A Concise Review on Classification of Sound using Deep Learning

Dr. Ajay Saini[1], Aashish Bhaskar2, Ayush Bhaskar[3], Ankita Bhaskar[4]
[1]Associate Professor, Arya Institute of Engineering and Technology,Jaipur, Rajasthan
[2, 3] Research Scholar, BITS Pilani, Rajasthan
[4] Research Scholar
[1]erajaysaini@gmail.com, [2]2020sc04100@wilp.bits-pilani.ac.in, [3]2021FC04271@wilp.bits-pilani.ac.in, [4]ankitabhaskar0309@gmail.com

Article Info
Volume 83
Page Number: 85 – 95
Publication Issue:
November/December 2020

Article History
Article Received: *25 October 2020*
Revised *22 November 2020*
Accepted: *10 December 2020*
Publication: *31 December 2020*

ABSTRACT

Sound classification is a rapidly expanding field of study. Speech-processing apps like Amazon Alexa, Google Home, Siri, and others are essential for supporting us in our daily lives. Other uses of sound categorization include surveillance using sounds such as gunshot detection, which can assist law enforcement authorities in dispatching assistance as soon as such behaviour is discovered.

The goal of this study is to build a machine learning model to correctly classify a set of urban sounds from a sound recording. A machine learning model was created and trained to classify ten various types of sounds found in an urban setting.

Keyword*: Sound Classification, Deep Learning, CNN, ANN*

1. INTRODUCTION

Deep learning is a machine learning technique that allows computers to learn by example in the same way that humans do. Deep learning is a critical component of self-driving automobiles, allowing them to detect a stop sign or discriminate between a pedestrian and a lamppost. It enables voice control in consumer electronics such as phones, tablets, televisions, and hands-free speakers. Deep learning has gotten a lot of press recently, and with good cause. It's accomplishing accomplishments that were previously unattainable. Deep learning (also known as deep structured learning) is part of a broader family of machine learning methods based on artificial neural networks with representation learning. Learning can be supervised, semi-supervised or unsupervised.

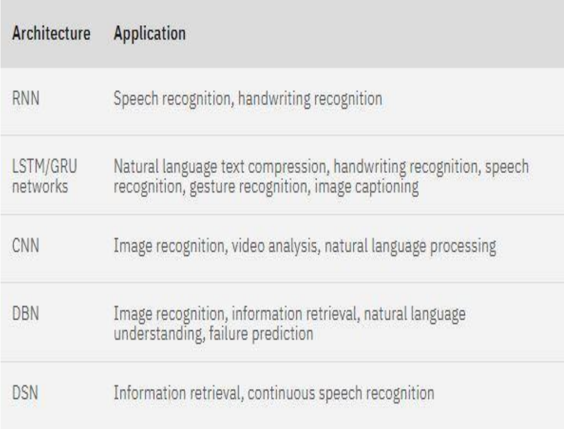
Deep learning architectures such as deep neural networks, deep belief networks, recurrent neural networks and convolutional neural networks have been applied to fields including computer vision, speech recognition, natural language processing, audio recognition, social network filtering, machine translation, bioinformatics, drug design, medical image analysis, material inspection and board game programs, where they have produced results comparable to and in some cases surpassing human expert performance. We study Neural networks have been used for implementing language models since the early 2000s.LSTM helped to improve machine translation and language modeling.

Other key techniques in this field are negative sampling and word embedding. Word embedding, such as word2vec, can be thought of as a representational layer in a deep learning architecture that transforms an atomic word into a positional representation of the word relative to other words in the dataset; the position is representedas a point in a vector space. Using word embedding as an RNN input layer allows the network to parse sentences and phrases using an effective compositional vector grammar. A compositional vector grammar can be thought of as probabilistic context free grammar (PCFG) implemented by an RNN. Recursive auto-encoders built a top word embeddings can assess sentence similarity and detect paraphrasing. Deep neural architectures provide the best results for constituency parsing, sentiment analysis, information retrieval, spoken language understanding, machine translation, contextual entity linking, writing style recognition, Text classification and others.

