Attention Enhanced Alexnet For Learning Salient Features in Radar Micro-Doppler Signatures

Shelly Vishwakarma*, Wenda Li*, Chong Tang*, Fangzhan Shi*, Raviraj Adve†, Kevin Chetty*
*Department of Security and Crime Science, University College London, UK
† Department of Electrical and Computer Engineering, University of Toronto, Canada

***Abstract*—This work introduces an attention mechanism that can be integrated into any standard convolution neural network (CNN) to improve model sensitivity and prediction accuracy with minimal computational overhead. We introduce the attention mechanism in a lightweight network- Alexnet and evaluate its classification performance for human micro-Doppler signatures. We show that the Alexnet model trained with an attention module can implicitly learn to highlight the salient regions in the radar signatures whilst suppressing the irrelevant background regions and consistently improve the network predictions by more than 4% in most cases. We further provide network visualizations through class activation mapping, providing better insights into how the predictions are made.**

***Index Terms*—Radar Sensing, Attention Networks, Deep Learning, Micro-Doppler Signatures, Human Activity Recognition**

I. Introduction

In recent years, deep convolutional neural networks (DCNNs) have become the state-of-the-art method for classifying human micro-Doppler signatures [1]. DCNNs can jointly learn informative features and classification boundaries, resulting in them being an order of magnitude faster than traditional approaches that use additional feature extraction algorithms. The success of DCNNs is attributed to the ever-increasing processing speeds of computers, greater availability of digitally recorded data, and almost unlimited memory capacity.

Unlike the vision community, radar researchers are constrained by the limited availability of open radar databases. Therefore, researchers have used different DCNN initialization methods for micro-Doppler classification with low training sample support. One such method is using a transfer learning technique where pre-trained networks from optical imagery (such as AlexNet, VGGNet, GoogleNet) are trained with a limited radar data set [2], [3]. However, the performance using low-weight networks such as Alexnet remained sub-optimal, possibly due to the low interpretability of the radar micro-Doppler signatures, especially at lower carrier frequencies.

To address this general problem, we propose a simple and yet effective solution, called attention mechanism [4]–[6]. The attention mechanism can automatically localize and highlight the salient regions of interest in the radar micro-Doppler signatures. In addition, it can improve model sensitivity and accuracy by suppressing feature activations in irrelevant regions. The attention modules are highly flexible and can be integrated with any existing DCNN architecture without introducing significant computational overhead in model parameters. The attention-enhanced DCNN (AE-DCNN) can be trained similarly to any standard DCNN network.

Given an intermediate feature map, our proposed attention module jointly utilizes the global features computed at the network's last layer to highlight salient local regions of interest at intermediate layers. Since the attention module uses global features to refine intermediate layer features, we termed this as a global spatial attention module (GSAM). The attention refined features from the intermediate layers are then aggregated with global features to yield the final predictions. In this work, we incorporate GSAM into a lightweight network- Alexnet, to demonstrate its effectiveness in automatically localizing the object of interest and improving the overall classification performance. We choose to evaluate our implementation on a publically available radar dataset in [7]. The dataset has been acquired using three synchronized RF sensors at three frequencies- 10GHz, 24GHz, and 77GHz. It comprises radar micro-Doppler signatures corresponding to eleven human activities of daily life. The results show that AE-Alexnet consistently improves prediction accuracy across different datasets while achieving performance better than complex state-of-the-art DCNN models such as Resnet and VGG.

Attention mechanisms have been commonly used in natural language processing (NLP) tasks such as image captioning, and machine translation [4]–[6]. In computer vision, it has been applied to -Image classification, Image Segmentation, and Image captioning [8], [9]. Attention models have also been exploited for medical report generation, and medical image classification [10]. In the context of radar image analysis, attention models have been exploited for synthetic aperture radar (SAR) image segmentation and classification problems [11]. More recently, it has been used for the classification of the high range resolution profile of radar targets [12]. However, our work uses a more complex cascaded network architecture for the desired task. Only a handful of works use attention mechanisms. To the best of the authors knowledge there is no literature available on using attention mechanisms to classify micro-Doppler signatures. Therefore, this work proposes one of the first used cases of attention mechanism in a feed-forward CNN model applied to a radar micro-Doppler signature classifi-

