Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 112 (2017) 2048–2056

Procedia Computer Science

www.elsevier.com/locate/procedia

International Conference on Knowledge Based and Intelligent Information and Engineering Systems, KES2017, 6-8 September 2017, Marseille, France

Classifying environmental sounds using image recognition networks

Venkatesh Boddapati[a], Andrej Petef[b], Jim Rasmusson[b], Lars Lundberg[a,*]

[a]*Department of Computer Science and Engineering, Blekinge Institute of Technology, 371 79 Karlskrona, Sweden*
[b]*Sony Mobile Communications AB, Mobilvägen, 221 88 Lund, Sweden*

Abstract

Automatic classification of environmental sounds, such as dog barking and glass breaking, is becoming increasingly interesting, especially for mobile devices. Most mobile devices contain both cameras and microphones, and companies that develop mobile devices would like to provide functionality for classifying both videos/images and sounds. In order to reduce the development costs one would like to use the same technology for both of these classification tasks. One way of achieving this is to represent environmental sounds as images, and use an image classification neural network when classifying images as well as sounds. In this paper we consider the classification accuracy for different image representations (Spectrogram, MFCC, and CRP) of environmental sounds. We evaluate the accuracy for environmental sounds in three publicly available datasets, using two well-known convolutional deep neural networks for image recognition (AlexNet and GoogLeNet). Our experiments show that we obtain good classification accuracy for the three datasets.

© 2017 The Authors. Published by Elsevier B.V.
Peer-review under responsibility of KES International

Keywords: Deep Learning; Convolutional Neural Networks; Environmental Sound Classification; Image Classification; GPU Processing.

1. Introduction

Automated classification of environmental sounds, like dog barking and siren, can be used in applications such as remote surveillance and home automation. An interesting application is the use of home monitoring equipment which identifies different sounds produced in a domestic/interior environment and alerts the user accordingly.

* Corresponding author. Tel.: +46-455-385833.
E-mail address: lars.lundberg@bth.se

1877-0509 © 2017 The Authors. Published by Elsevier B.V.
Peer-review under responsibility of KES International
10.1016/j.procs.2017.08.250

Examples of such sounds are baby crying, air conditioner, and glass breaking. The recognition of such domestic sounds, if implemented on a mobile device, can lead to new and important applications.

Environmental sounds consist of various non-human sounds (excluding music) in normal day-to-day life. During the past few years, many attempts to recognize environmental sounds have been made. Presently, there is an increasing focus on classifying environmental sounds using deep learning techniques[1,2,3,4,5]; the improvements in the field of image classification in recent years are leading researchers to start using images when classifying sounds.

The important difference between speech/music and environmental sound is that the former are strongly structured and clearly demarcated whereas the latter have no common structure[6]. This causes it to be a whole new problem. There is possibly an elegant solution in deep learning since deep neural networks have been proven to be able to handle vast amounts of data and model complex features due to advances in computing power, including the more general use of GPUs.

Wang et al[7] discuss the efficiency of Gabor-based non uniform scale frequency map that combines Principle Component Analysis and Linear Discriminate Analysis to extract features from the sound samples followed by classification using Support Vector Machines (SVMs); a high classification accuracy was reported. Lu et al[8] conclude that SVM provides more accurate classification of environmental sounds than *k*-Nearest Neighbor (kNN) and Gaussian Mixture Model (GMM).

The most common deep learning based approach for classification of sounds is to convert the audio file to an image, and then use a neural network to process the image. Mostafa et al[9] perform classification of music using Probabilistic Neural Network with satisfactory results. Most sound classification approaches use supervised pattern recognition. However, Zhang and Schuller[10] voice the problem that manual labelling of datasets is very costly and they recommend semi-supervised learning as a better solution. McLoughlin et al[6] state that classification of sound in realistic noisy environments is challenging and propose a deep neural network as a viable solution. Piczak[11] and Zhang et al[1] both convey the idea that convolutional neural networks have the best accuracy rates on Spectrogram analysis.

As summarized by Chachada et al[5] there are three broad ways of processing environmental sounds for classification purposes: 1) Framing-based where the audio signals are separated into frames using a Hamming window. Then the features are extracted from each frame and classified separately. 2) Sub-framing based processing where the frames are further subdivided and each frame is classified based on the majority voting of the sub-frames. 3) Sequential processing where the audio signals are divided into segments of typically 30 ms with 50% overlap. The classifier then classifies the features extracted from these segments.

