Research on dynamic routing algorithm based on gaussian mixture model

Yuzhan Huang*

University of Electronic Science and Technology of China, Chengdu, China

Abstract. In this paper, based on the method of environmental sound detection, a neural network model based on capsule network and Gaussian mixture model is proposed. The model proposed in this paper mainly aims at the disadvantages of dynamic routing algorithm in the capsule network, and proposes a dynamic routing algorithm based on Gaussian mixture model. The improved dynamic routing algorithm assumes that the characteristics of the data conform to the multi-dimensional Gaussian distribution, so the model can learn the distribution of data features by building distribution functions of different classes. The information entropy is used as the activation value of the salient degree of the feature. Through experiments, the accuracy of the proposed algorithm on Urbansound8K data set is more than 92%, which is 4.8% higher than the original algorithm.

Keywords: Capsule net, Dynamic routing algorithm, Gaussian mixture model.

1 Introduction

With the rapid development of artificial intelligence, deep learning has quickly replaced the traditional machine learning method with its powerful fitting ability, and has become one of the tools widely used by people. From face recognition to intelligent payment, from intelligent translation to text generation, deep learning can be seen everywhere. Of course, with the rise of deep learning, environmental acoustic event detection is closely integrated with deep learning.

The integration of deep learning and acoustics began with a convolutional neural network (CNN). Hershey et al. used a variety of models based on CNN architecture to recognize acoustic signals and achieved good results [1]. However, acoustic signals have a strong context relationship. Traditional CNN extracts local features through convolution check, and lacks a mechanism to infer local features to global position relations. Therefore, CNN's performance in time series data is not as good as its performance in picture data.To solve these problems, Keren et al. proposed a convolutional recursive neural network (CRNN) structure, which combined the advantages of recurrent neural network (RNN).

* Corresponding author: 201822120428@std.uestc.edu.cn

Therefore, it achieves better effect than CNN in timing signal [2]. Chew et al. proposed the long-short term memory (LSTM), which achieved excellent results through the mixed architecture with other models [3].

Capsule network was proposed by Hinton [4] in 2017. Different from scalar neurons in traditional neural network architecture, Capsule network uses vector as its basic unit, which is called Capsule. Therefore, compared with traditional neural network, capsule network has richer dimension of feature space and can carry more feature information. Similar to distributed representation, features represented by capsule can express more kinds of similarity, which is difficult for a single digit in traditional neuron to do. The second characteristic that differentiates capsule networks from traditional neural networks is the dynamic routing algorithm. Through the dynamic routing algorithm, the capsule network can combine the low-level features into the high-level features in the way of clustering, that is, the capsule network can represent the input into several feature vectors, and then cluster these feature vectors layer by layer, in order to complete the abstraction process of features. In the capsule network, the significance of features is reflected in the module length of vectors, so the Loss function is Margin Loss. In addition, since the sum of feature vectors is involved in the dynamic routing process, a Squash function is needed to compress the result's module length to between [0,1] to represent the probability. In this paper, a Gaussian mixture model (GMM) is proposed as a dynamic routing algorithm, used to replace the dynamic routing algorithm between the bottom capsule and the top capsule in the original capsule network, so that the model can learn the distribution of data, by building the distribution function of different classes, and the interpretability and classification ability of the model can be improved.

2 Improvement of capsule network

2.1 Capsule network

In the capsule network, the lower capsule and the higher capsule are connected through the weight matrix, and the weight matrix is adjusted according to the consistency degree of the lower capsule and the higher capsule. This process is the dynamic routing algorithm. As shown in Fig. 1[5], the green framed part represents a capsule unit, each capsule represents a feature, and the vector represents the salience of the feature. The output of the bottom capsule $\boldsymbol{u}_i$ is first multiplied by the transformation matrix $\boldsymbol{W}_{ij}$ to obtain the prediction vector $\hat{\boldsymbol{u}}_{j|i}$.

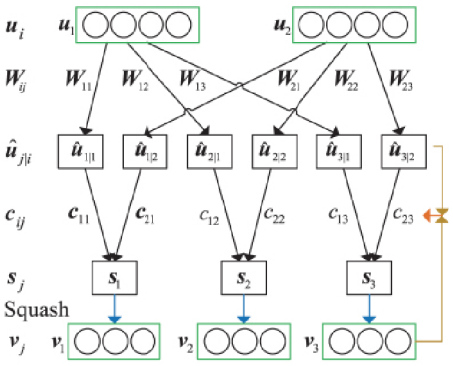

Fig. 1. Connection between the bottom capsule and the upper capsule.

