DCASE2018, task 2 – General-purpose tagging of Freesound audio with AudioSet labels

Huang Xie, 281685, [huang.xie@tuni.fi](mailto:huang.xie@tuni.fi)

Liang Fang, 281684, [liang.fang@tuni.fi](mailto:liang.fang@tuni.fi)

**Abstract**

In this report we describe two systems designed for DCASE2018 task 2. The first one is the baseline system provided by DCASE2018 organizer. We train it from scratch and obtain a final mAP@3 at 0.7635 over all classes, which is better than the original reported mAP@3 at ~0.7. Based on the baseline system, we replace the 3-layer CNN model with a VGG-like CNN model and get an improved final mAP@3 at 0.7981 over all labels.

**Introduction**

In DCASE2018, the objective of task 2 is to annotate Freesound audio clips with 41 labels from AudioSet. In the training dataset, there are 9473 audio clips, and each is annotated with only one of these 41 labels. For evaluation, a manually verified test dataset composed of 1600 samples is provided. The classification results are measured with mAP@3.

**Feature extraction**

The feature extraction method for the baseline system is used for our experiments. All audio recordings (44.1 KHz mono) are divided into overlapping windows of size 0.25s with a hop of 0.125s. Log-mel spectrogram features are extracted by applying STFT with window size of 25ms, window step of 10ms and 64 mel bands to each audio window.

**Methodology**

1. Baseline system

The 3-layer CNN used in the baseline system is presented in Figure 1 a).

1. VGG-like system

To improve the baseline system, we define a variant of the VGG model (Configuration A) by replacing the 1000-wide FC with a 41-wide FC at the end. We use the same input features as in the baseline system.

As shown in Figure 1 b), the VGG-like CNN contains 5 stacks of convolutional layers and 2 fully-connected layers. All the filters in convolutional layers are 5-by-5, and each convolutional layer is followed by batch-normalization and ReLU. 2-by-2 max-pooling layers are used in this model. Then two 4096-wide fully-connected layers are followed.

**Evaluation**

1. Dataset

For this given challenge task, a training dataset contains 9437 audio clips, each of which is annotated with a single AudioSet label, is provided for training. The test dataset contains 1600 samples which have been manually verified. To evaluate the classification performance, mAP@3 is used as evaluation metric.

We follow the prediction method used in baseline system. For each audio clip, we divide each clip into overlapping 0.25s-wide windows with a hop of 0.125s. Then, we obtain 41 predictions for each 0.25s-wide window and average all the window-level predictions to obtain a clip-level prediction. Finally, we output the top-3 predicted labels and calculate mAP@3.

