ATTENTION-BASED CONVOLUTIONAL NEURAL NETWORK FOR AUDIO EVENT CLASSIFICATION WITH FEATURE TRANSFER LEARNING

Tianxiang Chen tchen@pindrop.com

Udit Gupta ugupta@pindrop.com

ABSTRACT

Audio event classification is an urgent Content based Information Retrieval (CBIR) unsolved problem with numerous applications that it can benefit. This paper is explaining Pindrop's submission to the "Making Sense of Sound" challenge. In this submission we address the challenge of classifying audio excerpts based on their origin by using Convolutional Neural Networks with feature transfer learning. We use pre-trained VGGish network to extract feature embeddings. Our results show a remarkable improvement of the baseline system with achieving average recall rate of up to 92% across all the classes.

Index Terms— Making Sense of Sounds, Audio Event Recognition, Attention Network, Feature Transfer Learning, CNN, DCASE

1. METHODOLOGY

A total of four algorithms for audio event classification are provided as the part of the submission. Each of these algorithms apply attention-network mechanism and multi-class learning approaches to increase the generalizability of the models. In this project, we use similar attention mechanism introduced by Kong, Qiuqiang et al[1]. We also make use of the feature embeddings from the provided VGGish[2] pre-trained model in order to jump start training convergence which is trained using Google AudioSet[3] with 70M videos and an average of 4.6 minutes per video. The pre-trained model is downloaded from [4]. The process is described in detail in the remainder of this document where we describe the methodology for our best performed system (CBIR_2).

1.1. Audio Pre-processing

The MSOS Audio data Set contains one thousand five hundred audio clips taken from *Freesound data base*[5], the *ESC-50 dataset*[6], and the *Cambridge-MT Multitrack Download Library*. Each audio clip is five second long with a sample-rate of 44.1 KHz. In our method, we first downsample the audio files to 16 KHz in order to fit the VGGish pre-trained model. Each of the audio clips are divided into 5 non-overlapped frames of 960ms. Then, for each frame a time-frequency representation is calculated using log-mel energy filter banks with the window size 25ms and 10ms overlap and 64 frequency bands. Post-extraction all the frames are stacked together and are provided as inputs for the neural network.

1.2. Data Augmentation

Data augmentation was considered for one of the submitted systems(CBIR_3). This augmentation was performed using techniques similar to the ones described in [8]. Additional audio samples were created by mixing two audio samples from the same class with random delay, gain and equalization parameters. Care was taken that the data augmentation was performed using samples only from the training set so that the audio samples in the validation and testing sets remain totally unseen prior to prediction.

1.3. System Description

Figure 1 shows the overall structure of the best performed system. As shown in Figure 1., the input to the proposed neural network is the 480×64 mbe. The training samples are fed into two CNN based neural networks at the same time. The first CNN network initialized with weights from pre-trained VGGish model and freeze the weights during training. The weights of second network are random initialized. We use a shallow VGG architecture for each network which contains 8 CNN layers. We use a 3×3 filter for each CNN layer and padded the output with zeros to keep the output shape as same as the input. A max-pooling layer is performed on both time and frequency axis after CNN layer. An soft-max attention layer is applied after CNN layer. The attention feature from both two networks are then concatenate together and further fed into a fully connected layer followed by ReLU non-linearity activation and dropout rate of 0.2. Batch-Normalization[9] layer is also been added after each FC layer to reduce the chance of over-fitting. Finial classification layers is connected to the FC layer, with 5 output units and soft-max activation. An auxiliary task is also added to the classification layer with 91 output units for 91 different sound events. We train it on a GPU using the Adam Optimizer[10].