The probability that $\boldsymbol{u}_1$ belongs to the upper capsule $\boldsymbol{v}_j$ is equal to

$$\left(p_{1|1}, p_{2|1}\right)=\frac{1}{S_1}\left(e^{\langle \boldsymbol{u}_1, \boldsymbol{v}_1\rangle}, e^{\langle \boldsymbol{u}_1, \boldsymbol{v}_2\rangle}\right) \tag{1}$$

where the symbol $\langle \bullet \rangle$ represents the inner product and $S_i=\sum e^{\langle \boldsymbol{u}_i, \boldsymbol{v}_j\rangle}$. Considering the total i capsules in the lower layer, and all features are weighted average as input to the upper capsule $\boldsymbol{v}_j$,So the input of the upper capsule is ultimately determined to be $\boldsymbol{v}_{in}=\left(p_{1|i}\boldsymbol{u}_i, p_{2|i}\boldsymbol{u}_i\right)$. In order to make the model have the ability to represent nonlinear features, nonlinearity is usually introduced at the activation function level. Capsule net selected the squash function as the activation function, the expression of which is

$$squash(\boldsymbol{x})=\frac{\|\boldsymbol{x}\|^2}{1+\|\boldsymbol{x}\|^2}\bullet\frac{\boldsymbol{x}}{\|\boldsymbol{x}\|} \tag{2}$$

The final dynamic routing algorithm flow is shown in algorithm 1:

algorithm 1: dynamic routing algorithm

INPUT: $\hat{\boldsymbol{u}}_{j|i}$,The number of iterations r, The layer number of l

OUTPUT: $\boldsymbol{v}_j$

1. For all capsules in layer L and layer L +1: Initialize: $b_{ij}\leftarrow 0$

For iterates r times do:

2. For all capsules I in the L layer: $\boldsymbol{p}_i\leftarrow soft\,max(\boldsymbol{b}_i)$

3. For all capsules J in the L+1 layer: $\boldsymbol{s}_j\leftarrow\sum_i p_{j|i}\hat{\boldsymbol{u}}_{j|i}$

4. For all capsules J in the L+1 layer: $\boldsymbol{v}_j\leftarrow squash(\boldsymbol{s}_j)$

5. For all Capsule I in the Layer L and for all Capsule J in the Layer L+1: $b_{ij}\leftarrow b_{ij}+\langle u_i, v_j\rangle$

End for

Return $\boldsymbol{v}_j$;

where, the input R is the hyper-parameter. If the value is too large, the model will fall into over-fitting, else if the value is too small, the model will not converge. In the original text, the value is generally 3. After r iterations of the dynamic routing algorithm, similar capsule units obtain more underlying weights, and the corresponding vector modulus length in the capsule is also larger.

2.2 Improved capsule network

What was used in the original capsule model was actually K-means clustering. One of the major drawbacks of k-means is that it uses the mean of the cluster center, therefore, the K-means algorithm in the capsule network is replaced with GMM in this paper. In GMM, each capsule can have multiple classifications, so if a capsule is located in the middle of multiple overlapping clusters, its class can be defined simply by the soft-max function. GMM assumes that the data points is Gaussian distribution, for the existing vector

$\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots\ldots, \boldsymbol{x}_n$, and hope to find the distribution $p(\boldsymbol{x})$ they meet. Assume that the data can be divided into several categories, and each category will be studied separately, thus $p(\boldsymbol{x}) = \sum_{i=1}^{n} p(i) p(\boldsymbol{x} \mid i)$, where *i* represents the category, with a value of 1,2... , *n*. Since $p(i)$ doesn't depend on *x*, it can be thought of as a constant distribution, $p(i) = \pi_i$, and $p(\boldsymbol{x} \mid i)$ represents the probability distribution in class *i*. This probability distribution is generally assumed to be multidimensional normal distribution:

$$N(\boldsymbol{x}; \boldsymbol{\mu}_i, \boldsymbol{C}_i) = \frac{1}{(2\pi)^{\frac{d}{2}} (det\, \boldsymbol{C}_i)^{\frac{1}{2}}} exp\left(-\frac{1}{2} (\boldsymbol{x} - \boldsymbol{\mu}_i)^T \boldsymbol{C}^{-1} (\boldsymbol{x} - \boldsymbol{\mu}_i) \right) \tag{3}$$

where *d* is the number of components of the vector $\boldsymbol{x}$. Let $p(\boldsymbol{x} \mid i) = N(\boldsymbol{x}; \boldsymbol{\mu}_i, \boldsymbol{C}_i)$, according to Bayes' formula,

$$p(i \mid \boldsymbol{x}) = \frac{p(\boldsymbol{x} \mid i) p(i)}{p(\boldsymbol{x})} = \frac{\pi_i N(\boldsymbol{x}; \boldsymbol{\mu}_i, \boldsymbol{C}_i)}{\sum_{i=1}^{k} \pi_i N(\boldsymbol{x}; \boldsymbol{\mu}_i, \boldsymbol{C}_i)} \tag{4}$$

