Scalable Computing: Practice and Experience, ISSN 1895-1767, http://www.scpe.org
© 2025 SCPE. Volume 26, Issues 2, pp. 619–629, DOI 10.12694/scpe.v26i2.3892

RESEARCH ON INTENTION RECOGNITION METHODS BASED ON DEEP LEARNING

QIANG LI*, FENG ZHAO†, PAN GAO‡, HUANHUAN LI§, AND LINFENG YE¶

Abstract. In order to improve the accuracy of intelligent speech interaction robots, the author proposes a deep learning based intention recognition method for research. By introducing the GloveBibGRU-Self attention classification prediction model, an intention recognition function module is constructed, and the ROS distributed architecture is adopted to integrate the system functional modules, achieving intelligent voice interaction between humans and machines. The simulation results show that the speech intention recognition using the proposed method has higher accuracy. Compared with the intention recognition methods based on DCNN model, CNN-LSTM model, and GRU Self attention model constructed unidirectionally, the recognition accuracy is higher than 8. 02%, 4. 06%, and 2. 13%, respectively, and has better recognition effect, In terms of feature extraction, the training time of BiGRU is shortened by four times compared to traditional extraction methods based on BiLSTM models, resulting in higher training efficiency. According to the experimental findings, the speech interaction system developed utilizing the suggested intention recognition method maintains a high level of accuracy and efficiency in understanding user English speech commands. With an average accuracy rate of 89.72% and recognition times consistently below 0.35 seconds, it is evident that the proposed method is applicable for real-world speech interactions. The intent recognition method based on Glove2BiGRU-Self attention can be applied to English speech interaction in intelligent speech robots.

Key words: Artificial intelligence, Intention recognition, Predictive classification, Deep learning, English voice interaction

1. Introduction. The intersection of human-machine collaboration has emerged as a prominent research focus within the intelligent robotics domain, finding widespread applications across entertainment services, advanced manufacturing, military operations, medical rehabilitation, and beyond. As artificial intelligence technology advances rapidly, it has catalyzed significant transformations in societal dynamics. The seamless integration of AI and robotics technology, aimed at enhancing human-machine synergy, has evolved into a pivotal trend shaping the trajectory of robotics development [1].

Experts and scholars in the field of robotics believe that human-machine collaboration can be seen as a necessary attribute of new industrial robots, and the research and development of collaborative robot technology is the focus of robotics technology. In order to reduce collaboration risks in human-machine collaboration, collaborative robots often use flexible driving mechanisms to improve the performance of contact based human-machine interaction, and the robot body also uses lightweight components. However, although passive flexible structures have certain flexibility and can ensure a certain level of human-machine cooperation safety, they still have a certain degree of stiffness, and robots cannot adjust their own flexibility. In many fields such as entertainment services, advanced manufacturing, military, medical rehabilitation, etc., there is a very high demand for flexibility in contact based human-machine interaction. Only passive flexibility is difficult to meet complex human-machine cooperation tasks [2]. If robots can autonomously adjust flexibility according to task requirements and interaction environments, the performance of human-machine collaboration will be greatly improved. This requires robots to have a certain level of cognitive ability towards the external environment, among which identifying the movement intentions of collaborators is extremely important. We hope that robots have a certain ability to recognize the movement intentions of collaborators. According to task requirements, robots should not only passively follow human movements, but also actively approach human movement intentions, improving the smoothness and comfort of human-computer interaction [3]. Although human-machine

*State Grid Information & Telecommunication Group Co.,Ltd.,Beijing 102211,China (Corresponding author, QiangLi58@c163.com)
†State Grid Information & Telecommunication Group Co.,Ltd.,Beijing 102211,China (FengZhao83@126.com)
‡State Grid Information & Telecommunication Group Co.,Ltd.,Beijing 102211,China (PanGao137@163.com)
§State Grid Information & Telecommunication Group Co.,Ltd.,Beijing 102211,China (HuanhuanLi7@126.com)
¶State Grid Information & Telecommunication Group Co.,Ltd.,Beijing 102211,China (LinfengYe7@163.com)

collaboration technology has received increasing attention from scholars and has achieved certain research results, problems such as difficult recognition of human motion intentions, poor flexibility of robot motion, and low efficiency of human-machine collaboration still exist.

