Sensors & Transducers, Vol. 249, Issue 2, February 2021, pp. 17-24

www.sensorsportal.com

Sensors & Transducers

Published by IFSA Publishing, S. L., 2021
http://www.sensorsportal.com

Real Time Object Detection with Noisy Sensors Using Deep Learning

[1, 2] Aditya Singh, [1] Pratyush Kumar, [1] Vedansh Priyadarshi, [1] Yash More, [1] Aishwarya Praveen Das, [1] Bertrand Kwibuka and [1, *] Debayan Gupta

[1] Ashoka University, Rajiv Gandhi Education City, 131029, Sonepat, India
[2] Carnegie Mellon University, 5000 Forbes Ave, 15213, Pittsburgh, PA United States
Tel.:+91 9909869367
E-mail: debayan.gupta@ashoka.edu.in

Received: 10 December 2020 /Accepted: 15 January 2020 /Published: 28 February 2021

Abstract: In this paper, we introduce a first of its kind, radio-signal based object detection system for controlled environments, which substitutes complex signal processing and expensive hardware with deep learning networks to detect patterns from low-quality, inexpensive sensors. Our system operates in the less crowded low-frequency range of 433 MHz in contrast to existing RF-based sensing methods and uses mini-Doppler maps generated from raw I/Q data, thereby allowing us to use cheap, off-the-shelf software defined radios. We demonstrate that our system is versatile enough to handle occlusions and is also sensitive to multiple objects; additionally, it does not use visual data and hence is not hampered by bad lighting. The core of our system is a VGG-16 based CNN architecture trained on the mini-Doppler maps. We achieve an accuracy of 0.96 on a binary classification task of detecting the presence or absence of an object in an enclosed space. Furthermore, we observe that our system shows promise for more complicated detection algorithms as it is able to successfully differentiate between the presence of a single object and two identical objects placed together. Our results indicate that convolutional networks can learn features important enough from spectrograms that enable it to distinguish the presence of objects, thereby eliminating the need of sophisticated signal processing methods to do the same.

Keywords: Deep Learning, CNN, Object detection, Software defined radio (SDR).

1. Introduction

Object detection has attracted high research attention in the past few years due to the remarkable results in the field of deep learning. Neural networks have proven to be good feature extractors and are used in several applications such as human-computer interaction, traffic-monitoring, healthcare and autonomous driving. Deep learning models have achieved high accuracy in diagnostic tasks across multiple medical modalities, and object detection techniques have helped flag organ damage via access to just its corresponding images [12]. These applications often rely on deep learning agents to supervise decision making, perception and understanding of user behavior in different scenarios/environments.

Advancements in the field of deep learning have also led to various applications in robotic systems such as Spot Mini and Atlas by Boston Dynamics. Deep learning combined with video surveillance can be used to perform crowd scene analysis, crowd density estimation, crowd segmentation and detection, which can be critical for disaster management [27].

Estimation and monitoring of crowds can also help ensure practice of social distancing norms amid the current coronavirus pandemic[1]

Additionally, smart cities and buildings have been able to optimize energy consumption based on the number of people in a certain area. An example of this is Google DeepMind[2]. DeepMind has been able to cut down Google's data center cooling bill by 40 % by using an ensemble of neural networks to predict future temperature and pressure of the data center and recommend actions to maintain optimal system settings [13].

However, most of these detection and monitoring approaches have several inherent limitations. They:

1) Require multiple cameras to be installed in an area to achieve adequate performance and coverage which results in high deployment cost,

2) Suffer from the same limitations as humans, i.e, cameras cannot perceive visual information in the dark or see through walls and occlusions. Additionally, in the age of comprehensive facial recognition technology, visual surveillance approaches pose major questions challenging the right to privacy of the public. Fortunately, visible light is just one end of the frequency spectrum. Recent advances in wireless research have shown that certain radio frequency signals can traverse through walls/occlusions and span across dark and smoky environments which are difficult to monitor [2]. Radio waves are also reflected off of the human body [35] (and any other conducting material). Furthermore, since typical WiFi systems operate in the radio wave channel, their ubiquity is an added advantage for building RF-based detection systems. These properties together make radio waves a better alternative to visible light for human-body detection. Unfortunately, most existing research projects utilising RF-waves for human body detection focus on device-based active methods that require significant instrumentation of the person/object and the environment. This limits their scalability and robustness. Furthermore, such design also impinges on the privacy of the users. Thus, there has been a push in research to build device-free methods that can overcome these restrictions. Recent advances in device-free methods have been promising. Although they rely on using expensive, state-of-the-art hardware operating in the overcrowded high frequency WiFi channel [1, 18, 36].