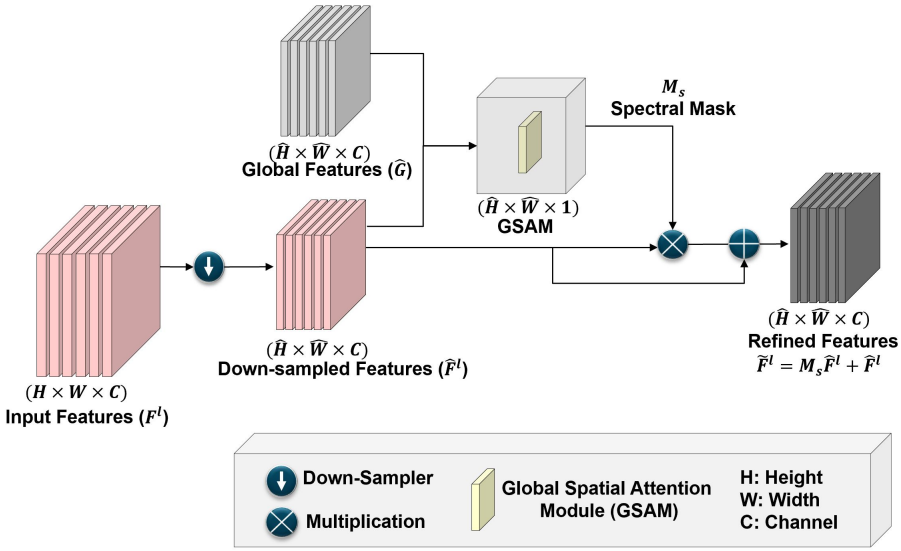

Fig. 1: Schematic of the proposed attention module. The intermediate feature maps F^l are adaptively refined and scaled with spatial attention mask M_s. Spatial regions are selected by jointly analysing the global features $\hat{G}$, and the intermediate layer features F^l. The idea is to attend the features on a regional basis that are most relevant for the human activity recognition task.

cation. The modified network is lightweight and end-to-end trainable.

II. Methodology

This section introduces the proposed attention mechanism that can be incorporated into any existing CNN architecture to improve its performance; however, we chose a lightweight Alexnet to be our base architecture in this work.

A. Alexnet

A standard AlexNet model contains eight layers; the first five are convolutional layers, the last three are fully connected layers, followed by a classification layer with a SoftMax activation function [13]. Finally, max-pooling layers follow the first two layers.

This work demonstrates that improved performance can be achieved by integrating attention modules in standard Alexnet architecture. Furthermore, it does not require multiple additional layers or the training of multiple models. Instead, it progressively suppresses feature responses in irrelevant background regions without the requirement to crop a region of interest and enhances the response by putting more weight on the most crucial spatial structural information in the radar micro-Doppler signatures.

B. Attention Enhanced Alexnet (AE-Alexnet)

1) Global Spatial Attention Module

Fig.1 presents the proposed global spatial attention module (GSAM). Given the feature maps $F^l \in \mathbb{R}^{H \times W \times C}$ at chosen intermediate layer $l \in 1,2,..,L$, GSAM computes a two-dimensional spatial attention mask M_s, where the entries of $M_s \in [0,1]$, in order to identify salient local information in the feature maps F^l and prune feature responses to suppress the information in the irrelevant regions. It does so by jointly utilizing the feature maps at the last convolutional layer (global features $\hat{G}$) and the feature maps F^l at any intermediate layer l. The deeper layers encode global information from a large spatial context to identify the location of the target objects in the images and model their relationship at a global scale. Therefore, these global features can provide flexibility regarding focusing on a regional basis and disambiguate irrelevant feature content present in intermediate layer features F_l. Here, H, W, and C are the feature maps' height, width, and the number of channels at any layer l.

