Overview of Capsule Neural Networks

Zengguo Sun[1,2], Guodong Zhao[2], Rafał Scherer[3], Wei Wei[4*], Marcin Woźniak[5]
[1] Key Laboratory of Modern Teaching Technology, Ministry of Education, Xi'an Shaanxi, China
[2] School of Computer Science, Shaanxi Normal University, China
[3] Dep. of Intelligent Computer Systems, Czestochowa University of Technology, Poland
[4] School of Computer Science and Engineering, Xi'an University of Technology, China
[5] Faculty of Applied Mathematics, Silesian University of Technology, Poland
sunzg@snnu.edu.cn, donl@snnu.edu.cn, rafal.scherer@pcz.pl, weiwei@xaut.edu.cn, marcin.wozniak@polsl.pl

Abstract

As a vector transmission network structure, the capsule neural network has been one of the research hotspots in deep learning since it was proposed in 2017. In this paper, the latest research progress of capsule networks is analyzed and summarized. Firstly, we summarize the shortcomings of convolutional neural networks and introduce the basic concept of capsule network. Secondly, we analyze and summarize the improvements in the dynamic routing mechanism and network structure of the capsule network in recent years and the combination of the capsule network with other network structures. Finally, we compile the applications of capsule network in many fields, including computer vision, natural language, and speech processing. Our purpose in writing this article is to provide methods and means that can be used for reference in the research and practical applications of capsule networks.

Keywords: Capsule network, Dynamic routing Mechanism, Convolutional neural network, Deep learning

1 Introduction

In recent years, the development of deep learning has received extensive attention. The essence of deep learning is to simulate the human brain to analyze data and obtain the optimal decision-making mechanism through a large amount of data training to complete specific tasks. One of the most representative network models in deep learning is the Convolutional Neural Network (CNN), which makes full use of the local features contained in the data itself by combining the three characteristics of sensing local area, sharing weights, and pooling [1]. The pool structure greatly improves the efficiency of network operation, but when down-sampling the feature map, a lot of effective information will be lost.

In 2017, Sabour et al. first proposed the concept of Capsule Network (CapsuleNetwork, CapsNet) [2]. The original intention of CapsNet is to overcome the shortcomings of CNN's low data transmission efficiency due to scalar transmission and to overcome the catastrophic consequences of CNN pooling operations. Pooling operation abandons a large amount of spatial location information, which makes the network unable to accurately identify the subtle changes in the spatial position of entities, and cannot accurately learn the position association between entities [3-5]. Capsnet encapsulates multiple neurons representing the same entity to form a matrix and makes full use of the size and direction of multi-dimensional vectors to transmit information so that the number of parameters of CapsNet is significantly reduced compared with CNN. At the same time, CapsNet can fully learn the internal spatial relationship and overall structural relationship of the data and capture the hidden features in the data so that CapsNet can maintain a high accuracy rate even with less training data [6-9]. Finally, CapsNet uses a transformation matrix to realize viewpoint invariance, which expands the network's ability to understand data from two-dimensional space to high-dimensional space [10].

Based on the unique advantages of CapsNet, researchers have conducted in-depth research on the improvement of dynamic routing mechanism, the improvement of network structure, and the combination of capsule network with other network structures. They hope that CapsNet can also be applied to more applications on the basis of improving CapsNet performance. In this paper, we focus on the theory, improvement, and application of CapsNet. Through these introductions, we hope to help readers understand related work methods and ideas and inspire new research ideas.

2 Analysis of the Advantages and Disadvantages of Traditional CNN

As one of the classic network models of deep learning, convolutional neural networks (CNN) are widely used in various fields of deep learning [11]. However, the traditional CNNs also exposed many problems. For example, the traditional CNN using pooling operation will discard the information about the precise position of the entity in the region, and its scalar output will lead to low expression ability of the observed data [12-18]. In addition, CNN needs many data to train the network model, and the traditional CNN architecture ignores the spatial hierarchy between objects [19-27].

The pooling operation in CNN aims to obtain spatial invariant features by reducing the resolution of the feature surface, reducing the number of neurons by reducing the number of connections between convolutional layers, and reducing the amount of calculation of the network model [8, 24]. However, the pooling operation discards a large amount of useful information when performing down-sampling transmission feature information. Take max-pooling, which is commonly used in pooling operations, as an example. Suppose that a filter with a size of 2×2 is used, and the input

* Corresponding Author: Wei Wei; E-mail: weiwei@xaut.edu.cn
DOI: 10.53106/160792642022012301004

data is down-sampled with a step size of 2, and a maximum value is taken from the 2×2 numbers to be transferred to the next layer of the network, while the remaining feature information is discarded. The abandoned data information contains the exact position information of the entity in the region, as well as the spatial relationship information such as the perspective, size, and direction between features [4, 7]. Therefore, the pooling operation has serious defects.

Figure 1 is a schematic diagram of CNN neurons. The neuron receives the input scalar from other neurons, multiplies the scalar weight by the input scalar, then sums the results, and then transfers the sum to a non-linear activation function (such as sigmoid, tanh, ReLU) to generate an output scalar, which is used as the input variable of the next layer. The structure of a scalar neuron is simple, but the amount of information carried by a single neuron is small, and the ability to express and describe data features is low, so more neurons are needed to describe the features of the same entity. Therefore, CNN needs more parameters to describe data features [3, 18, 21].

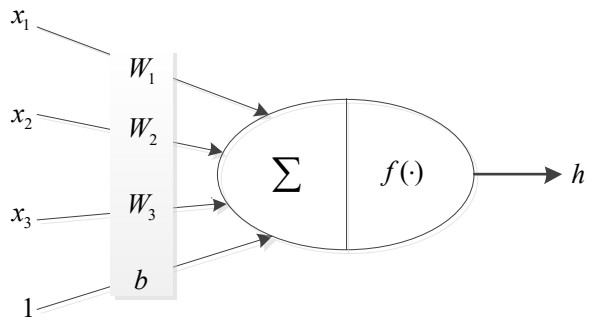

Figure 1. Scalar neuron

In addition to the disadvantages brought by feature operation, traditional CNN also has its structural disadvantages. The traditional CNN architecture ignores the hierarchical structure between layers (similar to what exists in the human brain), which limits their modeling ability. CNN tries to cover up such limitations by using a large number of training data, which makes the traditional CNN need more prior data for training. Moreover, traditional CNN also has the following disadvantages. First, it cannot separate overlapping adjacent objects. Secondly, each layer of the architecture contains some disordered neurons, which makes it difficult for some internal structures to execute. Finally, CNN cannot infer the geometric information of the data, so it is not robust to new perspectives.

Of course, CNN reduces the number of weights to be trained and reduces the computational complexity of the network through weight sharing [2]. This is the advantage of CNN structure, which can be used by other network structures based on CNN.

In summary, the feature operation of the CNN structure and its advantages and disadvantages are shown in Table 1.

In order to solve the defects of network model caused by traditional CNN pooling operation and scalar transmission, and make full use of the advantages of CNN architecture, scholars introduced the concept of capsule network.

Table 1. CNN feature operation and its advantages and disadvantages

Feature operations	Advantages	Disadvantages
Pooling operation	Reducing the dimension of the hidden layer in the middle and the amount of computation in subsequent layers.	Discarding information such as the precise position of the entity and the spatial relationship of the features.
Scalar transmission	Simple structure and easy to understand.	Carrying less information for single neuron. Unable to communicate the relationship between neurons. Low data feature expression ability.
Specific network structure	Simple structure.	Difficult to organize the internal structure of network layer; Ignoring the structural relationship between layers; Needing more prior data; Unable to separate overlapping adjacent objects; Not robust to new perspectives.
Weight sharing	Reducing the computational complexity of the network and the number of weights to be trained.	