Here, we present a Deep Learning-based portable object-detection system, which uses low-frequency radio waves (433 MHz, via inexpensive software radios), designed to operate through occlusions [6].

We perform the following experiments to test its versatility:

- Detect an object placed in line-of-sight of Receiver and Transceiver;
- Differentiate between the presence of single object and two identical objects kept together.

Our experiments suggest that deep neural networks can be successfully trained to carry out the otherwise intractable tasks of both single and multi-object detection using low-frequency radio waves. Our approach is partly inspired by the mechanisms used by certain owls who can use very low frequency waves (around 20 kHz) to localise their prey during night-time and through occlusions. Due to its portability and cheap deployment costs, our system can be a crucial step in the direction of building scalable, robust and cost-efficient surveillance systems.

2. Related Works

Recent applications of ANNs to signal processing show that CNNs are able to outperform decade old feature search algorithms in radio modulation recognition.

Timothy, *et al.* [19] achieved state of the art accuracy for radio modulation recognition by training a CNN on raw I/Q samples (WiFi enabled SDR operating at 900 MHz)collected across 11 different modulations (8 digital and 3 analog), and normalized to unit variance. Even for high signal to noise ratio, the CNN achieved a better accuracy over a large coverage area. They were able to improve the accuracy by leveraging deep residual networks and LSTM [14] which performed better on visual and time-series data [20].

Zhao, *et al*. [36] introduced RF-pose 3D which provides a significant improvement in RF-based sensing by incorporating ANNs. RF-pose 3D takes the 4D RF signal captured by an FMCW radio (window of 3 seconds) as input, similar to the radio used by RF-capture and WiTrack (5.4-7.2 GHz) [3, 4]. The model operates using a regular softmax loss and is able to predict the location of each key point in space as the voxel with the highest score. Their system is capable of accurately tracking the motion and presence of a single person. To scale this to multiple people, the authors operate on the output of the CNN (in the horizontal plane) from an intermediate layer (feature map).The resulting network is able to accurately detect the 3D skeleton of multiple people and simple actions (walking, sitting, standing) over a range of 40 feet in multiple environments.

Li, *et al*. [18] further augmented this system to detect multiple actions by performing multimodal training using the obtained 3D skeleton from RF-pose 3D and existing vision-based action recognition data sets. The spatial-temporal attention module trained on the multimodal data is able to perform action recognition on visual and RF-based data. Further, they found that transferring knowledge related to action recognition across modalities is able to empirically improve performance regardless of whether the skeletons are generated from RF or vision-based data. This system, called RF-action is able to accurately

[1]https://www.who.int/emergencies/diseases/novel-coronavirus-2019/advice-for-public

[2]https://deepmind.com/blog/article/deepmind-ai-reduces-google-data-centre-cooling-bill-40

recognise actions and interactions (29 single actions and 6 inter-actions) of multiple humans through walls and in the dark (up to 40 feet) and represents a significant improvement inaction recognition capabilities.

Wang, *et al.* [30] presented WiHear which enables to hear any "talks" without using any specific devices. WiHear detects and analyzes radio reflections due to mouth movements. WiHear solves the problem of recognising micro-movement of lips by using partial multipath effects and wavelet packet transformation. WiHear is also able to listen to multiple people talking by leveraging MIMO technology. They tested this on USRP N210 and commercial WiFi router. They achieve detection accuracy of 91 % on average for a single individual speaking no more than six words and up to 74 % for no more than three people talking simultaneously. Wang, *et al*. [29] also showed that a wearable device can be exploited to understand a user's fine-grained hand movements at millimeter level, which enables attackers to reveal users' secret keys entries.

Al-qaness, *et al.* [5] proposed a device-free CSI-based human activity recognition system which exploits CSI of an indoor ubiquitous wireless device. They described recent research in device-free human sensing technologies and proposed an effective method of CSI filtering, pattern segmentation using machine learning methods, and micro-activity classification. They evaluated the proposed method in an indoor environment and took useless noise into consideration. Their model was able to classify nine human micro-activities. They evaluated their model in various scenarios with ten people and their model reached high accuracy. There are some limitations to their methods as well. It can't track movements of two or more people.

Bahl, *et al.* [8] used RSSI for pose estimation. Presence Of human body within wireless range causes signal attenuation, leading to a spike in RSSI measurement. Thus, RSSI has been widely used in various human-based activity recognition, for instance, device-free indoor localisation [32], density estimation of crowd in a room [31].The ease of calculating RSSI makes it more useful across devices. But RSS measurements are not effective in indoor environment due to several reasons [21]. As RSS measurements, being calculated over a multi-path channel, are prone to effects of multipath fading, and presence of reflecting objects in the environment surrounding leads to the receiver seeing multiple copies of the transmitted signal superimposed on each other leading to difference in attenuation. Moreover, RSSI can only detect limited types of human activities due to single path loss.

