FUDMA Journal of Sciences (FJS)
ISSN online: 2616-1370
ISSN print: 2645 - 2944
Vol. 8 No. 3, June, (Special Issue) 2024, pp 137 - 144
DOI: https://doi.org/10.33003/fjs-2024-0803-2399

AN ENHANCED AUDIO EVENT DETECTION WITH ATTENTION NEURAL NETWORKS

***Muhammad Sadisu Isah, Obunadike, G. N. and Mukhtar Abubakar**

Department of Computer Science, Federal University Dutsin - Ma. Katsina State, Nigeria

*Corresponding authors' email: sadinmalam@gmail.com Phone: +2347017171707

ABSTRACT

Multimedia recordings are very vital in the aspect of audio event detection with attention neural networks and it's a task of recognizing an audio event in an audio recording. The aim of the proposed work is to improve the development of existing audio/sound event detection in continuous streams and audio recording. Also, to compute the classes of audio events such as gunshots, screaming, door slamming, bell ringing, coffee, bird singing etc from an audio recording and also to estimate the onset and offset of these acoustic events. In this work, the propose system is going to use the modern machine learning methods called attention neural networks. The enhancement in the quality of audio event detection is achieved using an attention neural network based approach. Different activation functions that include RELU, LeakyReLU, ReLU6, ELU, and Swish known as SILU were investigated, the performance of the models using the above mentioned activation functions and Evaluate the performance of the baseline system using the different activation functions and compare the performance with the results of the existing studied papers were presented and discussed. As discussed and shown in this research, Swish network achieved mAP of 0.361432 and dprime of 2.642 outperformed the ReLU network and D-prime, from the baseline paper even though they both achieved the same AUC using the same architecture with 1024 hidden units. Using the feature level attention model, Swish activation function with mAP of 0.370 outperformed ReLU with mAP 0.369 in the baseline paper. Swish performance is superior compared to the other activation functions networks investigated in this research.

Keywords: Audio Event detection, Attention Neural Network, Deep learning, RELU, LeakyReLU, ReLU6, ELU, Swish known as SILU

INTRODUCTION

Audio event detection is crucial for description of multimedia recordings and the task entails recognizing on-set and off-set of an audio event in a recording. Audio event detection refers to the recognition of audio events and the estimation of their temporal and special locations in an audio recording. Because of the significance and widespread application of sound, machines are required to assist humans with tasks involving hearing. Audio event classification is one of the most important hearing skills. Assigning one or more descriptive labels, such as "male speech" or "plucked string instruments," is part of this work to a particular audio signal. The artificial intelligence (AI) industries may find many uses for this task in applications involving the perception of sound, including speech recognition, music information retrieval, smart homes, public surveillance, sound print, and multimedia content analysis.

Human life is one of the computer acoustic environments that involve a variety of sounds such as music, speech, gunshot, door slamming and others. This is also referred to as the general field of "Machine learning" (Gemmeke, et al 2017), it is obvious that recognizing human speech is paramount when developing a system that is meant to pass a variety of hearing tasks at the level of the human audio system (Ear). However, such a task would not be achieved as long as the system is unable to identify non speech sounds as human.

There are many scenarios that audio event detection finds its application such as listening for sound of gunshots (Valenzise, et al 2007), monitoring traffic for sounds of accidents, surveillance, screams (Maron and Lozano-P´erez 1998), door slamming, bell ringing, coffee, bird singing etc which might designate abnormal significant activity. It is also applicable in cases like context recognition (Eronen, et al 2006), Wildlife monitoring (Briggs e*t al* 2017) and (Health monitoring system) Lighting control systems are essential in modern building automation and smart homes, efficiently managing illumination to enhance energy conservation and user comfort. This project tackles energy consumption challenges in hospital buildings by introducing Intelligent Lighting Control Systems (ILCS) that take natural light and occupancy into account, driven by Artificial Neural Networks (ANN) and machine learning algorithm (Okomba et al, 2024).

There exist numerous challenges for audio event detection system that can be employed in real life environment; this is because researchers' progress on audio event detection has been inactive until recent years. These challenges for audio event detection include but are not limited to environment noise, lack of structure, intra-class variability, defined ambiguity, overlapping sound events.

Some of the quoted challenges are not only specific to audio event detection and they can be applied into the machine learning tasks such as musical transcription, musical genre classification and ASR. In recent years, several international challenges have been organized to address some of the mentioned challenges such as DCASE 2013 Dennis *et al*, (2011), DCASE 2016 Maron *et al* (1986) challenges.

Related works

Audio events detection also refers to acoustic events detection; it's a task of reorganizing the offsets and onsets of an audio event in an audio recording. These acoustic events occur as physical events such as bird singing, door slamming, bell ringing, crowd cheering, and produce signature sounds that carry rich information about the audio events. These acoustics events are the actual units of audio and are temporally bound to the onset and offset of the physical events (Briggs et al, 2012).