3. Basic Concepts of Capsule Network

The original CapsNet is developed on the basis of CNN, so it makes full use of the advantages of CNN and overcomes many structural defects of CNN at the same time. Firstly, CapsNet only takes convolution operation as the first layer of the network, making full use of the advantages of CNN in feature extraction and generating a series of feature channels. Secondly, CapsNet constructs multi-dimensional vectors by grouping multiple characteristic channels of the convolution filter. Thirdly, CapsNet uses an affine transformation to connect the primary capsule and secondary capsules. These transformations can learn the relationship between the part and the whole in the data instead of detecting the independent features by filtering in different regions of the image. Finally, CapsNet transform weights are not optimized by regular back-propagation but by dynamic routing algorithm [28]. This section analyzes and summarizes the reasons for the unique advantages of CapsNet through the introduction of CapsNet's capsule model, dynamic routing, and network structure.

3.1 The Capsule

A capsule is a group of neurons nested in a layer, whose output represents different attributes of the same entity, such as direction, size, and posture. The purpose of introducing the capsule is to encapsulate a large amount of posture information (such as position, direction, zoom, and tilt) and other instantiation parameters (such as color and texture) of different parts or fragments of the object [20]. Each layer of CapsNet is composed of multiple capsules, which are a collection of vector neurons to represent a certain type of

feature of the entity. The output information includes the probability of the type of the object and the status information of the object (such as position, direction, size, deformation, speed, color, etc.). The parameters output from the low-level capsule will be converted into the prediction of the entity state by the high-level capsule. If the prediction is consistent, the parameter will be received [29].

It can be seen from Figure 1 that, in scalar neurons, the artificial neuron first weights the input scalars, and then sums the weighted input scalars. This makes it impossible for us to know the spatial hierarchical relationship of each neuron. Different from scalar neurons, vector neurons encapsulate the information that needs to be carried, and uses weight matrix to store the spatial hierarchical relationships and other relationships of neurons. The vector neuron model is shown in Figure 2. Firstly, the size and direction of input vector are encapsulated into a prediction vector, which is $u\{1, 2, 3\}$ in Figure 2. We define the probability of the existence of the objects detected by the child nodes as the length of these vectors, and define some internal states of the detected objects as the direction of the vectors. At the same time, the spatial hierarchical relationship and other relationships between the low-level features detected by the low-level capsule and the high-level feature of the high-level capsule are encoded. The weight matrix $\boldsymbol{W}\{1_j, 2_j, 3_j\}$ obtained by coding is used to process the prediction vector into a new input vector $\hat{u}_{\{j|1,\, j|2,\, j|3\ \ldots\}}$. Secondly, the input vector $\hat{u}_{\{j|1,\, j|2,\, j|3\ \ldots\}}$ is multiplied by coupling coefficient. Vector dimensionalities depend on the input data size. The calculation method of coupling coefficient is shown in the following section. Thirdly, the vector **s** is obtained by summing the input vectors after setting the coupling coefficient. Finally, the vector compression function is used to convert the vector **s** into a vector $\mathbf{v}_j$ as the output of the capsule of this layer.

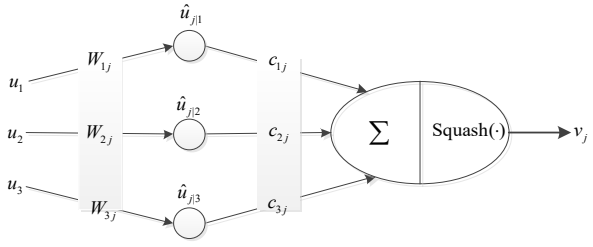

Figure 2. Vector neuron model

The capsule model adopts a parse tree structure, and each active capsule corresponds to each node on the parse tree one-to-one. The low-level capsules (child nodes) are used to detect the low-level feature information, and high-level capsules (parent nodes) are used to receive output information from child nodes.

3.2 Dynamic Routing

The basic CapsNet is trained by the dynamic routing algorithm proposed in [2]. Dynamic routing is performed between two consecutive capsule layers to update coupling coefficients. These coupling coefficients determine how the low-level capsules (assuming L1 layer) send their inputs to higher-level capsules (assumed L2 layer) that agree with the input [30].

The sum of the coupling coefficients between the capsules i and all the capsules in the l layer is 1, which is determined by "routing softmax", b_{ij} is defined as the logarithm of the prior probability of coupling capsule i and capsule j, and its initial value is set to 0. Then the coupling coefficient c_{ij} determined by the iterative dynamic routing process is determined by equation (1).

$$c_{ij} = \frac{exp\ (b_{ij})}{\sum_k exp(b_{ik})} \tag{1}$$

For all capsules except the first layer capsule, the total input s_j of the capsule is calculated by weighted sum of all "prediction vector" $\hat{\mathbf{u}}_{j|i}$ in l layer capsule. The coupling coefficient c_{ij} is the weight. The "prediction vector" $\hat{\mathbf{u}}_{j|i}$ is obtained by multiplying the original input $\mathbf{u}_i$ of the capsule in the l +1 layer by the weight matrix $\mathbf{W}_{ij}$.

$$s_j = \sum_i c_{ij}\ \hat{\mathbf{u}}_{j|i}\ ,\ \hat{\mathbf{u}}_{j|i} = \mathbf{W}_{ij}\,\mathbf{u}_i \tag{2}$$

Capsnet uses the length of the output vector of the capsule to represent the probability of the existence of the entity represented by the capsule. Therefore, the non-linear "squashing" function (3) is used to ensure that the short vector is reduced to the vector length close to 0, and the long vector is reduced to the vector length slightly less than 1. Where $\mathbf{v}_j$ is the vector output of capsule j.

$$\mathbf{v}_j = \frac{\|s_j\|^2}{1+\|s_j\|^2}\ \frac{s_j}{\|s_j\|} \tag{3}$$

The logarithm b_{ij} of the prior probability can be distinguished and learned as all other weights at the same time. They depend on the location and type of the two capsules and not on the current input image. Then, the logarithm B2 of the initial prior probability is iteratively optimized by measuring the consistency between the output $\mathbf{v}_j$ of each capsule j in l+1 layer and the prediction $\hat{\mathbf{u}}_{j|i}$ made by capsule i.

$$b_{ij} = b_{ij} + \hat{\mathbf{u}}_{j|i}\,.\,\mathbf{v}_j \tag{4}$$

Equation (4) updates b_{ij} with the dot product of $\hat{\mathbf{u}}_{j|i}$ and $\mathbf{v}_j$, which is the essence of dynamic routing. Among them, $\hat{\mathbf{u}}_{j|i}$ is the "personal" prediction of the l layer output $\mathbf{u}_i$ for the capsule in the l+1 layer, while $\mathbf{v}_j$ is the "consensus" prediction of all output **u** of the l layer for the capsule of l+1 layer.

Unlike CNN, CapsNet uses dynamic routing instead of pooling, so it makes full use of the information related to the current task. The capsule model can be considered as a parse tree, because each active capsule selects the next layer of the capsule as the parent capsule in the parse tree [2]. Therefore, compared with the max/average pooling method used in CNN, CapsNet can better handle different visual stimuli and provide better viewpoint invariance [29].

3.3 Basic Architecture of Capsnet

Taking the capsule network [2] first proposed by Hinton team as an example, the basic architecture of CapsNet is

shown in Figure 3. This model is very shallow, with only two convolution layers and one fully connected layer. The Conv1 layer converts the pixel intensity of input data into the activities of local feature detector by convolution, and uses ReLU as the activation function. The PrimaryCapsules layer is the basic layer of multi-dimensional entities. It splices the instantiated parameters together to form the analytic tree of the overall relationship. The DigitCaps layer uses dynamic routing to receive the output of each capsule in the PrimaryCapsules layer, and designs a capsule for each entity.