Zhao, *et al.* [34] proposed a wireless system called EQ-Radio that infers people emotions from the radio signals that bounce off their bodies. EQ-Radio tried to segment these RF reflections to the individual heartbeats. It had no idea how the heartbeats look like in reflected signals. There are two additional issues that they faced while classifying various human emotions. First, the reflected signals from human bodies were noisy. Second, they were operating in a low Signal to Interference and Noise Ratio (SINR) where there was too much interference due to breathing. EQ-Radio gave results very close to the ECG-based models.

Pu, *et al*. [24] proposed a Doppler shift based gesture recognition system, WiSee which is able to recognise 9 hand/leg gestures based on the unique Doppler shift profiles extracted from wireless signals. Due to different body movements with respect to the SDR and transmitter, there is a unique positive and negative Doppler shift pattern for different activities. WiSee has achieved the highest accuracy of 94 %.

Although the above systems show promising results, they use expensive hardware which operate on high frequency radio waves (2.0 GHz ≤ f ≤ 7.2 GHz). In the following sections we explain our approach presenting relevant differences since we deploy a low-cost, low-frequency setup which is capable of high accuracy detection.

3. Convolutional Neural Network

Convolutional Neural Networks (CNNs) have shown outstanding performance in understanding images [10], videos [17] and audios [7] related tasks not limited to object detection, colorization, image synthesis, pose recognition. Below we briefly describe the important layer types in CNN that are relevant to this paper. Deep Neural Networks (DNNs): DNNs are known to compute any continuous function to a desired level of precision by adding sufficient numbers of neurons in a hidden layer. Each unit of neuron in deep neural networks computes the weighted sum of inputs and the result is passed through an activation function (like softmax, ReLU etc.) to introduce non-linearity. Mathematically, the value of a neuron ${a_i}^n$ at the n-th layer can be calculated using $\sigma(\sum_j {w_{ij}}^n {a_j}^{n-1})$, where ${a_j}^{n-1}$ are the neurons from the previous layer, ${w^n}_{ij}$ are the weights and σ(·) introduces non-linearity. CNNs: In case of RGB images, CNNs work on 3D tensors consisting of two spatial dimensions, height and width, and a channel dimension. There is a chain of nonlinear filters which extract the abstract features from a low level RGB image like edges, boundaries, and finally the actual object. Unlike vanilla neural networks, neurons are partially connected to neurons in the previous layer. Due to this property of CNNs, they are easily able to learn sparse features present in the data. CNN consist of sequence of layers $f_1, f_2, \ldots$, and f_l. Each layer acts as a function which takes x_{l-1} tensor as an input and gives x_l tensor as output. For layer l, we can write this function as $x_l = f^l(x_{l-1})$. The functions $f^l s$ are non-linear local and translation invariant operators. The value of neurons in any layer l can be calculated using convolution operation of the weights kernels and tensors value in the previous layer, that is $x^l = \sigma(f^l * x_{l-1})$, where $*$ is the convolution

operator, f^l refers to the weight kernel at layer l, and x^l is the computed value of neurons in layer l. The CNN training can be summarised using the Algorithm 1.

Algorithm 1: CNN Training

Input: Tensors: $c, x_0, w_1, w_2, \ldots, w_{l-1}$
Output: Minimize the value of L, where *loss* is a loss function
for $k \leftarrow 1$ to l **do**
$$x_k \leftarrow f_k(x_{k-1}, w_k)$$
end for
$$L = loss(x_l, c)$$

Below we briefly explain the building blocks of CNNs:

1. Convolution Layer: If the input in l^{th} layer is an order 3 tensor of size $H^l \times W^l \times D^l$, then convolution kernel tensor of size $H \times W \times D^l$ overlay on top of the input tensor at a location. We compute the product of corresponding elements in all the D^l channels and sum the HWD^l product to get the convolution result at that location. The kernel is moved from left-to-right and top-to-bottom to complete the convolution operation.

2. Rectified Linear Unit (ReLU) Layer: The ReLU function performs truncation operation individually for each element in the input tensor:

$$y_{i,j,d} = max(0, x^l{}_{i,j,d}),$$

where $0 < i < H^l = H^{l+1}$, $0 < j < W^l = W^{l+1}$, and $0 < d < D^l = D^{l+1}$.