(Kwak and Chung, 2020) Proposed using derivative features for sound event detection based on deep neural networks. As input to the networks, they used log-mel-filterbank and its first and second derivative features for each frame of the audio signal. Two deep neural networks were used to evaluate the effectiveness of these derivative features.

There is an exponential increase in online multimedia sharing such as YouTube that has given rise to millions of hours of multimedia data. This is a large amount of non-annotated data which audio event detection techniques can be used to extract. In order to ease audio event detection research in this direction, Audioset Kong *et al* (2019) which is a large-scale dataset was published by Google in 2017.

Goetze, et al (2012) proposed, remote monitoring of bird sounds in the wild which involves the detection of bird sound can be developed Adavanne, and Virtanen (2017), and it can also be employed in the healthcare sector for the automatic detection of cough sounds in patients.

Deep learning is well-suited to develop representations for structured data and accelerate the exploration of chemical space, which means it has the potential to completely transform quantum chemistry. Despite the fact that convolutional neural networks are now the best option for images, audio and video data, Molecules contain atoms that are not confined to a grid. Rather, their exact placements hold physical information that is crucial and would be lost if they were discretized. Therefore, instead of requiring the data to be placed on a grid in order to simulate local correlations, researchers suggest using continuous filter convolutional layers. We use SchNet, a modern deep-learning architecture that models quantum interactions in molecules, to apply those layers. We derive a combined model that adheres to basic quantum chemistry laws for both total energy and interatomic forces. Predictions of rotationally invariant energy are included. and a potential energy surface which is smooth and differentiable. We achieve state-of-the-art performance for molecular dynamics trajectories and equilibrium molecule benchmarks with our design. Lastly, we present a more difficult benchmark that includes structural and chemical differences and points the way for further research. (Schütt, et al 2017).

Based on the research conducted the existing state of the art baseline system employed two activation function which are sigmoid and ReLU, while the proposed dissertation is going to employed different activation function such as LeakyReLU, ReLU, ReLU6, ELU, and Swish/SILU to provide easier optimization and better solution from closer initialization.

Machine Learning for Audio Event Detection

Machine learning is a branch of computer science whose primary aims is to design machine models that are capable of learning directly from given data which is (INPUT). In machine learning there are many types of learning this including supervised, unsupervised, reinforcement and many more. Machine learning systems are more convenient for certain tasks that involve processing large amounts of data or are challenging to develop as an algorithm by a software engineer. Audio event detection system is a typical example of such a challenging task. Hence, by understanding how machine learning systems for audio event detection process audio data; it is also possible to better understand how humans perceive audio events.

The aim of an audio detection system is to identify two aspects about a target event: (1) To identify where the event happens within a continuous audio signal and also (2) to identify its identity. The outputs of the audio event detection system expected are shown in figure 2.2. The viewpoint from the algorithm is, two dominant trends have been observed in previous works. The first trend depends on the conventional ASR framework (Schilit et al, 1994) and the second trend is detection by classification scheme (Radhakrishnan et al, 2005).

Automatic speech recognition (ASR) Framework

The audio event detection task is referred to as analogous to the continuous speech recognition one, the ASR framework has been employed for the event detection task (Schilit, et al 1994). The ASR framework, the audio events are treated like words in speech. This approach has been generally used in previous challenges, such as CLEAR 2006 (Dennis, et al 2011), CLEAR 2007 (Radhakrishnan, et al 2005) and DCASE 2013 (Adavanne, et al 2017).

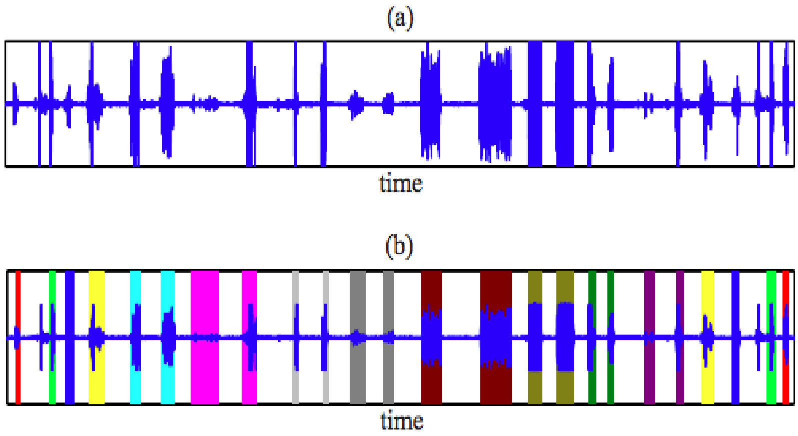

Figure 1: Audio event detection from two audio signals: (a) the input audio signal and (b) the output of the Oracle AED. The colors stand for various interest categories for events. Reproduced the figure from (Goodfellow, et al 2013).