2.ARCHITECTURE OF DEEP LEARNING

Connectionist architectures have existed for more than 70 years, but new architectures and graphical processing units (GPUs) brought them to the forefront of artificial intelligence. The last two decades gave us deep learning architectures, which greatly expanded the number and type of problems neural networks can address. This article introduces five of the most popular deep learning architectures—recurrent neural networks (RNNs), long short-term memory (LSTM)/gated recurrent unit (GRU), convolutional neural networks (CNNs), deep belief networks (DBN), and deep stacking networks (DSNs)—and then explores open source software options for deep learning.

Deep learning isn't a single strategy, but rather a collection of algorithms and topologies that can be used to solve a wide range of problems. While deep learning is not a new concept, it is exploding due to the convergence of deeply layered neural networks and the use of GPUs to expedite their execution. This expansion has also been fueled by big data. Because supervised learning algorithms (those that train neural networks with example data and reward them based on their success) are used in deep learning, the more data available, the better. The structures and algorithms utilised in deep learning are numerous and diverse. This section looks at five deep learning architectures that have been developed over the last 20 years. Notably, LSTM and CNN are two of the oldest approaches in this list but also two of the most used in various applications.



Architecture	Application
RNN	Speech recognition, handwriting recognition
LSTM/GRU networks	Natural language text compression, handwriting recognition, speech recognition, gesture recognition, image captioning
CNN	Image recognition, video analysis, natural language processing
DBN	Image recognition, information retrieval, natural language understanding, failure prediction
DSN	Information retrieval, continuous speech recognition

Figure 1 Different Architectures and their applications for Deep learning

3 Application of Deep Learning

- • **Toxicology and drug discovery -** A substantial percentage of potential medications are rejected by regulators. Insufficient efficacy (on-target impact), unwanted interactions (off-target effects), or unanticipated hazardous consequences are several reasons for failure. Deep learning has been used to anticipate bimolecular targets, off-targets, and hazardous consequences of environmental chemicals in foods, home items, and pharmaceuticals.
- • **Customer relationship management (CRM) -** Deep reinforcement learning was used to estimate the value of potential direct marketing actions, which were characterised in terms of RFM variables. Customer lifetime value was discovered to be a natural interpretation of the estimated value function.
- **Recommendation systems -** For content-based music and journal recommendations, recommendation systems have used deep learning to extract meaningful features for a latent factor model. For learning user preferences across many domains, multi-view deep learning has been used. The methodology improves recommendations in a variety of tasks by combining a collaborative and content-based approach.
- **Bioinformatics -** An autoencoder ANN was used in bioinformatics, to predict gene ontology annotations and gene-function relationships.
- **Medical Image Analysis -** Deep learning has been shown to produce competitive results in medical application such as cancer cell classification, lesion detection, organ segmentation and image enhancement.
- **Mobile advertising -** Finding the appropriate mobile audience for mobile advertising is always challenging, since many data points must be considered and analyzed before a target segment can be created and used in ad serving by any ad server. Deep learning has been used to interpret large, many-dimensioned advertising datasets. Many data points are collected during the request/serve/click internet advertising cycle. This information can form the basis of machine learning to improve ad selection.
- **Image restoration -** Deep learning has been

used to solve inverse problems including denoising, super-resolution, inpainting, and film colorization with great success. Learning approaches such as "Shrinkage Fields for Effective Image Restoration," which trains on an image dataset, and "Deep Image Prior," which trains on the image that needs restoration, are examples of these applications.

- **Financial fraud detection** - Deep learning is being successfully applied to financial fraud detection and anti-money laundering. "Deep anti-money laundering detection system can spot and recognize relationships and similarities between data and, further down the road, learn to detect anomalies or classify and predict specific events". The solution leverages both supervised learning techniques, such as the classification of suspicious transactions, and unsupervised learning, e.g. anomaly detection.
- **Military** - The United States Department of Defense applied deep learning to train robots in new tasks through observation.

