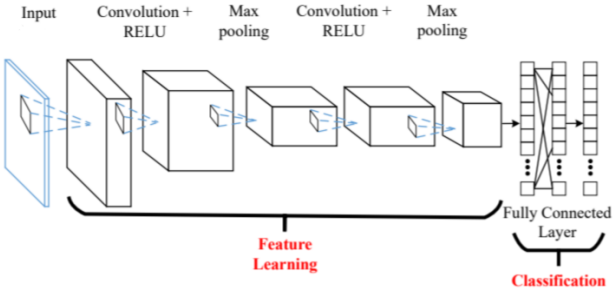
E. K-Nearest Neighbors

In K-Nearest Neighbors algorithm, each data point is assigned to the class with highest frequency among its K nearest neighbors using a distance function. We pick $K = 5$ in this case as it achieves the best result.

V. Results and Discussion

We use accuracy as the primary evaluation metric for both deep learning and machine learning models, which is the number of correctly predicted test samples divided by the total number of test samples.

Fig. 5. Example of a CNN model [10]



A. Results

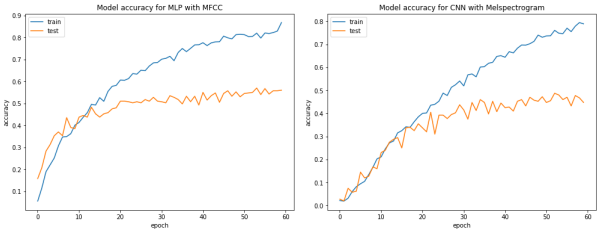
1) Deep learning models: For both MLP and CNN models, we train the data set for 60 epochs with batch size 32. We combine the two models with two features and listed the train test accuracy in Table I below.

Out of the 3 deep learning models, MLP model with MFCC features provide better accuracy of 56%. On the other hand, Melspectrogram features achieve better test accuracy than MFCC features on CNN despite Melspectrogram have lower training accuracy, implying that CNN model using Melspectrogram as features is less overfitted than that using MFCC as features.

TABLE I
Train Test accuracy of DL models

Model	Feature	Train Accuracy	Test Accuracy
MLP	MFCC	86.81%	56.00%
CNN	Melspectrogram	78.87%	44.75%
CNN	MFCC	87.12%	42.75%

Fig. 6. Accuracy of MLP with MFCC and CNN with Melspectrogram



2) Machine learning models: From Table II, we can see that random forests achieve much higher accuracy than SVM and KNN. Especially when using MFCC and Melspectrogram combined features, the random forests model reaches the accuracy of about 62%, which is the highest among all models.

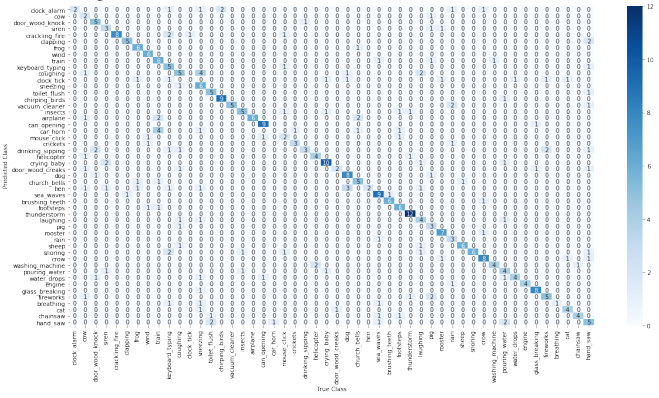
B. Discussion

In terms of models, one possible reason CNN models do not perform well is that CNNs depend on large training data while we only have 2000 audio clips in total that are categorized into 50 classes. This could also be the main cause of overfitting in MLP and CNN models as we already tried various overfitting

TABLE II
TEST ACCURACY OF ML MODELS

Model	Feature	Test Accuracy
RF	MFCC+Melspectrogram	62.50%
RF	Melspectrogram	58.00%
RF	MFCC	55.25%
SVM	MFCC	42.00%
KNN	MFCC	38.00%

Fig. 7. Confusion matrix for RF with combined features



reduction approaches such as dropout layer, L2 regularization, batch normalization, early stopping criterion, but none effectively reduces overfitting. We also tried to categorize the audios into 5 broad categories, but the result became even worse. We think this is because many audios under a same broad category are very distinct, for example, the sounds of clock alarm and mouse clip are both categorized as interior sounds, and thus increasing the difficulty of classification.

