INTERSPEECH 2019: Show & Tell Contribution
September 15–19, 2019, Graz, Austria

Robust Keyword Spotting via Recycle-Pooling for Mobile Game

Shounan An[1], *Youngsoo Kim*[1], *Hu Xu*[1], *Jinwoo Lee*[1], *Myungwoo Lee*[2], *Insoo Oh*[3]

[1]Game Dev. AI Team, NARC, Netmarble, South Korea
[2]Game Contents AI Team, NARC, Netmarble, South Korea
[3]Magellan Division, NARC, Netmarble, South Korea

{ethan.an, kimys, joseph.suh, hiall21, or35, ioh}@netmarble.com

Abstract

We present an effective method to solve a small-footprint keyword spotting (KWS) task via deep neural network for mobile game. Our goal is to improve the accuracy of KWS in various environments. To this end, we propose a new neural network layer named recycle-pooling. Extensive experiments indicate that our recycle-pooling based convolutional neural network (RP-CNN) indeed improves the performance of KWS in both clean and noisy data for mobile game. We will perform live demonstration of RP-CNN based KWS integrated into a full-sized, production-quality mobile game *A3: Still Alive*, which is one of the major games from Netmarble this year and will be available on market soon.

Index Terms: keyword spotting, recycle-pooling, convolutional neural network, mobile games

1. Introduction

With the rapid development of mobile game industry, the needs of speech based user interface is growing fast for various interactive scenarios such as driving car, cooking at home and any hands busy situations. Recently deep learning based methods show breakthrough progress for small-footprint keyword spotting (KWS) tasks [1]. Consequently KWS technology already has been deployed in real life products e.g., Amazon echo, Apple homepod and Google home. However, in literature there is no concrete work to use KWS in mobile game as speech based user interface.

The difficulties to deploy KWS in mobile game are that users could play game in any noisy places, which requires KWS system to be robust to noisy data. To this end, we propose a new neural network layer named recycle-pooling, which shows improved accuracy especially to noisy data. The useful behaviour of our recycle-pooling based convolutional neural network (RP-CNN) is summarized as follows.

- To the best of our knowledge, this is the first work try to resolve KWS tasks for mobile game. We integrated RP-CNN into our game *A3: Still Alive*, which will be released soon [2] (Figure 1) .
- Our RP-CNN shows better performance for both clean and noisy data, which make it possible to deploy KWS in mobile game products.

2. RP-CNN

In this section, we will first introduce our feature extraction method for RP-CNN, then recycle-pooling layer is explained in detail.

Figure 1: *A3: Still Alive with our RP-CNN based KWS system.*

2.1. Feature extraction

We choose mel-frequency cepstral coefficients (MFCC) [3] for 1 second of PCM voice signal data with 16kHz sampling rate as input feature to RP-CNN. Voice signal was converted to spectrogram using short-time Fourier transform (STFT) for each 512 samples window with hop size of 128 samples, and Hann function [4] was applied as a window function. 128-filter mel filterbank with 8kHz maximum frequency was used to get mel-spectrogram from the output of STFT. We pick 40 coefficients from discrete cosine transform (DCT) result of the mel-spectrogram which forms a single MFCC feature frame. Finally the input to RP-CNN is the 122×40 MFCC feature extracted from 1 second signal (Figure 2).

2.2. Recycle-Pooling layer

The architecture of RP-CNN is given in Figure 2. While the conventional CNN [5] based models usually use max-pooling, we use recycle-pooling for our RP-CNN. Max-pooling is well-known operation in deep learning based classification and detection especially in computer vision tasks. However in a speech signal processing point of view, max-pooling gets rid of important features for KWS. For example if we apply 2×2 kernel for max-pooling with stride 2, then 75% of features, which may contain import information would be dumped. Conversely there is no information loss in recycle-pooling operation. Recycle-pooling layer reshapes the higher resolution features into the lower resolution features via spatial re-arrangement. Therefore, recycle-pooling works as multi-resolution operation, which may result in accuracy improvement. As shown in Figure 2, recycle-pooling reshapes $122 \times 40 \times 64$ feature map into $61 \times 20 \times 256$.

3. Mobile SDK

The system diagram of our KWS for both training procedures and mobile SDK for inference is illustrated in Figure 3. The SDK mainly consists of three parts, the first one is a module for converting voice signal from microphone of the device to