4 .CNN

A CNN is a multilayer neural network that was biologically inspired by the animal visual cortex. The architecture is particularly useful in image-processing applications. The first CNNwas created by Yann LeCun; at the time, the architecture focused on handwritten character recognition, such as postal code interpretation. As a deep network, early layers recognize features (such as edges), and later layers recombine these features into higher-level attributes of the input.

The LeNet CNN architecture consists of various layers that perform feature extraction and classification (see the following image). The input image is separated into receptive fields, which feed into a convolutional layer, which extracts features from it. Pooling is the following stage, which decreases the dimensionality of the retrieved features (by down-sampling) while keeping the most significant data (typically through max pooling). The data is subsequently sent into a fully connected multilayer perceptron after another convolution and pooling stage. This network's final output layer is a collection of nodes that identify image features (in this case, a node per identified number). Back-propagation is used to train the network.

Deep processing layers, convolutions, pooling, and a fully linked classification layer opened the door to a slew of new deep learning neural network applications. The CNN has been successfully applied to video identification and numerous tasks within natural language processing in addition to image processing.

Recent applications of CNNs and LSTMs produced image and video captioning systems in which an image or video is summarized in natural language. The CNN implements the image or video processing, and the LSTM is trained to convert the CNN output into natural language

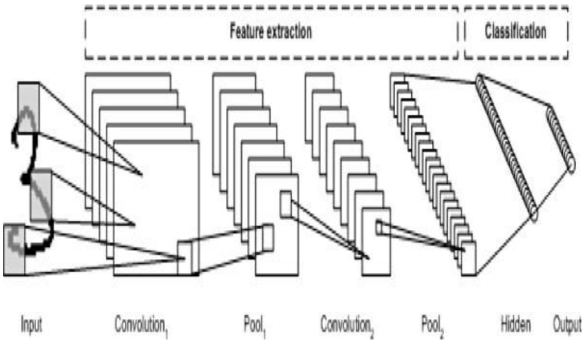


Figure 2 Architecture of CNN

5.SOUND CLASSIFICATION

Automatic environmental sound classification is a growing area of research with numerous real world applications. Whilst there is a large body of research in related audio fields such as speech and music, work on the classification of environmental sounds is comparatively scarce.

Likewise, observing the recent advancements in the field of image classification where convolution neural networks are used to to classify images with high accuracy and at scale, it begs the question of the applicability of these techniques in other domains, such as sound classification.

Sound recognition technologies are used for:

- Music recognition
- Speech recognition
- the automatic alarm detection and identification for surveillance, monitoring systems,based on the acoustic environment
- the assistance to disabled or elderly people affected in their hearing capabilities
- identifying species of fish and mammals in acoustical oceanography
- smart safety solutions[buzzword]

In monitoring and security solutions[buzzword], an important contribution to alarm detection and alarm verification can be supplied, using sound recognition techniques. In particular, these methods could be helpful for intrusion detection in places like offices, stores, private homes or for the supervision of public premises exposed to person aggression. In all these cases, a recognition system can report about a danger or distress event. It could further identify sounds like glass break, doorbells, smoke detector alarms, red alerts, human screams, baby cries and others. Sometimes, the alarm is triggered by other of detectors (e.g. temperature or video- based) and the sound recognizer would be associated to these other modalities, to verify the alarm, with the purpose of decreasing the global false alarm detection rate.

Solutions based on a sound recognition technology can offer assistance to disabled and elderly people affected in hearing capabilities, helping them to keep or recover some independence in their daily occupations.