In standard CNN architectures, the feature-map is gradually down-sampled to capture sufficiently large receptive fields. Therefore, the resulting spatial resolution of each layer might be different. To generate the attention mask M_s, we can either up-sample the global feature maps $\hat{G} \in \mathbb{R}^{\hat{H} \times \hat{W} \times C}$ to match to the intermediate feature maps' F^l spatial resolution $(H \times W)$. Since the spatial resolution of the feature maps might differ from layer to layer, the spatial grid re-sampling of the input feature maps F^l is performed to obtain feature maps $\hat{F}^l$ of the size equivalent to global features $\hat{G} \in \mathbb{R}^{\hat{H} \times \hat{W} \times C}$, where, $\hat{H}$, $\hat{W}$ and C are the height, width and number of channels of $\hat{G}$. The output of GSAM is $\tilde{F}^l = M_s \hat{F}^l$, where each feature map $\hat{F}^l \in \mathbb{R}^{\hat{\mathbb{H}} \times \hat{\mathbb{W}} \times \mathbb{C}}$ is scaled by the 2D spatial attention map $M_s \in \mathbb{R}^{\hat{\mathbb{H}} \times \hat{\mathbb{W}}}$.

The detailed global attention process is depicted in Fig.2. To compute the spatial attention, we first apply two pooling operations- average and max-pooling along the channel axis of both, the intermediate feature maps $\hat{F}^l$ and the global feature maps $\hat{G}$. The operations result in the generation of four efficient feature maps $\hat{F}^l_{Avg}$, $\hat{G}_{Avg}$, $\hat{F}^l_{Max}$, $\hat{G}_{Max}$ each of size $\hat{H} \times \hat{W}$. The average-pooled features $\hat{F}^l_{Avg}$ and $\hat{G}_{Avg}$ are added together and passed through a non-linear

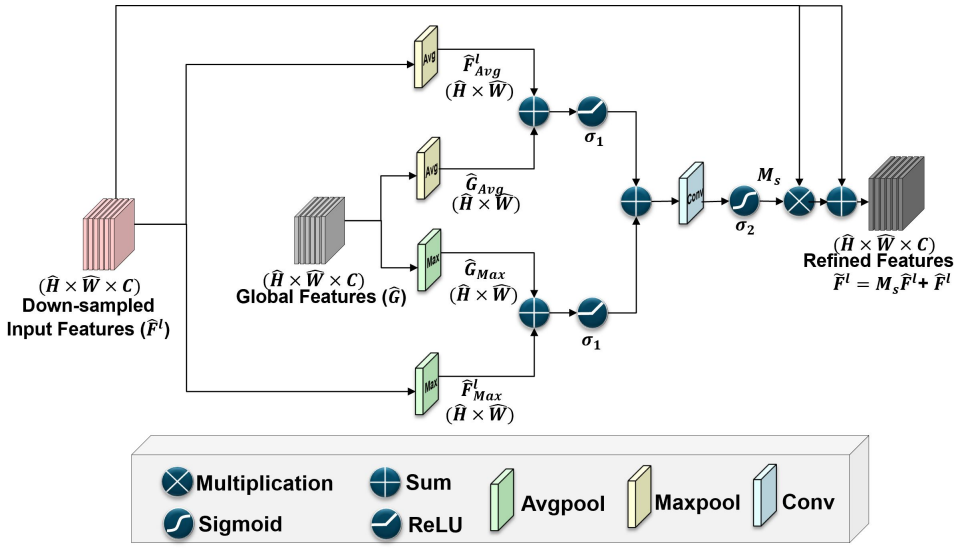

Fig. 2: The architecture of global spatial attention module (GSAM). As illustrated, GSAM jointly utilizes intermediate feature maps and the global features maps to compute the sum of the respective max-pooled and average-pooled features along the channel axis. The resulting 2D spatial maps are added and forwarded to a convolution layer. The range of the 2D-spatial attention mask M_s is restricted between [0, 1] through an element-wise sigmoid operation.