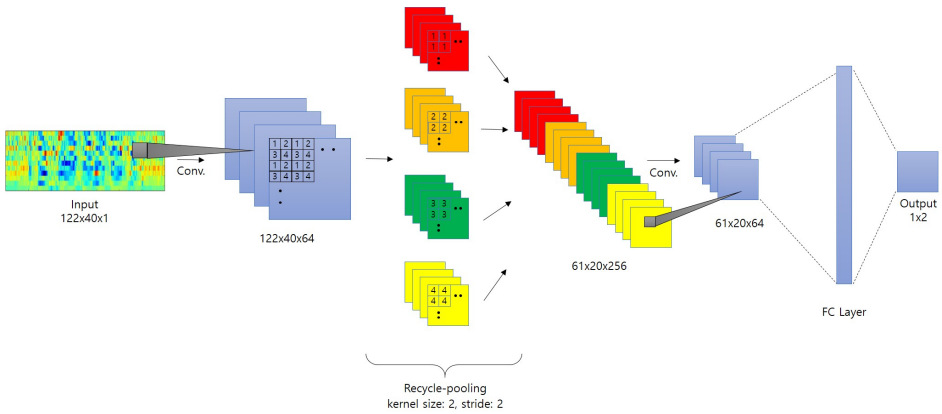


Figure 2: *The network architecture of RP-CNN for KWS tasks.*

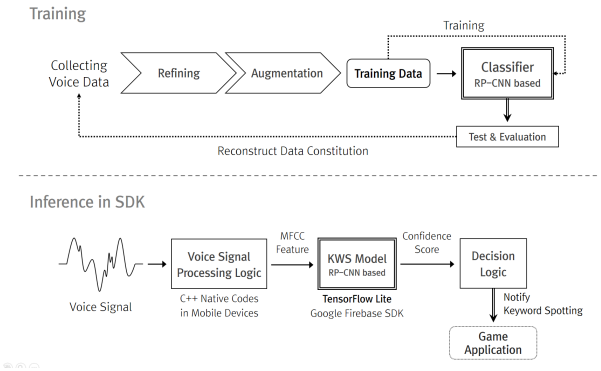
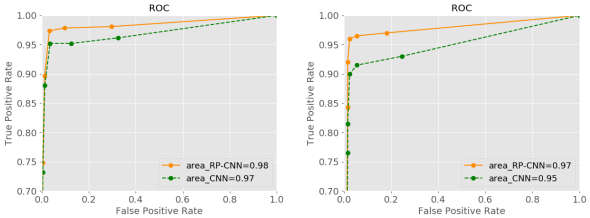


Figure 3: *System diagram of our KWS for mobile game.*



(a) *Results on clean data* (b) *Results on noisy data*

Figure 4: *ROC curves of the conventional CNN vs RP-CNN.*

MFCC features through fast Fourier transform (FFT), applying mel-filter bank and DCT. In the second part we choose TensorFlow Lite [6] from Google's Firebase SDK as our inference engine for its optimized operations designed for mobile devices. And the last one is the averaged multi-frame prediction scheme based on confidence scores generated from RP-CNN.

4. Experiments

Our keyword "narc ya" pronounced as [ˈnaːrkjaː] and we collected positive dataset in various environments including office, home, cafeteria, driving car etc. For negative dataset, we collected data from conversations on the radio and television. We also put some utterances which only contain partial of the keyword as hard negative samples. Finally, the training dataset consists of 50,000 positive samples from 109 people and 80,000 negative samples. We randomly select 10% as validation dataset. Models are trained 60 epochs with 256 batch size. During training, for the first 56 epochs, we set the learning rate as the 10^{-3} and for the rest epochs we set the learning rate as 10^{-4}.

For the test dataset, we additionally collected about 400 utterances from 13 people as positive samples and about the same size for negative samples both in clean and noisy environments. The performance of our RP-CNN and conventional CNN model is measured by receiver operating characteristics (ROC) curve, with the measurements of false positive rate (FPR) and true positive rate (TPR). The conventional CNN uses exactly the same architecture as RP-CNN and the only difference is to use max-pooling instead of recycle-pooling. Higher the area under the curve (AUC) value means better performance. As we can see from Figure 4, our RP-CNN achieves better performance than the conventional CNN model.

5. Conclusions

In this paper, we have presented a new neural network layer named recycle-pooling and RP-CNN for KWS tasks, which shows improved accuracy especially to noisy data. Experiments confirm the useful behaviour of RP-CNN. We will do live KWS demonstration integrated in our game *A3: Still Alive* on the spot.

6. References

[1] T. N. Sainath and C. Parada, "Convolutional neural networks for small-footprint keyword spotting," in *Proceedings of Interspeech*, 2015, pp. 1478–1482.

[2] "A3: Still alive," https://a3.netmarble.com/.

[3] Md.Sahidullah and G. Saha, "Design, analysis and experimental evaluation of block based transformation in mfcc computation for speaker recognition," *Speech Communication*, vol. 54, no. 4, pp. 543–565, 2009.

[4] F. J. Harris, "On the use of windows for harmonic analysis with the discrete fourier transform," *Proceedings of the IEEE*, vol. 66, no. 1, pp. 51–83, 1978.

[5] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Proceedings of Advances in Neural Information Processing Systems*, 2012, pp. 1097–1105.

[6] "Tensorflow lite," https://www.tensorflow.org/lite.