There are only a handful of companies who are working on sound recognition technology:

- AbiliSense (checking sounds of the home and city environment).
- Audio Analytic (AI company who's "Embedded sound recognition AI software gives consumer technology, such as smart speakers, hearables, smart home tech, mobile phones and automotive, a sense of hearing." has sound recognition software that makes consumer products more intelligent).
- OtoSense (checking sounds of engines).
- Wavio -software and product innovation company providing sound recognition solutions to clients such as product manufacturers, organizations, and government inclusive of accessibility for Deaf & hard of hearing, protected by sound recognition patents covering sound

recognition to notify users of detected sounds automatically

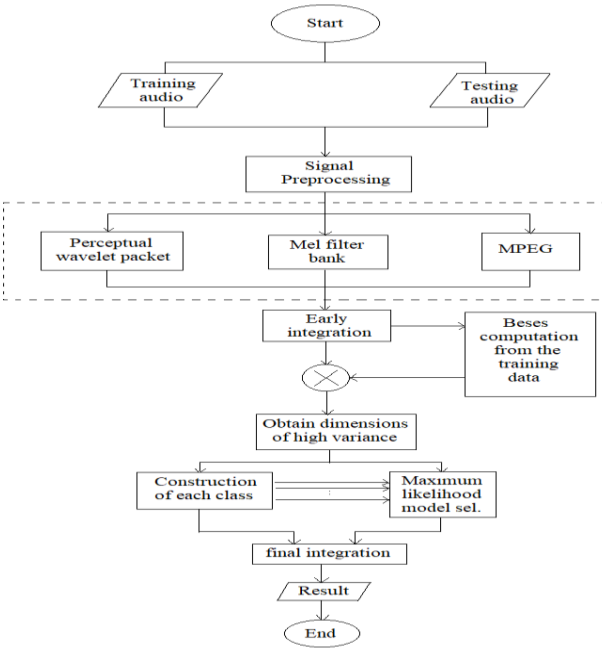

Start
Training audio
Testing audio
Signal Preprocessing
Perceptual wavelet packet
Mel filter bank
MPEG
Early integration
Beses computation from the training data
Obtain dimensions of high variance
Construction of each class
Maximum likelihood model sel.
final integration
Result
End

Figure 3 Data Flow Diagram Classification Process with deep learning

The data used for this research review paper is Kaggle's Urban Sound Classification dataset uploaded sound excerpts of less than or equal to 6 seconds in.wav format. These files have sounds of eight categories:

a) Air Conditioner.
b) Car Horn.
c) Children Playing.
d) Dog bark.
e) Drilling.
f) Engine Idling.
g) Gun Shot.
h) Jackhammer.
i) Siren.
j) Street Music.

In our research dataset is divided into Train and Test folder and comes with 2 .CSV file that contains two columns for Train dataset:

- ID: A distinct ID of each sound file.
- Class: Classification of sound file

The Test dataset .CSV file contains only the ID column, representing distinct ID of sound files in Test folder, the Class is to be predicted by the algorithm.

In deep learning, a computer model learns to perform classification tasks directly from images, text, or sound. Deep learning models can achieve state-of-the-art accuracy, sometimes exceeding human-level performance. Models are trained by using a large set of labeled data and neural network architectures that contain many layers

6.FEATURE EXTRACTION

The most crucial step in creating a machine learning model is feature extraction. Selecting relevant attributes aids the model's performance. We translate the time-domain signal to the mel-frequency scale, which is extensively used in speech processing, to classify sound. Here's a quick rundown of some of the terms used in sound processing.

Cepstrum

A cepstrum is the result of taking the inverse Fourier transform (IFT) of the alogarithm of the estimated spectrum of a signal.

Mel-frequency cepstrum (MFC)

In sound processing, the mel-frequency cepstrum (MFC) is a representation of the short-term power spectrum of a sound, based on a linear cosine transform of a log power spectrum on a nonlinear mel scale of frequency.

