SAR image classification method based on improved capsule network

Zezhou Yang*, Shaodi Jing
College of Information Engineering, Jiangxi University of Science and Technology, Ganzhou, 341000, China
dazhou1228@126.com
6720180562@mail.jxust.edu.cn

Abstract. Aiming at the problem that the convolutional neural network is difficult to recognize the posture transformation of the image, which requires a large number of label samples for training. This paper proposes an improved synthetic aperture radar image classification method based on the combination of dilated convolution and capsule network. We use dilated convolution to expand the receptive field while reducing the amount of calculation, and use the capsule network to improve the classification effect when there are fewer training samples. The experimental results show that the proposed method can still achieve high classification performance after the training sample is small and the test sample undergoes pose transformation, and at the same time, it has a low computational load.

1. Introduction
Synthetic aperture radar (SAR) uses the principle of synthetic aperture to achieve high-resolution microwave image. It has become an important method of earth observation because of its characteristics such as full-time, large amplitude, and high resolution. In recent years, with the rapid development of deep learning technology, convolutional neural networks (CNN) have made great progress in the field of computer vision [1-3]. Compared with traditional classification methods, the convolution filter in CNN can automatically detect important features in images [4], and has good generalization performance. Ding et al. [5] applied deep CNN to SAR image recognition, and achieves good recognition results through data augmentation, but it relies on a large amount of label data for training. Hollósi J et al. [6] pointed out that when the test data is very close to the training data, CNN can show exceptional results. However, if the input test image is transformed, the efficiency of the CNN will be greatly reduced. Therefore, a new method is needed to solve the problem of maintaining high performance when the test data by posture transformation.

A new method to solve the above problem is called Capsule Network (CapsNet) [7], first proposed by Sabour et al. They designed a novel multi-layered learning model, which can solve the problems of CNN's inability to recognize the posture of the image and requires a large number of training samples. Based on the combination of dilated convolution and capsule network, this paper proposes a SAR image classification method based on improved capsule network, which is called ICapsNet. The proposed method deepens the depth of the capsule network and can obtain deeper features of the input image. In the convolutional layer, dilated convolution [8] is used to expand the receptive field to extract a larger range of features without increasing the amount of calculation. Finally, the output of the dilated convolution is used as the input of the capsule layer, the scalar neuron is converted into a
set of vector neurons, and the parameters are updated through the dynamic routing algorithm, so that the network sufficient understands the feature information in the image, so as to obtain higher classification performance.

2. Basic Principles and Proposed Methods

2.1. Dilated convolution

Dilated convolution is to add hole to the convolution kernel of ordinary convolution to achieve the purpose of increasing the receptive field. Different from ordinary convolution, dilated convolution has an extra hyperparameter called dilation rate. The dilation rate refers to the interval between each element in the convolution kernel. The hole in dilated convolution is filled with 0. Assuming that the dilation rate is r and the size of the convolution kernel is k, the size of the dilated convolution kernel is r*(k-1)+1.

Fig.1 shows the difference between ordinary convolution and dilated convolution. Fig. 1(a) shows the process of performing a convolution with ordinary convolution, and its receptive field is 3x3. Fig.1(b) shows the process of performing a convolution with dilated convolution, and its receptive field is 5x5.

![Fig.1 The difference between ordinary convolution and dilated convolution.](image)

2.2. Capsule Network

The capsule network was proposed by Hinton, which is mainly used to solve the shortcomings of convolutional neural networks. A capsule is defined as a vector composed of a group of neurons. The length of the vector gives the estimated probability of an object in the image, and the direction represents the attributes of the object. Vector neurons have more abundant feature expression capabilities than scalar neurons.

![Fig.2 How the capsule works.](image)

The working principle of the capsule is mainly divided into three steps, as shown in Fig 2. First, the vector $u_i$ in the main capsule is multiplied by the matrix $W_y$ to get the shallow capsule, namely

$$\hat{u}_{ji} = W_y u_i$$  \hspace{1cm} (1)

Then add the weights of the shallow capsules $\hat{u}_{ji}$ to get the capsule $s_j$, the weighting coefficient is $c_y$, that is

$$s_j = \sum_i c_y \hat{u}_{ji}$$  \hspace{1cm} (2)
Different weighting coefficients $c_{ij}$ represent different degrees of importance of low-level features to high-level features. The larger the value, the more important the shallow capsule is to the capsule. The weighting coefficient $c_{ij}$ is obtained by $h_{ij}$ taking Softmax function. Initially, $h_{ij}$ is 0, which means that the importance of each low-level feature to the high-level feature is the same. As the training progresses, the parameters are updated. The core of the dynamic routing algorithm lies in the parameter $h_{ij}$ update method, $h_{ij} = b_{ij} + \hat{u}_{ji} v_i$. When updating the parameters, the low-level features and output features are comprehensively considered.

