AREA66 - SYSTEM DESCRIPTION

Rohit Saxena

TCS Research

1. SYSTEM DESCRIPTION

The MCE 2018 task consists of multi-target speaker detection and identification for the given i-vector extracted from raw audio. The dataset for the task consists of i-vectors from 41,845 utterances in training set and 8,631 utterances in the development set both the non-Blacklist and Blacklist speakers. The first subtask is to verify whether a given speech segment is from blacklist cohort or not. The second subtask is to identify who it is among blacklist speaker set, if the input segment turns out as a blacklist.

We formulate this problem as a multi-task classification problem where one task performs detection of blackist speaker and the other task identify who it is among the blackist speaker. Both the task are then jointly learned with the help of hard parameter sharing in neural network.

1.1. Model Description

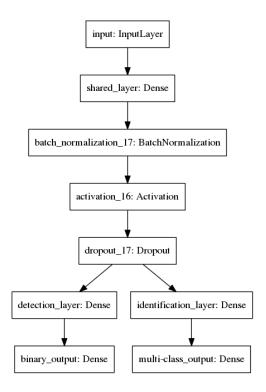


Fig. 1. Multi-target Detector baseline for MCE 2018

Our model is a multi-task feed-forward neural netowrk trained jointly for both detection and identification task. Given an i-vector (600 dimension), the model predicts both the score to determine whether the speaker is blacklist or not (Top-S detector) and the class probabilities (Top-1 detector) to identify the specific blacklist speaker. The network takes the input i-vector and it is passed to a shared Fully Connected (FC) layer which learns the common representation of the input. It is followed by a Batch Normalization layer to reduce the internal covariate shift. The network is then branched to learn task specific representation as shown in the Figure 1. Each branch has a FC layer to learn the task specific representation followed by its respective loss function.

The loss function for blacklist speaker detection branch is formulated as binary cross-entropy:

$$\mathcal{L}_D(q) = \sum_{i=1}^{2} -\hat{q}_i \log(q_i)$$
 (1)

The loss function for identification of blacklist speaker branch is formulated as multiclass cross-entropy:

$$\mathcal{L}_{I}(r) = \sum_{i=1}^{K} -\hat{r}_{i} \log(r_{i}), \qquad (2)$$

where K is the number of blacklist speakers and a class for non-blacklist speaker.

The joint loss function for learning detection and identification together is formulated as follows:

$$\mathcal{L}_{J}(\cdot) = (1 - \beta) \mathcal{L}_{D}(\cdot) + \beta \mathcal{L}_{I}(\cdot), \tag{3}$$

where β is the hyperparameter which determines the weight of the loss functions.

The hyperparameters are as follows: *learning rate* = 0.001, *optimizer* = Adam, *regularizer* = Early Stopping and Dropout.

1.2. Results

Model achieved Top-S detector EER of 1.5% and Top-1 detector EER of 35.0%.