2. Literature Review. The specific content of interaction intention recognition research varies for robots in different fields. For industrial robots, interaction intention recognition focuses on inferring the tasks that people will be performing, such as recognizing the intention of others to grab a certain item or operate the machine. For humanoid robots such as Sophia and Nadine, in order to improve the quality of human-computer interaction, it is necessary for robots to recognize human social intentions. The author focuses on how to enable humanoid robots to recognize human social intentions, namely interactive intentions [4]. According to different feature fusion methods, human-computer interaction intention recognition methods can be divided into rule-based methods and data-driven methods. Lou, H. et al. introduced a novel hybrid model named SACL, which combines self-attention mechanism, convolutional neural network (CNN), and long short-term memory network (LSTM). They began by collecting a dataset of common passenger travel issues in civil aviation airports using web crawlers. After preprocessing the data, they generated a word vector matrix. The model then utilizes a serial structure of CNN and attention mechanism to capture local information of the problem from the word vector matrix. Simultaneously, LSTM captures the global structural information of the text. By concatenating these two feature vectors, the model obtains a comprehensive representation of the text, which is fed into a fully connected neural network and softmax layer for text classification. The SACL model was trained using the Gradient Descent Method (GDM) and compared its performance with four other models. Results indicate that the proposed SACL model effectively identifies passengers' intentions in asking questions, which holds significant implications for enhancing the accuracy and efficiency of answer extraction in civil aviation airport passenger quality assurance systems [5]. Pan, Y. et al. introduced a novel robot teaching system that relies on detecting robot contact states and recognizing human motion intentions. This innovative system can accurately identify the contact status of the robot's hand joints and extract motion intention cues from surface electromyographic signals of the human body to guide the robot's movements. Furthermore, a dedicated module for robot motion mode selection has been developed, enabling the control of the robot's motion along a single axis, in a linear trajectory, or for repositioning purposes. Experimental findings demonstrate that the system effectively facilitates online robot teaching across three distinct motion modes[6]. Wang et al. introduced a deep learning approach to assess the condition of substation switchgear. This method leverages image data captured by the robot's optical camera, employing deep learning algorithms for analysis and detection. Initially, the method curates a dataset comprising images of substation switchgear for model training. Subsequently, utilizing the Yolov3 object detection network, an automatic recognition model is constructed to evaluate the status of substation switchgear. Experimental outcomes indicate that this method achieves an impressive accuracy rate of 9% in automatically identifying the condition of substation switchgear [7].

On the basis of the above research, the author proposes an intention recognition method based on the Glove2BiGRU-Self attention classification prediction model. The model is introduced to construct an intention recognition function module, and through the integration of different functional modules, intelligent human-machine speech interaction of robots is achieved.

3. Research Method.

3.1. Intention recognition. Intent recognition is the process in speech interaction systems where a robot determines the user's intention based on user instructions and provides timely feedback during communication between the user and the machine[8]. It typically includes steps such as preprocessing, text vectorization, feature extraction, and feature classification. The general process of intent recognition is shown in Figure 3.1:

As shown in the figure, intention recognition in robot intelligent speech interaction is a process of constructing a model for classification prediction based on manually selected speech data, feature processing, and then constructing a model for classification prediction. It belongs to a type of classification prediction model. Designers typically use machine learning and deep learning methods to train models. Compared to machine learning, deep learning takes less time and produces better learning results. Therefore, deep learning is adopted as the author's training method.