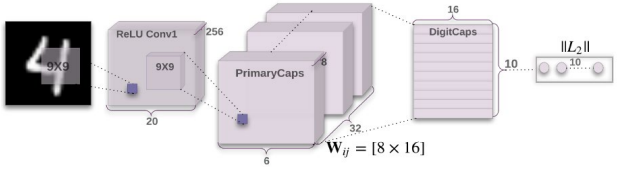

Figure 3. Simple CapsNet with 3-layer structure

With its unique neuron organization, routing and network architecture, CapsNet has its unique advantages. Because CapsNet uses a capsule formed by a collection of vector neurons as the basic unit of the network, a single capsule carries more and richer feature information, so the number of parameters of the network is less than that of CNN. These neurons that make up the capsule can organize some internal structures, which makes CapsNet more suitable for high-dimensional feature extraction. Moreover, the transformation matrix in the capsule model can learn to represent the relationship between the part and the whole in the data, and provide viewpoint invariance for the network. The use of dynamic routing mechanism makes full use of all the feature information carried by low-level capsules, so that CapsNet can maintain a high accuracy rate even with less training data, and enables CapsNet to capture the hidden features in the data, and the captured feature information is more abundant [8, 23-24, 29]. Moreover, the dynamic routing mechanism uses the relationship between "part" and "whole" to update the coupling coefficient, so that CapsNet can fully learn the internal spatial relationship and overall structure relationship of the data. This behavior enable CapsNet to learn how to infer attitude parameters from images [7, 28]. Finally, the special network architecture of CapsNet makes CapsNet more suitable for migration and new tasks than traditional CNN [31-32].

In summary, the characteristic operation and advantages of CapsNet are shown in Table 2.

Table 2. Characteristic operations and advantages of CapsNet

Characteristic operations	Advantages
Capsule model	Improving the transmission efficiency of neurons; Suitable for extracting high-dimensional features; Reducing the number of network parameters; Making the network view-invariant.
Dynamic routing mechanism	Less training data, high accuracy; Ability to capture hidden features, learn internal spatial relations and overall structural relations and infer attitude parameters.
CapsNet network architecture	Better suitable for migration and new tasks.

4. Improvement of CapsNet

CapsNet is increasingly showing its unique advantages in the field of deep learning. For example, CapsNet can maintain a high accuracy rate even with less training data, can capture hidden features in the data, and capture more feature information. However, the basic CapsNet has some shortcomings [10, 33]. For example, the training time of CapsNet using dynamic routing algorithm is longer; CapsNet is relatively weak in extracting local features. Therefore, scholars have improved CapsNet from the aspects of dynamic routing mechanism and network model.

4.1 Improvements of Dynamic Routing Mechanism

CapsNet is different from the single scalar output in traditional CNN. Traditional CNN uses a pooling operation to perform down-sampling, so that neurons remain unchanged for viewpoint changes, while capsules hope to retain information to achieve viewpoint translation equivalence, similar to the perception systems. Therefore, replace the pooling operation with a dynamic routing operation, and send the output of lower-level capsules (such as capsules representing entities such as nose, mouth, ears, etc.) as input to the parent capsule (such as the capsule representing the face) representing the part-whole relationship to achieve translation equivalence [12]. Dynamic routing has been proved to be an effective method with high generalization ability and less parameters. However, CapsNet relies on intensive clustering calculations in the inference process, so the training time of the network is relatively long [34,35]. In order to make up for the shortcomings of the dynamic routing mechanism, scholars mainly put forward their own suggestions on improving the training speed of dynamic routing mechanism, improving the routing accuracy of dynamic routing mechanism and reducing the number of parameters of dynamic routing mechanism.

4.1.1 Improvements in improving the training speed of the dynamic routing mechanism

In 2018, Aryan Mobiny et al. [10] proposed a consistent dynamic routing mechanism to accelerate CapsNet. Specifically, all capsules corresponding to the same pixel in the PrimaryCaps layer are forced to have the same routing coefficient. This strategy will greatly reduce the number of routing coefficients. For example, the original CapsNet has 32 capsules in each pixel position of the PrimaryCaps layer. The author only places one capsule at each pixel position, reducing the number of routing coefficients to 1/32 of the original. In order to compensate for the reduction in the number of capsules, the author increased the size of each capsule to 256 dimensions instead of 8 dimensions of the original capsule. Experimental results show that the proposed structure has similar classification accuracy and is 3 times faster than the CapsNet algorithm.

In 2019, Suofei Zhang et al. [35] extended existing routing methods within the framework of weighted kernel density estimation, and proposed a fast routing method with different optimization strategies: a fast routing algorithm based on Mean-shift. Mean-shift is a typical clustering

method based on weighted kernel density estimation framework for feature analysis and related visual tasks. The author uses a mean-shift-based routing algorithm to update the weights in a gradient descent manner, and discards the combination of "individual prediction" and "overall prediction" by Sabour et al., in order to improve the training speed at the expense of the network's ability to predict attitude parameters. The method proposed by the author increases the time efficiency of routing by nearly 40%.

4.1.2 Improvements to improve the routing accuracy of the dynamic routing mechanism

In 2018, Hinton et al. [29] introduced a new iterative routing process between capsule layers based on the Expectation Maximization (EM) algorithm. The author uses EM algorithm to update the weight of each image iteratively, so that the output of each capsule is routed to the next layer of capsules. The author uses EM algorithm to iterate between each pair of adjacent capsule layers to perform back propagation, so that the transformation matrix can be trained discriminatively, and the characteristic information routed between the capsule layers is more realistic. In order to improve the routing accuracy, the author improves the basic CapsNet proposed by Sabour et al. in two aspects. Firstly, the non-linearity achieved in the entire capsule layer is performed by the EM algorithm; secondly, backpropagation is performed between adjacent layers so that the transformation matrix can be trained discriminatively by backpropagating through the unrolled iterations of EM between each pair of adjacent capsule layers. Compared with the performance of CNN on the smallNORB dataset (5.2% test error rate achieved with 4.2M parameters), the trainable parameters of the method proposed by the author are only 68K, but the test error rate is 2.2%.

In 2018, Wei Zhao et al. [34] proposed a text classification strategy that stabilizes the dynamic routing process, eliminates noise interference, and improves routing accuracy. The author tries to iteratively correct the coupling coefficient by using the probability of the previous layer of capsules in the next layer of capsules. Compared with the update method of the coupling coefficient proposed by Sabour et al., the author introduces the probability of the existence of the prediction vector and the logarithm of the prior probability to jointly determine the update of the coupling coefficient. The author shows that the network model has achieved better results than CNN when tested on the 4 data sets of MR, Subj, CR and AG's.

4.1.3 Improvements in reducing the number of parameters of the dynamic routing mechanism

In 2018, Rodney Lalonde et al. rewritten the dynamic routing algorithm in two ways, thus solving the problem of the large number of parameters in CapsNet [18]. First of all, the author chose to let the network route only the sub-capsules in the user-defined kernel to the parent node instead of routing every sub-capsule to the parent node. Secondly, the transformation matrices are shared for each member of the grid within a capsule type but are not shared across capsule types. Because the same transformation matrix is shared at all spatial positions of a given capsule type, the total number of parameters to be learned can be significantly reduced.