Detection by classification

In this regard, the detection systems stick to the detection-by-classification methods including two classified: (1) the first classifier is trained to differentiate the target events from the background (2) the second classifier then classifies the target events into separate groups of interest. These two learned classifiers are adopted to detect audio events in audio recording or stream in a sliding window manner as shown in figure 2. The foreground/background classifier is trained first to reject background segments at each time.

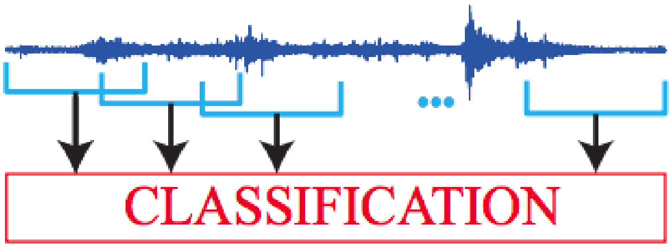

Figure 2: The sliding window-style audio event detection using a detection-by-classification method. Reproduced the figure from (Goetze, et al 2012).

Various Activation Functions Studied
During the course of study of audio event detection with attention neural networks, various activation functions like RELU, LeakyReLU, ReLU6, ELU, and Swish known as SILU were investigated.

Rectilinear Units Activation Function (ReLU)
ReLU Rectilinear Units Activation Function alleviates the vanishing gradient problems as the derivative of ReLU is either 0 or 1 not between 0 to 1. Whenever x is greater than 0, it will be 1 and if x is less than or equal to 0, it will be 0. However, whenever the value of x is 0 it leads to another problem called dying ReLU or dead activation.

$$f(x) = \text{Max}(0,x) \quad (1)$$

Leaky Rectilinear Activation Function (LeakyReLU)
The Leaky Relu activation function applies the element wise function as mathematically shown in equation (2.2) and figure 2.7.

$$f(x)=\max(0.01*x , x) \quad (2)$$

Rectilinear 6 Activation Function (ReLU6)
Also, ReLU6 proposed in (Alex Krizhevsky, 2012), it's a modified version of the Relu proposed in (Kong, et al 2019). They cap the units at 6, as this enables the model to learn sparse features earlier.

Exponential Linear Units (ELU)
The exponential linear unit (ELU) activation function, it's an activation function that converges cost to zero faster and more accurate results are produced. ELU has an extra positive constant known as alpha as different to the rest of the activation functions. ELU function and Relu are very similar except ELU has negative inputs. But, Relu smooths sharply while ELU smooths slowly until the alpha is equal to the output and is mathematically represented in the equation (3).

$$ELU(x) = \begin{cases} x, & \text{if } x >= 0 \\ \alpha * (\exp(x) - 1), & \text{if } x < 0 \end{cases} \quad (3)$$

Where alpha (α) is a small positive constant (usually a value between 0 and 1).

Swish Activation Function (SILU)
The Swish, a self-gated activation function also known as SILU proposed in (Ramachandran, et al 2018), is a non-monotonic and smooth function as mathematically defined in equation (4). Swish activation function, unlike Relu, is non-monotonic and smooth, but also, similar to Relu, as, is also unbounded above and bounded below. Swish activation function also differ from softplus, Relu as due to its non-monotonicity feature it gives negative outputs for a small negative inputs.

$$\mathbf{f(x) = x \cdot sigmoid(x)} \quad (4)$$

MATERIAL AND METHOD
Audioset Dataset
Audioset, which contains 2,084,320 of 10 second of audio clips sample taken from YouTube videos with a hierarchical ontology of 527 classes in the recent version (VI), different activation functions such LeakyRelu, ReLU6, and Swish were to be investigated and used to evaluate the performance of audio event detection with attention neural networks. Audioset is a collection of different sounds, because audioset is multi-labeled, each audio clip may have multiple sound classes. At least one label appears on every audio clip. There are 896,045 of audio data clips with single sound class, followed by 684,166 of audio data clips with two different sound classes, out of a total of 2,084,320 audio data clips; there are only 4,661 audio samples with seven labels. Audioset provides bottleneck features extracted from the bottleneck layer of a VGGish CNN, which was pre trained on 70 million of audio data clips from YouTube 100m dataset (Gemmeke, et al 2017) as shown in figure 3.