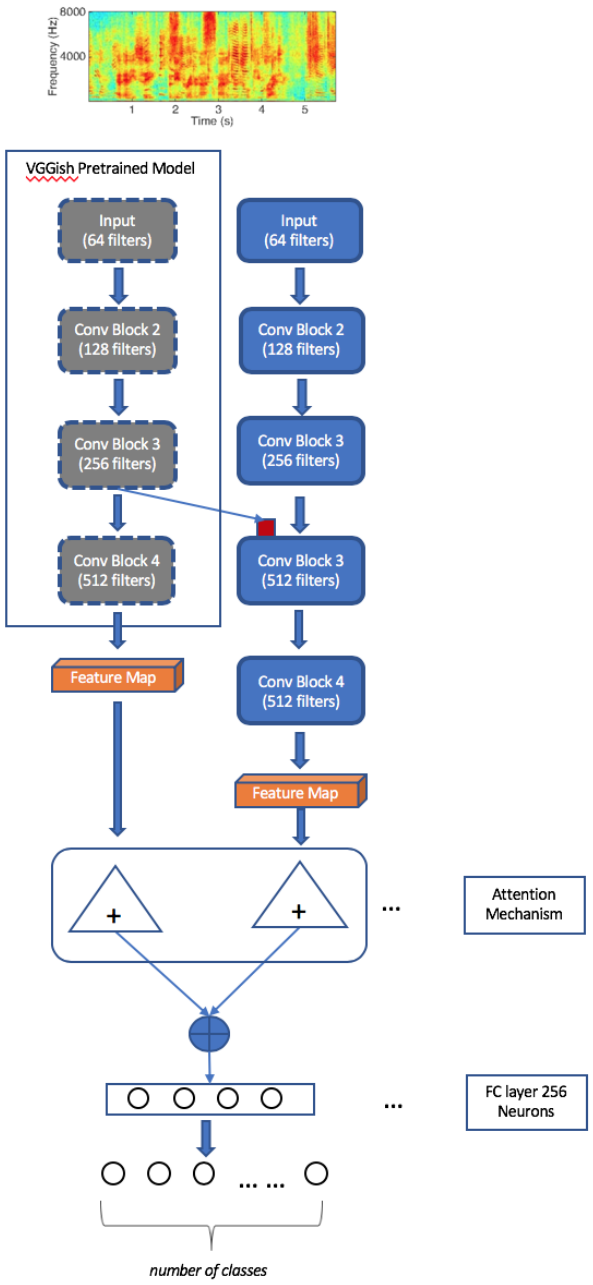

Fig. 1. Proposed model for audio event classification. The input is Log-Mel filter bank of size 480×64. The input layer and Conv Block 2 have 1 fully convolutional layer followed by a maxpooling layer of size 2×2, stirde (2, 2). The following Conv Blocks have 2 fully convolutional layers followed by a maxpooling layer. The red square block is the adapter[7], which is a convolutional layer with 1×1 filter

.

2. RESULTS

We perform a random split on released Development set to created train and validation sets. The test size is 20% of the total Development set. Average precision, recall and F1-socre are used to evaluate model performance. Table 1 shows the finial performance of each system. We made the final decision based on F1-score macro. As shown in Table 1, data augmentation doesn't show improvement to the overall performance. *CNN+Multi-Attention model* has the best overall performance which achieved 0.92 F1-score.

System ID	Name	Precision	Recall	F1
CBIR_1	CNN+Attention	0.91	0.90	0.91
CBIR_2	CNN+Muti-Attention	0.92	0.92	0.91
CBIR_3	CNN+Augmentation	0.91	0.90	0.90
CBIR_4	CNN+FC	0.85	0.85	0.85

Table 1. Classification results of all submitted systems

3. REFERENCES

[1] Qiuqiang Kong, Yong Xu, Wenwu Wang, and Mark D Plumbley, "Audio set classification with attention model: A probabilistic perspective," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 316–320.

[2] 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 *Acoustics, Speech and Signal Processing (ICASSP), 2017 IEEE International Conference on*. IEEE, 2017, pp. 131–135.

[3] 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 *Acoustics, Speech and Signal Processing (ICASSP), 2017 IEEE International Conference on*. IEEE, 2017, pp. 776–780.

[4] DTaoo, "Vggish: A vgg-like audio classification model," https://github.com/DTaoo/VGGish, 2017.

[5] Frederic Font, Gerard Roma, and Xavier Serra, "Freesound technical demo," in *Proceedings of the 21st ACM international conference on Multimedia*. ACM, 2013, pp. 411–412.

[6] Karol J Piczak, "Esc: Dataset for environmental sound classification," in *Proceedings of the 23rd ACM international conference on Multimedia*. ACM, 2015, pp. 1015–1018.

[7] Andrei A Rusu, Neil C Rabinowitz, Guillaume Desjardins, Hubert Soyer, James Kirkpatrick, Koray Kavukcuoglu, Razvan Pascanu, and Raia Had-

sell, "Progressive neural networks," *arXiv preprint arXiv:1606.04671*, 2016.

[8] Jan Schlüter and Thomas Grill, "Exploring data augmentation for improved singing voice detection with neural networks.," in *ISMIR*, 2015, pp. 121–126.

[9] Sergey Ioffe and Christian Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," *arXiv preprint arXiv:1502.03167*, 2015.

[10] Diederik P Kingma and Jimmy Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.