Exploring Deep Neural Networks for Audio Interaction Recognition

Sokolo D. Matsolo

[1] University of Johannesburg, Auckland Park, Johannesburg, South Africa
[2] 222206953@student.uj.ac.za

Abstract. In this paper, we investigate the application of deep neural networks (DNNs) to audio-based interaction identification. The EPIC-SOUNDS dataset, established for audio detection and classification in kitchen-recorded videos, is used in this study. The study describes the difficulties in identifying audio events, with a focus on problems like dataset constraints and noise interference, especially for brief audio events. Convolutional neural networks (CNNs) with overlapping layers and learning rate step decay properties in particular are shown to have several benefits when it comes to audio categorization. Additionally, information about the experimental setup—including data balancing and neural network architecture—is provided. The results are also reported in the publication, with a focus on performance measures and training settings. Ultimately, the findings demonstrate the potential of DNNs for audio interaction recognition and suggest avenues for further research.

Keywords: Deep neural networks · audio interaction recognition · convolutional neural networks · EPIC-SOUNDS

1 Introduction

A fascinating and difficult problem in data classification solutions has sparked a wave of creative thought in various fields. This problem has sparked the creation of innovative ideas that cut across numerous industries. Significant research efforts have focused on the field of audio event categorization and detection (AEC/D), revealing a deeply challenging yet compelling subject as our technological landscape continues to change [10]. The main objective of AEC/D is the recognition of various audio signals, ranging from spoken words to background noise, in large, unstructured audio streams. The concept of "Audio-Based Interaction Recognition," which focuses on identifying and annotating interactions predominantly driven by audio cues, emerged out of AEC/D. A wide range of sound forms can be decoded, categorized, and understood by algorithms because of the application of modern approaches in audio-based interaction recognition to real-world businesses [14]. Innovative contributions like these have led to practical solutions made specifically for different infrastructures. For instance, they play a key role in the creation of systems that analyze the sound quality (SQ) of electric vehicles (EVs), improve pedestrian security, and help the blind [13].

There are two distinct streams in the human auditory system, according to convincing research from neuroscience. These two streams are known as the dorsal stream, which is in charge of locating sound-emitting objects, and the ventral stream, which specializes in identifying them. The ventral stream excels in spectral resolution, enabling accurate object identification, while the dorsal stream operates at a higher temporal resolution and functions at a higher sampling rate, facilitating the task of object localization[9]. Using this evidence as the driving force for designing deep audio learning architectures. It has become obvious that event identification has many benefits, but it also faces ongoing difficulties brought on by issues like noise interference and overlap, which could affect the accuracy of study findings. Also, dealing with the lack of datasets, especially for audio events lasting under 2 milliseconds, is a significant barrier [11]. This can be seen in a number of audio-based event classification datasets, such as AudioSet [7], [3], and the more recent EPIC-SOUNDS [8]. These datasets display a variety of complexity, including the presence of audio segments with variable time lengths linked to distinct activities and the inclusion of background noises that correspond to the events' connected behaviors.

Deep learning emerged primarily in the field of image processing and then gained popularity in a variety of other fields, such as recommendation systems, natural language processing, audio processing, and even drug discovery. Standard approaches to audio data handling, like Gaussian mixture models and non-negative matrix factorization, performed worse than deep learning techniques, especially when large amounts of data were involved [12]. A lot of deep learning techniques have been borrowed from image processing, but it's important to understand the minor distinctions between image and audio classification, which call for a specific focus on audio. In contrast to visuals, which are two-dimensional, audio represents a one-dimensional sequence that develops over time. While audio is frequently transformed into two-dimensional time-frequency representations for analysis, the horizontal and vertical axes in visuals are not the same as these temporal and frequency features. While audio requires sequential analysis as it develops over time, images capture a single moment that is frequently viewed in its entirety or in fragments with less temporal structure. As a result of these distinct characteristics, methods designed expressly for audio processing have been developed.

