A Hybrid Vision Transformer-based Capsule Network for Radar Automatic Modulation Recognition

Abdulrahman Al-Malahi [1], HanCong Feng [1], KaiLi Jiang [1], Bin Tang [1]

[1] School of Information and Communication Engineering, University of Electronic Science and Technology of China, Chengdu, China.

Abstract. In radar Automatic Modulation Recognition (AMR), Single neural networks fail to achieve a satisfactory recognition accuracy especially in low SNR conditions. In this paper, we propose a novel AMR framework called A Hybrid Vision Transformer-based Capsule Network (HVTCN) that integrates Vision Transformers (ViT) with Capsule Networks (CapsNet) to enhance recognition performance. The ViT extracts global dependencies from radar spectrograms, while the CapsNet maintains spatial relationships, improving classification accuracy. Evaluations on benchmark datasets demonstrate superior performance under varying signal conditions.

Keywords: Radar modulation recognition; radar sorting; combined neural network.

1. Introduction

The ability of designed systems for Automatic Modulation Recognition (AMR) to categorize the received signals into modulation categories is vital for interference management, cognitive radio applications, and electronic intelligence (ELINT) [1]. Traditional AMR methods primarily rely on manually designed features derived from time-domain, frequency-domain, and time-frequency representations. These approaches, however, struggle with generalization across diverse and dynamic signal environments, making them suboptimal in real-world applications [2].

Artificial intelligence (AI), particularly deep learning techniques, offers promising solutions to these challenges. Neural networks, which are highly effective in learning complex representations, have been successfully applied to radar signal processing tasks. In particular, deep neural networks, convolutional neural networks (CNNs) [3], and recurrent neural networks (RNNs) have shown significant improvements in modulation classification accuracy and robustness [4]. However, these models still face limitations in capturing long-range dependencies and preserving feature hierarchies [5,6].

To address these challenges many researchers have proposed different techniques [7-9] including incorporating two neural network architectures, which achieved better accuracy than single networks. In this paper, we propose a novel hybrid neural network architecture for radar AMR. The proposed HVTCN model combines the Vision Transformer and Capsule Networks to leverage their respective strengths. ViT’ s has a powerful ability to capture long-range dependencies [10-12], while CapsNet works offer an alternative to traditional CNNs by maintaining hierarchical spatial relationships [13, 14]. This property is particularly valuable in AMR, where modulation patterns can undergo spatial transformations due to channel variations. By combining the strengths of ViT and CapsNet, we introduce a novel AMR framework that effectively learns robust representations from radar spectrograms.

2. Problem Formulation

Radar AMR is formulated as a supervised classification problem. Given a received radar signal $x(t)$, our objective is to learn a mapping function $f(x)$ that classifies the modulation type from a predefined set $M = \{M_1, M_2, \ldots, M_n\}$. The input signal is typically represented as a spectrogram $S(x)$ obtained via the Short-Time Fourier Transform (STFT):

$$S_x(t,f) = \int x(\tau) w(\tau - t) e^{-j2\pi f \tau} d\tau \quad (1)$$

where $w(\tau)$ is the window function. The classification function f is learned through a neural network that extracts spectral-temporal features from $S_x(t, f)$.

3. Proposed Architecture

As Fig. 1 shows, our proposed architecture integrates a Vision Transformer for feature extraction with a Capsule Network for classification. The data flow through the system begins with the transformation of raw signals into spectrograms, which are then processed by the ViT to extract global dependencies. The extracted feature embeddings are subsequently passed to the CapsNet, which preserves spatial relationships and refines the classification decision.