3. Batch Normalization (BN) Layer [16]: BN smooths the input layer by recentering and rescaling the input layer and can be represented mathematically as,

$$BN_{\gamma^k \beta^k}(x^k) = \gamma^k x^k + \beta^k,$$

where x^k is the input tensor, and γ^k and β^k are learnt during the training process. It also helps in mitigating the problem of internal covariate shift.

4. Dropout Layer: Dropout is a regularization technique used to ignore a percentage of randomly selected neurons during training to improve generalisation performance of the network. Dropout means temporarily removing hidden and visible units from the network. If the model performance is not improved after a new epoch, then the dropped units are retained.

5. Softmax Layer: This layer is used to convert the vectors of actual scores of classes to a vector of probability values. One more important property of the softmax operator is that it is shift invariant. The softmax loss is used to measure the difference between the class scores $s = \{s_C\}^k{}_{C=1}$ and the target label C^*as follow:

$$L_{softmax}(s, C^*) = -log \frac{e^{s_{C^*}}}{\sum_C e^{s_C}},$$

where s_{C^*} is the score prediction of the target class. Finally, the CNN computes class scores for new input and predicts it as the class with the highest score, as shown in Fig. 1.

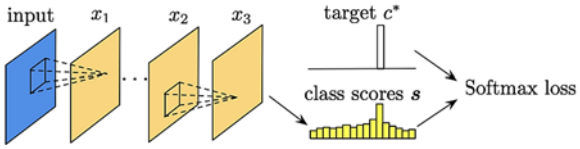

Fig. 1. CNN architecture for classification task, where x_1, x_2 and x_3 are feature maps, s shows the class scores, and c^* refers to the label. The Softmax loss function compares the score vector with the label.

4. System Setup and Data Collection

Our setup comprises a hardware platform (to send and receive signals), and software (to control the transmission and run detection algorithms). Our system uses a Transmitter (Tx) made out of a Raspberry Pi 3 and a receiver (Rx) based on a Realtek RTL SDR along with an off-the-shelf dipole antenna. We use the RTLSDR and SCIPY Python libraries for interfacing and transmitting narrowband FM waves at 433 MHz. We perform three basic experiments to examine whether our system can detect a single object and scale to multiple such objects. For the first experiment we train a Convolutional Neural Network (CNN) to detect the presence and absence of an object at close range. For the second experiment we double the range of our system setup and examine the impact on detection. Finally, for the third experiment we examine whether our setup is capable of distinguishing the presence of two similar objects from the presence of a single one. Three datasets, namely D_A, D_B and D_C are collected for each of the three experiments correspondingly.

D_A, D_B and D_C are collected in a closed room (8.5 m×5.5 m). For D_A, Tx and Rx are placed 1 m apart; the distance is doubled for D_B. In both cases the object (a 1 L water bottle) is placed exactly midway. We sample the received signal in raw I/Q form, with each recorded sample being 0.5 s. D_A is sampled at 2.6 MS/s, whereas D_B is recorded at a reduced 2 MS/s to decrease computational overhead. 2000 samples are collected in DA, and 6000 in DB. For DC, Tx and Rx are placed 2 m apart (as in DB. Data is collected for two cases, first the object is placed exactly midway as done in D_A and DB, for the second case two of the same bottles are placed 20 cm apart in line of sight of the antenna. For each of these cases 4500 samples are collected with each recorded sample being 0.5 s and a

sample rate of 2 MS/s. The setup used for collecting D_C has been shown in Fig. 2.

For D_C, Tx and Rx are placed 2 m apart (as in D_B. Data is collected for two cases, first the object is placed exactly midway as done in D_A and D_B, for the second case two of the same bottles are placed 20 cm apart in line of sight of the antenna. For each of these cases 4500 samples are collected with each recorded sample being 0.5 s and a sample rate of 2 MS/s. The setup used for collecting D_C has been shown in Fig. 2.

Fig. 2. System setup for detection of 2 identical bottled kept together, Transceiver and Receiver are placed 2 m apart.

5. Preprocessing and Model

After the three sets of data are collected, the data is operated on in three different representations. The first is a raw time series radio signal, and the second is a power spectral density function generated for our samples as in [22]. For the final representation we apply Short-Time Fourier Transform (STFT) to each sample to convert it into a spectrogram representation in the frequency-time domain (50 % overlap in each chunk) [18, 36]. We obtain 1,000 and 3,000 samples for each case (object, no object) in D_A and D_B respectively. In D_c we obtain 4,500 samples for each case (one object, two objects 20 cm apart).