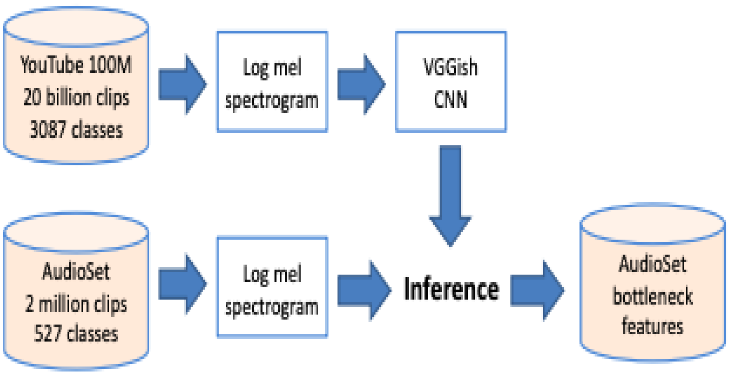

Figure 3: Representation of bottleneck features extraction. Reproduced the figure from (Kong, et al 2019)

Baseline Architecture

The average pooling, max pooling, single level attention, multi-level attention and feature level attention models were trained using anaconda virtual environment using condor jobs. To train the models, Tensorflow and Pytorch were used. Also, to extract the audioset embedding feature, vggish model and vggish pca were used during the training. The models are trained for 50,000 iterations in total during the training for each model. However, the prediction of 5 models ranging from 10,000 - 50,000 number of iterations were averaged as the final prediction to stabilise and ensemble the results, which the prediction randomness caused by the entire model can be reduced.

Evaluation Criterion

In proposed dissertation, the evaluation criterion used in Kong *et al* (2019), precision and recall will be adopted as well as false positive rate. (P) Precision defined in Eqn 5, (R) Recall defined in Eqn 6 and false positive rate (FPR) is defined in Eqn 7.

$$P = TP/ (TP + FP) \quad (5)$$

$$R = TP/ (TP + FN) \quad (6)$$

$$FPR = FP/ (FP + TN) \quad (7)$$

True positives (TP) show a situation where the reference and the system prediction it shows an event to be active, False negatives (FN), are where the reference shows an event is active but the system prediction shows an event is inactive, A false positive (FP) is where the system prediction shows an event is active but reference shows it is inactive; and finally, True negative (TN) occurs where the reference and system prediction shows an event is inactive. Precision, recall and false

RESULTS AND DISCUSSION

Performances of the Models with Different Activation Functions

In this research, the proposed models in (Kong et al, 2019), were used using different activation functions such as *RELU, LeakyReLU, ReLU6, ELU, and Swish known as SILU* are explored for modifying the proposed models. The performance of the proposed baseline system in using different activation functions over the audioset dataset is presented in terms of mean average precision (mAP), area under the curve (AUC) and d-prime in table 1, table 2, and table 3, respectively and their corresponding bar chart representation in figure 4, 5 and 6, respectively. Also, different hidden units that include 1024 and 2048 were investigated using the feature level attention model in (Kong at al, 2019) with the above mentioned activation functions, and the results are shown in table 4. and bar chart represented in figure 7.

Table1: Mean average precision results obtained using different activation functions

MODEL	ReLU	LeakyReLU	ReLU6	ELU	Swish
Average pooling	0.300	0.296	0.300	0.292	0.294
Max pooling	0.292	0.288	0.294	0.280	0.291
Single attention	0.337	0.333	0.337	0.323	0.331
Multi attention	0.357	0.351	0.335	0.343	0.352
Feature level attention	0.361	0.359	0.359	0.352	0.361