Finally, a nonlinear squash function is used for the capsule $s_j$ to obtain the deep capsule $v_j$, namely

$$v_j = \frac{s_j}{1 + \|s_j\|^2}$$

### 2.3. Improved Capsule Network

This paper proposes an improved capsule network model. As shown in Fig. 3. The network model is mainly divided into dilated convolutional layer and capsule layer. The dilated convolution layer extracts the information features in the image through 5 layers of dilated convolution, and then inputs the extracted features to the capsule network layer. The capsule network layer includes a primary capsule and a digital capsule. These two capsule layers convert the scalar neurons in the convolutional layer into vector output, so that the extracted features can be sufficient expressed.

In the improved capsule network, the size of the convolution kernel is 5×5, and the ReLU activation function is used after each dilated convolution layer. The propagation process of the improved capsule network and the parameters of each layer are as follows:

1. Perform dilated convolution of a 64×64 SAR image with 16 convolution kernels, where the stride $s=1$ and the dilation rate $r=2$, then the output of the first dilated convolution layer 16 feature maps of 64×64;
2. Using 32 strides $s=1$, dilated rate $r=2$ convolution kernel to convolution the output of the previous layer, and then through the max pooling operation with a strides $s=2$, the output obtained is 32 feature map of 32×32;
3. Using 64 strides $s=1$, dilated rate $r=2$ convolution kernel to convolution the output of the previous layer, the output obtained is 64 feature map of 32×32;
4. Using 128 strides $s=1$, dilated rate $r=2$ convolution kernel to convolution the output of the previous layer, and then through the max pooling operation with a strides $s=2$, the output obtained is 128 feature map of $16 \times 16$;

5. Using 256 strides $s=1$, dilated rate $r=2$ convolution kernel to convolution the output of the previous layer, and then through the max pooling operation with a strides $s=2$, the output obtained is 256 feature map of $8 \times 8$;

6. Next is the main capsule layer. After $32 \times 8$ stride $s=2$ convolution kernel convolution, the output is $32 \times 8$ feature maps of $2 \times 2$, which are expanded into $32 \times 2 \times 2$ elements, namely 128 capsules, each capsule is a vector with dimension 8;

7. The last layer is the digital capsule layer, there are 10 capsules in total, and each capsule is a vector with dimension 16. The output of the primary capsule layer is converted into a digital capsule layer through a conversion matrix and a dynamic routing algorithm;

8. Finally, use the margin loss function to continuously train the weights of each layer in the improved capsule network, so that the network achieves better performance.

3. Test Results and Discussions

For the method based on the improved capsule network, the MSTAR data set is used for method verification. The MSTAR dataset contains 10 types of vehicle targets: 2S1, BMP2, BRDM2, BTR60, BTR70, D7, T62, T72, ZIL131, ZSU234. Fig 4 shows the SAR images of the MSTAR dataset and their corresponding optical image. The sizes of various images of the original MSTAR data are different. In order to facilitate processing, this paper will uniformly crop them to 64x64 pixels.

![Fig. 4 Optical images and SAR images of 10 types of targets in the MSTAR dataset.](image)

In order to verify the effectiveness of the method based on the improved capsule network, the method was compared with the method based on the fully convolutional neural network (FCNN) and the method based on the capsule network (CapsNet). Among them, the method based on CapsNet adopts two convolution kernel sizes for experiments. The reason is that a large convolution kernel has a larger receptive field, and the extracted feature information is richer, which helps to improve the classification performance of the network. However, the disadvantage of this method is that it increases training parameters and reduces the efficiency of the model. The training parameters corresponding to each method are shown in Table 1. Obviously, the training parameters of the proposed method are the same as those of the 5x5 convolution kernel CapsNet, with 1,991,136 parameters, and the training parameters are tied at the least. The training parameters of FCNN are about 2.6 times that of ICapsNet, and the training parameters of the 7x7 convolution kernel CapsNet are about 1.5 times of ICapsNet.

| Table 1 Training parameters of the four methods |
|-----------------------------------------------|
| 5x5 CapsNet | 7x7 CapsNet |
| 1,991,136    | 2,986,704    |
3.1. Different proportions of training samples
Different proportions of training samples (20%, 40%, 60%, 80% and 100% of training samples are taken respectively) and all test samples. The horizontal axis of Fig.5 represents the proportion of training samples, and the ordinate shows the change in the average correct recognition rate of the four recognition methods as the training samples decrease. It can be analyzed from the Fig.5 that the accuracy of the four methods will decrease with the reduction of training samples. However, the accuracy of ICapsNet decreases slowly, and the method based on FCNN decreases very quickly. When the training sample is reduced to 20%, the accuracy of ICapsNet is as high as 91.71%, while the accuracy of the other three methods from high to low are 82.06%, 78.35%, and 72.82%. This shows that ICapsNet can still have a higher recognition rate under the condition of less training data, which reflects the robustness of the model.

