Master Thesis Defense Ravali Nalla

Date: Thursday, 23.01.2020
Time: 09:30 am
Place: Fraunhofer FKIE, Wachtberg

Title: Evaluation of Deep Learning Technique for Impulse Sound Classification in 2D Representation

Abstract:

Impulse sound classification has several civil and security-related applications. Convolutional Neural Networks(CNNs) have been proven very effective in image classification and show promise for audio applications as well. So, we would like to apply image classification networks from vision domain to the audio domain for impulse sound classification task by treating audio segments as images. We use a ballistic sound dataset collected by Fraunhofer Institute researchers for impulse sound classification tasks.

For using audio segments as images we extract spectrograms of them and treat them as images for the training. A convolutional 2D network specifically for impulse sound classification task has been designed and then a convolutional 1D network with similar architecture is created. We compare the performance of these networks to see how beneficial it is to use spectrograms instead of audio raw data. We observe that the accuracy achieved using spectrogram data representation is 1.3% less than raw data representation but with a 40% gain in the number of parameters used, which can be considered as a good trade-off. Later, the performance of this network with our selected image classifier networks is compared. From state-of-the-art image classification networks, VGGNet-16, Inception v3, Inception-ResNet-v2, NASNet, ResNeXt and Efficient Net are chosen as candidate architectures. Applying image data augmentation techniques to spectrograms does not make sense, so we apply spectrogram data augmentation techniques: time masking and frequency masking and then evaluate the effect of augmentation techniques on the network performance. The image classification networks selected are trained from scratch on the ballistic sound dataset and their performance is evaluated on different test sets. The same chosen classification networks are trained on a ballistic sound dataset with ImageNet pre-trained weights to study if the pre-trained weights help the training. We compare and evaluate the performance of networks trained from scratch and networks trained with pre-trained weights. ResNeXt-50 performs best with an averaged F1-score of 80 with 23.1M parameters.

We also study the effect of unsupervised learning by using Autoencoders, in which the data is first trained to learn unsupervised features and use these Autoencoder weights as an initial point for the classifier. But, we have seen that using autoencoders does not give a boost to the classifier's performance.