The use of mobile devices is ubiquitous, and there are a number of applications that would benefit from being able to identify both sounds and objects. Previous studies[12] have shown that, for environmental sound classification, the ratio (classification performance)/(computational cost) is more favourable for deep neural networks compared to both Gaussian Mixture Models (GMM) and Support Vector machines (SVM). Since this ratio is particularly important on mobile devices with limited processing and battery capacity, evaluating deep neural networks for sound (and image) classification on mobile devices is very relevant. As discussed above, sound classification often uses images in the form of Spectrograms etc. However, many of the deep learning networks used for such classification are designed specifically for Spectrograms and other visual representations of sound, and the images they use are often extremely rectangular (e.g., 96 x 1366 pixels[13]). Such image recognition networks cannot be effectively used for normal image recognition, e.g., when identifying different objects in pictures or videos taken by the camera in the mobile device. Most mobile devices will have image recognition of normal (almost) quadratic pictures, and the developers of such devices will use a deep learning image recognition network designed for that purpose, e.g., AlexNet[14] and GoogLeNet[15]. Using the same image recognition network also for classifying environmental sounds would have a number of practical advantages, such as reuse of software and not having to maintain competence in the tuning of different kinds of learning networks. However, there is a research gap concerning the performance and usefulness of deep neural networks, designed for normal object recognition, when it comes to classify Spectrograms and similar sound related images. The aim of this study is to explore that research gap, and see if the same deep neural network can be used for classification of images as well as for environmental sounds.

The rest of this paper is structured in the following way. Section 2 describes the datasets used, and Section 3 the method used in our experiments. The results from the experiments are presented in Section 4. Section 5 compares our results with other studies and discusses practical aspects. The conclusions from our study are presented in Section 6.

2. Datasets

Three publicly available datasets were selected for evaluation of the models: *ESC-50*[16], *ESC-10*[16] and *UrbanSound8K*[17].

The ESC-50 dataset is a collection of 2000 short (5 seconds) environmental recordings comprising 50 equally sized classes of sound events in 5 major groups (animals, natural soundscapes and water sounds, human non-speech sounds, interior/domestic sounds, and exterior/urban noises) prearranged into 5 folds for cross-validation. ESC-10 is a less complex standardized subset with 10 equally sized classes (400 recordings) selected from the ESC-50 dataset (dog bark, rain, sea waves, baby cry, clock tick, person sneeze, helicopter, chainsaw, rooster, fire crackling). UrbanSound8K is a collection of 8732 short (less than 4 seconds) excerpts of various urban sound sources (air conditioner, car horn, playing children, dog bark, drilling, engine idling, gun shot, jackhammer, siren, street music) prearranged into 10 folds.

Since the size of the ESC datasets is too small to use for deep learning, modifications were made to the original data in the form of time-stretching. Piczak[11] expanded the training sound samples by adding random delays and class dependent time stretching to the original recordings. In that study, the number of variations of each sound file in the original dataset were 10 and 4 for the ESC-10 and ESC-50 datasets, respectively. We have used a similar approach, and for ESC-10 and ESC-50 each original audio file was used to produced six additional audio files with varying degrees of time-stretching applied. The signal processing toolbox in Matlab provides a function that resamples an audio file with a new sampling interval which is a factor of the original one. A factor larger than one speeds up the audio and decreases the length of the audio. We used the factors: 0.6, 0.75, 0.9, 1.1, 1.25 and 1.4. As a result, there are seven times as many audio files in the expanded datasets.