For the mean vector

$$\boldsymbol{\mu}_i = \int p(\boldsymbol{x} \mid i) \boldsymbol{x} d\boldsymbol{x} = \int p(\boldsymbol{x}) \frac{p(i \mid \boldsymbol{x})}{p(i)} \boldsymbol{x} d\boldsymbol{x} = E\left[\frac{p(i \mid \boldsymbol{X})}{p(i)} \boldsymbol{X} \right] \tag{5}$$

where $E(\bullet)$ represents the mathematical expectation of all samples,thus

$$\boldsymbol{\mu}_i = \frac{1}{n} \sum_{j=1}^{n} \frac{p(i \mid \boldsymbol{x}_j)}{p(i)} \boldsymbol{x}_j = \frac{1}{\pi_i n} \sum_{j=1}^{n} p(i \mid \boldsymbol{x}_j) \boldsymbol{x}_j \tag{6}$$

For the covariance matrix $\boldsymbol{C}_i$,

$$\boldsymbol{C}_i = \frac{1}{\pi_i n} \sum_{j=1}^{n} p(i \mid \boldsymbol{x}_j) (\boldsymbol{x}_j - \boldsymbol{\mu}_i) (\boldsymbol{x}_j - \boldsymbol{\mu}_i)^T \tag{7}$$

And because $\pi_i = p(i) = \int p(i \mid \boldsymbol{x}) p(\boldsymbol{x}) d\boldsymbol{x} = E[p(i \mid \boldsymbol{X})]$, therefore,

$$\pi_i = \frac{1}{n} \sum_{j=1}^{n} p(i \mid \boldsymbol{x}_j) \tag{8}$$

Theoretically, it is necessary to solve the equations composed of (4),(6),(7) and (8). However, such a large system of equations is difficult to solve, so in practical application, an iterative process is used to solve the approximate solution. The iteration process is as follows:

$$p(i \mid \boldsymbol{x}_j) \leftarrow \frac{\pi_i N(\boldsymbol{x}_j; \boldsymbol{u}_i, \boldsymbol{C}_i)}{\sum_{i=1}^{k} \pi_i N(\boldsymbol{x}_j; \boldsymbol{u}_i, \boldsymbol{C}_i)} \tag{9}$$

$$\boldsymbol{\mu}_i \leftarrow \frac{1}{\sum_{j=1}^{n} p(i \mid \boldsymbol{x}_j)} \sum_{j=1}^{n} p(i \mid \boldsymbol{x}_j) \boldsymbol{x}_j \tag{10}$$

$$\boldsymbol{C}_i \leftarrow \frac{1}{\sum_{j=1}^{n} p(i \mid \boldsymbol{x}_j)} \sum_{j=1}^{n} p(i \mid \boldsymbol{x}_j)(\boldsymbol{x}_j - \boldsymbol{\mu}_i)(\boldsymbol{x}_j - \boldsymbol{\mu}_i)^T \tag{11}$$

$$\pi_i \leftarrow \frac{1}{n} \sum_{j=1}^{n} p(i \mid \boldsymbol{x}_j) \tag{12}$$

In the original Capsule net, the significance of the feature is represented by the normalized modulus length of the vector. However, the essence of GMM is based on weighted Euclidean distance. The center vector obtained by clustering with Euclidean distance is the weighted average of vectors within the class. Therefore, an additional activation value needs to be added to identify the significance of the corresponding feature. In this paper, information entropy is selected as the activation value of the model. The activation value a_i is :

$$a_i = sigmoid\left(\frac{d}{2} + \frac{d}{2} ln 2\pi + \sum_{l=1}^{d} ln \boldsymbol{\sigma}_i^l \right) \tag{13}$$

