ACOUSTIC SCENE CLASSIFICATION BY FEED FORWARD NEURAL NETWORK WITH CLASS DEPENDENT ATTENTION MECHANISM

Jia-Ming Liu, Hui-Hui Wang, Mingyu You, Ruiwei Zhao

Department of Control Science and Engineering,
Tongji University,
Shanghai, 201804, China
james.liu.n1@gmail.com

Guo-Zheng Li

Data Center of Traditional Chinese Medicine,
China Academy of Chinese Medical Science,
Beijing, 100700, China

ABSTRACT

In the acoustic scene classification task, we proposed a novel attention mechanism embedded to feed forward networks. On top of a shared input layer, 15 separated attention modules are calculated for each class, and output 15 class dependent feature vectors. Then the feature vectors are mapped to class labels by 15 subnetworks. A softmax layer is employed on the very top of the network. In our experiments, the default feature, MFCC and mel filterbank with delta and acceleration, is used to represent each segment. We split each 30s audio recording into 1s segments and calculate label for the segment, then output the most frequent label for the 30s recording. The best single neural network could get 77.4% cross validation accuracy without further feature engineering and any data augmentation. We train 5 models with MFCC features and 5 models with mel filterbank features, then make an ensemble with majority vote, getting a 78.6% final cross validation result. For submission, the 10 models are retrained with full dataset. And, the final submission is a majority vote ensemble of the 10 models' outputs.

Index Terms— acoustic scene classification, deep learning, attention mechanism, feed forward network

1. INTRODUCTION

Acoustic scene classification task is to classify audio recordings into predefined classes[1]. We tried to solve this problem by deep neural networks. A novel attention mechanism structure is proposed for this task. The rest of this report is organised as following: Section 2 described and explained the structure of the proposed network. Section 3 gave the details of ensemble. And Section 4 shows the results from cross validatoin. In section 5, we draw conclusions.

2. FEED FORWARD NETWORKS WITH CLASS DEPENDENT ATTENTION MECHANISM

In the dataset, each recording clip lasts 30 seconds. To formulate the problem, we define $X = x_1, x_2, ..., x_T$ to be the acoustic feature, and Y is the label associated to the recording. In the task, not all of the frames are crucial for classification. For example, there are non-informative silence parts in almost all the classes. It's not very reasonable to segment recordings into small pieces and classify the pieces to classes directly. The classifier will try to classify every piece, informative or not, to their associated label. Those non-informative pieces would potentially damage the classifier.

Inspired by[2], we proposed a novel feed forward structure with class dependent attention mechanism. The feed forward mechanism

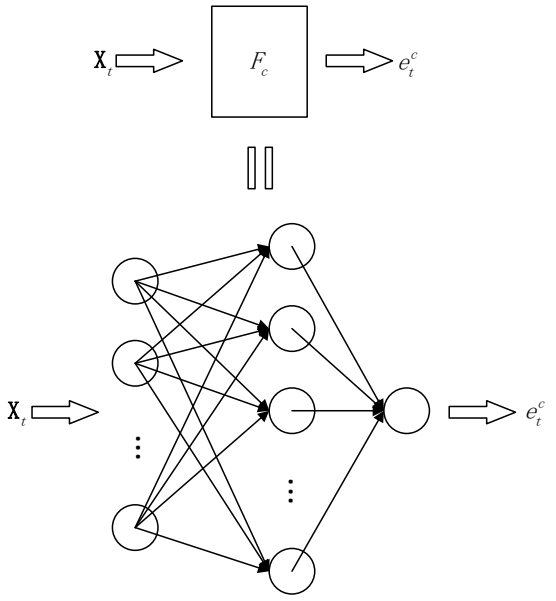

Figure 1: Attention Layer

is a weighted average of the feature space, where the weights are also learned from the feature space. A attention layer is a small feed forward network that maps feature vector to its "importance", as shown in Fig. 1. The reason we name our attention layer "class dependent attention" is because each class has a independent attention layer associated with them. This idea is motivated by the fact that different acoustic scenes have different characteristic events. For example, when a classifier is used to distinguish whether a scene is "restaurant", it will make a decision mainly depending on if the scene contains the impact sound between dishes and knives/forks. When to recognise a home scene, the classifier should focus more on water tap sounds. By introducing this mechanism, we hope the events could give themselves reasonable weights for different classes and form different "context" vector V_c for class c. The normalized importance α_t^c of x_t is calculated by Equ. 1 and Equ. 2