Mel-Frequency Cepstral Coefficient (MFCC)

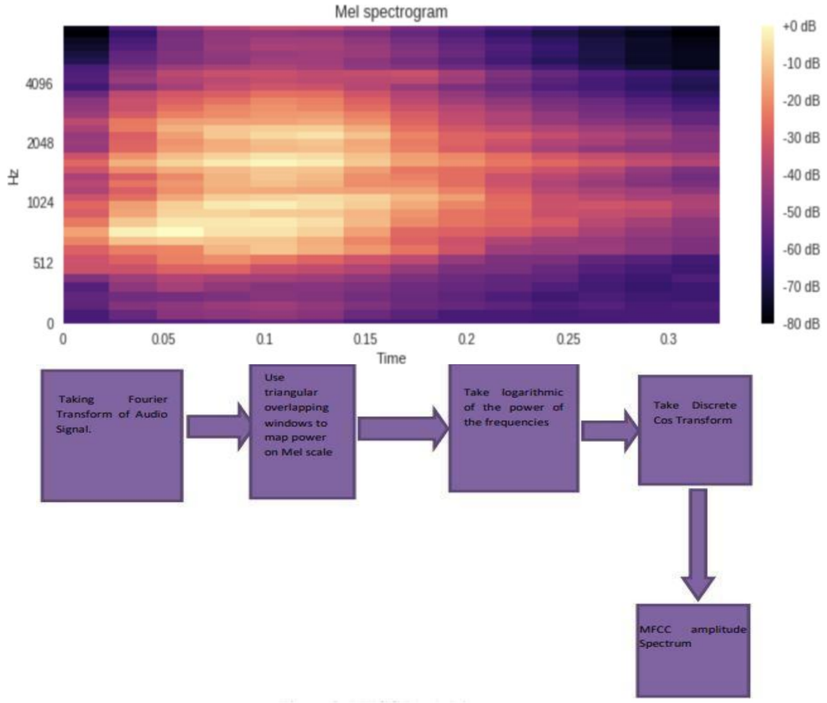

Mel spectrogram
4096
2048
Hz
1024
512
0
0
0.05
0.1
0.15
0.2
0.25
0.3
Time
+0 dB
-10 dB
-20 dB
-30 dB
-40 dB
-50 dB
-60 dB
-70 dB
-80 dB
Taking Fourier Transform of Audio Signal.
Use triangular overlapping windows to map power on Mel scale
Take logarithmic of the power of the frequencies
Take Discrete Cos Transform
MFCC amplitude Spectrum

Figure 4 Block diagram of MFCC

The MFCCs (mel-frequency cepstral coefficients) are the coefficients that make up an MFC. They're made up of a cepstral representation of the audio clip (a nonlinear "spectrum-of-a-spectrum"). The distinction between the cepstrum and the mel-frequency cepstrum is that the frequency bands in the MFC are evenly separated on the mel scale, which more closely approximates the human auditory system's response than the linearly spread frequency bands used in the standard cepstrum. This frequency warping can help with sound representation in audio compression, for example.

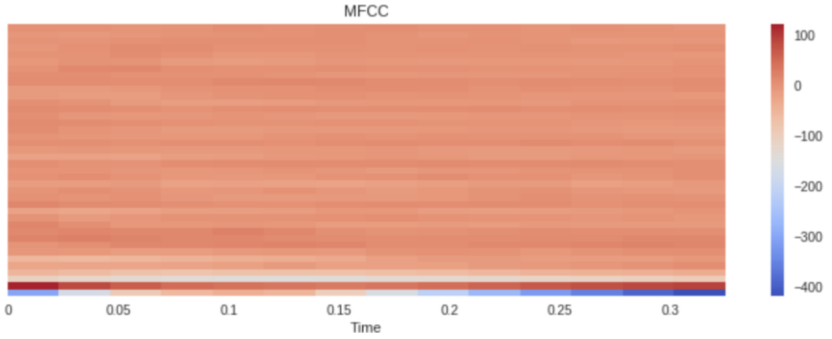


Figure 5 MFCC of a dog bark

Melspectrogram

An object of type MelSpectrogram represents an acoustic time-frequency representation of a sound: the power spectral density P(f, t).It is sampled into a number of points around equally spaced times ti and frequencies fj (on a Mel frequency scale).