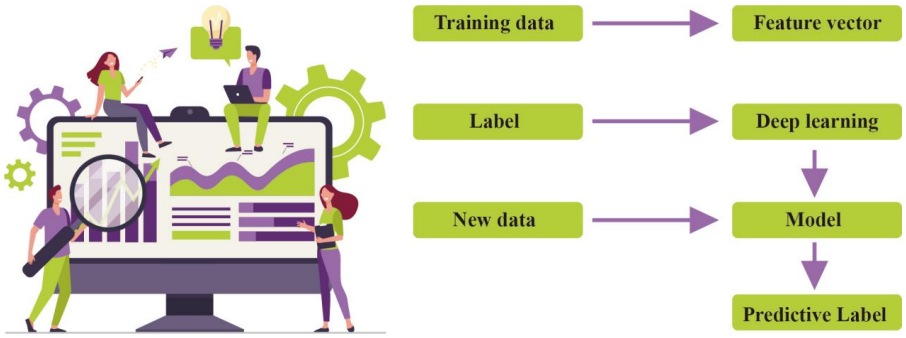

Fig. 3.1: Consciousness recognition processing flow

3.2. Glove_BiGRU_Self-attention model. Traditional intention recognition methods based on template matching or artificial feature sets have high costs and low scalability. In order to enhance the model's understanding performance of user intentions, a user intention classification method based on Glove2BiGRU-Self attention is proposed to help the model learn speech features more fully and quickly. The commonly used deep learning algorithm for consciousness classification model is LSTM, which can solve the long order dependency problem of RNN. However, in computation, due to the design of many variables and parameters, it consumes a considerable amount of time and resources [9]. Therefore, using GRU instead of LSTM, GRU is a variant of LSTM, which reduces one gate compared to LSTM, has a simpler structure, and still has good learning performance, which can significantly save training time and improve model learning efficiency. The user awareness recognition classification model based on Glove-BiGRU-Self attention mainly includes five layers, namely input layer, Glove layer, BIGRU layer, Self attention layer, and Softmax layer.

In the input layer, stuttering segmentation is used to process the user instructions converted into text. For words that are not in the dictionary, an HMM model based on Chinese character word formation ability is used for segmentation. Then, a stuttering segmentation machine is used for segmentation. Finally, all segmentation results are input into the Glove layer for semantic feature representation; Using Glove tool to convert the segmented words in the input layer into word vectors, and utilizing the semantic properties between word vectors to solve the word vectors; In the BiGRU layer, feature extraction is performed on the transformed word vectors. In order to make the information contained in each node more complete, bidirectional GRU is used to model the speech data in both forward and backward directions, making the output of this layer more global; In order to improve the efficiency of feature extraction, Self attention is introduced to calculate different weights in the sample sequence features of the BiGRU model, automatically learning more important semantic features; Input sentence vectors containing word level feature weights into the Softmax layer, and use the Softmax classifier for text classification to obtain the final intent category labels, achieving intent recognition of user instructions [10]. The specific calculations for each step of the GloveBibGRU-Self attention classification model are as follows:

By using a stutterer to segment instruction statements, Glove tool converts each word into a corresponding word vector. Using the Euclidean distance formula to calculate the semantic similarity between two word vectors, the Euclidean distance formula is as follows 3.1:

$$d = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2} \tag{3.1}$$

Construct a co-occurrence matrix X based on the number of times word i and adjacent word j appear together in the context window, X_{ij}, as matrix elements. According to the decay function Decay=1/d, it can be inferred that the distance between two words and the weight of the semantic relationship between words in the total count are inversely proportional. The farther the distance, the smaller the weight [11]. Construct an approximate relationship between word vectors and co-occurrence matrices, and solve word vectors w_i^T and w_j

based on their relationship.

$$w_i^T w_j + b_i + b_j = log(X_{ij}) \tag{3.2}$$

Among them, b_i and b_j are the bias terms of w_i^T and w_j.

Input the obtained word vector into the BiGRU model, set the model update gate and reset gate to z_i and r_i, respectively, and select the input information based on the reset gate. The value calculation of the reset gate at time t is shown in equation 3.3:

$$r_t = \sigma(W_r x_t + U_r h_{t-1}) \tag{3.3}$$

Among them, x_t is the input at time t; W_r represents resetting door permissions; h_{t-1} represents the hidden layer output at time t-1, σ is the Sigmad activation function.

The value of the update gate determines which new information can be retained in the cellular state. Firstly, the update gate forgets and updates the previous and added information, then, the tanh layer creates a new candidate value ht based on the reset gate value, and finally combines the above two parts to achieve the update between the new and old meta cells ht and h_{t-1}. The specific steps are shown in equations 3.4 to 3.6:

$$z_t = \sigma(W_z x_t + U_z h_{t-1}) \tag{3.4}$$

$$h_t = tanh(W_h x_t + r_t * U_h h_{t-1}) \tag{3.5}$$

$$h_t = (1 - z_t) * h_{t-1} + z_t * h_t \tag{3.6}$$

Among them, the weight of the update gate is W_z, z_t is the value of the update gate at time t, and ht-1 is the output of the hidden layer at time t-1; X_t is input at time t.