In 2018, Hao Ren et al. [36] proposed a routing algorithm based on K-means clustering theory. The author regards the K-means routing between the L1 layer capsule and the L2 layer capsule as a K-means clustering process. The L2 layer capsule is the cluster center of the L1 layer capsule. Given multiple capsules, K-means clustering is to find centers of clusters related to them, and minimize the loss function. The main difference between K-means routing and Sabour et al.'s dynamic routing is that the logarithm of the prior probability is replaced by the new logarithm, instead of being replaced by the new logarithm of the prior probability plus the old logarithm. After each iteration, the new parameter information completely replaces the log of prior probability and other historical information. After the iterative update of the dynamic routing of Sabour et al., some historical parameters will be reserved for the next iteration update [2]. Therefore, it can be considered that the improvement reduces the number of network parameters at the expense of partial accuracy.

4.1.4 Other improvements to the dynamic routing mechanism

In 2018, Zhenhua Chen et al. [37] proposed to embed the parameters in the dynamic routing mechanism with all other parameters in the neural network into the optimization process, making the coupling coefficients in the routing process trainable completely. Whether in CapsNet proposed by Sabour et al. [2] or another CapsNet proposed by Hinton et al. [29], the number of routing iterations must be set manually by testing. Therefore, the author proposes to incorporate the routing process of capsule into the whole optimization process, thereby eliminating the setting of routing times and ensuring convergence.

In 2019, Mohammed Amer et al. [38] proposed a new capsule network model PathCapsNet, which uses fan-in protocol routing for information transfer. For PathCapsNet, the author uses another form of dynamic routing protocol. The protocol calculates the weight of the contribution of all capsules from PrimaryCapsules to a specific DigitCaps capsule, so that the sum of the weights is a normalized probability of 1.0. In the dynamic routing of the traditional capsule network, the sum of the weights of a specific PrimaryCapsule capsule to each DigitCapsule is a normalized probability of 1.0. The author's test results on the MNIST dataset show that CapsNet can increase the test error from 0.48% to 0.42% by using the routing protocol proposed by the author. Moreover, through the reasonable coordination of network depth, Max pooling, DropCircuit regularization and new protocol routing technology, the network can obtain better results than the traditional CapsNet [2], and further significantly reduce the number of parameters.

4.2. Improvements of Network Structure

Compared with deep CNNs, CapsNet has achieved good results in terms of shallow structure and considerable parameter savings. However, the lack of depth may limit the expressiveness of the network. Moreover, the number of parameters of the first convolutional layer in CapsNet is large, which will increase the number of overall parameters of the CapsNet network significantly [2, 38]. In addition, the original CapsNet model architecture was built specifically for

MNIST (a relatively low-dimensional data set). Its feature learning ability for complex high-dimensional data sets needs to be improved [21]. By improving the network depth of the original CapsNet and expanding the multi-path network, the performance of the network is improved and the number of parameters is reduced.

In 2018, Sameera Ramasinghe et al. [11] proposed several improvements to the CapsNet structure. The author proposes a related module to learn dataset-wise priority scheme, instead of capturing the priority of each data point separately. The module can prioritize the prediction according to the initial capsule and effectively predict the decision capsule. The author feeds back the main capsule prediction to a trainable end-to-end conditional random field module to learn the interdependence between attributes. The difference from the original CapsNet network structure is that the network structure proposed by the author adds a conditional random field module between the PrimaryCapsule layer and the DigitCapsule layer to provide the prediction of the capsule routing position, and then adds related modules to sort the priority of the capsules. The author shows that compared with the original structure, the improved model structure has increased the accuracy of multi-label classification by more than 33%.

In 2018, Congying Xia et al. [39] applied the capsule network to text modeling, extracted and aggregated semantics from the utterance in a layered manner, and proposed two capsule-based network structures INTENTCAPSNET and INTENTCAPSNET-ZSL. The former extracts semantic features from semantics and aggregates them to distinguish semantic intentions. The latter discerns new intentions from existing intentions through knowledge transfer. The INTENTCAPSNET structure can be understood as the structure from the input data to the DigitCapsule layer in the original CapsNet. The difference from the original CapsNet network structure is that the network structure proposed by the author adds a Zero-shot DetectionCaps layer to distinguish emerging intentions after the DetectionCaps layer of aggregated semantic features, which can be understood as adding a new capsule layer after the DigitCapsule layer of the original CapsNet network structure. In the inference process, the Zero-shot DetectionCaps capsule layer adjusts the existing intent and emerging intent in the DetectionCaps layer to obtain the activation vector for the emerging intent for zero-sample intent detection.

In 2018, Canqun Xiang et al. [12] proposed a two-stage multi-scale structure that replaced the Conv1 layer in the original CapsNet network structure. The multi-scale structure proposed by the author is shown in Figure 4. The unit consists of two stages. In the first stage, the high-level features, middle-level features and original features of the data are extracted through three-level branches. In the second stage, the hierarchical structure of features is encoded by the multi-dimensional primary capsule. Encode the high, medium, and low-level features obtained in the first stage respectively to obtain 12-dimensional, 8-dimensional, and 4-dimensional capsules. Through the use of three branches, a multi-dimensional PrimaryCapsule is obtained. Compared with the original CapsNet network structure, the network structure proposed by the author is more detailed in extracting the initial features of the data, and can extract richer original feature information. The author's experimental results show that the accuracy of the proposed network structure is improved in FashionMNIST and CIFAR10 dataset classification tasks, and the number of parameters is only about 40% of the original CapsNet network structure.

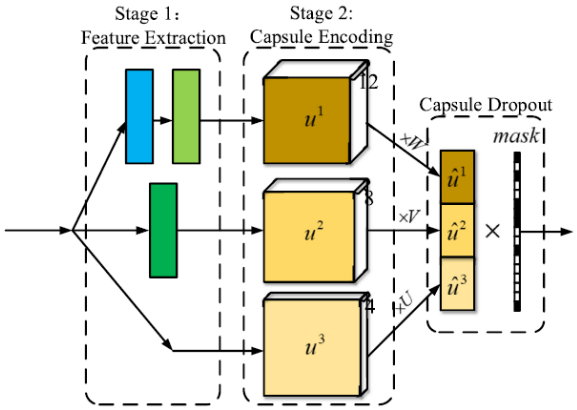

Figure 4. Multi-scale structural model

In 2019, Rosario et al. [25] proposed a multi-channel capsule network (MLCN). It is a separable and resource-efficient organization of the capsule network, allowing parallel processing while achieving high accuracy at a lower cost. In MLCN, the original CapsNet is split, and the primary capsules are divided into independent sets, called lanes. Each channel is responsible for learning different dimensions and different characteristics of the vectors, and uses CapsNet's protocol routing organization for training. Compared with the original CapsNet structure, MLCN requires a smaller or similar number of parameters to achieve the same accuracy, and can train MLCN faster than CapsNet [25].

In 2019, Mohammed Amer et al. [38] proposed PathCapsNet (PathCapsNet), which is a deeply parallel multi-path version of CapsNet. PathCapsNet shares the upper part of CapsNet, starting from the original capsule layer, passing through the DigitCaps layer, and ending with the reconstruction layer. However, PathCapsNet is fundamentally different from the original CapsNet in the way the capsule is constructed. In PathCapsNet, each primary capsule is composed of a deep CNN called a path. Therefore, the input features are input to different paths, and the output of each path includes a primary capsule. The author's results show that through the reasonable coordination of network depth, maximum pool, DropCircuit regularization, and new fan-in protocol routing technology, better results can be obtained than CapsNet, while further reducing the number of parameters [38]. The network structure of the proposed PathCapsNet is shown in Figure 5.