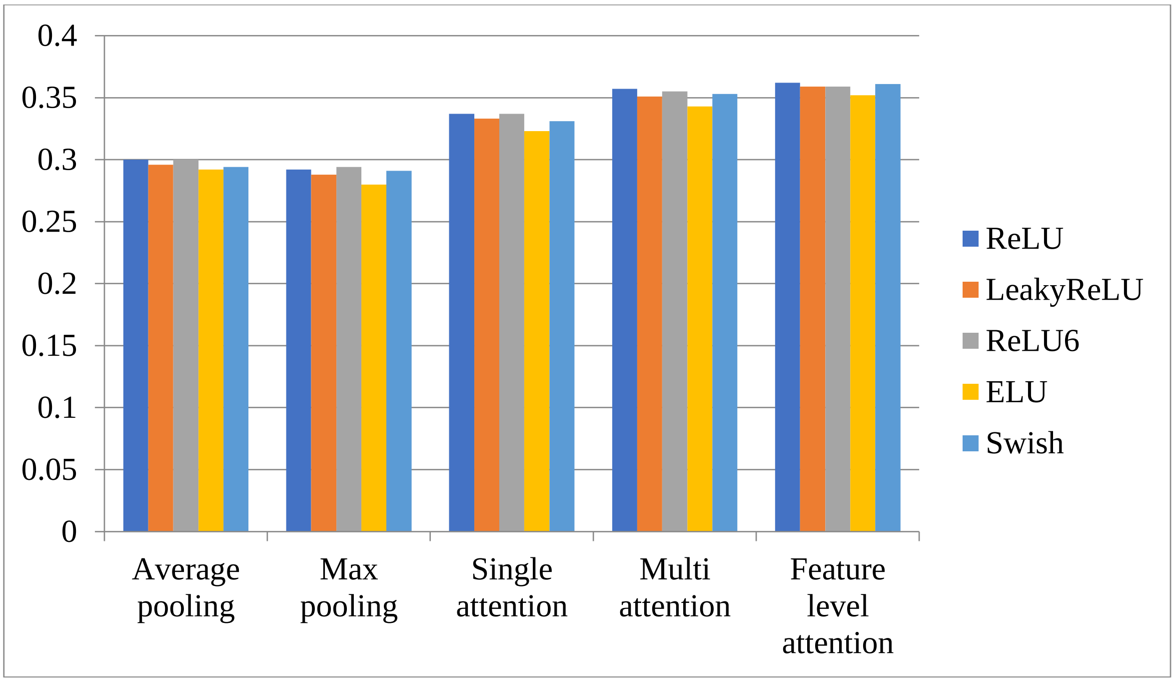

Figure 4: Chart representation of mean average precision results

Table 2: Area under curve results obtained using different activation functions

MODEL	ReLU	LeakyReLU	ReLU6	ELU	Swish
Average pooling	0.964	0.963	0.962	0.963	0.963
Max pooling	0.960	0.959	0.960	0.957	0.959
Single attention	0.968	0.967	0.967	0.966	0.967
Multi attention	0.968	0.968	0.969	0.965	0.968
Feature level attention	0.969	0.969	0.969	0.968	0.969

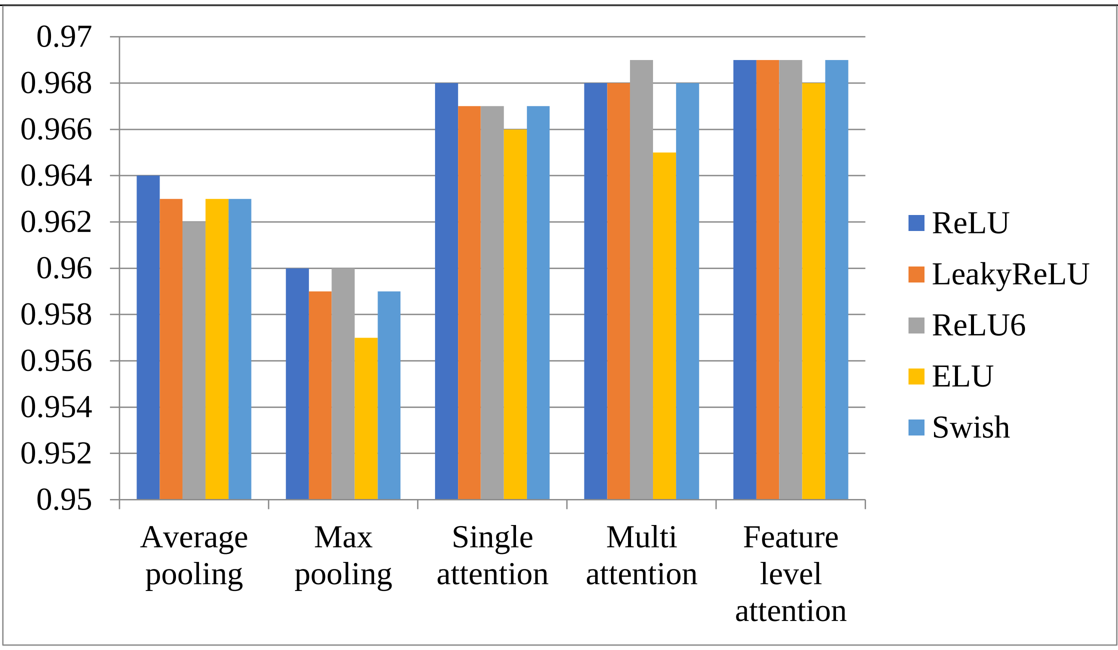

Figure 5: Chart representation of area under curve results

Table 3: d-prime results obtained using different activation functions

MODEL	ReLU	LeakyReLU	ReLU6	ELU	Swish
Average pooling	2.536	2.520	2.530	2.502	2.533
Max pooling	2.471	2.462	2.473	2.422	2.465
Single attention	2.612	2.602	2.606	2.574	2.607
Multi attention	2.621	2.619	2.631	2.568	2.618
Feature level attention	2.641	2.637	2.637	2.627	2.642