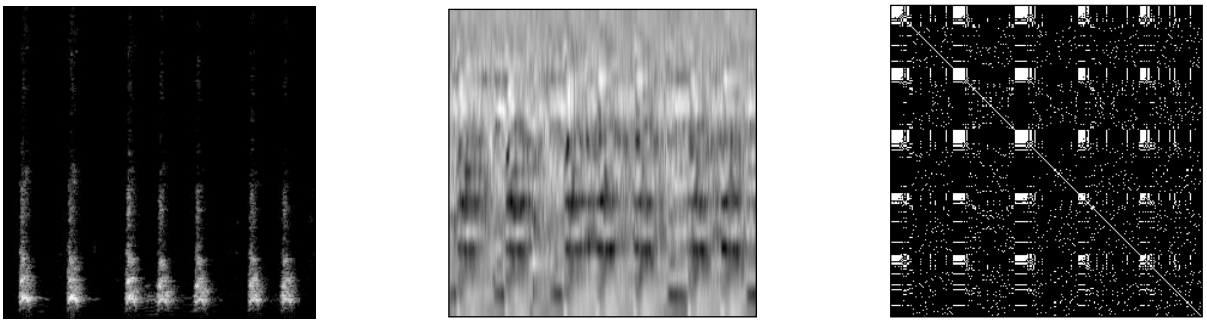
3. Method

3.1. Experimental setup

We used a desktop PC with 12 GB RAM, an Intel Core i7-960 (8 cores @3.20 GHz), and NVIDIA GeForce GTX 970 in the experiments. The PC was running Ubuntu 14.04 LTS, and we used Anaconda Python, and the deep learning frameworks Caffe, TensorFlow (with Keras), and the NVIDIA Deep Learning GPU training system (DIGITS).

The main idea in this paper is to investigate how well sounds can be classified using deep learning networks designed for normal object recognition in images. Audio can be represented in the form of visual images by converting it into Spectrogram, Mel-Frequency Cepstral Coefficients (MFCC), and Cross Recurrence Plot (CRP). Spectrogram is a representation of the energy in the spectrum of frequencies, of a sound, that varies with time. MFCC is a non-linear representation of the power spectrum of a sound adjusted to log scale. A CRP is a matrix visualization where each element represents the distance between the phase trajectories of a time series, such as an audio sample.

The input audio file is converted into a monophonic signal by adding together half the signal amplitudes of each channel in case it is stereophonic. The extraction of Spectrogram was done using in-built function of Matlab. MFCC was extracted using proprietary Matlab code, and CRP was extracted using the CRP toolbox in Matlab.



(a) (b) (c)

Fig. 1. The three types of extracted features (a) Spectrogram, (b) MFCC and (c) CRP of a single sound sample are shown.

The classification accuracy on the test set is chosen as the evaluation metric. The two main test sets, i.e., ESC-10 and ESC-50 are balanced with 40 sound files in each class, and the cost of a misclassification is assumed to be the same for all classes; in these cases accuracy captures the most relevant classification aspects. We used 5-fold cross validation, i.e., the deep network is trained on the training subset of the dataset, and the trained model is then used to predict the samples in the test subset of the dataset according to the 5-fold cross validation method.

We used two well-known deep learning image recognition networks: AlexNet and GoogLeNet. Both of these networks have competed successfully in the ImageNet Challenge (http://image-net.org/challenges/LSVRC/). Our implementations can be found on https://github.com/bkasvenkatesh/Classifying-Environmental-Sounds-with-Image-Networks.

As a baseline approach the feature extraction parameters and the network hyper-parameters were all initialized to the following common values: sampling rate: 32 kHz, frame length: 30ms, frame overlap percentage: 50%, training period: 50 epochs, base learning rate: 0.01, solver type: Stochastic Gradient Descent, learning rate change policy: Exponential decay, and gamma: 0.95. The size of the input image is 256 x 256 pixels with three color channels (RGB).

3.2. Experiment

3.2.1. Initial experiment

Recurrence plots (like CRP) have been used for classifying musical instrument sound[18], but not for classifying environmental sounds. Consequently, we expected that Spectrograms and MFCC were better at classifying environmental sounds than CRP images, but we were not sure. We therefore started by comparing the accuracy using Spectrograms, MFCC, and CPR images for the ESC-50 and ESC-10 datasets.

Unlike Spectrogram and MFCC images, which have many similarities, a CRP image is a totally different concept. The CRP toolbox for Matlab was used to generate the CRP images from the expanded ESC-50 and ESC-10 sound datasets. Computing the recurrence plot is a more complex and hence more time-taking process when compared to Spectrogram and MFCC generation. Hence it is suitable for shorter time series than the 5 seconds audio files in the datasets. In the initial trials of producing the recurrence plots for single audio clips it was observed that a 5 seconds clip took more than a day to process. However, a down-sampled (22.05 kHz) clip of less than 0.7 seconds length took around 3 seconds to process. The short clip had 15000 samples and, as a consequence, the resulting recurrence plot image had a resolution of 15000 x 15000 pixels.

If an even shorter clip having 256 samples were to be used to produce a plot image of size 256 x 256 then there would have been a major loss in the audio signal information. Hence it was decided that a short down-sampled audio clip containing 15000 samples was a reasonable compromise between audio information and computation time.