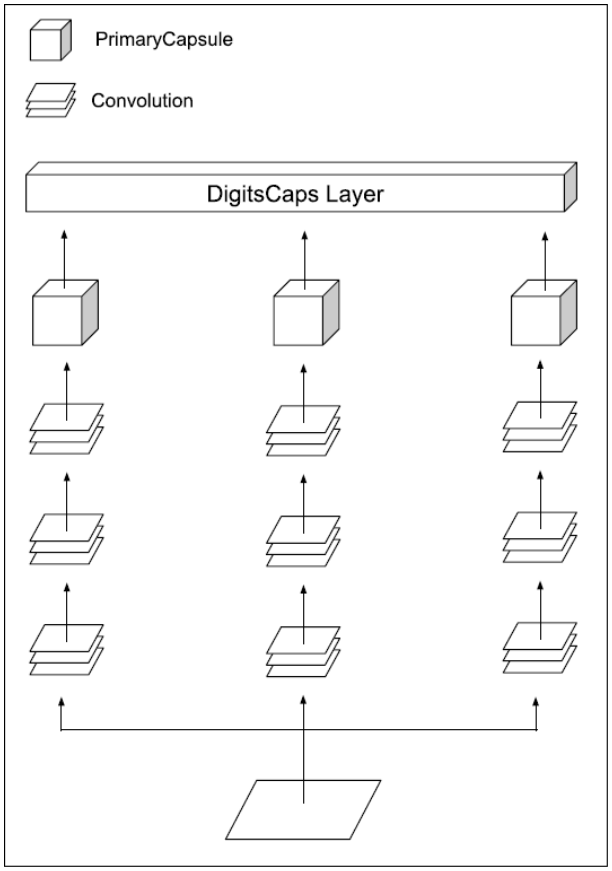

Figure 5. Network structure of PathCapsNet

4.3 Combination of Capsule and Other Models

In 2018, Adrien Deliege et al. [8] constructed a neural network called HitNet, which has a layer composed of capsules, called the hit or miss layer (HoM, which can be understood as the DigitCaps layer in the original CapsNet network structure)). All the features obtained by the HoM capsule can span the range, and can reach any value within this interval, regardless of other features. In terms of classification performance, one of the main advantages of the HoM layer is that it can be incorporated into any other network. This means that the HitNet sub-part used to calculate the feature map fully connected to the HoM can be replaced by any finer network to improve the performance of performing more complex tasks. The author's experiment proved that HitNet can achieve the latest performance of MNIST digital classification task with a shallower architecture, and HitNet is better than CapsNet's results on multiple data sets, and the speed is at least 10 times faster [8]. In 2018, the capsule projection layer proposed by Liheng Zhang et al. [40] is similar. Any network can be used to replace the network structure before the capsule projection layer. The performance of the network model with the capsule projection layer is also improved to varying degrees.

In 2018, Ayush Jaiswal et al. [42] proposed a Generative Adversarial Capsule Network (Cap-suleGAN), which uses CapsNets instead of standard Convolutional Neural Networks (CNNs) in the setting of Generative Adversarial Networks (GAN) as the framework of the discriminator, it also models the image data. Compared with the images generated by convolutional GAN, the images generated by CapsuleGAN are closer to real images and more diverse, so as to obtain better semi-supervised classification performance on the test data set [41].

In 2019, Assaf Hoogi et al. [33] combined the self-attention mechanism model with CapsNet and proposed the self-attention capsule network (SACN). The self-attention mechanism is responsible for accurately extracting the internal features of certain key positions, and CapsNet is responsible for analyzing the rotational spatial relationship between the features in the region. The initial features extracted from the two are combined to form PrimaryCapsules. The author's experimental results show that the proposed SACN improves the classification performance within and between different data sets significantly, and is better than the original CapsNet in terms of classification accuracy and robustness. Future research would lead to adopt other types of network training [65].

5 Application of Capsule Network

CapsNet developed from the basis of CNN, while inheriting the advantages of CNN, it also overcomes the shortcomings of CNN due to scalar transmission and pooling operations. CapsNet uses a collection of vector neurons creatively as the basic unit of the network, and the types of transferable features are more diverse. As an emerging network model, CapsNet has fewer parameters, smaller training set, richer feature types, and stronger ability to understand high-dimensional images compared with CNN. Therefore, CapsNet has been widely and successfully applied in various fields.

5.1 Applications in Visual Images

CapsNet was first used by Sabour, Hinton, etc. to recognize MNIST data sets of handwritten numeral images. The multi-layer capsule system after discriminative training has achieved the most advanced performance on MNIST, and is much better than convolutional networks in recognizing highly overlapping digits [2]. Later, Hinton et al. applied CapsNet to adversarial attack detection, which can detect adversarial images of three different data sets effectively [41]. The three data sets include the handwritten digital image MNIST data set, the fashionMNIST data set of clothes image, pants image, shoes image and bags image, and the street view number image SVHN data set. At the same time, CapsNets performs better than CNN in classification and detection tasks of the MNIST data set [37, 42-44] and the CIFAR-10 data set [7, 45] containing 10 categories with less training data. CapsNet can summarize complex objects well and perform well when the image is tilted or when the object is viewed from an unfamiliar angle. The paper [46] found that CapsNet is better at understanding clothes than ConvNets when viewing clothes at different rotation angles. In face recognition detection, although the internal representation of the detected entity is very complex, CapsNets can also learn good feature information from a small number of instances and converge faster [13]. The paper [47] extended the application range of CapsNet to Very Low Resolution (VLR) face image recognition. In the UCCS face database, 16×16 resolution images are used to match the corresponding 80×80 resolution image, the recognition rate of CapsNets reaches

more than 95%. CapsNet can overcome the spatial difference of detected objects, and achieves the latest accuracy of 97.6% on the German Traffic Sign Recognition Benchmark Data Set (GTSRB) [26]. The papers [48-49] use an improved dynamic routing algorithm to infer the position coordinates of the entity in the image. Compared with the convolution model, CapsNet can learn and derive the coordinates of the entities better. At the same time, it is found in paper [49] that CapsNet can also migrate the predicted movement information of the entity's position to data sets of different target types.

5.2 Applications in Remote Sensing Images

The paper [22] compares the recognition accuracy of CapsNet and other network models in public benchmark remote sensing image data sets under different training rates. Especially in the remote sensing image scene classification data set NWPU-RESISC45 created by Northwestern Polytechnic University (NWPU), CapsNet achieved 89.03% accuracy with 10% training rate, which was far higher than 76.19% accuracy rate of GoogLeNet and 76.47% accuracy rate of VGG-16. Kazi Aminul Islam team [50-51] used CapsNet to process multi-spectral satellite images, and through transfer learning, the seagrass features learned by CapsNet in a certain place can be applied to different observation points, and the seagrass in that place can still be classified and quantitatively analyzed. On the premise of insufficient training samples, CapsNet also has good performance in classification accuracy of hyperspectral images. Xue Wang et al. [6, 52] found that the accuracy of CapsNet in the classification of hyperspectral images of Indian pine trees reached 88.24%, which was the highest accuracy rate among all tested network models, while the accuracy rate of CNN under the same conditions was only 82.35%. The unique network structure of CapsNet not only has high classification accuracy for hyperspectral images with few training samples [53], but it is robust to the transformation of fineness of hyperspectral images and image rotation [16]. C.P. Schwegmann et al. [54] applied CapsNet to SAR image detection, using information such as the spatial position of the capsule to improve ship detection accuracy. The experimental results of the paper [54] show that CapsNet improves the accuracy of ships detection in SAR images to 91.03%, which is much higher than other network models.