A unidirectional GRU typically outputs the current state in order from front to back. But the input and output of a moment are often related to the state of both before and after the moment. Therefore, in order to make the node information more complete, a bidirectional GRU is used to establish the association between the node and the time before and after. Using GRU to model from front to back and from back to front respectively, the outputs $\overrightarrow{h_t}$ and $\overleftarrow{h_t}$ of the forward GRU and the backward GRU are calculated as equations 3.7 to 3.8:

$$\overrightarrow{h_t} = \overrightarrow{GRU}(x_i, \overrightarrow{h_{t-1}}) \tag{3.7}$$

$$\overleftarrow{h_t} = \overleftarrow{GRU}(x_i, \overleftarrow{h_{t-1}}) \tag{3.8}$$

The output of BiGRU is determined by two unidirectional GRUs, and the weighted sum is used to calculate the output ht of BiGRU at time t, as shown in equation 3.9:

$$h_t = [\overrightarrow{h_t}; \overleftarrow{h_t}] \tag{3.9}$$

If the frame rate of the speech sample is T, the final output H of the BiGRU layer is as shown in equation 3.10

$$H = [h_1, h_2, h_3, \cdots, h_T] \tag{3.10}$$

Due to the coherence of human language, in intention recognition tasks, nodes need to combine contextual information in order to fully express themselves. The comprehensive context information content BiGRU makes the extracted features of the model at this layer more globally representative, which is more conducive to the subsequent classification and recognition of the model [12].

In order to enable the model to quickly filter out valuable key information from a large amount of speech data and capture more direct semantic relationships, the fully connected output of the BiGRU layer is connected

to the input of the Self attention layer. The speech data is encoded based on the weight calculated by the Self attention layer. By calculating the different weights of each word, the model can focus on more important lexical segments when recognizing a sentence. The use of Self attention as the attention function results in a simpler and more efficient Self attention structure with lower computational costs compared to Attention. Using self attention to associate any two frames in speech through computation can make feature usage more efficient. Assuming the model weight is w_w and the bias is b_w, a multi-layer perception mechanism is used to obtain the hidden representation (u_{it}) of h_{it}, as shown in equation 3.11:

$$u_{it} = tanh(w_w h_{it} + b_w) \tag{3.11}$$

Calculate the similarity between the vector u_{it} and its context vector u_w to measure the importance of the word. Use the softmax function to calculate the normalized weight a_{it}, the a_{it} is related to the input state and uw at each moment, calculated as equation 3.12:

$$a_{it} = \frac{exp(u_{it}^T y_w)}{\sum_t exp(u_{it}^T u_w)} \tag{3.12}$$

u_w can be seen as a semantic representation of input, which is randomly initialized and learned together with the input during the training process. Finally, calculate the sentence vector based on the normalized weight a_{it}, as shown in equation 3.13:

$$S = \sum_{i=1}^{n} a_{it} h_{it} \tag{3.13}$$

The output sentence vector already contains various word level weight information of the input state. Build an intention recognition classifier in the Softmax layer, input the sentence vector output from the previous layer into the classifier for text classification. If the weight of the Softmax classifier is w_f and the bias coefficient is bf, then the probability y of the input x belonging to a certain intention category can be calculated as equation 3.14:

$$y = softmax(w_f s + b_f) \tag{3.14}$$

Among them, $y \in R^k$, k represents the total number of user intent categories.

Using a minimum loss function to update and optimize the parameters in the model, the optimization function is defined as equation 3.15:

$$L(y, p) = \sum_{t=1}^{N} \sum_{j=1}^{C} y_{ij} log(p_{ij}) \tag{3.15}$$

Among them, N and C represent the total number of intent recognition categories in the dataset. a_{it} is the intention category label of the i-th sample in the j-th category [13,14].

3.3. Overall design of intelligent English voice interaction system. The intelligent English voice interaction system refers to a human-computer interaction system that answers user questions or executes related actions based on user English voice commands. Based on understanding user instructions, the system utilizes databases or networks to crawl relevant information and answer user questions. The author adopts the ROS distributed architecture to integrate the functions of the robot intelligent voice system, which includes four main functional modules: voice wake-up, speech recognition, intention recognition, and speech synthesis. The specific architecture is shown in Figure 3.2.