$$e_t^c = F_c(x_t) \tag{1}$$

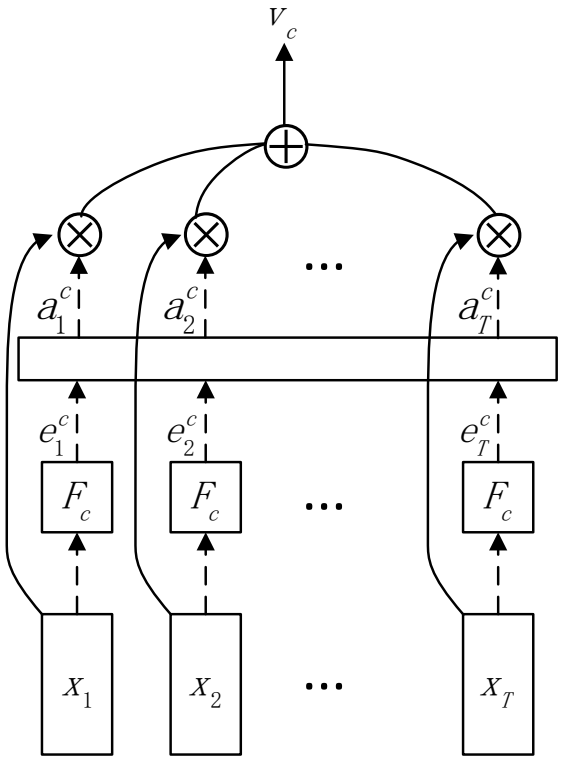

Figure 2: Generating context vector for class c

$$\alpha_t^c = exp(e_t^c) / \sum_{\tau=1}^{T} exp(e_\tau^c) \tag{2}$$

where F_c is attention layer for class c, e_t^c is the importance of x_t when classifying class c. α_t^c is the normalized importance of e_t^c where we followed softmax formula to amplify the difference between e. With α_t^c calculated, the context vector V_c can be easily derived by Equ. 3. The calculation process could be seen in Fig. 2

$$V_c = \sum_{t=1}^{T} \alpha_t^c * x_t \tag{3}$$

Since the context vectors are separated between classes, we built separated subnetworks S_c to further map V_c to it's corresponding class c. On the top of each S_c, we use a softmax layer to unify all the predicted classes (Fig. 3).

2.1. Ensemble

It is usual to boost classification performance by making an ensemble among multiple classifiers. To achieve a good ensemble, the classifiers should be both accurate and diverse. We trained 5 classifiers with default MFCC feature and 5 classifiers with mel filterbank feature. The predictions within MFCC and mel filterbanks are consistent, but they are diverse between features. So, the ensemble was made from a diverse and accurate set of classifiers in the end.

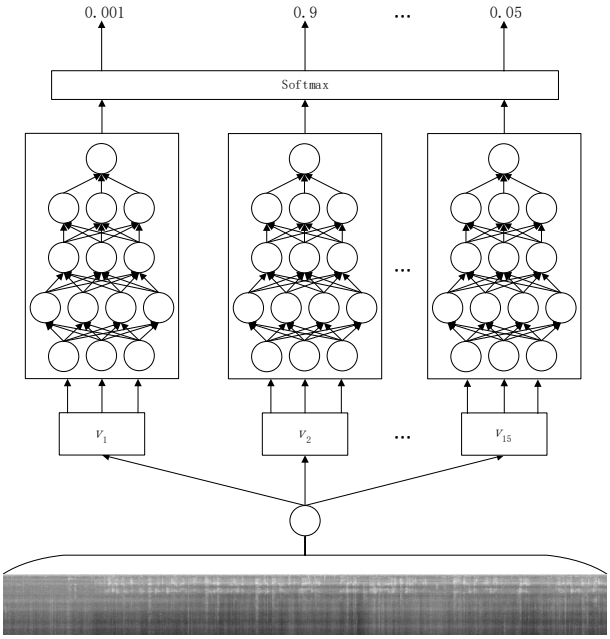

Figure 3: Classify a variate length recording

3. EXPERIMENTS

We trained networks with MFCC features and mel filterbank features with a context window of 15 frames. Each 30s audio is segmented into 100 frames. And each 100 frames sample was associated with a predicted label. The subnetworks S_c are all feed forward networks with 5 layers and 128 nodes for mel filterbank feature, and 6 layers with 128 nodes for MFCC feature. The activation functions is ReLU, and batch normalization is applied before each layer. The models are implemented by Lasagne [3].