Replace the distribution π_i with activation values a_i. Taking into account the iteration between the bottom capsule and the upper capsule, the threshold value of the *i* capsule in the L layer is represented by the symbol $a_i^{(l)}$. Each pair of indices (*i*,*j*) is assigned a weight matrix W_{ij}, therefore, the improved dynamic routing algorithm is shown in algorithm 2:

algorithm 2: Improved dynamic routing algorithm

INPUTS: $x_j^{(l)}$,The number of layers L, the activation value $a_i^{(l)}$ of layer L;

OUTPUTS: $x_j^{(l+1)}$

1.Perform a linear transformation of the inputs to all L-layer capsules $\boldsymbol{y}_j^{(l)} \leftarrow \boldsymbol{W}_{ij} \boldsymbol{x}_j^{(l)}$

2.For all L-layer capsules N, initialize $p_{ij} \leftarrow N\left(\boldsymbol{y}_j^{(l)}; \boldsymbol{u}_i, \boldsymbol{\sigma}_i^2\right)$

3.For all L-layer capsules N, initialize $R_{ij} \leftarrow \frac{a_i^l p_{ij}}{\sum_i a_i^l p_{ij}}$, $r_{ij} \leftarrow \frac{a_i^l R_{ij}}{\sum_j a_i^l R_{ij}}$

4.For all L +1 capsules M, $\boldsymbol{x}_j^{(l+1)} \leftarrow \sum_j r_{ij} \boldsymbol{x}_j^{(l)}$

5.For all L +1 capsules M, $\boldsymbol{\sigma}_i^2 \leftarrow \sum_j r_{ij} \left(\boldsymbol{x}_j^{(l)} - \boldsymbol{x}_j^{(l+1)} \right)^2$

6.For all L +1 capsules M,Calculate the activation value, $a_i^{(l+1)} \leftarrow sigmoid\left(\frac{d}{2} + \frac{d}{2} ln 2\pi + \sum_{l=1}^{d} ln \boldsymbol{\sigma}_i^l \right)$

3 Data processing and experiment

3.1 Data preprocessing

The data set used in this paper is Urbansound8K data set, which contains 8732 annotated sound fragments, all of which are less than or equal to 4 seconds in length, including 10 categories: air conditioning sound, car siren sound, children playing sound, dog barking, drilling sound, engine idling sound, gunfire, jackhammer, siren sound and street music sound. Typical signals for each category are shown in Figure 3. In this paper, all the original signals in the data set are preprocessed by calculating the logarithmic MEL spectrum, with a sampling rate of 32kHz, and the original signals less than 4 seconds in length are processed by 0 complement. The selected frame length is 1 frame per 1000 sampling points, the step size is 500 sampling points, and the number of MEL filters is set to 128. Therefore, the final MFCC feature map is 128*257, and the typical MFCC features of each type of signal are shown in Figure 2.

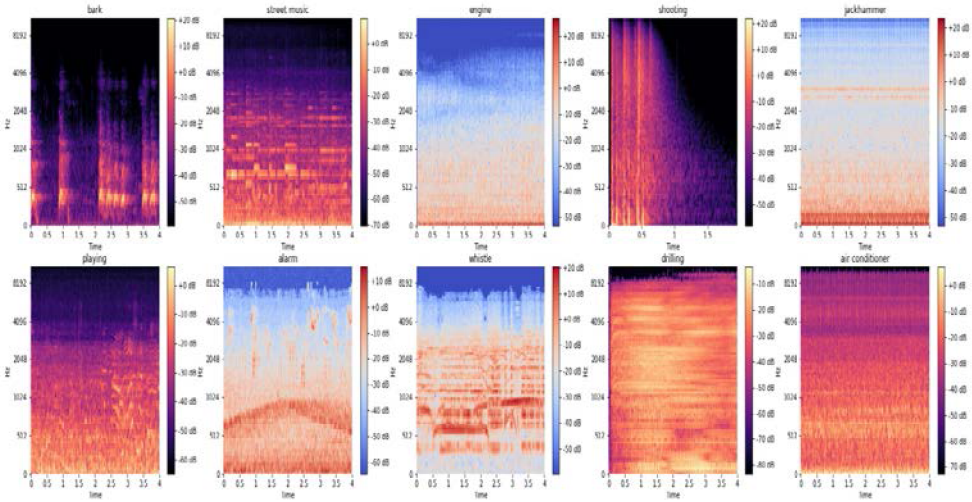

Fig. 2. Logarithmic MEL spectra of each typical type of original signal.

