AUTHOR GUIDELINES FOR DCASE 2019 CHALLENGE TECHNICAL REPORT

Technical Report

Kharin Alexander

MePhI University
laboratory of bionanophotonics., Kashirskoe sh.31 ,Moscow
Moscow 115409, Russia
S__asha@mail.ru

ABSTRACT

Multy-layer convolutional neural network with following Dense layer with 4.7 millions of parameters was used for training on mel-spectrograms of audio data. Such large number of parametersand small dataset (~9k samples without augumentation) leads to vulnerability of model to overfitting. Agumentation of audiofiles (i.e cropping of spectrograms) was not found wery effective way to get rid of overfitting. The following ways found to be reasonoble: standard Kfold technique with training on 5 Kfolds and averaging of the results and so-called 'noisy data annealing'. That method lies on sequential training of the model on general set for several epochs (30 in our case) followed by train - ing on poorly labeled, but larger dataset for 5 epochs. After several cycles we can observe significant reduction of the overfit - ting (lwrap scores 0.61 for base model, 0.66 for noisydata-annealed model). Such increase is caused by partial 'reset' of the trainable parameters during training on poorly-labelled set. The more set-specific parameters are, the higher is 'reset' rate, so such annealing enhances the significance of non-overfitting-responsible features and reduces the impact of highly dataset-specific features.

1. DATASET

I use the dataset provided by Detection and Classification of Acoustic Scenes and Events (DCASE) Challenge, DCASE Task 2 records. The audio data is labeled using a vocabulary of 80 labels from Google's AudioSet Ontology, covering diverse topics: Guitar and other Musical instruments, Percussion, Water, Digestive, Respiratory sounds, Human voice, Human locomotion, Hands, Human group actions, Insect, Domestic animals, Glass, Liquid, Motor vehicle (road), Mechanisms, Doors, and a variety of Domestic sounds.
Dataset contained 2 types of tagged audio records:
1st is 4970 hand-labelled sound records, 2nd is larger set (19815 records from the YFCC dataset) with worse labelling quality.

2. SOUND PREPROCESSING AND NETWORK ACRCHITECTURE

The audioclips were converted to mel spectrograms to get with 1/64 s per sample to get 128*128 spectrograms for each 2s. Data augumentation: during training procedure a random 128*128 chunk was taken for each spectrogram
Convolutional neural network with architecture, proposed in (https://www.kaggle.com/mhiro2/simple-2d-cnn-classifier-with-pytorch) was applied to the mels spectrograms of the sounds. It contains the sequence of 3x3 conv block, each of them contains 2 convolutional layers (see figure 1). That convolutional layers are ended by Dense layer and linear units.

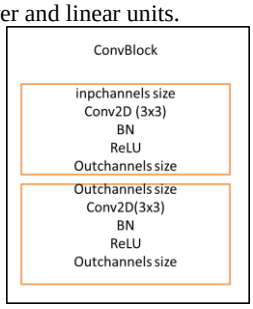

Figure 1: Composition of single conv2d block.
Overall architecture was the next:
spectrogram-convblock64channels-convblock128channels-convblock256channels-convblock512channels-globalmax-dense128-dense80.

3. LEARNING PROCEDURE AND NOISY DATA USAGE

Learning with performed with categorical crossenthropy loss on softmax of output layer with Adam optimizer (lr=0.001,).
1st mode – without annealing – learning the same architecture on 5 k-folds. The mean lwrap score is shown on figure 1. It is evident, that the model is rapidly falling into overfitting mode. The procedure of annealing is the training of poorly-labelled data for 5 epochs after each 30 epochs of training on hand-labelled dataset. That procedure leads to the reduction of the overfitting on validation subsets.

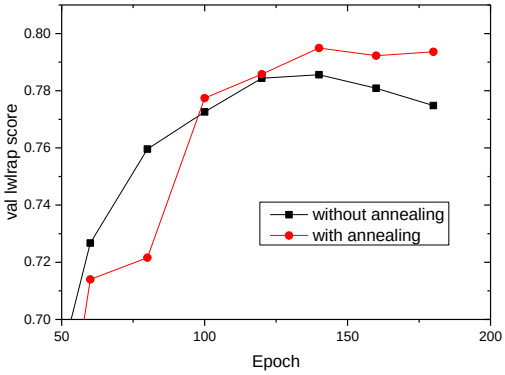

Figure 2: validation lwlrap score during training with and without annealing on noisy data.

The model checkpoints with the best lwlrap values on validation during training were saved and the mean output of 5 folds was used for final prediction.

4. CONCLUSION

I found the annealing on big amount of poorly-labelled data effective for overfitting reduction in current task