activation function σ_1 to focus on the informative regions in $\hat{F}^l$ relative to global information. The same process is repeated for the max-pooled features. The average-pooled and max-pooled features descriptors are finally added together and passed through a convolution layer to generate a global spatial attention map $M_s(F^l)$ encoding the regions to emphasize or suppress. In short, the global spatial attention map can be formulated as

$$M_s(F^l) = \sigma_2(f^{1\times1}(\sigma_1(\hat{F}^l_{Avg} + \hat{G}_{Avg}) + \sigma_1(\hat{F}^l_{Max} + \hat{G}_{Max}))) \quad (1)$$

Where $F^l_{Avg} = AvgPool(F^l)$, $\hat{G}_{Avg} = AvgPool(\hat{G})$, $F^l_{Max} = MaxPool(F^l)$, $\hat{G}_{Max} = MaxPool(\hat{G})$, and σ_2 is the normalisation function which can be sigmoid or softmax operation to restrict $M_s \in [0,1]$. $f^{1\times1}$ represents a convolution operation with the filter of size 1×1. However, we used element-wise sigmoid operation to normalise the spatial mask.

The final refined features $\tilde{F}^l$ are computed by the element-wise multiplication $\bigotimes$, of the attention mask M_s with the down-sampled intermediate layer feature maps $\hat{F}^l$ as shown below

$$\tilde{F}^l = M_s \bigotimes \hat{F}^l \quad (2)$$

During multiplication, the spatial attention map M_s of size $\hat{H}\times\hat{W}$ are copied along the channel dimension of the $\hat{F}^l \in \mathbb{R}^{\hat{\mathbb{H}}\times\hat{\mathbb{W}}\times\mathbb{C}}$ resulting in overall size of $\hat{H}\times\hat{W}\times C$.

2) GSAM Enhanced Alexnet For Classification

Fig.3 presents the attention-gated classification model of Alexnet. The proposed attention units are incorporated into the 2^{nd}, 3^{rd} and 4^{th} layer of the Alexnet to exploit local information present in these intermediate layers. We found that the attention maps are less effective if applied to the first layer feature maps as the first layer represents very low-level features that are not discriminative enough to require attention.

We use activation maps of the 5^{th} layer as our global features. Generally, the global feature maps must encode global spatial contextual information; it is usually obtained from the layer just before the final softmax layer. However, in Alexnet, the layer before the softmax layer is the fully-connected layer. In the context of radar micro-Doppler signatures, since most signatures of interest are highly localized, flattening may have the disadvantage of losing important spatial contextual information. Therefore, we consider the 5^{th} activation maps as our global features (right before any flattening is done).

The local feature maps at 2^{nd}, 3^{rd} and 4^{th} layers are passed through the GSAM along with the global feature maps to obtain the attention refined feature maps. Then, we aggregate these attention refined features and the global features together to yield the final predictions. In order to do so, we first compute the global average pooling along the spatial axis, resulting in a vector of length equal to the number of channels in refined feature maps. In addition, we also perform the global average pooling on the global feature maps. Subsequently, the average pooled features are concatenated and passed through two fully connected layers. Finally, a softmax operation is applied to the resulting flattened vector, and the entry with maximum activation is selected as the prediction.

III. Experimental Dataset Description and Results

A. Evaluation Datasets

We test the performance of the AE-Alexnet on the publically available radar dataset acquired from three synchronized RF sensors at the following three frequencies- 10GHz,