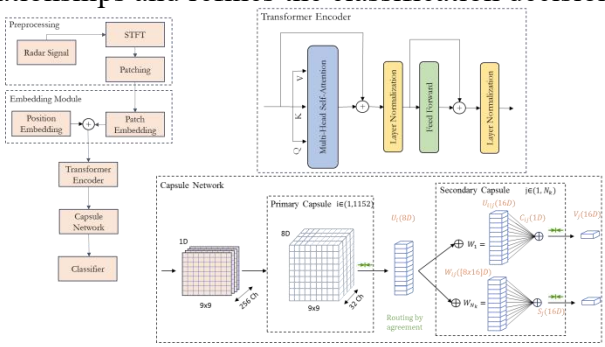

Figure. 1 Structure of the proposed HVTCN

3.1 Preprocessing

The received raw radar IQ signal is first converted into a spectrogram $X \in R^{H\times W}$ which is a 2D time-frequency representation of the signal obtained via STFT:

$$X(t, f) = \sum_n x[n]w[n - t]e^{-j2\pi fn} \quad (2)$$

where x[n] is the radar signal in time domain, w[n] is the window function, f is the frequency bin, and t is the time frame. Since transformers operate on sequences, the spectrogram is divided into $P \times P$ non-overlapping patches. Each patch is treated as a single vector:

$$x_p = \text{Reshape}(X) \in R^{N\times P^2} \quad (3)$$

where $N = (H \times W)/P^2$ is the total number of patches.

3.2 Embedding Module

Each patch vector x_p is then passed through a linear projection layer to obtain feature representations:

$$z_p = W_e x_p + b_e \quad (4)$$

where $W_e \in R^{d\times P^2}$ is a learnable weight matrix, b_e is the bias term, d is the embedding dimension. This transforms each patch into a d-dimensional feature vector. Since transformers lack convolutional inductive bias, positional encodings are added to the patch embeddings:

$$z_p^{'} = z_p + p_p \quad (5)$$

where p_p is the position encoding vector that ensures the network retains spatial information in a sequence.

3.3 Transformer Encoder

The core of ViT is the self-attention mechanism, which allows patches to attend to each other globally, Query, Key, and Value are computed as in 6, attention scores are computed as shown in 7, while attention output is calculated as in 8, where W_q, W_k, W_v are learnable projection matrices.

$$[Q, K, V] = [W_q z_p^{'}, W_k z_p^{'}, W_v z_p^{'}] \quad (6)$$

$$A = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right) \quad (7)$$

$$Z = AV \quad (8)$$

This process allows the network to capture long-range dependencies between patches. In multi-head self-attention, multiple attention heads are used, concatenated, and projected as in (9), where h is the number of attention heads.

$$\mathrm{MHSA}(z_p') = \mathrm{Concat}(Z_1, Z_2, \ldots, Z_h)W_o \quad (9)$$

The attention output is then passed through a feed-forward network (FFN):

$$z_{ffn} = \mathrm{ReLU}(W_1 Z + b_1)W_2 + b_2 \quad (10)$$

This output is then fed into linear layer to form the final ViT output feature representation which is then fed into the Capsule Network for further processing:

$$Z_{ViT} = \mathrm{LayerNorm}(z_{ffn}) \quad (11)$$

3.4 Capsule Network

Each feature vector corresponds to a part of the signal, capturing specific aspects of the modulation. Let the input to the primary capsule layer be represented as in (12), where N is the number of patches, and d is the dimensionality of the feature vectors:

$$S_j = Z_{ViT} \in R^{N \times d} \quad (12)$$

The capsules process the input embeddings and generate capsule vectors. A capsule v_i represents a group of neurons that encode information about a particular feature, such as pose or orientation. Each capsule output v_j is computed using dynamic routing between capsules. For each capsule j, v_j is obtained using squashing function by shrinking short vectors to a value near to 0 and long vectors to be near to 1 according to (13), where s_j is the total input.

$$v_j = \frac{\|s_j\|^2}{1+\|s_j\|^2}\frac{s_j}{\|s_j\|} \quad (13)$$

Let $\hat{u}_{i|j}$ represent the predicted output vectors from the previous layer capsules, for primary capsules, it represents the transformation matrix applied to the input Z_{ViT}. For capsule j, s_j is calculated according to (14), where c_{ij} are the routing coefficients, u_i is the capsule output in the layer below, and W_{ij} is a weight matrix.