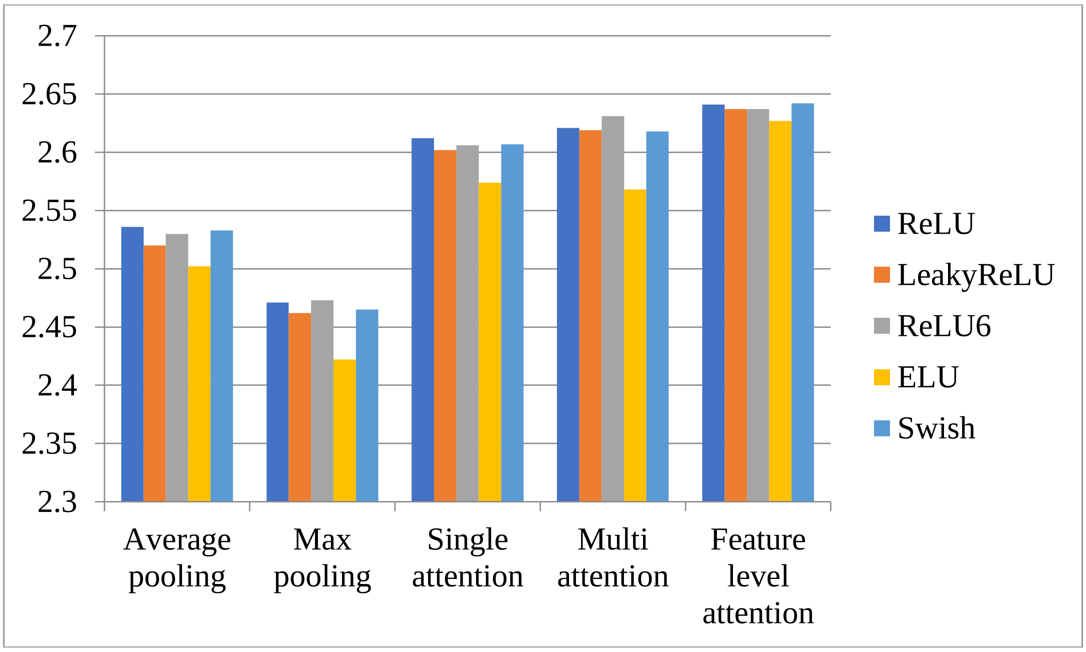

Figure 6: Chart representation of d-prime results

Also, from table 4.4 and figure 4.5 it can be observed that the performance of Swish network with mAP of 0.370 outperforming the ReLU network in (Kong et al, 2019) with mAP of 0.369 using 2048 hidden layers with the same architecture. However they both achieved approximately the same AUC of 0.969 and d-prime of 2.641. It can also be observed that the performance of the other activation function networks slightly differ from the first architecture with 1024 hidden layers.

Table 4: Mean average precision results obtained using feature level attention model with different activation functions and hidden units

Hidden unit	ReLU	LeakyReLU	ReLU6	ELU	Swish
1024	0.361	0.359	0.359	0.352	0.361
2048	0.369	0.367	0.368	0.365	0.370

