Transfer Learning limited edition sounds

Athanasios Agrafiotis s2029413

1 Problem

This paper presents how can you classify small audio records using convolutional neural networks. Different architectures of convolutional neural netwroks has been used until now for the same task with different datasets. This paper presents how can we classify sounds using a pretrained network of sixteen layers and known as VGG16. A dataset of 8 different sound was used to train our network.

Keywords: *Classification, Convolutional Neural Networks (CNN), Supervised learning, VGG16.*

1.1 Related work

In previous work a convolution neural network has been trained that could classify enviromental sounds [2]. The architecture of the network was simple with two convolutional layers. For the research three different dataset was used the ESC-50 with fifty balanced classes, the ESC-10 with ten classes and at last the Urban8K of 10 classes. The convolutional neural network performed well with 64% accuracy for the ESC-50, 84% accuracy for the ESC-10 and at last 74% accuracy for the urban8k dataset. This work consider as related because it has as feature selection the log-mel-spectogram of each sound.
Another research which has been conducted using convolutional neural network refered in [4]. The main purpose of the research is to gather a dataset of soundclips which can be used for speech recognition tasks. The audio clips are approximately between one second and contain single words in English language. Each of the words can be used as command in Iot or robotic applications. The dataset has evaluated in a convolutional neural network. The architecture of the convolutional neural network had 6 convolutional layers [1]. **A sample of this dataset was used to train our network with sounds which can missclassify easily(three-tree,on-one)** .

2 Convolutional Neural Networks

Convolutional neural network has been widely used for image and audio classification. The process of convolution is basic matrix multiplication and summation.

$$f(x) * g(x) \tag{1}$$

In a convolutional network the image enters a matrix, then a box which can named as a *receptive field* it iterates through the image with a step, the step is known as *stride* and can be different number according to the task. Inside the receptive field it has a number of weights which multiplied with the matrix of the image the weights of the image is known as *kernel filter*. The weights of the kernel filter are totally random and update by *backpropagation*. The multiplication and sumation of the image with the kernel has as result a *feature map*. The feature map represent the features that has extracted through the image. Is the edge detection of the image. Next is the process of pooling known as *downsampling* which iterate through the feature and pool the maximum numbers(max-pooling) or the average numbers(average-pooling). The last part of a convolutional network is the activation function which introduce non-linearities to the system.

$$f(x) \begin{cases} 0 & \text{for } x < 0 \\ x & \text{for } x => 0 \end{cases} \tag{2}$$

3 Feature Selection

As a feature selection the Mel-log spectogram of each sound extracted using librosa of scikit learn. The mel-log-spectogram is a representation of the signal using fourier transform, the fourier tansform represents the signal in amplitude and time frequency domain. The fourier transform applied using window each result represent a spectrum. The spectrum has the amplitude and the frequency of each window. The Mel-log spectogram reproduce the distridution of the spectrum mapping color according to the applitude of the signal. As the amplitude getting higher a darker is used. A representation of the mel-spectogram can observed in figure 2a.

3.1 VGG16

The VGG16 is model of convolutional network and it was named based the team which create it the 'Visual Geometry Group' the number of 16 is the number of layers. The VGG16 had the best performance in 2014 of the imagenet [3] for image localization and classification. The imagenet is a big scale dataset which has been used for image classfication competitions. You can observe the architecture in the figure 5 . What makes special VGG16 is that the weights of the convolutional network is opensource for everyone in the library of Keras. The pretrained weights of the network can be used as a **feature extraction**. The pretrained network gives two options the option of **fine tuning** and the option of **tranfer learning**.