5.3 Applications in Medical Images

The advanced capsule features in CapsNet include the semantic category, direction, position, and other features of image stimulation. The paper [23] uses these features extracted in CapsNet to reconstruct human fMRI (functional Magnetic Resonance Imaging) images and achieved good results. Aryan Mobiny et al. [10] found that the accuracy of two-dimensional lung cancer image screening is higher than the AlexNet network and ResNet-50 network when the number of training samples is less. When screening 3D lung cancer images, CapsNet is far ahead with an accuracy rate of 91.84%. The paper [33] found that CapsNet can successfully classify plaques in CT tumor lesion images with complex backgrounds. When processing medical image data sets such as brain magnetic resonance imaging (MRI) images, CapsNet's robustness to image rotation and affine transformation plays a very good role, effectively overcoming the problems of CNNs in processing brain tumor classification [19]. The paper [18] applied CapsNet to the segmentation of CT scan images of pathological lungs. The network model of the capsule structure was used to ensure segmentation accuracy and reduce the number of parameters of the U-Net structure by 95.4%. When Jiménez-Sánchez et al. [28] studied retinal image classification, they found that CapsNet can be trained with less data and is more robust in dealing with unbalanced class distributions, which makes CapsNet have a broad application prospect in the field of medical image.

5.4 Applications in Other Fields

Yequan Wang et al. [55] combined the recurrent neural network with CapsNet for human emotion analysis. The network model with the addition of the capsule structure can output words with emotional tendency reflecting the attributes of the capsule without using any language knowledge, and the emotional description of the data has high accuracy. Mohammad Taha Bahadori applies CapsNet to the diagnosis of medical clinical data. Compared with EM-Capsules and deep GRU networks, the S-Capsules network proposed by him has a faster learning speed and stronger generalization ability [56]. Because CapsNet can still classify each entity even when multiple detection entities overlap each other. According to this feature of CapsNet, the paper [57] applies CapsNet to the mixed speech data to extract a single voice event and achieves better results than other methods. The paper [58-60] applied CapsNet to sound detection successfully, and greatly reduced the over-fitting phenomenon. When Wei Zhao et al. [34] explored the application of CapsNet in text classification, they found that CapsNet's classification accuracy is more competitive than other models when converting single-label text classification to multi-label text classification. The paper [61] applies CapsNet to video object segmentation. The proposed method can segment multiple frames simultaneously on the basis of one reference frame and one segmentation mask, which greatly increases the segmentation speed of video objects. In addition, CapsNet is also widely used in road network traffic flow prediction [30], protein structure prediction [62], and camera-based UAV synchronous positioning [27]. CapsNet were used in [63] for detecting fake news. Their solution consists in two parallel networks and non-static word embedding, and is better than competitive approaches. In [64] CapsNet are used to remaining life estimation. It transpired that advantages of CapsNets from computer vision tasks are also valid in fault prognostics yielding better accuracy than CNNs.

6 Conclusion

In recent years, CNN, as a classic network model in the field of deep learning, has received more and more attention. However, CNN has its insurmountable flaws. For example, the use of the pooling sub-sampling method discards a lot of useful information, the network ignores the hierarchical structure between layers and needs a large number of training samples. As the next generation of the deep learning network model, CapsNet developed from the basis of CNN, inherited

the system advantages of CNN, and overcomes the shortcomings of CNN due to pooling operation and scalar transmission, and gradually becomes a new research hotspot in the field of deep learning.

This paper first analyzes the advantages and disadvantages of CNN, and on the basis of summarizing the advantages and disadvantages of CNN, leads to the second part of the paper, which is the basic concept of CapsNet. This paper introduces the uniqueness of CapsNet from three aspects: capsule model, dynamic routing mechanism, and CapsNet's network model. These operations of CapsNet enable CapsNet to maintain high accuracy even with less training data and can capture hidden features in the data. However, the basic CapsNet has some shortcomings. The third part of the paper summarizes the three aspects of improving dynamic routing mechanism, improvement of the network model, and the combination of capsules with other models. The similarities and differences between these improvements and the original CapsNet are compared, and the improvement direction and the benefits and disadvantages are analyzed. The fourth part of the paper lists the application examples of CapsNet in the field of visual image, remote sensing image, medical image, and other fields and summarizes the reasons why CapsNet has been successfully applied in these fields.

Although CapsNet has been widely used in various fields, its advantages do not mean that it can solve and improve all the problems in the past. For example, CapsNet often cannot reconstruct complex images accurately. This paper summarizes the structural characteristics of CapsNet and the improvement direction of CapsNet by researchers, hoping to sort out the development trend and context of CapsNet for readers and provide a reference for the development direction of CapsNet network.

Acknowledgements

This work was supported by the National Natural Science Foundation of China (No. 61102163), the Fundamental Research Funds for the Central Universities (No. GK201903085), the Key Laboratory of Land Satellite Remote Sensing Application Center, Ministry of Natural Resources of the People's Republic of China (No. KLSMNR-202004), and the State Key Laboratory of Geo-Information Engineering (No. SKLGIE2019-M-3-5).

References

[1] J. Y. Zhang, H. L. Wang, Y. Guo, X. Hu, Review of deep learning, *Application Research of Computers*, Vol. 35, No. 7, pp. 1921-1928, July, 2018.

[2] S. Sabour, N. Frosst, G. E. Hinton, Dynamic routing between capsules, *31st Conference on Neural Information Processing Systems (NIPS 2017)*, Long Beach, CA, USA, 2017, pp. 3856-3866.

[3] W. Y. Wang, H. C. Li, L. Pan, G. Yang, Q. Du, Hyperspectral image classification based on capsule network, *IEEE International Geoscience and Remote Sensing Symposium*, Valencia, Spain, 2018, pp. 3571-3574.

[4] M. E. Paoletti, J. M. Haut, R. Fernandez-Beltran, J. Plaza, A. Plaza, J. Li, F. Pla, Capsule Networks for Hyperspectral Image Classification, *IEEE Transactions on Geoscience and Remote Sensing*, Vol. 57, No. 4, pp. 2145-2160, April, 2019.

[5] Y. Luo, J. Zou, C. Yao, X. Zhao, T. Li, G. Bai, HSI-CNN: a novel convolution neural network for hyperspectral image, *International Conference on Audio, Language and Image Processing*, Shanghai, China, 2018, pp. 464-469.

[6] X. Wang, K. Tan, Y. Chen, CapsNet and Triple-GANs Towards Hyperspectral Classification, *Fifth International Workshop on Earth Observation and Remote Sensing Applications*, Xi'an, China, 2018, pp. 1-4.

[7] R. Mukhometzianov, J. Carrillo, CapsNet comparative performance evaluation for image classification, *arXiv preprint*, arXiv: 1805.11195, May, 2018.

[8] A. Deliège, A. Cioppa, M. V. Droogenbroeck, Hitnet: a neural network with capsules embedded in a hit-or-miss layer, extended with hybrid data augmentation and ghost capsules, *arXiv preprint*, arXiv: 1806.06519, June, 2018.

[9] S. Chaib, M. E. A. Larabi, Y. Gu, K. Bakhti, M. S. Karoui, Very High Resolution Image Scene Classification with Capsule Network, *IEEE International Geoscience and Remote Sensing Symposium*, Yokohama, Japan, 2019, pp. 3049-3052.

[10] A. Mobiny, H. V. Nguyen, Fast capsnet for lung cancer screening, *International Conference on Medical Image Computing and Computer-Assisted Intervention*, Granada, Spain, 2018, pp. 741-749.

[11] S. Ramasinghe, C. D. Athuraliya, S. H. Khan, A context-aware capsule network for multi-label classification, *Proceedings of the European Conference on Computer Vision*, Munich, Germany, 2018, pp. 546-554.

[12] C. Xiang, L. Zhang, Y. Tang, W. Zou, C. Xu, MS-CapsNet: A novel multi-scale capsule network, *IEEE Signal Processing Letters*, Vol. 25, No. 12, pp. 1850-1854, December, 2018.

[13] J. O. Neill, Siamese capsule networks, *arXiv preprint*, arXiv: 1805.07242, May, 2018.