The mel frequency scale is defined as: mel = 2595 * log10 (1 + hertz / 700**)**

Figure 6 Melspectrogram of dog bark

7.METHODOLOGY

For sound classification, we use Python-based library called "librosa" to generate the images of melspectrogram of each sound excerpt and use 2-D Convolutional Neural Network to makea sound classifier. First, we read the sound files and use the following function to create spectrogram images and save them in a folder.

For training our model, we load the saved melspectrogram images and convert them to 64x64 pixels before feeding in batches of 32 images to our neural network. This is shown in Python code below:

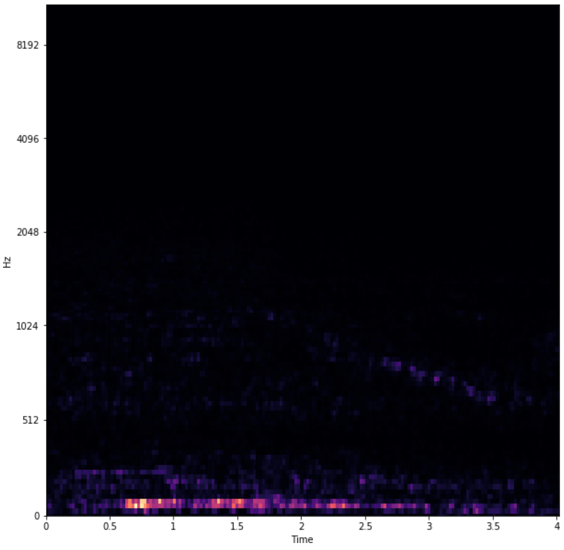


Figure 7 Melspectrogram of sound of siren.

For training our model, we load the saved melspectrogram images and convert them to 64x64 pixels before feeding in batches of 32 images to our neural network. It shows the neural network configuration for our sound classifier. We used Adam optimizer and categorical crossentropy loss as this is a multi-class classificationproblem.

8.RESULTS

At the end of 120 EPOCHS researcher, the following parameters are observed: Training loss: 0.3144, Training Accuracy: 88.51%, Validation Loss: 0.7127 – Validation Accuracy: 79.62%.

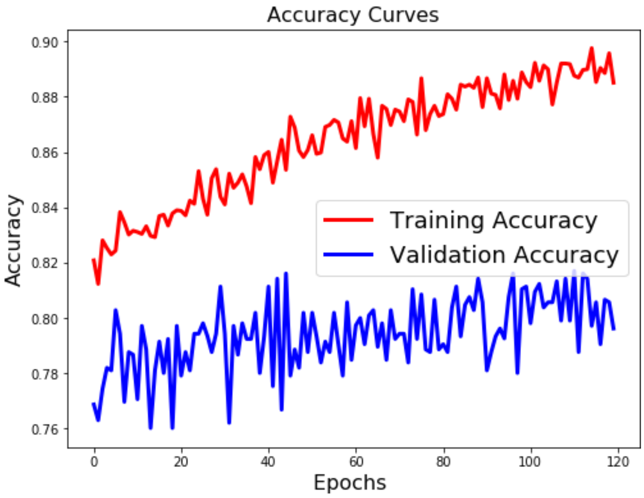


Figure 8 Training and validation Loss (left) and accuracy (right) curves of model

Features of Sound Classification

- Convolution neural networks can classify sound clips to a high degree of accuracy through the use of image representations.
- Complex networks with more filters in

later layers outperform simpler ones whenworking with similar images.

- The amount of data is key to improving classification accuracy, particularly with similar images
- It can be used for Music classification
- Understanding sounds in environment for wildlife prevention etc.
- Sound surveillance for terrorist activity response

9.CONCLUSION

Sound categorisation is crucial and may be applied to a variety of situations to make life easier. This study explains how to use a Convolutional Neural Network to classify sounds into ten different categories and how well it categorises sound recordings.

Future work on it could include sounds of individuals disagreeing, which could result in a fight, as well as gunshots or explosions. Installing such devices to notify authorities can assist law enforcement agencies in dispatching immediate assistance in the event of a disruption of law and order. Other uses for this project include music genre classification, recognising animal species for wildlife preservation, and so on. As a result, instead of relying on people to classify noises, a machine learning algorithm can do it flawlessly and make our lives easier.