We start with a baseline architecture as shown in Fig. 3 consisting of two convolution layers and two dense layers, then progressively vary the hyperparameters to analyze their effect on performance and arrive at the final model(s). The selected CNN model is based on the VGG-16 [26] architecture which has shown remarkable success in image recognition. It contains four convolutional layers, each followed by a pooling layer using SAME padding. ReLU activation functions are used in the convolutional layers to introduce non-linearity and a Glorot uniform kernel initializer is used to initialise the convolution layers. Two fully connected layers follow the stack of convolutional layers. The final dense layer uses a softmax activation function and the model is trained using binary cross-entropy loss via Stochastic Gradient Descent. For the model used on D_A, Batch Normalisation is done before the inputs are fed into the second convolutional layer and 40 % of the neurons are dropped after the final convolutional block to prevent overfitting. In contrast, for D_B, Batch Normalisation is done after every convolutional layer and the first dense layer, and 50 % of the neurons are dropped after the second and fourth convolutional layer.

The model trained on D_c is slightly different, it uses four convolutional layers with SAME padding and ReLU activations, with a pooling layer after the second and fourth convolutional layer. Two fully connected layers follow the stack of convolutional layers, Batch Normalisation is done after each convolutional layer and the first dense layer. The final dense layer uses a softmax activation function and 50 % of the neurons are dropped after the second and fourth convolutional layer and the first dense layer. We evaluate the performance of our CNN classifiers using 5-fold Cross Validation technique[3]

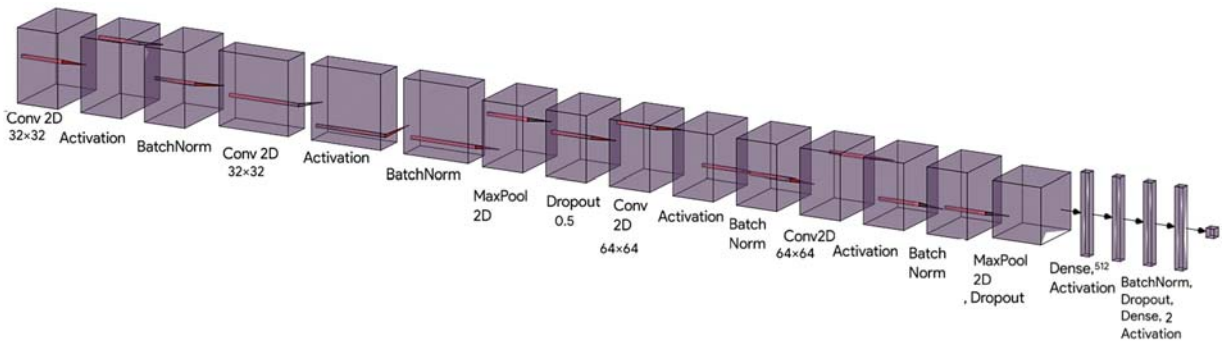

Fig. 3. Visualisation of the CNN architecture used for D_C.

[3]Our models are built using Tensorflow and trained on an NVIDIA RTX 2070

6. Observations

Training a three layered shallow network using the raw I/Q Data and PSD data did not yield useful results as the classifier was unable to generalize well on the test set. This Suggests that the input data was not an adequate feature for detecting the object (or was too noisy). However, we observed that training the CNN described above using the heatmaps of the reflected signals as input yielded a high accuracy on the test set. After training the CNN on D_A for 10 epochs, the network is able to achieve a test accuracy of 0.96 (average over 10 runs). Furthermore, an accuracy of 0.87 was observed on the test set after training and evaluating the CNN on D_B for 20 epochs; there is a decrease in inaccuracy as the distance between T_x and R_x (and the distance of the object from T_x and R_x) is doubled. Still, our system is able to perform high-accuracy object detection even after the range is increased.

Finally, training the described CNN on D_C resulted in a high testing accuracy of 0.99, suggesting that our system is not only capable of simple object detection, but can also be scaled to detect the presence of multiple similar objects. Given this initial success, it would be interesting to analyze the robustness of our low-frequency detection through further experiments. The direct next steps would be determining:

a) The exact functioning range of such a setup,

b) Whether the position of objects be localised to what precision? and

c) How many different/identical objects can be tracked simultaneously and whether our system can track human bodies.

The accuracy results are recorded in Table 1 and the confusion matrix for D_B and D_C is shown in Fig. 4 and Fig. 5 respectively. Furthermore, loss and accuracy with respect to D_A are plotted against the epochs and shown in Fig. 6.

This demonstrates that CNNs are an effective model for identifying and extracting relevant features to carry out such classification tasks.

Table 1. Comparison of performance of CNN on D_A, D_B (as the distance between T_x and R_x is doubled from 1 m to 2 m) and D_C.