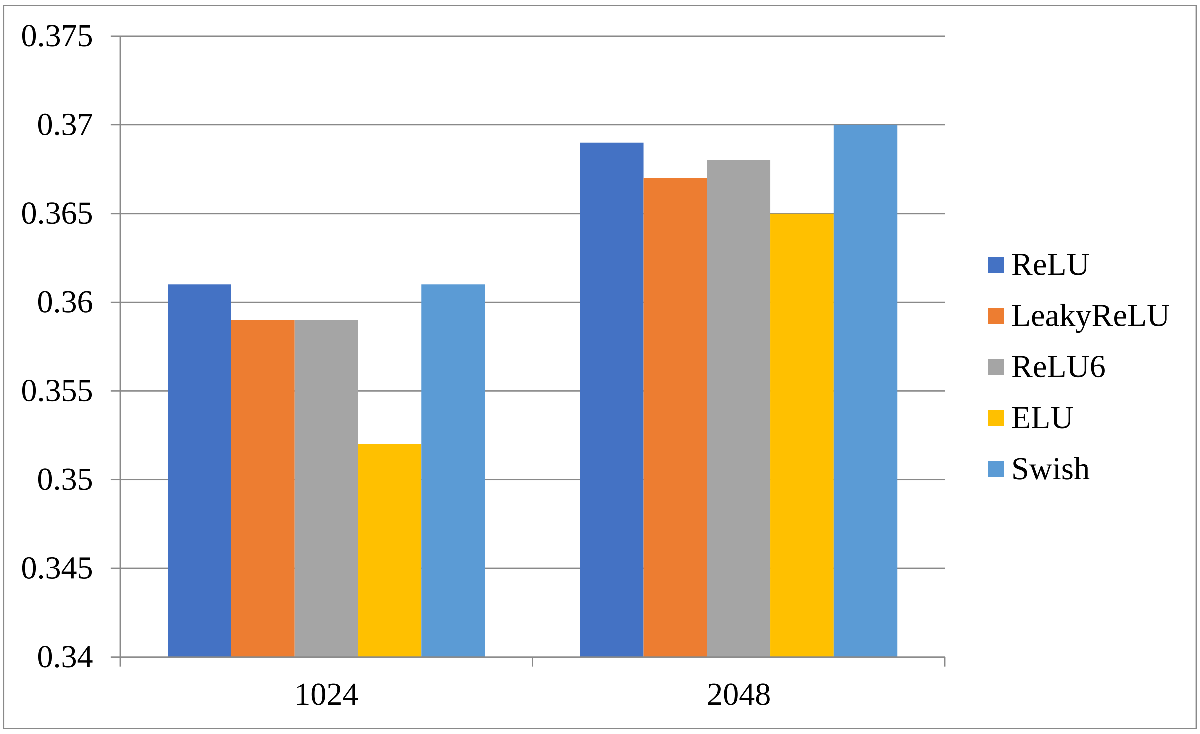

Figure 7: Chart representation of mean average precision results using feature level attention model with different activation functions and hidden units

Discussion

In this research, swish network achieved mAP of 0.361432 and dprime of 2.642 outperformed the ReLU network proposed in (Kong at al, 2019) achieved mAP of 0.360732 and d-prime of 2.641 even though they both achieved the same AUC using the same architecture with 1024 hidden units. Using the feature level attention model, Swish activation function with mAP of 0.370 outperformed ReLU with mAP 0.369 proposed in (Kong at el, 2019) and the rest of the activation functions proposed in this project using 2048 hidden units with the same architecture as well. Also, Swish performance is superior compared to the other activation functions networks investigated in this research. It is also observed during the training and convergence of ELU-networks is faster than the rest of the models. Also, ELU and LeakyReLU networks tend to overfit during the training and it is observed that due to smaller dataset size with 3 hours of training might be the possible reason for ELU overfitting.

CONCLUSION

In this report, brief introduction, background and context, were discussed, Also, related literature to neural networks have been investigated in this report as techniques in audio event detection have been employed very fast and replaced the traditional machine learning methods with neural networks approaches recently.

Audiodataset, baseline architecture and evaluation metrics were discussed. Also, swish network achieved mAP of 0.361432 and dprime of 2.642 outperformed the ReLU network in the baseline paper achieved mAP of 0.360732 and dprime of 2.641 even though they both achieved the same AUC. Also, Swish performance is superior compared to the other activation functions networks investigated in this research. Also, using the feature level attention model, Swish activation function with mAP of 0.370 outperformed ReLU with mAP 0.369 proposed in the baseline paper and the rest of the activation functions proposed in this project using 2048 hidden units. In this research work, different activation functions using the proposed models were investigated using an Audioset dataset. In order to make a general recommendation on which activation function is best for the models proposed, different datasets with different activation functions should be explored and investigated. Also, the onset and the offset of the acoustic events should be investigated.

REFERENCES

Adavanne, S. and Virtanen, T., "A report on sound event detection with different binaural features," DCASE2017 Challenge, Tech. Rep., September 2017.

Alex Krizhevsky, "Convolutional Deep Belief Networks on CIFAR-10"

Ankit Shah: DCASE 2017 CHALLENGE SETUP: TASKS, DATASETS AND BASELINE SYSTEM

Bello, J. P., Mydlarz, C., and Salamon, J., "Sound analysis in smart cities," in binaural features," DCASE2017 Challenge, Tech. Rep., September 2017.

Briggs F., B. Lakshminarayanan, L. Neal, X. Z. Fern, R. Raich, S. J. Hadley, A. S. Hadley,and M. G. Betts. Acoustic classification of multiple simultaneous bird species: A multi-instance multi-label approach. *The Journal of the Acoustical Society of America*, 131(6):4640–4650, 2012. for mobile robots using audio features," in IEEE International Conference on Multimedia

Chu, S., Narayanan, S., Kuo, C., and Mataric, M., "Where am I? scene recognition Computational Analysis of Sound Scenes and Events, Virtanen, T., Plumbley, M., and Ellis, D. P., Eds. Springer, 2018, pp. 373 – 397.

Dennis, J., Tran, H. D., and Li, H., "Spectrogram image feature for sound event classification in mismatched conditions," IEEE Signal Processing Letters, vol. 18, no. 2, pp. 130 – 133, 2011.

Elizalde .B et al. Experimentation on The DCASE Challenge 2016: Task 1 - Acoustic Scene Classification and Task 3 - Sound Event Detection in Real Life Audio. Tech. rep. Detection, Classification of Acoustic Scenes, and Events 2016, 2016.