2 Background

2.1 Audio-Based Interaction Recognition Challenge

Specifically, the research problem centers on the use of the EPIC-SOUNDS dataset, which is a useful tool for addressing problems associated with audio recognition and classification in egocentric videos. Annotating individual audio clips in the videos is a complex operation that fits into making EPIC-SOUNDS. These segments are to be recognized, and the annotation task is to describe the actions that produce the appropriate sounds. These audio descriptions are categorized to make it possible to identify actions that can only be distinguished by aural signals. Additionally, operations that include objects colliding are carefully marked with details about the materials involved—for example, placing glass objects on hardwood surfaces. By cross-referencing these annotations with visual labels, ambiguities in the dataset are successfully eliminated and correctness is ensured.

2.2 Deep Neural Network

Deep neural network (DNN) architectures consist of several hierarchical layers, each of which transforms the outputs of the preceding layer in a non-linear way. The network can make complex decisions because of its ability to extract complex patterns and move from low-level to high-level features because of its hierarchical representation. As evidenced by DenseNet's superior performance over ResNet in audio tagging, DNNs outperform other machine learning techniques in audio classification [1]. The effectiveness of DNNs in these applications was further demonstrated by another study, which found that a 4-layered CNN significantly improved accuracy in acoustic scene classification. A deep learning classification algorithm based on CNNs achieved up to 92% testing accuracy in music genre classification [4].

Our work is inspired by DNNs since it makes use of overlapping layers in deep convolutional neural networks (DCNNs) to improve deep representation understanding. A more thorough understanding of the fundamental dimension of the data representation and the relationship between channel overlap and generalization performance is made possible by overlapping layers [2]. Moreover, they have been demonstrated to improve the learning procedure for complex models such as one-hidden-layer CNNs [6]. Cross-layer neuron integration in DCNNs is advantageous for training deeper networks, resolving the vanishing gradient problem, and quickening the rate of convergence for classification tasks [16].

Furthermore, although many researchers use a static learning rate, recent work has revealed that learning rate step decay can adaptively fine-tune deep neural network training. By lowering overfitting and stabilizing the training process, these dynamic changes to the learning rate promote quicker optimization, more effective model training, and enhanced generalization [5]. Additionally, the advantage of periodic variations in learning rate step decay is that they reduce the possibility of convergence to local minima, allowing the optimization process to explore different areas of the loss landscape. Another important advantage is the fact that only one new parameter needs to be tuned in a training experiment while all other internal parameters are kept constant [15].

3 Experimental Setup

3.1 Dataset

The EPIC-SOUNDS dataset, a large collection of audio annotations that record the time and class labels in the audio tracks of egocentric videos from EPIC-KITCHENS-100, is utilized by this study. This dataset includes 44 different classes and 78.4k categorized segments of auditory events and actions. By annotating the audio-based exchanges that take place throughout the gathered 100 hours of uncut video material, the dataset expands upon this one. We can use labeled temporal timestamps for the training and validation split, but only unlabeled timestamps are used for the recognition test split. In addition, we release the free-form descriptions from the initial annotation process as well as the temporal timestamps for annotations that were unable to be placed into any of the 44 classes [8].

3.2 Implementation Details

For better model optimization, the network model needs to be trained. The multi-audio event detection system typically consists of three primary processes: model construction, model assessment, and data preparation.

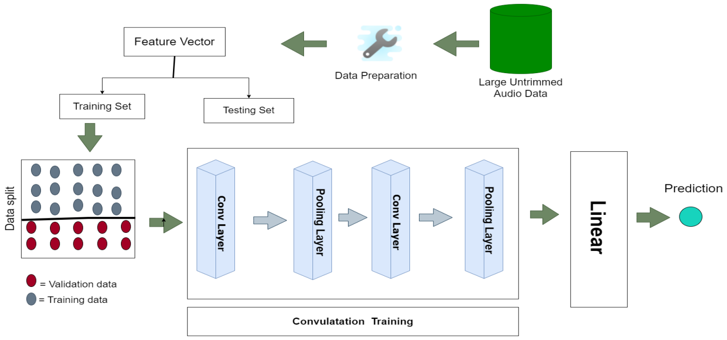

Fig. 1. A figure caption.