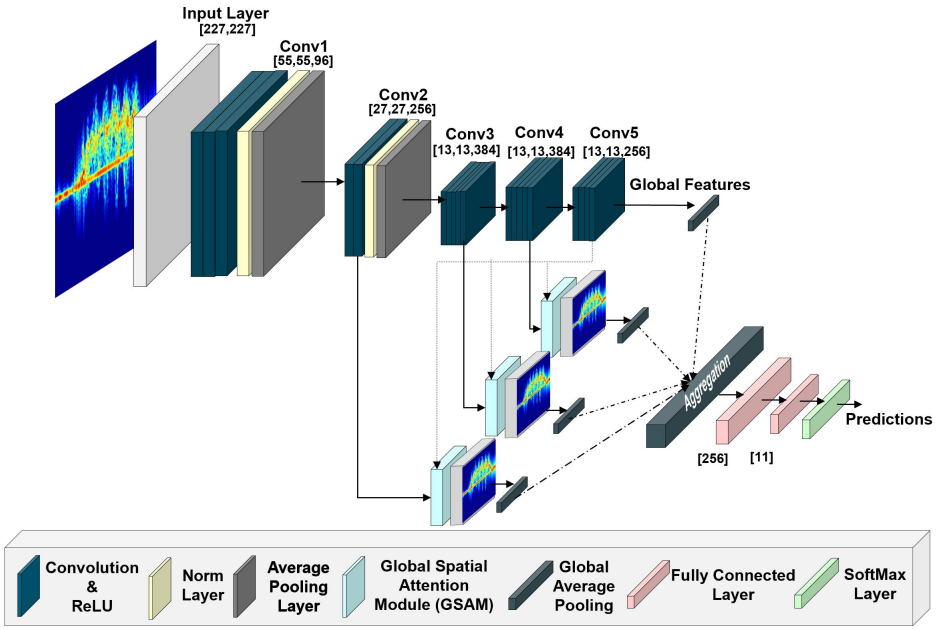

Fig. 3: The architecture of the proposed AE-Alexnet for human activity recognition.

TABLE I: Human Activity Dataset Description

Activity Number	Activity ID	Activity
1	WLKT	Walking towards the radar
2	WLKA	Walking away from the radar
3	PICK	Pick up an object from the ground
4	BEND	Bending
5	SIT	Sitting on a chair
6	KNEEL	Kneeling
7	CRWL	Crawling towards the radar
8	WTOES	Walking on both toes
9	LIMP	Limping with right leg stiff
10	SHSTEP	Walking with short steps
11	SCSSR	Scissor gait

24GHz, and 77GHz [7]. The experimental setup used for data acquisition placed all the sensors side-by-side at the height of 1-meter from the ground, with the test subject moving between 0.5m to 3m in front of the sensors. For the detailed explanation of the data acquisition protocol, we refer the readers to [7].

The dataset consists of micro-Doppler signatures of six participants of different heights, gender and ages groups performing eleven different activities of daily living, as listed in Table I. These activities are mainly inspired by intelligent home applications, where monitoring of daily living can help support non-intrusive health monitoring and enabling healthy living [14]. Each participant repeated these activities ten times, resulting in 60 radar signatures per class per sensor. Note that all the experiments are performed in line-of-sight conditions.

B. Network Training Parameter Settings

We empirically found the following parameter settings to be the most effective- optimization with adaptive moment estimation (ADAM) with an initial learning rate of α of 0.0001, gradient decay factor of $\beta_1 = 0.9$ and squared gradient decay factor of $\beta_2 = 0.99$. The learning rate is updated for every 100 epochs with a batch size fixed to 10. All the attention modules at intermediate layers are randomly initialized.

We perform training of our AE-Alexnet on Matlab 2020b, where all the variables are stored as 64-bit floats, with the following GPU configuration- GeForce GTX 1650 Ti, Compute Capability of '7.5' with a multi-processor count of 16.

C. AE-Alexenet Classification Results and Comparison to state-of-the-art Alexnet framework

We perform 5-Fold cross-validation on our dataset, where the entire dataset is split into five-folds, with each fold used as a testing set at some point. The first fold is used to test the model in the first iteration, and the rest are used to train the model. The second fold is used as the testing set in the second iteration, while the rest serve as the training set. This process is repeated until each fold of the 5-folds has been used as the testing set.

The 5-Fold classification results corresponding to the sensor dataset at 24GHz are presented in Table II. The performance difference over standard Alexnet is presented in the brackets. The highlighted values represent an improvement of over 1% compared to standard Alexnet. We used the following metrics for the class-wise classification performance evaluation: accuracy, precision, and recall. As we can observe, AE-Alexnet improves the results at all metric levels. It achieves higher precision and reduces the false-positive rate, likely because the attention mechanism suppresses irrelevant background in the radar signatures and forces the network to predict based on class-specific features. Moreover, we see that the precision improved by more than 5% in seven target classes which are significant enough to demonstrate the effectiveness of the attention

TABLE II: 5-Fold Class-wise classification performance for AE-Alexnet using micro-Doppler radar signature dataset acquired at 24GHz. The improvement over standard Alexnet is presented in the brackets. Values colored red highlights the improvement of more than 1%.