$$s_j = \sum_i c_{ij}\hat{u}_{i|j}, \quad \hat{u}_{i|j} = W_{ij}u_i \quad (14)$$

The routing coefficients c_{ij} are updated iteratively through a dynamic routing algorithm, where the network learns to assign a higher weight to more relevant capsules refining their outputs based on the agreement between them in consecutive layers [15]. The iterative procedure ensures that capsules with similar responses are grouped together. The routing coefficients are updated using (15):

$$c_{ij} = \frac{\exp(b_{ij})}{\sum_k \exp(b_{ik})} \quad (15)$$

where b_{ij} is the log prior for the routing coefficient, and it is updated as follows:

$$b_{ij} \leftarrow b_{ij} + v_j \cdot \hat{u}_{i|j} \quad (16)$$

The normalization applied by the squashing function ensures that the length of the output capsule vector v_j corresponds to the probability of the presence of a feature, while the direction encodes the feature's properties. The length of the capsule vector is used to make the final decision:

$$\hat{y} = \mathrm{argmax}(\| v_j \|) \quad (17)$$

where $\hat{y}$ represents the predicted modulation class, corresponding to the highest activation in the final capsule vector.

4. Experimental Results

4.1 Experimental Setup

To evaluate the performance of the proposed model, experiments are conducted using the RadioML2018.01a dataset. The model is trained using a train-test split, and its performance is measured using the accuracy metric. experiments are implemented on Google Colab environment with a single 12GB NVIDIA Tesla K80 GPU and a 64GB RAM, all models were trained on PyTorch. The ratios of training, validation, and test sets are 70%, 15%, and 15%, respectively.

4.2 Results and Comparison with State-of-the-Art Models

As Fig. 2-(a) shows, the proposed HVTCN model outperforms traditional networks such as ResNet and LSTM as well as recently proposed combined networks such as MCLDNN and others, achieving higher accuracy. Table 1 illustrates the average accuracy in different ranges of SNR, it is obvious that the proposed HVTCN model keeps the superiority over the other models in both high and low SNRs, which reflects its robustness in noisy environments, increasing the model generalization of use,

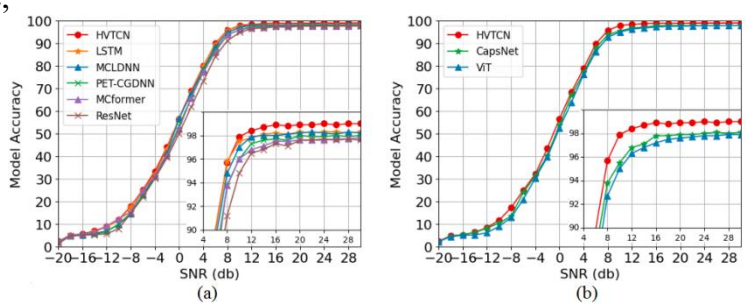

Figure. 2 Recognition performance results on RadioML2018.01a (a) comparative results (b) ablation results

Table 1. Recognition accuracy of different models

Model	Average accuracy		
	0≤SNR≤20	20<SNR≤30	All SNR
ResNet [3]	76.63	97.63	60.16
MCformer [8]	78.18	97.67	61.46
PET-CGDNN [9]	78.60	97.94	61.30
MCLDNN [7]	79.59	98.22	61.88
LSTM [4]	79.98	98.26	62.55
HVTCN	80.38	98.96	62.98

4.3 Ablation Experiments

Two separate ViT and CapsNet models were built and trained in order to illustrate the contribution of each of them on the hybrid model performance, the results of ablation experiments in Fig. 2-(b) demonstrate that the hybrid model shows a better performance than both separate networks, this superiority is kept along low and high SNRs.