Feature Extraction

The following steps describe how to use the Librosa package to produce a representation of the log-mel spectrogram. In order to extract audio features from a given audio array, we first identify the start and end instances. We scale these features by determining the mean over a number of time intervals after obtaining the MFCC through audio segmentation. The top 30,000 rows of metadata for every audio were extracted iteratively. During each loop, we retrieve important data such as participant IDs, video IDs, class labels, and start and finish samples. For each cycle, we attempt to obtain the audio data from the dataset using the provided video ID. The features list then has these extracted features appended to it, along with the matching class label. Because a lot of overlapping, Principal Component Analysis (PCA) was used to reduce the audio's original 128 Mel frequency bands to 35, making it easier for the neural network to learn.

Data Balancing

For training purposes, different sound categories have unique audio clip availability. For example, the "metal-only collision" categories have an excess of over 6,600 audio clips, yet the "background" category just has a handful of audio clips. There is a long-tail trend in the distribution of these audio snippets among sound classes. Training data is fed in little chunks during the training process. When a data balancing approach isn't used, audio clips are randomly chosen from the dataset. As a result, sound classes like "metal-only collision," which include a large quantity of training clips, are more likely to be included in the training batches. Sometimes all the data in a mini-batch comes from a single sound class. This can lead to the model overfitting to sound classes with more training data and underfitting to those with fewer training examples. To address this issue, we have developed a balanced sampling strategy for training the CNN. This strategy involves approximately equal representation of audio clips from all sound classes to form a mini-batch. It's worth noting that the term "approximately" is used because an audio clip may be associated with multiple tags.

3.3 Neural Network Architecture

A single Convolutional Neural Network is trained to accommodate recordings from the kitchen environment, but only the important subset of events is taken into account for each scene during the decoding stage. Additionally, both the training and decoding procedures incorporate a binary scene feature. The CNN architecture includes the following components:

- The network receives 40 frames of derivative-rich log Mel filter bank acoustic characteristics.
- The first 1D convolution layer has 80 filters with a size of 3×40 and a stride of 1×1, and it is followed by max-pooling with a pool shape of 4×3 and a stride of 1×3.
- In the second 1D convolution layer, there are 100 filters measuring 13 units in size, and these are succeeded by max-pooling, which employs a pool size of 13 and a pool stride of 13.
- With 1935 units per, the third and fourth layers are fully connected.
- The output layer utilizes sigmoid activation functions, with one unit assigned to each event and an additional unit dedicated to identifying instances of audio. This configuration allows for the classification of concurrent or overlapping events.

To improve network performance, all layers employ rectified linear units (ReLU). The first convolution layer receives a dropout at a rate of 0.15, while the two fully-connected layers receive a dropout at a rate of 0.3 in order to reduce overfitting. Additionally, L2 weight regularization with a 0.0001 little penalty is applied to all layers. And also the processed data is then normalized using batch normalization.

Implementing the Adam optimization method with an initial learning rate of 0.001, the network is trained to maximize cross-entropy. When the validation loss does not decrease for ten consecutive epochs, the training procedure is over. Stochastic gradient descent is used in to fine-tune the network while keeping the original training set of data and without changing the initial convolution layer's parameters. A limited comparison of a few designs is possible by monitoring the loss of the development data to fine-tune the model parameters, within time restrictions.

3.4 Analysis

In this research, our objective was to develop an audio recognition model primarily using CNN architecture, incorporating both fully connected and convolutional layers for deep learning. To evaluate the model's performance, we employed a set of key metrics. Accuracy provided an overall assessment of predictive correctness, while precision gauged the model's ability to precisely identify churn instances, ensuring dependable forecasts. Additionally, recall was pivotal in determining the model's effectiveness in capturing all actual churn cases, preventing the oversight of critical instances. Lastly, the F1 score, a significant metric, struck a balance between precision and recall, considering the essential trade-off between false positives and false negatives. Together, these metrics comprehensively assessed the predictive performance of our models.