The purpose of this study is to see how well CNN architecture can categorise sound signals based on spectrograms of the sound spectrum. Convolutional Neural Networks are commonly used to solve picture categorization issues. This research study demonstrates how deep neural architectures may be used to classify sounds. In comparison to direct sound classification, this strategy using CNN for sound classification using spectrograms reduced the amount of trainable parameters. In comparison to other existing approaches, we obtained Training Accuracy: 88.51 percent and Validation Accuracy: 79.62 percent in the experimental tests conducted using CNN. This approach provides promising results for the creation of sound categorization systems in key regions, according to the results of the trial. The possible question for future work is whether tensor deep stacking network could be efficiently used with CNN to classify the sound signals. The power of tensors can be utilized to train the network on high definition images instead of compressed images.

10.FUTURE SCOPE

Further experiments with this data and our feature learning algorithms could go in a variety of directions. There are a few basic techniques to enhance categorization accuracy without making major changes to our algorithms. A crucial initial step would be to further evaluate and confirm our dataset. Our folds were built on a variety of assumptions; we can divide our segments into four second slices with a two-second hop size, but we have no means of knowing for sure whether the annotated sound remains in the sliced segment after automatic slicing. We must demonstrate that every slice derived from a segment contains enough of the original sound to be classified. Unfortunately, this process will be far from easy; this would likely require an extensive manual effort in the form of crowd-sourced annotations. While this may not improve accuracy relative to the baseline system, it may increase the accuracy of the entire system, if negatives are being generated because the sound in question is not actually present in the labeled slice. In addition, we could try various other pre-processing techniques to try to improve our data, such as automatic gain control or other loudness normalizations.

Secondly, there is the issue of noise. Both the baseline and the feature learning algorithms perform quite poorly on our background salience slices. We saw several other approaches that people have used to build features robust to noise, and many of them would be easily transferable to the system we have already established. Next, there is the issue of temporal dynamics of sound, and how to handle classification using data from different timescales. In our work, we handled the time dimension by summarizing our features over the entire slice file. Due to the specific problems with noise in this field, we feel that in implementing a deep learning system for urban sound classification, it would be necessary to carefully design each layer from the ground up. This would allow us to construct a system tailored to our specific classification task for urban sounds.

REFERENCES

[1]. Bengio, Yoshua (2009). "Learning Deep Architectures for AI" (PDF). Foundations and Trends in Machine Learning. 2 (1): 1–127. CiteSeerX 10.1.1.701.9550. doi:10.1561/2200000006. Archived from the original (PDF) on 2016-03-04. Retrieved 2015-09-03.

[2]. Ciresan, D.; Meier, U.; Schmidhuber, J. (2012). "Multi-column deep neural networks for image classification". 2012 IEEE Conference on Computer Vision and Pattern Recognition. pp. 3642–3649. arXiv:1202.2745. doi:10.1109/cvpr.2012.6248110. ISBN 978-1-4673-1228-8.

[3]. https://www.mathworks.com/discovery/deep-learning.html

[4]. Graves, Alex; Eck, Douglas; Beringer, Nicole; Schmidhuber, Jürgen (2003). "Biologically Plausible Speech Recognition with LSTM Neural Nets" (PDF). 1st Intl. Workshop on Biologically Inspired Approaches to Advanced Information Technology, Bio-ADIT 2004, Lausanne, Switzerland. pp.

[5]. https://developer.ibm.com/technologies/artificial-intelligence/articles/cc-machine-learning-deep-learning-architectures/

[6]. Sound classification Retrieved from http://www.paroc.com/knowhow/sound/sound-classification, 2017.