Activity	Accuracy	Precision	Recall
WLKT	0.968 **(2.2)**	0.994 **(10.4)**	0.968 **(2.2)**
WLKA	0.995 (0.6)	0.989 (0.4)	0.995 (0.5)
PICK	0.922 **(5.2)**	0.931 **(7.4)**	0.922 **(5.2)**
BEND	0.948 **(6.7)**	0.929 **(8)**	0.948 **(6.8)**
SIT	0.964 **(4)**	0.969 (0.2)	0.964 **(4.1)**
KNEEL	0.959 **(4.6)**	0.955 **(7.1)**	0.959 **(4.6)**
CRWL	0.983 **(3.3)**	0.986 (-0.2)	0.983 **(3.3)**
WTOES	0.959 **(17.9)**	0.930 **(6.2)**	0.959 **(17.9)**
LIMP	0.954 **(9.7)**	0.939 (-0.8)	0.954 **(9.7)**
SHSTEP	0.932 **(4.1)**	0.946 **(10)**	0.931 **(4.1)**
SCSSR	0.947 **(2.8)**	0.962 **(11.4)**	0.947 **(2.9)**

mechanism. The class-wise results for 10GHz and 77GHz sensor datasets will be presented in a subsequent paper.

To further rigorously evaluate our attention module, we perform additional classification experiments for 2-Fold, 3-Fold, and 4-Fold partitions and compare its performance with the 5-Fold dataset. We follow the same protocol specified in the previous section and benchmark its performance with standard Alexnet. Table III summarizes our experimental results. The AE-Alexnet outperforms the baseline; a consistent improvement in performance can be observed across all the folds and all frequency sensor data. It shows that GSAM boosts the accuracy of baselines significantly for 5-Folds and favorably improves the performance of more challenging 3-Fold and 2-Fold scenarios where limited data is used for training (indicating low-training sample support). The results demonstrate that our proposed approach is powerful, showing the efficacy of a new attention mechanism that generates richer spatial feature descriptors with a quite small overhead in terms of parameters and computation. It motivated us to apply our proposed module GSAM to lightweight networks like Alexnet and demonstrate its great potential for applications on low-end devices.

D. AE-Alexnet Visualisations With Class Activation Mapping

We use class activation mapping (CAM) for the qualitative analysis and determine which part of the input signatures is responsible for network predictions. We compare the CAM visualization results of attention refined features with standard Alexnet features. The resulting visualization maps are presented in Fig.4-Fig.6 corresponding to target class WLKT at three different frequencies- 10GHz, 24GHz, and 77GHz, respectively. The first column presents the ground truth signatures, the second presents the CAM visualization of raw features, and finally, the third column presents the CAM visualization of attention refined features obtained at the following intermediate convolutional layers- 2^{nd}, 3^{rd} and 4^{th}.

In Fig.4-Fig.6, we can see that the CAM masks of AE-Alexnet cover the class-specific activity regions better than unrefined features. The network bases its classification on the entire signature, but the most decisive input comes from the red areas. It shows that GSAM can effectively calculate

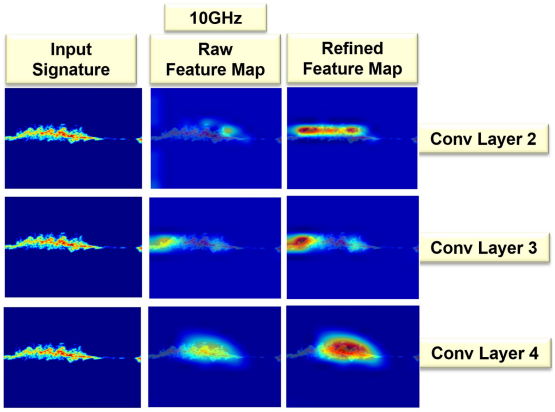

Fig. 4: Class activation mapping (CAM) visualisation result from AE-Alexnet for radar signatures classification at 10GHz. Red regions contribute the most. The detected region highly agrees with the object of interest.