As for the two features, Melspectrogram generally achieves better accuracy on test dataset. This might because Melspectrogram contains more information than MFCC as the computation of MFCCs involves taking logarithm and computing the Discrete Cosine Transform, and we only use the first 13 coefficients as features, thus some other features could get lost. Achieving the best accuracy with combined features is expected as the combined features contains more characteristics and provide a well-rounded representation of the original audio data.

From the confusion matrix for the best model as shown in Figure 7, we can see that some classes such as thunderstorms, sea waves, and crying baby are easier to classify, while other classes such as car horn and hen are more difficult to classify since they are easily confused with other classes like train and dog respectively.

VI. CONCLUSIONS AND FUTURE WORK

A. Conclusion

For this project, we experiment several deep learning and machine learning models with 2 different features and their combined features to classify 2000 environmental sound clips into 50 classes.We observe that random forests with combined features of MFCC and Melspectrogram produce the best test accuracy of 62.5%. Among the deep learning models, MLP with MFCC features achieves better accuracy of 56%.

B. Future Work

Due to the limitation of time, there are still many other models and feature extraction techniques worth trying. As CNNs have limited performance in dealing with time series data, we would like to implement RNN models such as LSTM to see if RNNs will have better results. Further, in order to solve the overfitting problem, we need to get a larger training dataset, possibly by combining the current ESC-50 with the UrbanSound8K dataset.

VII. CONTRIBUTIONS

Tong Lin: Worked for feature extractions of MFCC, models of SVM, Random forest, KNN and plot of confusion matrix.

Yuchen Zhang:Background research, worked for data preprocessing, feature extraction and paper writing.

Shiqi Liu: Extracted Melspectrogram and combined features, implemented the random forests, MLP, CNN model, tuned parameters to try to reduce overfitting.
All of us spent time on this project and wrote report.

REFERENCES

[1] Piczak K J . Environmental sound classification with convolutional neural networks[C]// 2015 IEEE 25th International Workshop on Machine Learning for Signal Processing (MLSP). IEEE, 2015.

[2] Jivitesh Sharma Environment Sound Classification using Multiple Feature Channels and Deep Convolutional Neural Networks // 2019 ResearchGate.

[3] Piczak, Karol. (2015). ESC: Dataset for Environmental Sound Classification. 1015-1018. 10.1145/2733373.2806390.

[4] GitHub. 2020. Karolpiczak/ESC-50. [online] Available at: ¡https://github.com/karolpiczak/ESC-50¿ [Accessed 14 June 2020].

[5] Seongkyu Mun, et al. "Deep Neural Network Bottleneck Features for Acoustic Event Recognition".*INTERSPEECH* 2016

[6] Wikipedia, "Multilayer perceptron." https://en.wikipedia.org/wiki/Multilayer_perceptron, 2020. (Accessed: 13 Jun, 2020).

[7] D. Svozil, V. Kvasnicka and J. Pospichal, "Introduction to multi-layer feed-forward neural networks", *Chemometrics and Intelligent Laboratory Systems*, vol.39, no.1, pp.43-62, 1997.

[8] V. Boddapati, "Classifying Environmental sounds with image networks", M.S. Thesis, Sch. of Comp. Science, Blekinge Institute of Technology, Sweden, 2017.

[9] Wikipedia, "Convolutional neural network." https://en.wikipedia.org/wiki/Convolutional_neural_network, 2020. (Accessed: 13 Jun, 2020).

[10] P. Kamencay, M. Benco, T. Mizdos and R. Radil, "A New Method for Face Recognition Using Convolutional Neural Network", *Advances in Electrical and Electronic Engineering*, vol. 15, no. 4, 2017.

[11] N. Dey, P. Mahalle, P. Shafi, V. Kimabahune and A. Hassanien, *Internet of Things, smart computing and technology.* 2020.

[12] E. Ricci, S. Rota Bulo, C. Snoek, O. Lanz, S. Messelodi and N. Sebe, *Image Analysis and Processing - ICIAP 2019.* Cham: Springer International Publishing AG, 2019.