In order to reduce a 5 seconds sound clips from the dataset to the size required to generate a CRP image, we used an audio event extraction technique similar to the one used by Zhang et al[1], where three high energy frames in the spectrogram are recognized and joined together to create an event-only clip. In our case, however, the five highest energy sound sample points are identified and a window of 3000 samples around each of the five high points is extracted from the clip and joined together creating a clip containing 15000 samples which is then used to produce the recurrence plot. The resulting images 15000 x 15000 images were then scaled down to the appropriate size of 256 x 256 using standard image compression. In this way the CRP image dataset is created and used to train the networks.

These initial results showed, that the accuracy using CRP was very low (see details in Section 4). Because of this we did not use CRP images in our main experiment described below, i.e., only Spectrograms and MFCC images were used in the main experiment.

3.2.2. Main experiment

The main experiment consisted of two steps. First, we investigated how the sampling rate of the audio files affected the classification accuracy. We considered three different sampling rates: 8 kHz, 16 kHz, and 32 kHz. This means that compared to 32 kHz, the high frequencies were filtered out in the case of 8 kHz and 16 kHz. Using Matlab, we obtained the corresponding Spectrogram and MFCC images for all the sound files in the (seven times expanded) ESC-50 and ESC-10 datasets. Based on these images, the classification accuracy for AlexNet and GoogLeNet were evaluated using 5-fold cross validation. We measured the average accuracy and the standard deviation for the five folds.

The experiment with the three frequencies showed that the best classification accuracy was obtained 16 kHz for ESC-50, and for 8 kHz for ESC-10. In the second step of the main experiment, we used the best frequencies when we investigated how the frame length affected the classification accuracy for the two datasets (see Section 3 for details). We considered four different frame lengths: 20 ms, 30 ms, 40 ms, and 50 ms. Since we use 50 frame overlap, we get 2 x 20 = 40 frames per second for 50 ms frame length, i.e., for a 5 seconds sound file we get 200 frames; each frame is represented as one pixel in the *x*-dimension. Consequently, for 50 ms frame length, the last 56 pixels in the *x*-dimension are blank. For 40 ms frame length we get 250 frames, leaving only 6 pixels blank in the *x*-dimension. For 20 ms frame length we get 500 frames, and, as a consequence of the 256 x 256 size of the image, only the first 2.5 seconds of the sound file can be represented in the Spectrogram and MFCC images for this frame length. Again, the classification accuracy for AlexNet and GoogLeNet were evaluated using 5-fold cross validation, and we measured the average accuracy and the standard deviation for the five folds. It turned out that the best accuracy was obtained for a frame length of 50 ms for the ESC-10 dataset and for 40 ms for the ESC-50 dataset (see Section 4 for details).

Using 5-fold cross validation, we evaluated accuracy for the UrbanSound8K dataset. The UrbanSound8K dataset contains 10 classes, just like ESC-10, and we used the optimal settings for ESC-10 (i.e., 8 kHz sampling rate and 50 ms frame length) when evaluating UrbanSound8K. The results are presented in Section 4.

3.2.3. Using color channels

The spectrogram, MFCC and CRP images are black and white. In this experiment, the Spectrogram, MFCC and CRP images were combined, into a single color image. The deep networks used in this study were designed to analyze color images.

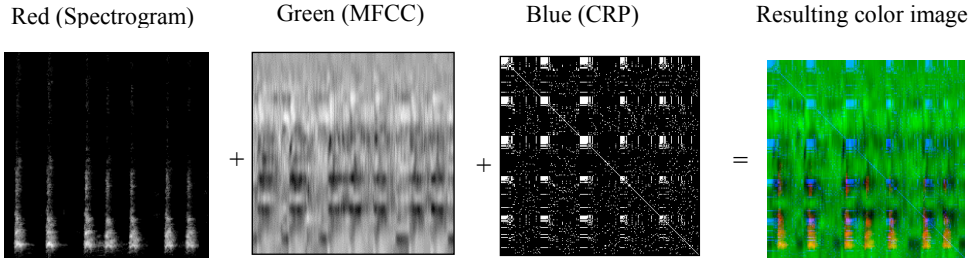


Fig. 2. The combination of three image types to obtain a single color image.