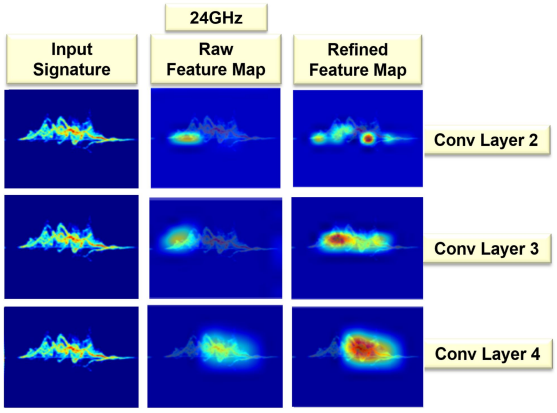

Fig. 5: Class activation mapping (CAM) visualisation result from AE-Alexnet for radar signatures classification at 24GHz.

the importance of the spatial locations in convolutional layers and detects the region that highly agrees with the object of interest. Although the attention map outlines the discriminant region, it does not necessarily coincide with the entire activity region.

IV. Discussion and Future Work

In this work, we propose an attention mechanism that jointly exploits the global information to highlight the local regions of interest in the intermediate network layers and suppress the background noise and the region of no-interest, thus significantly improving the classification performance compared to the standard network. The CAM visualization results show that the detected region that contributes the most to the predicted class highly agrees with the regions of interest. The proposed attention module could support explainable deep learning, a vital research area for automatic radar signature classification. In particular, we investigated several aspects, including-spatial attention mechanism, feature aggregation strategy with attention mechanism, and visualization techniques to give deeper insights into the predictions made by the

TABLE III: Multi-Fold AE-Alexnet classification results

Training	Testing	Baseline Accuracy Alexnet 5-Fold Accuracy	Modified Alexnet with Attention 5-Fold Accuracy	Modified Alexnet with Attention 4-Fold Accuracy	Modified Alexnet with Attention 3-Fold Accuracy	Modified Alexnet with Attention 2-Fold Accuracy
77GHz	77GHz	90.83	94.84 **(4.01)**	94.03**(3.2)**	94.23**(3.4)**	92.22**(1.39)**
24GHz	24GHz	90.15	95.74 **(4.91)**	94.85**(4.02)**	94.89**(4.06)**	92.83 **(2)**
10GHz	10GHz	90.93	93.89 **(3.06)**	92.65 **(1.82)**	91.75 **(0.92)**	90.71 **(-0.12)**

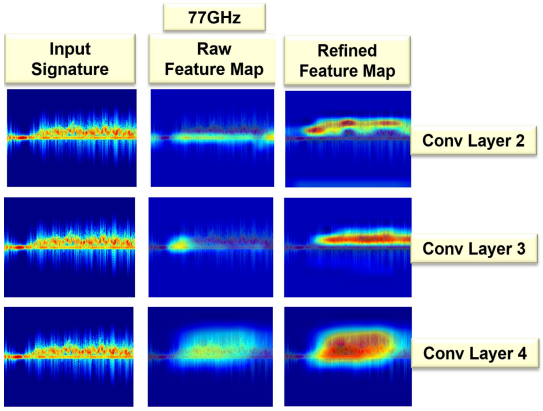

Fig. 6: Class activation mapping (CAM) visualisation result from AE-Alexnet for radar signatures classification at 77GHz.

network. However, many open research problems still need to be addressed.