Eronen A. J., V. T. Peltonen, J. T. Tuomi, A. P. Klapuri, S. Fagerlund, T. Sorsa, G. Lorho, and J. Huopaniemi. Audio-based context recognition. Audio, Speech, and Language Processing, IEEE Transactions on, 14(1):321–329, 2006.

Gemmeke, J. F., Ellis, D. P., Freedman, D., Jansen, A., Lawrence, W., Moore, R. C., Plakal, M., and Ritter, M., "Audio set: An ontology and human-labeled dataset for audio events," in IEEE International Conference on Acoustics, Speech and Signal Processing, 2017, pp. 776 – 780.

Goetze, S., Schroder, J., Gerlach, S., Hollosi, D., Appell, J.-E., and Wallhoff, F., "Acoustic monitoring and localization for social care," *Journal of Computing Science and Engineering*, vol. 6, no. 1, pp. 40 – 50, 2012.

Goodfellow, I. J., Warde-Farley, D., Mirza, M., Courville, A. C., and Bengio, Y.,"Maxout networks." ICML (3), vol. 28, pp. 1319 – 1327, 2013.

Han, K. J., Chandrashekaran, A., Kim, J., and Lane, I., "The capio 2017 conversational speech recognition system," arXiv preprint arXiv:1801.00059, 2017.

Hawkins, D. M., "The problem of overfitting," *Journal of chemical information and computer sciences*, vol. 44, no. 1, pp. 1 – 12, 2004.

Howard G., M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, H. Ada" MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications

Kao, C.-C., Wang, W., Sun, M., and Wang, C., "R-crnn: Region-based convolutional recurrent neural network for audio event detection," arXiv preprint arXiv:1808.06627, 2018.

Kingma D. P. and J. Ba, "Adam: A method for stochastic optimization," in International Conference on Learning Representations (ICLR), 2015.

Kong, Qiuqiang, Changsong Yu, Yong Xu, Turab Iqbal, Wenwu Wang, and Mark D. Plumbley. "Weakly Labelled AudioSet Tagging With Attention Neural Networks." IEEE/ACM Transactions on Audio, Speech, and Language Processing 27, no. 11 (2019): 1791-1802.

Kristof T. Schütt, Pieter-Jan Kindermans, Huziel E. Sauceda, Stefan Chmiela, Alexandre Tkatchenko, Klaus-Robert Müller, " SchNet: A continuous-filter convolutional neural network for modeling quantum interactions"

Lee Y, D. K. Han, and H. Ko. "Acoustic Signal Based Abnormal Event Detection in Indoor Environment Using Multiclass Adaboost". In: Proc. 2013 IEEE International Conference on Consumer Electronics (ICCE). 2013, pp. 322–323.

M. Abadi, A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, G. S. Corrado, A. Davis, J. Dean, M. Devin, et al. Tensorflow: Large-scale machine learning on heterogeneous systems, 2015. Software available from tensorflow. org, 1, 2015

Maron O and T. Lozano-P´erez. A framework for multiple-instance learning. Advances in neural information processing systems, pages 570–576, 1998.

OkombaN. (2024). Intelligent Lighting Control Systems for Energy Savings in Hospital Buildings Using Artificial Neural Networks. *FUDMA JOURNAL OF SCIENCES*, 8(2), 390 - 398. https://doi.org/10.33003/fjs-2024-0802-2320

Paulya L, H. Peela, S. Luoa, D. Hoggb and R. Fuentes, " Deeper Networks for Pavement Crack Detection". 34th International Symposium on Automation and Robotics in Construction (ISARC 2017)

Prajit Ramachandran, Barret Zoph, Quoc V. Le, " Searching for Activation Functions". [Submitted on 16 Oct 2017 (v1), last revised 27 Oct 2017 (this version, v2)]

Prajit Ramachandran, Barret Zoph, Quoc V. Le, " Swish: a Self-Gated Activation Function"

Radhakrishnan, R., Divakaran, A., and Smaragdis, A., "Audio analysis for surveillance applications," in IEEE Workshop on Applications of Signal Processing to Audio and Acoustics, 2005. IEEE, 2005, pp. 158 – 161.

Schilit, B., Adams, N., and Want, R., "Context-aware computing applications," in First Workshop on Mobile Computing Systems and Applications, 1994, pp. 85 – 90.

Srivastava N, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, “Dropout: A simple way to prevent neural networks from overfitting,” *The Journal of Machine Learning Research*, vol. 15, no. 1, pp. 1929–1958, 2014.

Valenzise G., L. Gerosa, M. Tagliasacchi, F. Antonacci, and A. Sarti. Scream and gunshot detection and localization for audio-surveillance systems. In Advanced Video and Signal Based Surveillance, 2007. AVSS 2007. IEEE Conference on, pages 21–26. IEEE, 2007.