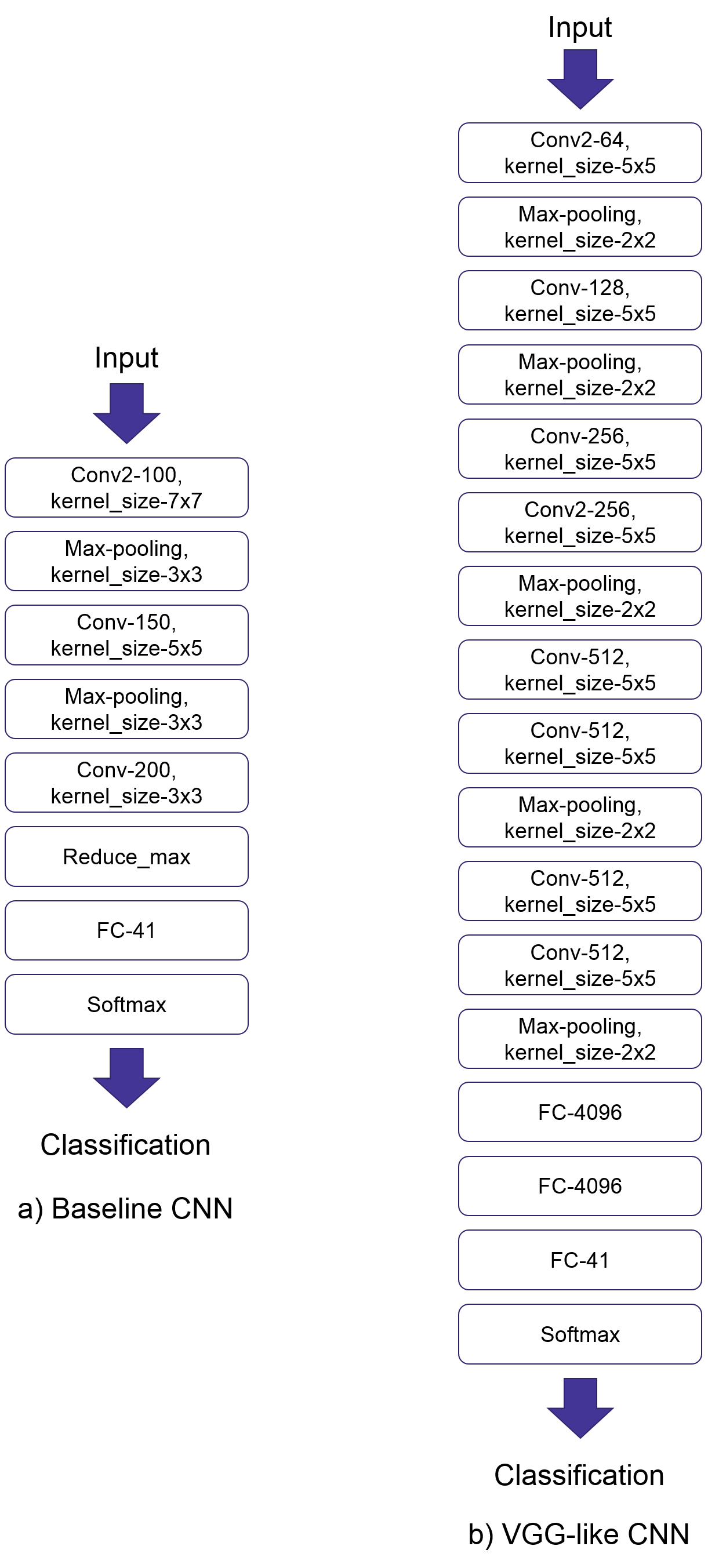


Figure 1. Baseline CNN and VGG-like CNN

1. Results

After training both systems for ~12 hours, we evaluate the final mAP@3s on the test dataset and obtain 0.7635 for the baseline system and 0.7981 for the VGG-like system, respectively. Figure 2 shows more details about mAP@3 for each label.

图片包含 物体

描述已自动生成

Figure 2. mAP@3 for the 41 labels

**Discussions**

In Figure 2, it shows that both systems result in a similar mAP@3 distribution over all labels. Particularly, they could not identify audio clips with label “Squeak” appropriately. For the baseline system, the statistical information of the predicted labels for “Squeak” audio clips are shown in Figure 3, and we can see that misclassification occurs frequently between “Squeak” and “Fart”. From the aspect of human sense, it would be understandable.

图片包含 屏幕截图

描述已自动生成

Figure 3. Prediction for Squeak audio clips

Furthermore, we notice that the baseline system achieves better performance than the reported one, which is 0.6943. Currently, we have not considered other optimization techniques but only extend the training period (~142K vs ~41K training steps). It seems that it could achieve even better result by involving a validation dataset. On the other hand, we change the baseline system by replacing the 3-layer CNN with a deeper VGG-like CNN. The new classification system achieves better result than the baseline system. It shows that the VGG models, which have been successful in computer vision, can also be used for audio classification tasks.

**Conclusions**

For the task 2 in DCASE2018 challenge, we train the baseline model and a VGG-like model from scratch. For the baseline system, the final mAP@3 on the test dataset is better than the original reported one (0.7635 vs 0.6943). By replacing the 3-layer CNN in the baseline system with a VGG-like CNN, we obtain an even better result than the baseline system (0.7981 vs 0.7635).

**References**

[1] E. Fonseca *et al.*, “General-purpose Tagging of Freesound Audio with AudioSet Labels: Task Description, Dataset, and Baseline,” no. 688382, 2018.

[2] K. Simonyan and A. Zisserman, “Very Deep Convolutional Networks for Large-Scale Image Recognition,” pp. 1–14, 2014.