5. Conclusions

This paper presents a novel hybrid approach for radar AMR called HVTCN, combining the Vision Transformer and Capsule Network. First, the time-frequency representations of radar signals are divided into non-overlapping patches before they are being processed by ViT module using self-attention mechanism which enables the model to capture global dependencies and learn more accurate representations of radar signals. CapsNet is then used to recognize the radar signals based on the features extracted by ViT. CapsNet preserves hierarchical relationships between features,

improving the robustness and generalization of the model. The dynamic routing mechanism in CapsNet ensures that the model is resistant to noise and can effectively handle complex, overlapping modulation types. Experimental comparative results demonstrated that the proposed model outperforms traditional single and recently proposed combined networks and shows robustness against noisy conditions. This approach offers an improvement in radar modulation recognition and has potential applications in both military and civilian radar systems.

References

[1] Zhang W, Xue K, Yao A, Sun Y. Automatic Modulation Recognition Based on Multimodal Information Processing: A New Approach and Application. Electronics, 2024, 13(22): 4568.

[2] Javadi S.H, Farina A. Radar Networks: A Review of Features and Challenges. Information Fusion, 2020, 61, 48–62.

[3] O’Shea T. J, Roy T, Clancy T.C. Over-the-air deep learning based radio signal classification. IEEE Journal Selected Topics Signal Processing, 2018, 12(1): 168–179.

[4] Rajendran S, et al. Deep learning models for wireless signal classification with distributed low-cost spectrum sensors. IEEE Transactions on Cognitive Communications and Networking, 2018, 4(3): 433–445.

[5] Papadopoulos, Konstantinos, Mohieddine Jelali. A Comparative Study on Recent Progress of Machine Learning-Based Human Activity Recognition with Radar. Applied Sciences, 2023, 13(23): 12728.

[6] C. Weber T, Felhauer, Peter M. Automatic modulation classification technique for radio monitoring. Electronic Letters, 2015, 51(10: 794-796.

[7] J. Xu, C. Luo, G. Parr, Y. Luo. A spatiotemporal multi-channel learning framework for automatic modulation recognition. IEEE Wireless Communication Letters, 2020, 9(10): 1629–1632.

[8] S. Hamidi-Rad, S. Jain. MCformer: A transformer based deep neural network for automatic modulation classification. IEEE Global Communication Conference (GLOBECOM), 2021, 1–6.

[9] F. Zhang, C. Luo, J. Xu, Y. Luo. An efficient deep learning model for automatic modulation recognition based on parameter estimation and transformation. IEEE Communication Letters, 2021, 25(10): 3287–3290.

[10] Cheng, Lei, Siyang Cao. Retentive Vision Transformer for Enhanced Radar Object Detection. arXiv, 2025, 2501.17977.

[11] Om Uparkar, Jyoti Bharti, R.K. Pateriya, Rajeev Kumar Gupta, Ashutosh Sharma. Vision Transformer Outperforms Deep Convolutional Neural Network-based Model in Classifying X-ray Images. Procedia Computer Science, 2023, 218: 2338-2349.

[12] Maurício, José, Inês Domingues, Jorge Bernardino. Comparing Vision Transformers and Convolutional Neural Networks for Image Classification: A Literature Review. Applied Sciences, 2023, 13(9): 5521.

[13] Haq, Mahmood Ul, Muhammad Athar Javed Sethi, Atiq Ur Rehman. Capsule Network with Its Limitation, Modification, and Applications—A Survey. Machine Learning and Knowledge Extraction, 2023, 5, (3): 891-921.

[14] SHI Biao, LI Yu Xia, YU Xhua, YAN Wang. Short-term load forecasting based on modified particle swarm optimizer and fuzzy neural network model. Systems Engineering-Theory and Practice, 2010, 30(1): 158-160.

[15] Sara Sabour, Nicholas Frosst, and Geoffrey E. Hinton. Dynamic routing between capsules. 31st International Conference on Neural Information Processing Systems (NIPS'17). New York: Red Hook, 3859–3869.