3.5 Tools and Libraries

In this research paper, we conducted the implementation of all the algorithms using Python 3 (version 3.11.2) and leveraged the capabilities of the OpenCV and Scikit-learn libraries within Jupyter notebooks. To create the Convolutional Neural Network (CNN) for detecting diabetic retinopathy, we employed TensorFlow Keras. The Jupyter notebooks were executed on virtual machines provided by Google Colaboratory, which had standard runtime configurations featuring Intel Xeon single-core processors running at 2.3 GHz and approximately 12.7 GB of available RAM. The training of the CNN model was carried out on the GPU runtime, using a single Tesla K80 GPU.

4 Results

Following the general approach, 30,600 audio were split into training and validation sets for evaluation purposes. All three models were implemented and tested using a 9,104 testing set, then results were recorded.

4.1 Testing Accuracy Metric

The number of epochs is set at 100, following the experimental setup section. Consequently, the various experimental results that develop from these different procedures are as follows:

Table 1. Convolutional neural network training parameters.

	Loss	Acc	Validation Loss	Validation Acc
Epoch = 25	1.2491	0.6013	1.4026	0.5790
Epoch = 50	0.9835	0.6824	1.1842	0.6483
Epoch = 75	0.6825	0.7766	1.0432	0.7195
Epoch =100	0.6394	0.7912	1.0310	0.7299

A continuous decline in the training loss is shown in Table 1. Similarly, the validation loss has been moving downward. Validation loss has stabilized and followed a gradual decrease, indicating the system's continued ability to handle the data. There is still work to be done before the neural network can fully fit the data. Furthermore, the final validation set is regularly detected with an accuracy of 0.7299%.

4.2 Performance Evaluation

The study also focuses on the calculation of segment-based precision, recalls, and F-scores for each class in kitchen audio events. The obtained results are recorded for the proposed model's training process.

Table 2. Segment-based metrics for kitchen audio events.

Class	Precision	Recall	F1-score
Background	0.93	0.99	0.96
Ceramic-only collision	0.93	1.00	0.96
Cut / chop	0.66	0.64	0.65
Footstep	0.83	0.90	0.86
Metal / ceramic collision	0.85	0.96	0.91
Metal / plastic collision	0.83	0.92	0.87
Metal / wood collision	0.81	0.91	0.86
Metal-only collision	0.34	0.28	0.31
Open / close	0.64	0.63	0.64
Plastic-only collision	0.66	0.61	0.63
Rustle	0.63	0.63	0.63
Scrub / scrape / scour / wipe	0.39	0.30	0.34
Slide object	0.76	0.74	0.75
Stir / mix / whisk	0.76	0.75	0.75
Water	0.73	0.71	0.72

Analyzing the results for each individual event, as shown in Table 2, we are able to identify that the model performs noticeably better in terms of classification when it deals with action events for which a significant amount of training data is available. This is most noticeable when events like "scrub/scrape/scour/wipe", the “metal-only collision”, “cut/chop”, and “rustle” have better results than events such as “metal/ceramic collision”, “background”, “ceramic-only collision”. However, the model’s performance at identifying audio events with limited training data is still compatible with the proposed model, suggesting that there are limitations in these instances.

5 Conclusion

We combined learning rate step decay with overlapping layers in this research. We demonstrated the importance of our approach using the EPIC-SOUNDS dataset, achieving an impressive result. During our participation, we investigated a number of novel approaches in addition to the effective strategies we had previously discussed. These techniques included experimenting with data reduction and sampling, lowering the batch size, adjusting the dropout ratio, and using min-max normalization as preprocessing. It’s crucial to remember that these strategies did result in better outcomes.

Our future study will investigate transfer learning using robust models like ResNet50 and the Slow-Fast model, which should produce better outcomes. Furthermore, we realize that a commonly used approach to model optimization is the periodic restart of learning rates. Our models can be made much more accurate by applying methods such as the Stochastic Gradient Descent with Warm Restarts (SGDWR) and the Cyclical Learning Rate Technique (CLR). During our post-competition investigations, we plan to explore these strategies in more detail.