3.2.4. Evaluating Convolutional Recurrent Neural Networks (CRNN)

Convolution recurrent neural networks (CRNN) is a combination of convolutional neural networks (CNN) and recurrent neural networks (RNN). The main advantage of recurrent neural networks is that they can preserve state, which is useful when processing sequences of data, e.g., in video classification[19]. We only use one image for each sound event, and it was therefore not obvious that combining CNN with RNN would improve accuracy. However, CRNNs have shown to be successful for detecting environmental sound events like gun shot, crying baby, and rain in situations where multiple simultaneous sound sources are mixed together[20] (a.k.a. polyphonic sound events). In those cases, CRNN outperformed both CNN and RNN. Therefore, we wanted to compare the classification performance of a convolution recurrent neural network with AlexNet and GoogLeNet.

We implemented a CRNN based on the design made Choi et al[13]. In order to benefit from the ability to preserve state in the CRNN, each image is fed into the network in a number of sequential steps in the same way as Choi et al did. Our implementations can be found on https://github.com/bkasvenkatesh/Classifying-Environmental-Sounds-with-Image-Networks. This network was implemented in a framework called TensorFlow[21]. The original network was designed to accommodate data image size of 96 x 1366 pixels, but we modified the implementation to the image size used in our study (256 x 256). In this experiments we used the same settings for learning rate and number parameters as Choi et al[13] (see Section 4 for details).

4. Results

Table 1 shows the result from the initial experiment. The table shows that the accuracy using CRP images was very low. As a consequence of this we decided to use only Spectrogram and MFCC images in the main experiment.

Table 1. Accuracy obtained in the initial experiment; average from 5-fold cross validation.

Dataset	AlexNet			GoogLeNet		
	Spectrogram	MFCC	CRP	Spectrogram	MFCC	CRP
ESC-10	78.4%	73.0%	28.6%	78.7%	75.9%	27.7%
ESC-50	63.2%	44.9%	12.7%	67.8%	49.1%	10.1%

Table 2 shows the classification accuracy for different sampling rates for AlexNet and GoogLeNet. The table shows that GoogLeNet has higher accuracy than AlexNet in most cases. Table 3 shows the sampling rates for which we obtained the highest average accuracies in Table 2. It turns out that 16 kHz is the best sampling rate for all cases except for ESC-10, Spectrograms, and AlexNet; in that case the best accuracy is obtained for 8 kHz sampling rate.

As mentioned before, the frame length used in Table 2 was 30 ms. In the second part of the main experiment we evaluated different frame lengths. Table 4 shows the classification accuracy for different frame lengths for AlexNet and GoogLeNet; in this part of the experiment we used the sampling frequencies listed in Table 3. Again, it turns out that GoogLeNet has higher accuracy than AlexNet in most cases.

Table 5 shows the highest accuracy values from Table 4, and the corresponding settings. As discussed in the previous section, we also measured the accuracy for the UrbanSound8K dataset. Since that dataset has the same number of classes (10) as ESC-10 we used the same settings for UrbanSound8K as for ESC-10.

Table 2. Accuracy for different sampling rates; average and (standard deviation) from 5-fold cross validation.

Dataset	AlexNet / Sampling rate			GoogLeNet /Sampling rate		
	8 kHz	16 kHz	32 kHz	8 kHz	16 kHz	32 kHz
ESC-10 Spectrograms	82.5% (0.4)	77.8% (0.9)	78.4% (0.6)	85.3% (0.7)	86.2% (0.7)	78.7% (0.8)
ESC-10 MFCC	13.7% (0.7)	77.4% (0.6)	73.0% (0.8)	10.3% (0.5)	80.2% (0.6)	75.9% (0.7)
ESC-50 Spectrograms	67.1% (1.1)	68.7% (0.8)	63.2% (0.9)	68.4% (1.9)	71.7% (0.9)	67.8% (1.6)
ESC-50 MFCC	2.7% (0.3)	46.5% (0.5)	44.9% (1.3)	2.1% (0.3)	53.5% (1.0)	49.1% (0.8)

Table 3. Sampling rates for which we obtained the best accuracy.

Dataset	AlexNet		GoogLeNet	
	Spectrograms	MFCC	Spectrograms	MFCC
ESC-10	8 kHz	16 kHz	16 kHz	16 kHz
ESC-50	16 kHz	16 kHz	16 kHz	16 kHz

Table 4. Accuracy for different frame lengths; average and (standard deviation) from 5-fold cross validation.