Dataset	Validation Accuracy	Testing Accuracy
D_A	0.94	0.96
D_B	0.91	0.87
D_C	0.98	0.99

7. Conclusions

This paper proposes that it is possible to perform simple object detection with high accuracy using extremely cheap radio hardware by leveraging CNNs. We demonstrate this by developing a proof-of-concept system to detect the presence of a water bottle based on spectrograms of the received signal. Additionally, we are also able to use our system to detect the presence of multiple identical objects. Critically, we do almost no explicit signal processing or feature engineering, relying on the neural network to do the heavy lifting. This indicates the possibility of doing complex, customized object detection without expense on human capital or classical signal processing.

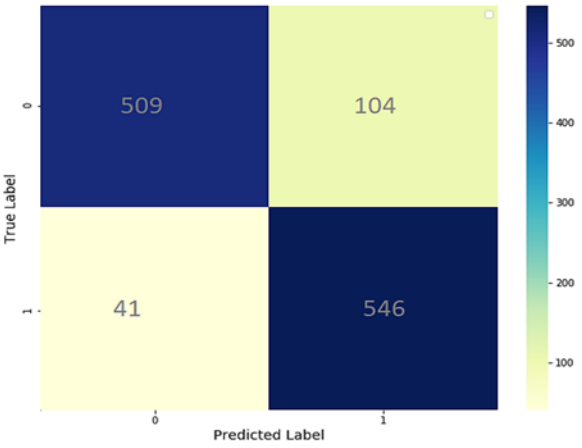

Fig. 4. Confusion matrix for D_B.

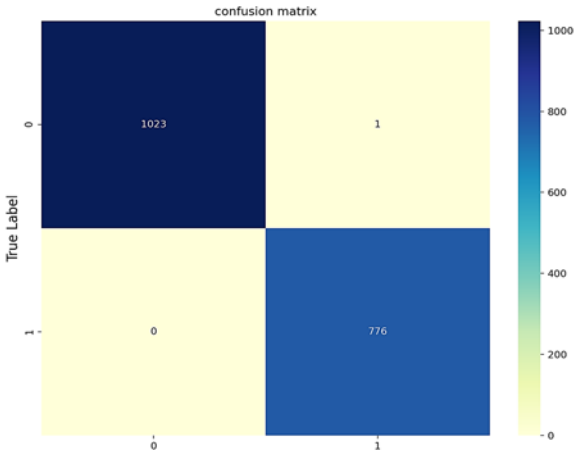

Fig. 5. Confusion matrix for D_C.

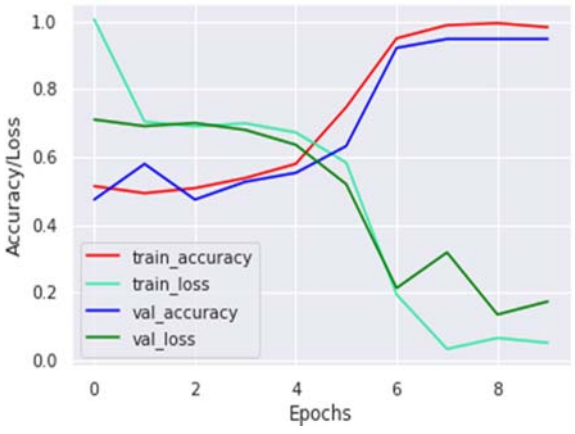

Fig. 6. Accuracy/Loss vs Epochs for D_A.

In the future, we plan to scale the current system by using multiple SDRs to emit an FMCW wave, to determine how the system works in terms of detecting and localizing multiple objects (and human bodies) when it is fed more information about the surroundings. Furthermore, we want to extend our work to perform material detection and classification tasks; additionally, we also want to test whether our low-frequency system can be used to localize objects in an indoor environment. We also want to perform far more complex classifications, reporting precise location of objects, calibrating item sizes, handling more noise, and adapting our system for new environments.

References

[1]. F. Adib, D. Katabi, See through walls with WiFi!, in *Proceedings of the Association for Computing Machinery Conference*, 2013, pp. 75-86.

[2]. F. Adib, Z. Kabelac, D. Katabi, R. C. Miller, 3D tracking via body radio reflections, in *Proceedings of the 11th USENIX Symposium on Networked Systems Design and Implementation (NSDI 14),* Seattle, WA, 2014, pp. 317–329.

[3]. F. Adib, C.-Y. Hsu, H. Mao, D. Katabi, F. Durand, Capturing the human figure through a wall, *Association for Computing Machinery Transactions on Graphics*, Vol. 34, Issue 6, Oct. 2015, pp. 1-13.