1) The attention refined feature vectors are aggregated with the global feature vector to yield the final predictions. However, to do so, a global average pooling operation is performed along the spatial axis of the AE feature maps, resulting in a vector of length equal to the number of channels at each intermediate layer. Future work will investigate the aggregation of average pooled features along the channel axis, resulting in a feature matrix of size equal to the spatial dimension of the input maps. This 2D matrix can be flattened to generate a feature vector and aggregate these vectors for investigating better network predictions.
2) In this work, we consider Alexnet to be our base architecture for investigating the performance of the attention mechanism. However, the literature suggests that the VGG-16 network seems to perform well over radar micro-Doppler signatures [7]. Therefore, we believe introducing an attention mechanism into VGG-16 or more complex networks like ResNet could significantly improve the classification performance. Moreover, we hope to investigate the most practical combination of intermediate layers to give the best performance in the case of deeper networks such as VGG-16, VGG-19, and Resnet.
3) The dataset used to investigate the network performance comprised some poor quality radar signatures that could have significantly influenced the network's performance. Future investigations will drop the poor quality signatures and use data augmentation schemes such as adding additive white Gaussian noise (AWGN) to increase the training support and improve the overall classification performance. Initial investigation into data augmentation schemes found that adding AWGN noise data to the training further improved the performance of AE-Alexnet by more than 1% for the 77GHz dataset. However, the detailed experiments are still under investigation and will form part of future research.

ACKNOWLEDGMENTS

This work was funded under the OPERA Project, the UK Engineering and Physical Sciences Research Council (EPSRC), Grant EP/R018677/1.

REFERENCES

[1] M. Amin, *Radar for indoor monitoring: Detection, classification, and assessment*. CRC Press, 2017.
[2] M. S. Seyfioglu, B. Erol, S. Z. Gurbuz, and M. G. Amin, "Diversified radar micro-doppler simulations as training data for deep residual neural networks," in *2018 IEEE radar Conference (radarConf18)*. IEEE, 2018, pp. 0612–0617.
[3] S. Z. Gurbuz and M. G. Amin, "Radar-based human-motion recognition with deep learning: Promising applications for indoor monitoring," *IEEE Signal Processing Magazine*, vol. 36, no. 4, pp. 16–28, 2019.
[4] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in neural information processing systems*, 2017, pp. 5998–6008.
[5] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," *arXiv preprint arXiv:1409.0473*, 2014.
[6] M.-T. Luong, H. Pham, and C. D. Manning, "Effective approaches to attention-based neural machine translation," *arXiv preprint arXiv:1508.04025*, 2015.
[7] S. Z. Gurbuz, M. M. Rahman, E. Kurtoglu, T. Macks, and F. Fioranelli, "Cross-frequency training with adversarial learning for radar micro-doppler signature classification (rising researcher)," in *Radar Sensor Technology XXIV*, vol. 11408. International Society for Optics and Photonics, 2020, p. 114080A.
[8] F. Wang, M. Jiang, C. Qian, S. Yang, C. Li, H. Zhang, X. Wang, and X. Tang, "Residual attention network for image classification," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 3156–3164.
[9] S. Jetley, N. A. Lord, N. Lee, and P. H. Torr, "Learn to pay attention," *arXiv preprint arXiv:1804.02391*, 2018.
[10] J. Schlemper, O. Oktay, M. Schaap, M. Heinrich, B. Kainz, B. Glocker, and D. Rueckert, "Attention gated networks: Learning to leverage salient regions in medical images," *Medical image analysis*, vol. 53, pp. 197–207, 2019.
[11] B. Shi, Q. Zhang, D. Wang, and Y. Li, "Synthetic aperture radar sar image target recognition algorithm based on attention mechanism," *IEEE Access*, vol. 9, pp. 140 512–140 524, 2021.
[12] M. Pan, A. Liu, Y. Yu, P. Wang, J. Li, Y. Liu, S. Lv, and H. Zhu, "Radar hrrp target recognition model based on a stacked cnn-bi-rnn with attention mechanism," *IEEE Transactions on Geoscience and Remote Sensing*, 2021.
[13] M. Z. Alom, T. M. Taha, C. Yakopcic, S. Westberg, P. Sidike, M. S. Nasrin, B. C. Van Esesn, A. A. S. Awwal, and V. K. Asari, "The history began from alexnet: A comprehensive survey on deep learning approaches," *arXiv preprint arXiv:1803.01164*, 2018.
[14] S. A. Shah and F. Fioranelli, "Rf sensing technologies for assisted daily living in healthcare: A comprehensive review," *IEEE Aerospace and Electronic Systems Magazine*, vol. 34, no. 11, pp. 26–44, 2019.