[14] T. Iesmantas, R. Alzbutas, Convolutional capsule network for classification of breast cancer histology images, *International Conference Image Analysis and Recognition*, Povoa de Varzim, Portugal, 2018, pp. 853-860.

[15] T. Tian, X. Liu, L. Wang, Remote Sensing Scene Classification Based on Res-Capsnet, *International Geoscience and Remote Sensing Symposium*, Yokohama, Japan, 2019, pp. 525-528.

[16] D. Wang, Q. Xu, Y. Xiao, J. Tang, B. Luo, Multi-scale Convolutional Capsule Network for Hyperspectral Image Classification, *Chinese Conference on Pattern Recognition and Computer Vision*, Xi'an, China, 2019, pp. 749-760.

[17] J. H. Fu, X. F. Wu, S. F. Zhang, Study on Characteristics of Capsule Network Based on Affine Transformation, *Journal of Signal Processing*, Vol. 34, No. 12, pp. 1508-1516, December, 2018.

[18] R. LaLonde, U. Bagci, Capsules for object segmentation, *arXiv preprint*, arXiv: 1804.04241, April, 2018.

[19] P. Afshar, A. Mohammadi, K. N. Plataniotis, Brain tumor type classification via capsule networks, 25th *IEEE International Conference on Image Processing*, Athens, Greece, 2018, pp. 3129-3133.

[20] A. Shahroudnejad, P. Afshar, K. N. Plataniotis, A. Mohammadi, Improved explainability of capsule networks: Relevance path by agreement, *Global Conference on Signal and Information Processing*, Anaheim, CA, USA, 2018, pp. 549-553.

[21] E. Xi, S. Bing, Y. Jin, Capsule network performance on complex data, *arXiv preprint*, arXiv: 1712.03480, December, 2017.

[22] W. Zhang, P. Tang, L. Zhao, Remote sensing image scene classification using CNN-CapsNet, *Remote Sensing*, Vol. 11, No. 5, Article No. 494, March, 2019.

[23] K. Qiao, C. Zhang, L. Wang, B. Yan, J. Chen, L. Zeng, L. Tong, Accurate reconstruction of image stimuli from human fMRI based on the decoding model with capsule network architecture, *arXiv preprint*, arXiv: 1801.00602, January, 2018.

[24] F. Deng, S. Pu, X. Chen, Y. Shi, T. Yuan, S. Pu, Hyperspectral image classification with capsule network using limited training samples, *Sensors*, Vol. 18, No. 9, Article No. 3153, September, 2018.

[25] V. M. Rosario, E. Borin, J. M. Breternitz, The multi-lane capsule network (mlcn), *arXiv preprint*, arXiv: 1902.08431, February, 2019.

[26] A. D. Kumar, Novel deep learning model for traffic sign detection using capsule networks, *arXiv preprint*, arXiv: 1805.04424, May, 2018.

[27] S. Prakash, G. Gu, Simultaneous Localization And Mapping with depth Prediction using Capsule Networks for UAVs, *arXiv preprint*, arXiv: 1808.05336, August, 2018.

[28] A. Jiménez-Sánchez, S. Albarqouni, D. Mateus, Capsule networks against medical imaging data challenges, *Intravascular Imaging and Computer Assisted Stenting and Large-Scale Annotation of Biomedical Data and Expert Label Synthesis*, Granada, Spain, 2018, pp. 150-160.

[29] G. E. Hinton, S. Sabour, N. Frosst, Matrix capsules with EM routing, *International Conference on Learning Representations*, Vancouver, Canada, 2018, pp. 1-15.

[30] Y. Kim, P. Wang, Y. Zhu, L. Mihaylova, A capsule network for traffic speed prediction in complex road networks, *Sensor Data Fusion: Trends, Solutions, Applications*, Bonn, Germany, 2018, pp. 1-6.

[31] P. Nair, R. Doshi, S. Keselj, Pushing the limits of capsule networks, *arXiv preprint*, arXiv: 2103.08074, March, 2021.

[32] A. C. Liu, J. Li, Z. Y. Ma, On learning and learned data representation by capsule networks, *IEEE Access*, Vol. 7, pp. 50808-50822, April, 2019.

[33] A. Hoogi, B. Wilcox, Y. Gupta, D. L. Rubin, Self-Attention Capsule Networks for Image Classification, *arXiv preprint*, arXiv: 1904.12483v1, pp. 1-9, April, 2019.

[34] W. Zhao, J. B. Ye, M. Yang, Z. Y. Lei, S. F. Zhang, Z. Zhao, Investigating capsule networks with dynamic routing for text classification, *arXiv preprint*, arXiv: 1804.00538, pp. 1-12, September, 2018.

[35] S. F. Zhang, Q. Zhou, X. F. Wu, Fast dynamic routing based on weighted kernel density estimation, in: H. Lu (Eds.), Cognitive Internet of Things: Frameworks, Tools and Applications. *ISAIR 2018. Studies in Computational Intelligence*, Vol. 810, Springer, Cham, 2018, pp. 301-309.

[36] H. Ren, H. Lu, Compositional coding capsule network with k-means routing for text classification, *arXiv preprint*, arXiv: 1810.09177, pp. 1-5, October, 2018.

[37] Z. H. Chen, D. Crandall, Generalized capsule networks with trainable routing procedure, *arXiv preprint*, arXiv: 1808.08692, pp. 1-4, August, 2018.

[38] M. Amer, T. Maul, Path capsule networks, *arXiv preprint*, arXiv: 1902.03760, pp. 1-10, February, 2019.

[39] C. Y. Xia, C. W. Zhang, X. H. Yan, Y. Chang, P. S. Yu, Zero-shot user intent detection via capsule neural networks, *arXiv preprint*, arXiv: 1809.00385, pp. 1-11, September, 2018.

[40] L. H. Zhang, M. Edraki, G. J. Qi, Cappronet: Deep feature learning via orthogonal projections onto capsule subspaces, *Advances in Neural Information Processing Systems*, Montreal, Canada, 2018, pp. 5819-5828.

[41] N. Frosst, S. Sabour, G. Hinton, DARCCC: Detecting adversaries by reconstruction from class conditional capsules, *arXiv preprint*, arXiv: 1811.06969, pp. 1-13, November, 2018.

[42] A. Jaiswal, W. AbdAlmageed, Y. Wu, P. Natarajan, Capsulegan: Generative adversarial capsule network, *Proceedings of the European Conference on Computer Vision*, Munich, Germany, 2018, pp. 526-535.

[43] F. Y. Chen, N. Chen, H. Y. Mao, H. L. Hu, Assessing four neural networks on handwritten digit recognition dataset (MNIST), *arXiv preprint*, arXiv: 1811.08278v1, pp. 1-4, November, 2018.

[44] S. S. R. Phaye, A. Sikka, A. Dhall, D. Bathula, Dense and diverse capsule networks: Making the capsules learn better, *arXiv preprint*, arXiv: 1805.04001, pp. 1-11, May, 2018.

[45] S. D. Lin, C. Q. Hong, Y. X. Chen, An Object recognition model combining capsule network and convolutional neural network, *Telecommunication Engineering*, Vol. 59, No. 9, pp. 987-994, September, 2019.

[46] M. Engelin, *CapsNet Comprehension of Objects in Different Rotational Views: A comparative study of capsule and convolutional networks*, Master's Thesis, KTH, School of Electrical Engineering and Computer Science (EECS), Stockholm, Sweden, 2018.

[47] M. Singh, S. Nagpal, R. Singh, M. Vatsa, Dual Directed Capsule Network for Very Low Resolution Image Recognition, *Proceedings of the IEEE International Conference on Computer Vision*, Seoul, Korea, 2019, pp. 340-349.