| Model          | FCNN  | CapsNet(5x5) | CapsNet(7x7) | ICapsNet |
|---------------|------|-------------|-------------|----------|
| Trainable parameters | 5,207,316 | 1,991,136  | 3,036,000  | 1,991,136 |

Fig.5 The change of the average correct recognition rate as the number of training samples decreases

3.2. Posture transformation on the test sample
The experiment in this section is not to do any processing on the training samples and to transform the test samples. The posture transformation mainly includes translation, rotation and scaling. Using S, T, R to define scaling, translation, and rotation, where n represents the original image is scaled to n times the original size, u and v represent the length of the left, right, and up and down translations, and θ represent the angle of image rotation. The final posture transformation can be obtained by multiplying the matrix of equation (6):

\[
\begin{bmatrix}
    x' \\
y' \\
1
\end{bmatrix} =
\begin{bmatrix}
    S \\
    T \\
    R
\end{bmatrix}
\begin{bmatrix}
x \\
y \\
1
\end{bmatrix}
\]

\[
\begin{bmatrix}
    n & 0 & 0 \\
0 & n & 0 \\
0 & 0 & 1
\end{bmatrix}
\begin{bmatrix}
    1 & 0 & u \\
0 & 1 & v \\
0 & 0 & 1
\end{bmatrix}
\begin{bmatrix}
    \cos \theta & -\sin \theta & 0 \\
\sin \theta & \cos \theta & 0 \\
0 & 0 & 1
\end{bmatrix}
\begin{bmatrix}
x \\
y \\
1
\end{bmatrix}
\]

In the experiment, u, v takes the value 10, n and the θ variable take different values. Table 2 shows the influence of the four groups of different posture transformations on the test sample on the recognition rate of the four methods. It can be seen that the scaling level n and the rotation angle θ have impact on the recognition rate, and are inversely proportional to the recognition rate. Meanwhile,
ICapsNet achieves the best recognition effect under the four pose changes, which verifies the robustness of ICapsNet.

| Transform method | Recognition rate (%) | FCNN | CapsNet (5x5) | CapsNet (7x7) | ICapsNet |
|------------------|----------------------|------|---------------|---------------|----------|
| $n=0.95, \theta=\pi/30$ | 91.84 | 96.00 | 97.15 | 98.27 |
| $n=0.9, \theta=\pi/30$ | 82.56 | 90.23 | 93.86 | 94.68 |
| $n=0.95, \theta=\pi/60$ | 92.70 | 96.33 | 96.99 | 98.06 |
| $n=0.9, \theta=\pi/60$ | 89.94 | 95.67 | 96.78 | 97.48 |

4. Conclusion
Combining the advantages of dilated convolution and capsule network, this paper proposes an improved SAR image classification method for capsule network. Using dilated convolution for image feature extraction, a larger receptive field can be obtained and thus richer feature information can be obtained without increasing the training parameters of the network. Using the capsule network can convert scalar neurons into vector neurons, thereby obtaining the relationship between the whole and part of the image, which is beneficial to the improvement of classification performance. Experimental results show that the method proposed in this paper can achieve better classification results under different data conditions. At the same time, the parameters of the network model are the least, which verifies the accuracy and robustness of the proposed method.

References
[1] Chen S Z, Wang H P, Xu F, et al. Target classification using the deep convolutional networks for SAR images[J]. IEEE Transactions on Geoscience and Remote Sensing, 2016, 54(8): 4806–4817.
[2] Zhang Z M, Wang H P, Xu F, et al. Complex-valued convolutional neural network and its application in polarimetric SAR image classification[J]. IEEE Transactions on Geoscience and Remote Sensing, 2017, 55(12): 7177–7188.
[3] Furukawa H. Deep learning for target classification from SAR imagery: Data augmentation and translation invariance[J]. IEICE Technical Report, 2017, 117(182): 13–17.
[4] Zhao Juanping, Guo Weiwei, Zhang Zenghui, et al. A coupled convolutional neural network for small and densely clustered ship detection in SAR images[J]. Science China Information Sciences, 2019, 62(4): 42301:1-16.
[5] Ding J , Chen B , Liu H , et al. Convolutional Neural Network With Data Augmentation for SAR Target Recognition[J]. IEEE Geoscience & Remote Sensing Letters, 2016:364-368.
[6] Hollósi J, Pozna C R. Improve the Accuracy of Neural Networks using Capsule Layers[C]2018 IEEE 18th International Symposium on Computational Intelligence and Informatics (CINTI). IEEE, 2018: 015-018.
[7] Sabour S, Frosst N, Hinton G E. Dynamic routing between capsules[C].Advances in neural information processing systems. 2017: 3856-3866.
[8] Yu F, Koltun V. Multi-Scale Context Aggregation by Dilated Convolutions[J]. 2015. arXiv preprint arXiv:1511.07122.