Dataset	AlexNet / Frame length				GoogLeNet / Frame length			
	20 ms	30 ms	40 ms	50 ms	20 ms	30 ms	40 ms	50 ms
ESC-10 Spect.	72.6% (0.4)	82.5% (0.4)	83.0% (0.6)	85.6% (0.6)	68.5% (0.4)	86.2% (0.7)	88.7% (0.8)	90.5% (0.5)
ESC-10 MFCC	73.1% (0.6)	77.4% (0.6)	68.1% (0.8)	72.3% (0.7)	76.1% (0.7)	80.2% (0.6)	76.2% (0.6)	75.4% (0.7)
ESC-50 Spect.	66.7% (0.7)	68.7% (0.8)	67.6% (0.7)	65.4% (0.9)	69.2% (0.7)	71.7% (0.9)	73.2% (0.5)	71.5% (0.8)
ESC-50 MFCC	45.5% (1.6)	44.9% (1.9)	45.7% (0.6)	44.7% (1.3)	47.4% (0.8)	53.1% (0.8)	50.3% (0.6)	46.7% (0.6)

Table 5. Highest accuracy values in Table 4 and the accuracy for UrbanSound8K.

Dataset	Settings	AlexNet	GoogLeNet
ESC-10	Spectrogram, 8 kHz sampling frequency	86% (50 ms frame length)	91% (50 ms frame length)
ESC-50	Spectrogram, 16 kHz sampling frequency	69% (30 ms frame length)	73% (40 ms frame length)
UrbanSound8K	Spectrogram, 8 kHz sampling frequency	90% (50 ms frame length)	93% (50 ms frame length)

Table 6 shows the accuracy when combing the Spectrogram, MFCC, and CRP images. We reused the CRP image from the initial experiment for the ESC-10 and ESC-50 dataset, and generated new CRP images for UrbanSound8K. The CRP image for each sound file was combined with the Spectrogram and MFCC images with the settings that generated the highest accuracy, e.g., we combine the CRP with the Spectrogram obtained for16 kHz sample rate and 40 ms frame length and the MFCC obtained for 16 kHz sample rate and 30 ms frame length for ESC-50 when we use GoogLeNet.

Table 6 shows that compared to using Spectrograms with the best settings (see Table 5) there is no gain of combining the Spectrogram, MFCC and CRP images, except for a small improvement for UrbanSound8K when using AlexNet. The images used in Table 5 is a subset (one out of three colour channels) of the images used in Table 6. However, although we have more information when doing the classifications in Table 6, the accuracy is not improved.

Table 7 shows the accuracy when using a convolutional recurrent neural network (CRNN) on the ESC-50 data set. The learning rate was set to 0.01 and the number of parameters is 500000 in this case. The table shows that the accuracy is only 60%, both when we only use Spectrograms and when we combine Spectrograms, MFCC, and CRP images. This accuracy is lower than the values obtained using AlexNet and GoogLeNet.

Table 6. Accuracy when combining spectrograms, MFCC, and CRP images into one color image.

Dataset	AlexNet	GoogLeNet
ESC-10	86%	86%
ESC-50	65%	73%
UrbanSound8K	92%	93%

Table 7. Accuracy using Convolutional Recurrent Neural Networks (CRNN).

Dataset	Spectrogram	Spectrogram, MFCC, and CRP combined
ESC-50	60.3%	60.0%

5. Analysis and Discussion

As discussed in the previous section, the best accuracy on UrbanSound8K dataset is 93% when trained on GoogLeNet using Spectrogram feature with a frame length of 50 ms and a sampling rate of 8 kHz; the best accuracy is 91% for ESC-10 and 73% for ESC-50. These three datasets have been used in a number of evaluations. In the study by Piczak[11] the best average accuracy was 81% for ESC-10, 74% for UrbanSound8K, and 65% for ESC-50. A rather shallow convolutional neural network with four layers was used in that study. Salamon and Bello[22] also use a rather shallow convolutional neural network with five layers. They only used UrbanSound8K in their study and the best average accuracy they got was 79%. UrbanSound8K is also used in an evaluation by Ye et al[23]. They obtained a best average accuracy of 78% using a Mixture of Expert models (MoE). The ESC-10 dataset was used in study by Pillos et al[24]. They obtained a best average accuracy of 74.5% using a multi-layer perceptron and MFCC images. The ESC-10 dataset was also used in an evaluation done by Hertel et al[25]. They used a deep neural network with 14 layers and achieved a best average accuracy of 89.9%.