[48] W. T. Liu, E. Barsoum, J. D. Owens, Object Localization with a Weakly Supervised CapsNet, *arXiv preprint*, arXiv: 1805.07706v3, pp. 1-12, December, 2019.

[49] W. T. Liu, E. Barsoum, J. D. Owens, Object localization and motion transfer learning with capsules, *arXiv preprint*, arXiv: 1805.07706v1, pp. 1-9, May, 2018.

[50] K. A. Islam, D. Pérez, V. Hill, B. Schaeffer, R. Zimmerman, J. Li, Seagrass detection in coastal water through deep capsule networks, *Chinese Conference on*

Pattern Recognition and Computer Vision, Guangzhou, China, 2018, pp. 320-331.

[51] D. Pérez, K. Islam, V. Hill, R. Zimmerman, B. Schaeffer, J. Li, Deepcoast: Quantifying seagrass distribution in coastal water through deep capsule networks, *Chinese Conference on Pattern Recognition and Computer Vision*, Guangzhou, China, 2018, pp. 404-416.

[52] X. Wang, K. Tan, Q. Du, Y. Chen, P. Du, Caps-TripleGAN: GAN-Assisted CapsNet for Hyperspectral Image Classification, *IEEE Transactions on Geoscience and Remote Sensing*, Vol. 57, No. 9, pp. 7232-7245, September, 2019.

[53] Y. Ma, Z. Z. Zheng, Z. Q. Guo, F. Mou, F. R. Zhou, R. Kong, A. K. Hou, M. C. Zhu, Y. He, J. Ren, H. X. Chen, Z. G. Liu, G. Q. Zhou, J. Li, Classification Based on Capsule Network with Hyperspectral Image, *IGARSS 2019-2019 IEEE International Geoscience and Remote Sensing Symposium*, Yokohama, Japan, 2019, pp. 2750-2753.

[54] C. P. Schwegmann, W. Kleynhans, B. P. Salmon, L. W. Mdakane, R. G. V. Meyer, Synthetic aperture radar ship detection using capsule networks, IGARSS 2018-2018 *IEEE International Geoscience and Remote Sensing Symposium*, Valencia, Spain, 2018, pp. 725-728.

[55] Y. Q. Wang, A. X. Sun, J. L. Han, Y. Liu, X. Y. Zhu, Sentiment analysis by capsules, *Proceedings of the 2018 world wide web conference*, Lyon, France, 2018, pp. 1165-1174.

[56] M. T. Bahadori, Spectral capsule networks, 6th *International Conference on Learning Representations (ICLR 2018 Workshop)*, Vancouver, BC, Canada, 2018, pp. 1-5.

[57] Y. M. Liu, J. Tang, Y. Song, L. R. Dai, A capsule based approach for polyphonic sound event detection, *2018 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference*, Honolulu, Hawaii, USA, 2018, pp. 1853-1857.

[58] T. Iqbal, Y. Xu, Q. Q. Kong, W. W. Wang, Capsule routing for sound event detection, *2018 26th European Signal Processing Conference*, Rome, Italy, 2018, pp. 2255-2259.

[59] F. Vesperini, L. Gabrielli, E. Principi, S. Squartini, Polyphonic sound event detection by using capsule neural networks, *IEEE Journal of Selected Topics in Signal Processing*, Vol. 13, No. 2, pp. 310-322, May, 2019.

[60] J. J. Wang, S. N. Ji, L. Cui, J. Xia, Q. Yang, Domestic activity recognition based on attention capsule network, *Acta Automatica Sinica*, Vol. 45, No. 11, pp. 2199-2204, November, 2019.

[61] K. Duarte, Y. S. Rawat, M. Shah, CapsuleVOS: Semi-Supervised Video Object Segmentation Using Capsule Routing, *Proceedings of the IEEE International Conference on Computer Vision*, Seoul, Korea, 2019, pp. 8480-8489.

[62] C. Fang, Y. Shang, D. Xu, Improving protein gamma-turn prediction using inception capsule networks, *Scientific reports*, Vol. 8, pp. 1-12, October, 2018.

[63] M. H. Goldani, S. Momtazi, R. Safabakhsh, Detecting fake news with capsule neural networks, *Applied Soft Computing*, Vol. 101, Article No. 106991, March, 2021.

[64] A. Ruiz-Tagle Palazuelos, E. L. Droguett, R. Pascual, A novel deep capsule neural network for remaining useful life estimation, *Proceedings of the Institution of Mechanical Engineers, Part O: Journal of Risk and Reliability*, Vol. 234, No. 1, pp. 151-167, February, 2020.

[65] J. Bilski, B. Kowalczyk, A. Marchlewska, J. M. Zurada, Local Levenberg-Marquardt Algorithm for Learning Feedforwad Neural Networks, *Journal of Artificial Intelligence and Soft Computing Research*, Vol. 10, No. 4, pp. 299-316, October, 2020.

Biographies

Zengguo Sun was born in Xi'an, Shaanxi, China, in 1980. He received his bachelor's degree in Computer Science and Technology in 2003, and PhD degree in Control Science and Engineering in 2010, all from the Xi'an Jiaotong University. He was a visiting scholar in Pennsylvania State University from 2007 to 2008. Now he is an Associate Professor in Shaanxi Normal University. His research interests are in radar image processing.

Guodong Zhao was born in Fuping, Shaanxi, China, in 1995. He received his bachelor's degree in Information Management and Information System from Inner Mongolia University, in 2018. He is currently pursuing the master's degree in Computer Software and Theory with Shaanxi Normal University. His research interests are in radar image processing.

Rafał Scherer is a full professor at the Czestochowa University of Technology, Poland. His research focuses on developing new methods in neural networks, computer vision, computational intelligence and data mining, ensembling methods in machine learning, content-based image indexing. He authored more than 130 research papers and two books: on multiple classification techniques (2012) and computer vision methods (2020) published by Springer.

Wei Wei (SM'17) received the M.S. and Ph.D. degrees from Xi'an Jiaotong University, Xi'an, China, in 2005 and 2011, respectively. He is currently an Associate Professor with the School of Computer Science and Engineering, Xi'an University of Technology, Xi'an. He ran many funded research projects as principal investigator and technical

members. His current research interests include the area of wireless networks, wireless sensor networks application, image processing, mobile computing, distributed computing, and pervasive computing, Internet of Things, and sensor data clouds. He has published around 100 research papers in international conferences and journals. Dr. Wei is a Senior Member of the China Computer Federation. He is an Editorial Board Member of the Future Generation Computer System, the IEEE Access, Ad Hoc & Sensor Wireless Sensor Network, The Institute of Electronics, Information and Communication Engineers, and KSII Transactions on Internet and Information Systems. He is a TPC member of many conferences and a regular Reviewer of the IEEE Transactions on Parallel and Distributed Systems, the IEEE Transactions on Image Processing, the IEEE Transactions on Mobile Computing, the IEEE Transactions on Wireless Communications, the Journal of Network and Computer Applications, and many other Elsevier journals.

Marcin Woźniak received diplomas in applied mathematics and computational intelligence. He is an Assoc. Professor at Institute of Mathematics of the Silesian University of Technology in Gliwice, Poland. In his scientific career, he was visiting University of Wrzburg, Germany, University of Lund, Sweden and University of Catania, Italy. His main scientific interests are neural networks with their applications together with various aspects of applied computational intelligence. He is a scientific supervisor in editions of “the Diamond Grant” and “The Best of the Best” programs for highly gifted students from the Polish Ministry of Science and Higher Education. Marcin Woźniak served as an editor for various special issues of IEEE ACCESS, Sensors, Frontiers in Human Neursocience, etc., and as an organizer or a session chair at various international conferences and symposiums, including IEEE SSCI, ICAISC, WorldCIST.