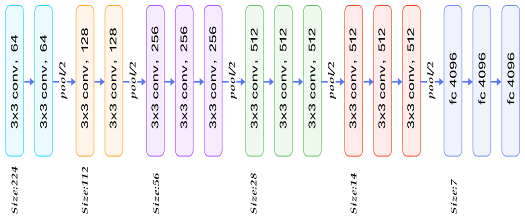


Figure 1: VGG16-architecture

3.1.1 Transfer learning

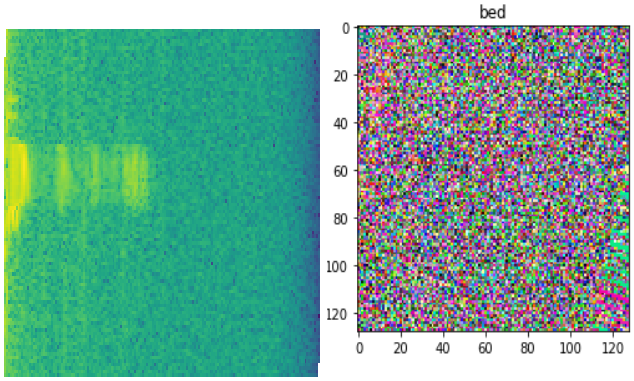
With transfer learning we can use the weights of the pretrained model to extract the feature. Then we can test different models to classify our data. In this research the feature which selected was the mel-log-spectogram in figure 2a and in 2b the result after extracting the features using VGG16 pretrained weights. The process of transfer learning is a feedforward process less time consuming because we do not have to train our network and change the weight with backpropagation.

3.1.2 Fine tuning

For the process of fine tuning is to change the architecture of the pretrained network. In this case we can add more layers at the end of the network. Fine tuning is more time consuming method because each of the new layers has a new weights which had to be trained again with backpropagation.

3.2 Data augmentation

Data augmentation is really important for a convolutional neural network. The process is actually helps with **generalization** which for each which pass through the generator it reproduces with different attributes like angle,shadow,reflect,zoom. All the new images is belong to the same class so it help get for information for each class. The Keras has it's own library which called *image_generator*.



(a) Mel-log-spectogram bed

(b) Bed result after file transfering

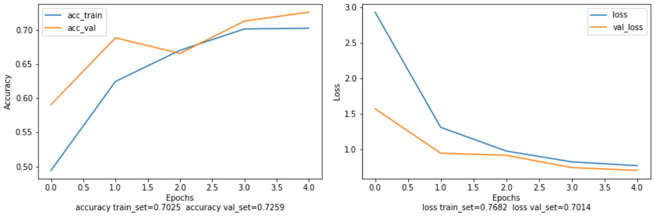
Figure 2: Bed after extracting the features

3.3 Research

In first stage of the research the a smaller dataset was in order to evaluate the preprocess. **Transfer learning was** used and **logistic regression** to classify the three sounds. The test was conducted in a smaller dataset which display in table 1. The image size was 50x50 the number of the epochs was **5** and at last the data augmentation which was used :rotation_range,width_shift_range,height_shift_range,shear_range, zoom_rannege,horizontal_flip,vertical_flip for training set and validation set.

Audio Word	Train
bed	1542
yes	2140
zero	2139

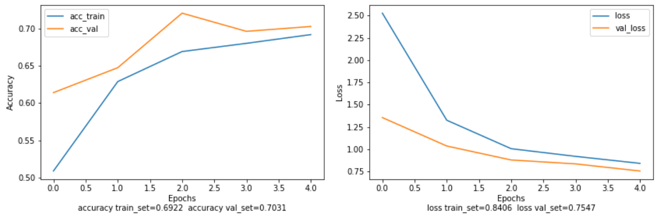
Table 1: smaller dataset



(a) Accuracy

(b) loss

Figure 3: Accuracy and loss Adam lr=0.001 epsilon=1e-6



(a) Accuracy (b) loss

Figure 4: Accuracy and loss RMS lr=0.001 epsilon=1e-6

Adam optimizer performed better for the same task and the loss of the train set was lower than the rms optimizer. The accuracy that we got so far was **72.5%** in the validation set. For this limited dataset the transfer learning performed well but the sounds was totally different of each other. In the next epxeriment the sounds that will be used will be more identical to each other.

3.4 Classify eight classes

The optimizer with the best performance was used in a the bigger dataset. The prerformance of the logistic regression depicted in 6a and in 6b. The validation dataset does not perform well because not data augmentation was used on this set. The dataset has not tested in a previous research show as a baseline it consider the number of observations of the higher class divided by the number of observations **13.9**. The words of this dataset cannot be easily distinguish from each other even by humans.