3.2 Model structure

The preprocessed MFCC features were used as samples to train the improved capsule network. In the model, the mainstream CNN baseline model, the original capsule model and the improved capsule network model were selected. According to the dimension of MFCC features, the layers of the capsule network were as follows: the first layer was the capsule layer, with a total of 257 capsules; The second layer is the capsule layer with a total of 512 capsules, the third and fourth layers are the capsule layer with 128 capsules; The fifth layer is the dropout layer with a parameter of 0.3. Since sound signals have a strong context relationship, this paper tests the performance of the models which added a layer of LSTM structure to. The last layer of all models is a fully connected layer of size 10, used to map the output to 1-10. The loss function of all models is Cross Entropy, the optimization algorithm is Adam, the learning rate is 0.0001, the batch size is 128, and a total of 50 rounds are trained.

3.3 Analysis of experimental results

Accuracy test was conducted with the trained model and validation set data. The results of each model on validation set were shown in Table 1, and the accuracy of each model on test set was shown in Table 2. The accuracy of the Capsule-GMM-LSTM is higher than that of CNN-LSTM and Capsule-LSTM model.CAPS-GMM-LSTM model has the highest average accuracy on both the verification set and the test set, reaching 94.7% and 92.2%, respectively. The three types with the highest misjudgment rate were "air conditioning sound", "street music" and "children playing", mainly because "air conditioning sound" is similar to white noise, and the model lacks noise removal mechanism, so it is easy to confuse air conditioning sound and white noise. However, the characteristics of "children's play" and "street music" are not so obvious as other types, and the context relationship is more significant. Therefore, the model is required to have a strong ability to deal with the temporal relationship.

Table 1. Accuracy rate of each model on verification set.

Model type	CNN-LSTM	Capsule-LSTM	Capsule-GMM-LSTM
air condition/%	81.4	85.6	**90.7**
horns/%	96.2	97.0	**99.1**
children play/%	75.1	79.5	**89.6**
barking/%	84.5	88.1	**95.3**
drilling/%	93.3	96.7	**99.7**
engine/%	94.6	97.2	**99.8**
gunfire/%	91.4	95.3	**96.8**
jackhammer/%	86.2	90.8	**94.2**
sirens/%	83.9	87.3	**95.5**
street music/%	76.8	80.4	**86.4**
average/%	86.3	89.8	**94.7**

Table 2. Accuracy rate of each model on test set.

Model type	CNN-LSTM	Capsule-LSTM	Capsule-GMM-LSTM
air condition/%	76.3	81.6	**87.4**
horns/%	94.1	96.1	**97.9**
children play/%	69.9	74.3	**87.2**
barking/%	81.5	85.9	**91.7**
drilling/%	90.0	94.4	**98.3**
engine/%	92.7	96.7	**99.1**
gunfire/%	88.6	94.2	**93.2**
jackhammer/%	83.5	88.2	**89.8**
sirens/%	79.4	85.0	**93.7**
street music/%	70.7	77.2	**83.9**
average/%	82.7	87.4	**92.2**

3.4 Summarize

In this paper, a dynamic routing algorithm based on GMM is proposed, and an improved capsule network model is used to classify audio events. In view of the difficulty of traditional neural network in extracting the characteristics of temporal relationship in audio signal, capsule unit is used to represent a single frame, and the improved dynamic routing algorithm is used to cluster the relationship between frame and event. Compared with traditional neural network, the relationship between frame and event is more accurately characterized. Through experimental verification, the accuracy of the model using the

improved algorithm reaches 94.7% on the verification set and 92.2% on the test set, respectively, which proves the effectiveness of the method.

References

1. Sophiya E, Jothilakshmi S. Large scale data based audio scene classification [J]. International Journal of Speech Technology, 2018, 21(4): 825-836.
2. Keren G, Schuller B. Convolutional RNN: An Enhanced model for extracting features from sequential data [A]. 2016 International Joint Conference on Neural Networks [C]. Canada: IEEE, 2016.3412-3419.
3. Chew J, Sun Y, Jayasinghe L, et al. DCASE 2018 Challenge: Solution for task 5 [R]. DCASE2018 Challenge, Tech. Rep, 2018.
4. Sabour S, Frosst N, Hinton G E. Dynamic routing between capsules [A]. Advanced in Neural information processing systems [C]. US:NIPS, 2017.3856-3866.
5. Qihang Zheng, Zhangquan Wang, Banteng L, et al. Research on Family Activity Recognition Method Based on Additive Distance Capsule Network [J].Acta Electronica Sinica,2020,48(8):1580-1586.