VIII. Response to Reviews

Critique by group 12

This talk was talking about how to use deep learning algorithm to classify environment sounds into 50 categories. They explained the details of their design and the decision reasons behind them clearly from feature extraction to the model design. They have distributed their presentation evenly so that it is easy to see that all of them have contributed in this project. However, the code demo video has quite low resolution, maybe you have compressed the video too much. Nevertheless, the code demo has full of images to show the current progress of the model, which makes the demo vividly. They have tried a lot of the combinations of the models, which are interesting. However, if they could make more analysis on the models about their features and what made the prediction good would be helpful for the audiences.
Our Response: Thanks!
Some improvements/unclear:
- Could explain more about the possible reasons why random forest and MFCC would have a better result.
Our Response: We revised our model and feature extraction process, and now the random forest with combined features has the best accuracy. We explained in our paper that the reason CNN cannot out-perform random forest is because of not large enough dataset. We tried several ways to reduce overfitting but none works effectively.
- If the data for 50 categories is too small, why don't you classify them just into 5 categories instead.
Our Response: take 5 major categories as big group, when we look all categories closer, it can be seen that all categories in each big group didn't share lots of similarity and categories between each big group may have little difference. for example, sounds from cow and sheep are not very similar, while they all belongs to animals. Considering this situation, we think the result will be bad. We have tried this method and result shows that the accuracy is quite low.
- Should have higher resolution for demo video
Our Response: The resolution would be better if you download our video rather than play online.
- Beside the mel-frequency transformation, is there any other audio pre-processing that worthes trying?
Our Response: Yes there are, like GFCC and CQT, while MFCC is what we chose for this project.
- The ending is missing.
Our Response: Same as above, this could be fixed if you download our video
- If there is a goal about what kind of experiments you are trying would be a plus.
Our Response: thanks for the suggestion
- How much percentage of that 13 coefficients cover the whole coefficients?
Our Response: We compute the first 20 coefficients and only take the first 13 ones.
- Could mention about the optimizer configuration since it sometimes helps training the model
Our Response: Optimiaer used for MLP and CNN are both adam, which has good performance generally.

Critique by group 13

In this final project, group 40 chooses sound classification as their topic which aims to classify environmental sounds with proper machine learning algorithms. They use different features and different models on the ESC (Environmental Sound Classification) dataset. Detailed explanations about the models (e.g., MLP and CNN) and the results comparison are given. It's motivated that they combine some features together which our group should try later. And the code demos are very vivid for those corresponding pictures shown.
Our Response: Thanks!
Some improvements/unclear:
- Maybe they could discuss the reasons why they should use deep learning methods or traditional methods firstly. I believe discussing that first will help a lot in the feature extraction, model choosing and so on.
Our Response: We originally tried to use traditional methods like random forest, svm, knn as baseline methods, and then use DL methods to try to improve the accuracy. But lastly it turns out RF with combined feature has highest accuracy of 62%.
- In their results, it seems accuracy on the test set is not that good. But on the contrary, accuracy on the training set is much higher. First, there should be some model adjusting and so on to avoid the overfitting problem. Second, maybe they could to find more details about their dataset. Because for our group, we get bad accuracy results on the test set first on our project and adjusting the models and features won't help at all. But after we look deeper into the dataset "Urbansound8K". It says it must use 10-fold cross validation and we can't shuffle the data. Then we finally fix the problem.
Our Response:
- In their results, some results of traditional methods are not worse than the deep learning methods. Maybe they could discuss that more or improve the accuracy results later and then explain the results again.
Our Response: Yes we discussed in paper why CNN cannot out-perform the result of random forests, which is largely due to the size of dataset and consequently the problem of overfitting.

Critiques by group 24.

The presentation of group 40 explained the dataset (ESC) and the corresponding features (MFCC) they used to do classification. They give a detailed analysis of their two models including MLP and CNN. They also used random forest, SVM and KNN to do classification and give detailed results and analysis with some clear plots.
- I think they give the presentation in a rational and logical way, including the introduction of dataset, which features are used, and how to implement the model, finally the comparison of the results.
- I think they give a detailed introduction of dataset and

feature extraction, including how to categorize dataset into 5 categories and why choosing 13 coefficients of MFCC as features. Some improvements:

- The best accuracy is not high enough, which is 52 among all model's accuracy. I think it has a large space to improve.

Our Response: Revised random forest with combined features have accuracy 62.5%, and we think it is satisfactory for a classification problem with 50 categories.

- And among all models, there exist overfitting. And I believe some methods could be used to reduce overfitting to make the results better.

Our Response: We already tried L2 regularization, dropout layer, batch normalization, early stopping conditions, but sometimes the results get worse and the problem of overfitting still exists.

- In the results slide, could you explain why the test accuracy of RF models decrease when also include Mel as features?

Our Response: In revised version, combined features have highest accuracy.

- From my personal point, I think the title of the third slide is improper, because in this slide you guys mainly talk about what you did rather than the background.

Our Response: That was a brief intro of the project, we did that based on the template