4. RESULTS AND DISCUSSION

For a single network, the cross validation accuracy was 75.5% in average, best accuracy reached 77.4% (however, accuracy over 77% appeared only once). The ultimate ensemble model obtained 78.6% accuracy (Fig. 4).

We used only the default MFCC and melfilterbank features in our experiments. The model should be more powerful if we could find a better feature representation. Moreover, the models got overfitting very quickly. The training cost (categorical cross entropy) went down to 1e-5 in less than 10 epochs, validation errors for each 100 frames stayed in around 30%. This may be caused by shortage of data. We believe the models could show their potential if more data are provided.

5. CONCLUSION

Attention mechanism is used in many research areas. Models get capability to "focus" on the most related parts in input data. In this report, we proposed a novel attention mechanism structure based on low-complexity feed forward networks. In the case of acoustic scene classification, the new model could let the network focus on

File-wise evaluation, over 4 folds

Scene label	Nref : Nsys	Accuracy	Fold1	Fold2	Fold3	Fold4
beach	78 : 73	84.5 %	100.0 %	90.5 %	89.5 %	57.9 %
bus	78 : 62	68.3 %	73.7 %	35.0 %	89.5 %	75.0 %
cafe/restaurant	78 : 66	64.1 %	66.7 %	57.9 %	66.7 %	65.0 %
car	78 : 73	87.2 %	75.0 %	100.0 %	78.9 %	95.0 %
city_center	78 : 78	94.9 %	100.0 %	94.7 %	89.5 %	95.5 %
forest_path	78 : 83	93.8 %	81.0 %	100.0 %	94.4 %	100.0 %
grocery_store	78 : 93	88.3 %	89.5 %	95.2 %	100.0 %	68.4 %
home	78 : 71	75.3 %	100.0 %	50.0 %	90.0 %	61.1 %
library	78 : 69	64.0 %	47.6 %	33.3 %	75.0 %	100.0 %
metro_station	78 : 101	94.7 %	78.9 %	100.0 %	100.0 %	100.0 %
office	78 : 76	93.1 %	94.7 %	100.0 %	88.9 %	88.9 %
park	78 : 54	52.2 %	60.0 %	38.9 %	80.0 %	30.0 %
residential_area	78 : 97	73.1 %	89.5 %	23.8 %	89.5 %	89.5 %
train	78 : 67	58.1 %	61.1 %	63.2 %	30.4 %	77.8 %
tram	78 : 107	86.7 %	100.0 %	83.3 %	63.6 %	100.0 %
Overall accuracy	1170 : 1170	78.6 %	81.2 %	71.1 %	81.7 %	80.3 %

Figure 4: Final ensemble result

different representative sounds for different classes. In the future, we hope the attention mechanism could be applied on more network structures like convolutional neural networks. However, more data are required to train more complexed models.

6. ACKNOWLEDGMENT

This work was supported by the Natural Science Foundation of China under grant no. 61273305.

7. REFERENCES

[1] T. Heittola, "Acoustic scene classification." [Online]. Available: http://www.cs.tut.fi/sgn/arg/dcase2016/task-acoustic-scene-classification

[2] C. Raffel and D. P. W. Ellis, "Feed-Forward Networks with Attention Can Solve Some Long-Term Memory Problems," *arXiv:1512.08756 [cs]*, Dec. 2015, arXiv: 1512.08756. [Online]. Available: http://arxiv.org/abs/1512.08756

[3] S. Dieleman, M. Heilman, J. Kelly, M. Thoma, D. K. Rasul, E. Battenberg, H. Weideman, S. K. Snderby, instagibbs, Britefury, C. Raffel, J. Degrave, peterderivaz, Jon, J. D. Fauw, diogo149, D. Nouri, J. Schlter, D. Maturana, CongLiu, E. Olson, B. McFee, and takacsg84, "Lasagne: First release." Aug. 2015, dOI: 10.5281/zenodo.27878. [Online]. Available: http://zenodo.org/record/27878