[4]. F. Adib, Z. Kabelac, D. Katabi, Multi-person localization via RF body reflections, in *Proceedings of the 12th USENIX Symposium on Networked Systems Design and Implementation (NSDI 15)*, Oakland, CA, pp. 279–292, May 2015.

[5]. M. A. A. Al-qaness, Device-free human micro-activity recognition method using WiFi signals, *Geospatial Information Science*, Vol. 22, Issue 2, 2019, pp. 128–137.

[6]. P. Ali-Rantala, L. Ukkonen, L. Sydanheimo, M. Keskilammi, M. Kivikoski, Different kinds of walls and their effect on the attenuation of radiowaves indoors, in *Proceedings of the IEEE Antennas and Propagation Society International Symposium (IEEE-APS)*, Vol. 3, 2003, pp. 1020–1023.

[7]. Y. Aytar, C. Vondrick, A. Torralba, SoundNet: Learning sound representations from unlabeled video, in *Proceedings of the 30th International Conference on Neural Information Processing Systems (NIPS'16),* Red Hook, NY, USA, 2016, pp. 892–900.

[8]. P. Bahl, V. N. Padmanabhan, Radar: an in-building RF-based user location and tracking system, in *Proceedings of the IEEE Conference on Computer Communications. Nineteenth Annual Joint Conference of the IEEE Computer and Communications Societies (INFOCOM'2000),* Vol. 2, 2000, pp. 775–784.

[9]. G. Benton, M. Finzi, P. Izmailov, A. G. Wilson, Learning invariances in neural networks, in *Proceedings of the 34th Conference on Neural Information Processing Systems (NeurIPS 2020),* Vancouver, Canada, 2020.

[10]. J. Deng, W. Dong, R. Socher, L. Li, Kai Li, Li Fei-Fei, ImageNet: A large-scale hierarchical image database, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2009, pp. 248–255.

[11]. J. Ehrich, E. Molloy, M. Mantovani, Conceptual design of future children's hospitals in europe: Planning, building, merging, and closing hospitals, *The Journal of Pediatrics*, Vol. 182, Issue 12, 2016, pp. 411-412.

[12]. A. Esteva, A. Robicquet, B. Ramsundar, V. Kuleshov, M. DePristo, K. Chou, C. Cui, G. Corrado, S. Thrun, J. Dean, A guide to deep learning in healthcare. *Nature Medicine*, Vol. 25, Issue 1, January 2019, pp. 24–29.

[13]. R. Evans, J. Gao, DeepMind AI reduces google data centre cooling bill by 40 %, Jul 2016. https://deepmind.com/blog/article/deepmind-ai-reduces-google-data-centre-cooling-bill-40.

[14]. S. Hochreiter, J. Schmidhuber, Long short-term memory, *Neural Computation*, Vol. 9, Issue 8, Nov. 1997, pp. 1735–1780.

[15]. F. Hu, Y. Deng, W. Saad, M. Bennis, A. H. Aghvami, Cellular-connected wireless virtual reality: Requirements, challenges, and solutions, *IEEE Communications Magazine*, Vol. 58, Issue 5, 2020, pp. 105–111.

[16]. S. Ioffe, C. Szegedy, Batch normalization: Accelerating deep network training by reducing internal covariate shift, in *Proceedings of the 32nd International Conference on Machine Learning (ICML'15),* Vol. 37, 2015, pp. 448–456.

[17]. K. Kang, H. Li, J. Yan, X. Zeng, B. Yang, T. Xiao, C. Zhang, Z. Wang, R. Wang, X. Wang, W. Ouyang, T-CNN: Tubelets with convolutional neural networks for object detection from videos, *IEEE Transactions on Circuits and Systems for Video Technology*, Vol. 28, Issue 10, 2018, pp. 2896–2907.

[18]. T. Li, L. Fan, M. Zhao, Y. Liu, D. Katabi, Making the invisible visible: Action recognition through walls and occlusions, in *Proceedings of the Institute of Electrical and Electronics Engineers- International Conference on Computer Vision (IEEE/CVF- ICCV)*, 2019, pp. 872–881.

[19]. T. J. O'Shea, J. Corgan, T. C. Clancy, Convolutional radio modulation recognition networks, in Jayne C., Iliadis L. (eds) Engineering Applications of Neural Networks. EANN 2016. Communications in Computer and Information Science, Vol. 629. *Springer,* 2016.

[20]. T. J. O'Shea, T. Roy, T. C. Clancy, Over the air deep learning based radio signal classification, *IEEE Journal of Selected Topics in Signal Processing*, Vol. 12, Issue 1, Feb. 2018, pp. 168 – 179.