Based on the related work discussed in the previous paragraph and on the results reported by us, it is clear that the depth of the neural network is crucial; deeper networks give higher accuracy. The most obvious case are the results for ESC-10. Piczak used four layers and got 81% accuracy. AlexNet that was used in our study has eight layers and, for AlexNet, we got an accuracy of 86% for ESC-10, Hertel et al used a CNN with 14 layers and got an accuracy of

89.9%, and we got 91% accuracy using GoogLeNet that has 22 layers. The same trend is also visible for the UrbanSound8K measurements.

Some of the operations done in our experiments were very time consuming. Producing Spectrogram images for the expanded ESC-50 dataset took approximately 5 minutes; the time for producing the corresponding MFCC images was the same, i.e., 5 minutes. However, producing the CRP images took 24 hours. Training AlexNet for 50 epochs on ESC-50 took 34 minutes; the corresponding training took 60 minutes on GoogLeNet. Training 50 epochs on our convolutional recurrent neural network implementation took 48 hours. Classifying 80 images only took about 10 seconds independent on the network (AlexNet, GoogLeNet, or CRNN).

Previous studies have shown that convolutional recurrent neural networks obtain high classification performance when different environmental sounds are mixed with each other[20]. We did not consider the additional challenge that different sounds, such as dog bark and gun shot, can occur at the same time; when sounds are not mixed our study shows that convolutional recurrent neural networks (CRNN) have relatively low accuracy (compare tables 5 and 7). However, we believe that CRNN could be more useful if each sound event was represented as a sequence of images, e.g., by chopping up long Spectrograms[13] to more quadratic images that can be handled by the same image recognition networks that the mobile device uses for normal image recognition.

In our experiments, we used a GPU to accelerate the classification tasks. Classification of sound (and images) is an important functionality in mobile devices. The GPU hardware available in modern mobile devices is becoming very powerful, thus making it realistic to use GPU hardware to accelerate classification also for mobile devices.

6. Conclusions

We have shown that deep convolutional neural networks, which are designed specifically for object recognition in images, can be successfully trained to classify spectral images of environmental sounds. This makes it possible to use the same technology for both object and sound recognition and classification. Most mobile devices contain both cameras and microphones, and companies that develop mobile devices would like to provide functionality for classifying both videos/images and sounds; using the same technology for both these classification tasks would reduce the development cost significantly.

In our main experiment we evaluated different sampling rates, frame lengths, and two deep convolutional neural networks (AlexNet and GoogLeNet) using three publicly available datasets. The best possible classification accuracies on the ESC-50, ESC-10, and UrbanSound8K datasets were 73%, 91%, and 93% respectively with GoogLeNet. GoogLeNet had higher classification accuracy than AlexNet in most of the case that we investigated. We believe that the main reason for this is that GoogLeNet is considerable deeper than AlexNet (22 compared to 8 layers).

Our experiments also showed that the use of convolutional recurrent neural networks did not result in high accuracy.

We evaluated the classification accuracy for three image representations of sound – Spectrograms, Mel-Frequency Cepstral Coefficients (MFCC), and Cross Recurrence Plot (CRP). In most cases, we obtained the highest classification accuracy when using Spectrograms. The prediction accuracy for CRP was very low. Combining Spectrograms, MFCC, and CRP as different color channels of the same image did not improve the classification accuracy.

All the programs used in our experiments are available on the Internet, making it possible to repeat our experiments or do the same experiments on other datasets.

References

1. H. Zhang, I. McLoughlin, and Y. Song, “Robust sound event recognition using convolutional neural networks,” in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP),* 2015, pp. 559–563.
2. J.-C. Wang, H.-P. Lee, J.-F. Wang, and C.-B. Lin, “Robust Environmental Sound Recognition for Home Automation,” in *IEEE Transactions on Automation Science and Engineering*, vol. 5, no. 1, pp. 25–31, Jan. 2008.
3. O. Gencoglu, T. Virtanen, and H. Huttunen, “Recognition of acoustic events using deep neural networks,” in *Signal Processing Conference (EUSIPCO), 2014 Proceedings of the 22nd European*, 2014, pp. 506–510.
4. D. Barchiesi, D. Giannoulis, D. Stowell, and M. D. Plumbley, “Acoustic Scene Classification: Classifying environments from the sounds they produce,” in *IEEE Signal Processing Magazine*, vol. 32, no. 3, pp. 16–34, May 2015.
5. S. Chachada and C.-C. Jay Kuo, “Environmental Sound Recognition: A Survey,” in *Proceedings of the 2013 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference*, 2013.