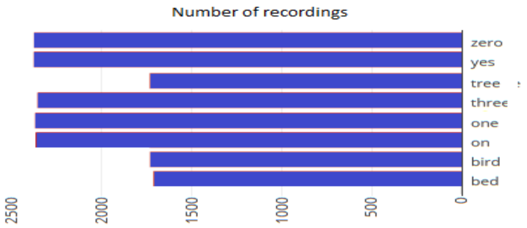
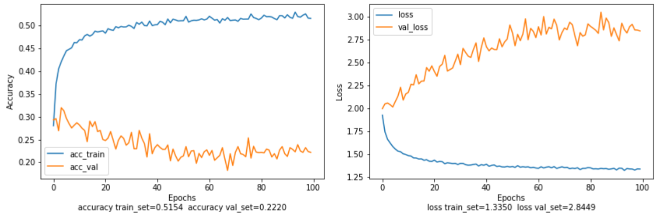


Figure 5: number of records



(a) Accuracy (b) loss

Figure 6: Accuracy and loss Adam lr=0.001 epsilon=1e-6

Bellow the most missclassified sounds is demonstrated. In table 2 you can observe that the most missclassified sound was the bird with tree and the word on with yes. The best score has the sound of bed which classified most correctly.

	three	one	on	yes	tree	bird	bed	zero
three	**103**	13	23	8	14	1	7	2
one	12	**93**	34	12	19	2	0	1
on	15	35	**89**	38	31	0	15	13
yes	4	11	25	**120**	39	1	18	19
tree	9	11	15	27	**123**	19	15	16
bird	4	3	5	4	49	**86**	10	12
bed	10	1	8	15	20	2	**159**	22
zero	9	5	14	26	36	2	21	**124**

Table 2: confusion matrix

	precision	recall	f1-score
three	0.69	0.57	0.62
one	0.47	0.53	0.50
on	0.36	0.41	0.38
yes	0.54	0.59	0.56
tree	0.42	0.37	0.39
bird	0.63	0.56	0.60
bed	0.64	0.68	0.66
zero	0.57	0.55	0.56
avg/total	0.53	0.53	.053

3.5 Conclusion

The research which had as main purpose to classify the dataset of elimited edition using a pretrained network known as vgg16. It has as result good performance with 2 classes. The training repeated using 8 classes with poorest results and the performance was evaluated using a confusion matrix. In the end using all the classes which correspond to 30 the network did not performed well Figure 7.

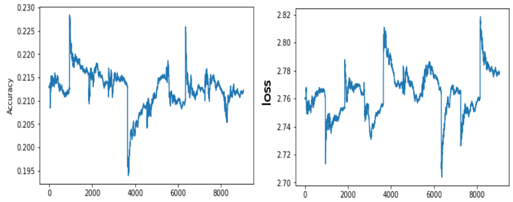


Figure 7: loss-accuracy-30classes-**iterations**

References

[1] Ossama Abdel-Hamid et al. “Convolutional Neural Networks for Speech Recognition”. In: *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 22 (Oct. 2014), pp. 1533–1545. URL: https://www.microsoft.com/en-us/research/publication/convolutional-neural-networks-for-speech-recognition-2/.

[2] K. J. Piczak. “Environmental sound classification with convolutional neural networks”. In: *2015 IEEE 25th International Workshop on Machine Learning for Signal Processing (MLSP)*. Sept. 2015, pp. 1–6. DOI: 10.1109/MLSP.2015.7324337.

[3] Karen Simonyan and Andrew Zisserman. “Very Deep Convolutional Networks for Large-Scale Image Recognition”. In: *CoRR* abs/1409.1556 (2014). arXiv: 1409.1556. URL: http://arxiv.org/abs/1409.1556.

[4] Pete Warden. “Speech Commands: A Dataset for Limited-Vocabulary Speech Recognition”. In: *CoRR* abs/1804.03209 (2018). arXiv: 1804.03209. URL: http://arxiv.org/abs/1804.03209.