[21]. A. Parameswaran, M. Iftekhar Husain, S. Upadhyaya, Is RSSI a reliable parameter in sensor localization algorithms: an experimental study, *Computer Science*, 2009.

[22]. A. Pérez-Zapata, A. F. Cardona-Escobar, J. A. Jaramillo-Garzón, G. M. Díaz, Deep convolutional neural networks and power spectral density features for motor imagery classification of EEG signals, in *Proceedings of the International Conference on Augmented Cognition (AC)*, *Springer International Publishing*, 2018, pp. 158–169.

[23]. A. F. Pérez-Zapata, A. F. Cardona-Escobar, J. A. Jaramillo-Garzón, G. M. Díaz, Deep convolutional neural networks and power spectral density features for motor imagery classification of EEG signals, in D. D. Schmorrow and C. M. Fidopiastis, editors, *Augmented Cognition: Intelligent Technologies*, *Springer International Publishing*, 2018, pp. 158-169.

[24]. Q. Pu, S. Gupta, S. Gollakota, S. Patel, Whole-home gesture recognition using wireless signals, in *Proceedings of the 19th Annual International Conference on Mobile Computing & Networking (MobiCom'13)*, New York, NY, USA, 2013, pp. 27–38.

[25]. K. Shibata, H. Yamamoto, People crowd density estimation system using deep learning for radio wave sensing of cellular communication, in *Proceedings of the International Conference on Artificial Intelligence in Information and Communication (ICAIIC)*, 2019, pp. 143–148.

[26]. K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition, in *Proceedings of the International Conference on Learning Representations*, 2015.

[27]. G. Sreenu, M. A. S. Durai, Intelligent video surveillance: a review through deep learning techniques for crowd analysis, *Journal of Big Data*, Vol. 6, Issue 1, June 2019.

[28]. N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, R. Salakhutdinov, Dropout: A simple way to prevent neural networks from overfitting, *J. Mach. Learn. Res.*, Vol. 15, Issue 1, Jan. 2014, pp. 1929–1958.

[29]. C. Wang, X. Guo, Y. Wang, Y. Chen, B. Liu, Friend or foe?: your wearable devices reveal your personal PIN, in *Proceedings of the 11th ACM on Asia Conference on Computer and Communications Security (ASIA CCS'16)*, New York, NY, USA, 2016, pp. 189–200. https://doi.org/10. 1145/2897845.2897847.

[30]. G. Wang, Y. Zou, Z. Zhou, K. Wu, L. M. Ni, We can hear you with Wi-Fi!, *IEEE Transactions on Mobile Computing*, Vol. 15, Issue 11, 2016, pp. 2907–2920.

[31]. C. Xu, B. Firner, R. S. Moore, Y. Zhang, W. Trappe, R. Howard, F. Zhang, N. An, SCPL: Indoor device-free multi-subject counting and localization using radio signal strength, in *Proceedings of the 12th International Conference on Information Processing in Sensor Networks (IPSN'13)*, New York, NY, USA, 2013, pp. 79–90.

[32]. D. Zhang, Y. Liu, L. M. Ni, RASS: A real-time, accurate and scalable system for tracking transceiver-free objects, in *Proceedings of the IEEE International Conference on Pervasive Computing and Communications (PerCom)*, 2011, pp. 197–204.

[33]. M. Zhao, S. Yue, D. Katabi, T. S. Jaakkola, M. T. Bianchi, Learning sleep stages from radio signals: A conditional adversarial architecture, in *Proceedings of the 34th International Conference on Machine Learning (ICML'2017)*, Vol. 70, 2017, pp. 4100–4109.

[34]. M. Zhao, F. Adib, D. Katabi, Emotion recognition using wireless signals, *Commun. ACM*, Vol. 61, Issue 9, Aug. 2018, pp. 91–100.

[35]. M. Zhao, T. Li, M. A. Alsheikh, Y. Tian, H. Zhao, A. Torralba, D. Katabi, Through-wall human pose estimation using radio signals, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018, pp. 7356–7365.

[36]. M. Zhao, Y. Tian, H. Zhao, M. A. Alsheikh, T. Li, R. Hristov, Z. Kabelac, D. Katabi, A. Torralba, Rf-based 3D skeletons, in *Proceedings of the Association for Computing Machinery Conference: Special Interest Group on Data Communication (ACM-SIGCOMM)*, 2018, pp. 267–281.

[37]. Z. Zou, Q. Chen, I. Uysal, L. Zheng, Radio frequency identification enabled wireless sensing for intelligent food logistics, *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, Vol. 372, Issue 2017, June 2014, 20130313.