6. I. McLoughlin, H. Zhang, Z. Xie, Y. Song, and W. Xiao, "Robust Sound Event Classification Using Deep Neural Networks," in *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 23, no. 3, pp. 540–552, Mar. 2015.
7. J.-C. Wang, C.-H. Lin, B.-W. Chen, and M.-K. Tsai, "Gabor-Based Nonuniform Scale-Frequency Map for Environmental Sound Classification in Home Automation," in *IEEE Transactions on Automation Science and Engineering*, vol. 11, no. 2, pp. 607–613, Apr. 2014.
8. L. Lu, H.-J. Zhang, and S. Z. Li, "Content-based audio classification and segmentation by using support vector machines," in *Multimedia Systems*, vol. 8, no. 6, pp. 482–492, Apr. 2003.
9. M. M. Mostafa and N. Billor, "Recognition of Western style musical genres using machine learning techniques," in *Expert Systems with Applications*, vol. 36, no. 8, pp. 11378–11389, Oct. 2009.
10. Z. Zhang and B. Schuller, "Semi-supervised learning helps in sound event classification," in *Acoustics, Speech and Signal Processing (ICASSP), 2012 IEEE International Conference on*, 2012, pp. 333–336.
11. K. J. Piczak, "Environmental sound classification with convolutional neural networks," in *Proceedings of the IEEE 25th International Workshop on Machine Learning for Signal Processing (MLSP),* 2015, pp. 1–6.
12. S. Sigtia, A. M. Stark, S. Krstuloviv, M. D. Plumbley, "Automatic environmental sound recognition: performance versus computational cost," in *IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP)*, 11/2016, Vol. 24, No. 11.
13. K. Choi, G. Fazekas, M. Sandler, and K. Cho, "Convolutional Recurrent Neural Networks for Music Classification," *arXiv preprint arXiv:1609.04243*, 2016.
14. A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, 2012, pp. 1097–1105.
15. C. Szegedy et al., "Going deeper with convolutions," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 1–9.
16. K. J. Piczak, "ESC: Dataset for environmental sound classification," in *Proceedings of the ACM International Conference on Multimedia*. ACM, 2015.
17. J. Salamon, C. Jacoby, and J. P. Bello, "A dataset and taxonomy for urban sound research," in *Proceedings of the ACM International Conference on Multimedia*. ACM, 2014, pp. 1041–1044.
18. T. Park and T. Lee, "Musical instrument sound classification with deep convolutional neural network using feature fusion approach," *arXiv preprint arXiv:1512.07370*, 2015.
19. Z. Xu, J. Hu, and W. Deng, "Recurrent Convolutional Neural Network for Video Classification," in *Proceedings of the IEEE International Conference on Multimedia and Expo (ICME)*, 2016.
20. E. Cakir, G. Parascandolo, T. Heittola, H. Huttunen, and T. Virtanen, "Convolutional recurrent neural networks for polyphonic sound event detection," in *IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP)*, 01/2017, Vol. 25, No. 1.
21. M. Abadi et al., "Tensorflow: Large-scale machine learning on heterogeneous distributed systems," *arXiv preprint arXiv:1603.04467*, 2016.
22. J. Salamon and J. P. Bello,"Deep convolutional neural networks and data augumentation for environmental sound classification," *arXiv preprint arXiv:1608.04363*, 2016.
23. J. Ye, T. Kobayashi, and M. Murakawa, "Urban sound event classification based on local and global features aggregation," *in Applied Acoustics, 117,* 2017, pp. 246-256.
24. A. Pillos, K. Alghamidi, N. Alzamel, V. Pavlov, and S. Machanavajhala, "A real-time environmental sound recognition system for the Android OS," in *Proceedings of Detection and Classification of Acoustic Scenes and Events,* September 2016, Budapest.
25. L. Hertel, H. Phan, and A. Mertins, "Comparing time and frequency domain for audio event recognition using deep learning," in *Proceedings of the IEEE International Joint Conference on Neural Networks (IJCNN)*, July 2016, Vancouver, Canada.