References

1. Wenhao Bian, Jie Wang, Bojin Zhuang, Jiankui Yang, Shaojun Wang, and Jing Xiao. Audio-based music classification with densenet and data augmentation. In *Pacific Rim International Conference on Artificial Intelligence*, 2019.
2. David Bonet, Antonio Ortega, Javier Ruiz-Hidalgo, and Sarath Shekkizhar. Channel redundancy and overlap in convolutional neural networks with channel-wise nnk graphs. *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4328–4332, 2021.
3. Honglie Chen, Weidi Xie, Andrea Vedaldi, and Andrew Zisserman. Vggsound: A large-scale audio-visual dataset. *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 721–725, 2020.
4. Diwen Deng, Yiwu Gu, and Yiyi Zhu. Comparison of multiple machine learning algorithms for music genre classification. *Applied and Computational Engineering*, 2023.
5. Yimin Ding. The impact of learning rate decay and periodical learning rate restart on artificial neural network. *Proceedings of the 2021 2nd International Conference on Artificial Intelligence in Electronics Engineering*, 2021.
6. Simon Shaolei Du and Surbhi Goel. Improved learning of one-hidden-layer convolutional neural networks with overlaps. *ArXiv*, abs/1805.07798, 2018.

7. Jort F. Gemmeke, Daniel P. W. Ellis, Dylan Freedman, Aren Jansen, Wade Lawrence, R. Channing Moore, Manoj Plakal, and Marvin Ritter. Audio set: An ontology and human-labeled dataset for audio events. *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 776–780, 2017.
8. Jaesung Huh, Jacob Chalk, Evangelos Kazakos, Dima Damen, and Andrew Zisserman. EPIC-SOUNDS: A Large-Scale Dataset of Actions that Sound. In *IEEE International Conference on Acoustics, Speech, Signal Processing (ICASSP)*, 2023.
9. Evangelos Kazakos, Arsha Nagrani, Andrew Zisserman, and Dima Damen. Slow-fast auditory streams for audio recognition. *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 855–859, 2021.
10. Huy Phan, Philipp Koch, Fabrice Katzberg, Marco Maaß, Radoslaw Mazur, Ian Mcloughlin, and Alfred Mertins. What makes audio event detection harder than classification? 09 2017.
11. Arjun Prashanth, S. Jayalakshmi, and Vedhapriyavadhana Rajamani. A review of deep learning techniques in audio event recognition (aer) applications. *Multimedia Tools and Applications*, pages 1–15, 06 2023.
12. Hendrik Purwins, Bo Li, Tuomas Virtanen, Jan Schluter, Shuo-Yiin Chang, and Tara Sainath. Deep learning for audio signal processing. *IEEE Journal of Selected Topics in Signal Processing*, 13(2):206–219, may 2019.
13. Sneha Singh, Sarah Payne, and Paul Jennings. Toward a methodology for assessing electric vehicle exterior sounds. *Intelligent Transportation Systems, IEEE Transactions on*, 15:1790–1800, 08 2014.
14. Dan Stowell, Dimitrios Giannoulis, Emmanouil Benetos, Mathieu Lagrange, and Mark D. Plumbley. Detection and classification of acoustic scenes and events. *IEEE Transactions on Multimedia*, 17(10):1733–1746, 2015.
15. Kang Wang, Yong Dou, Tao Sun, Peng Qiao, and Dong Wen. An automatic learning rate decay strategy for stochastic gradient descent optimization methods in neural networks. *International Journal of Intelligent Systems*, 37:7334 – 7355, 2022.
16. Zeng Yu, Tianrui Li, Guangchun Luo, Hamido Fujita, Ning Yu, and Yi Pan. Convolutional networks with cross-layer neurons for image recognition. *Inf. Sci.*, 433-434:241–254, 2018.