[7]. https://research.steinhardt.nyu.edu/scmsAdmin/media/users/ec109/MTT-14-01-013.pdf

[8]. Koul, Deepali, and Satish Kumar Alaria. "A new palm print recognition approach by using PCA & Gabor filter." *International Journal on Future Revolution in Computer Science & Communication Engineering* 4.4 (2018): 38-45.

[9]. Piyusha and Satish Kumar Alaria” Feature Based Sentiment Analysis on Movie Review Using SentiWordNet”,

International Journal on Recent and Innovation Trends in Computing and Communication (ISSN: 2321-8169), Vol. 4 Issue-9 (2016): 12-15.

[10]. Dogiwal, Sanwta Ram, et al. "Image Preprocessing Methods in Image Recognition." *International Journal of Computers and Distributed Systems* 1.3 (2012): 96-99.

[11]. [11] R. Radhakrishnan, A. Divakaran, and A. Smaragdis, "Audio analysis for surveillance applications," in IEEE Workshop on Applications of Signal Processing to Audio and Acoustics, 2005. IEEE, 2005, pp. 158–161. [Online]. Available: https://doi.org/10.1109/ASPAA.2005.1540194

[12]. [12] M. Cristani, M. Bicego, and V. Murino, "Audio-visual event recognition in surveillance video sequences," IEEE Transactions on Multimedia, vol. 9, no. 2, pp. 257–267, 2007. [Online]. Available: https://doi.org/10.1109/TMM.2006.886263

[13]. [13] S. Chu, S. Narayanan, C.-C. J. Kuo, and M. J. Mataric, "Where am i? scene recognition for mobile robots using audio features," in 2006 IEEE International conference on multimedia and expo. IEEE, 2006, pp. 885– 888. [Online]. Available: https://doi.org/10.1109/ICME.2006.262661

[14]. [14] R. Bardeli, D. Wolff, F. Kurth, M. Koch, K.-H. Tauchert, and K.-H. Frommolt, "Detecting bird sounds in a complex acoustic environment and application to bioacoustic monitoring," Pattern Recognition Letters, vol. 31, no. 12, pp. 1524–1534, 2010. [Online]. Available: https://doi.org/10.1016/j.patrec.2009.09.014

[15]. [15] C. Mydlarz, J. Salamon, and J. P. Bello, "The implementation of low-cost urban acoustic monitoring devices," Applied Acoustics, vol. 117, pp. 207–218, 2017. [Online]. Available: https://doi.org/10.1016/j.apacoust.2016.06.010

[16]. [16] D. Steele, J. Krijnders, and C. Guastavino, "The sensor city initiative: cognitive sensors for soundscape transformations," GIS Ostrava, pp. 1– 8, 2013.

[17]. [17] V. Davidovski, "Exponential innovation through digital transformation," in Proceedings of the 3rd International Conference on Applications in Information Technology. ACM, 2018, pp. 3–5. [Online]. Available: https://doi.org/10.1145/3274856.3274858

[18]. [18] F. Tappero, R. M. Alsina-Pages, L. Duboc, and F. Al ` ´ıas, "Leveraging urban sounds: A commodity multi-microphone hardware approach for sound recognition," in Multidisciplinary Digital Publishing Institute Proceedings, vol. 4, no. 1, 2019, p. 55. [Online]. Available: https://doi.org/10.3390/ecsa-5-05756

[19]. [19] E. Pyshkin, "Designing human-centric applications: Transdisciplinary connections with examples," in 2017 3rd IEEE International Conference on Cybernetics (CYBCONF). IEEE, 2017, pp. 1–6. [Online]. Available: https://doi.org/10.1109/CYBConf.2017.7985774

[20]. [20] E. Pyshkin and A. Kuznetsov, "Approaches for web search user interfaces-how to improve the search quality for various types of information," JoC, vol. 1, no. 1, pp. 1–8, 2010. [Online]. Available: https://www.earticle.net/Article/A188181

[21]. [21] M. B. Dias, "Navpal: Technology solutions for enhancing urban navigation for blind travelers," tech. report CMU-RI-TR-21, Robotics Institute, Carnegie Mellon University, 2014.