From the framework diagram, it can be seen that the ROS Master is the core of the entire system, and the four modules communicate with the ROS Master in both directions for unified control. The specific workflow of the entire system is as follows: When the user issues a voice command, the microphone is used to collect the voice signal of the command and determine whether it belongs to the device wake-up word. When the instruction

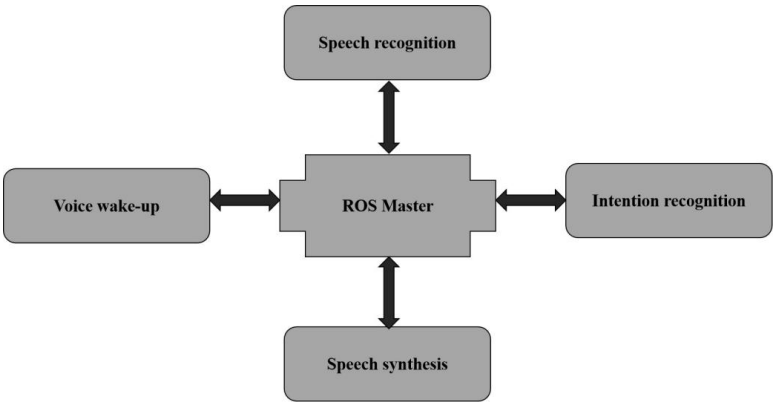

Fig. 3.2: Communication framework diagram of intelligent voice interaction system functional modules

content contains wake-up words, the device enters the working state. When the user raises a question or issues control instructions, the device will match the data information of the local intelligent question answering library through AIML when in offline state, and perform offline command word recognition. If the recognized command word is an input question, the device will search for answers in the question answering library, and the voice service of the voice cloud platform will synthesize the voice service, convert the text information into speech, and then play it; If the recognized command word is a control command, the device will execute the required action based on the command content, achieving offline voice interaction of the device. When the device is in network mode, the search range of voice command recognition will expand to the entire voice cloud platform, which will analyze the user's intentions. If the identified content involves third-party applications, the system will translate it into structured information, send it to relevant applications, process it, and return the processing results to the voice cloud platform in a structured form. If the returned information is judged as a control command, move according to the content of the command; If the returned message is judged as a question, the voice cloud platform will send the returned text reply back to the device, and the voice synthesis service of the voice cloud platform will synthesize the voice and play the result [15].

4. Result analysis.

4.1. Experimental Environment Construction. The experiment is based on ROS integrated intelligent voice interaction robot's various functional modules. ROS is a universal software framework for robots, which combines individually designed functional modules in a loosely coupled manner using a distributed structure. It is often used in the development and application of multi node and multi task robots. Due to the need for ROS to run on Linux, Ubuntu16. 04 in the Linux operating system and the corresponding ROS version Kinetic were chosen as the experimental testing platform, and they were installed and configured separately on the Linux system.

4.2. Data Sources and Preprocessing. The experiment selected the Frames English conversation dataset for intention recognition and training of the BiGRU Self attention model. Frames is a complex artificial dataset for natural language understanding research launched by a deep learning company, Maluuba, in recent years. It is mainly used in areas such as machine reading comprehension, intelligent question answering, and text mining. It contains 19 986 question and answer pairs, mostly based on text recording. 6429 text records were selected as the total sample for the experiment, and the total sample was annotated using five categories: Encyclopedia, chat, Q&A, search, and news. The number of samples in each category was 1089, 1971, 1431, 1018, and 920, respectively. Divide the total sample into three parts: Training set, validation set, and test set according to 7:2:1 [16].

4.3. Evaluation indicators. In this experiment, recognition accuracy was selected as the evaluation index for evaluating the intention recognition effect of classification algorithms, and the calculation method is

Table 4.1: Classification recognition results of models on different categories of file texts (unit:%)

Model	Encyclopedia	Chatting type	Q&A category	Search class	News	Average
DCNN	85.17	77.21	78.42	82.30	81.6	80.85
CNN-LSTM	86.32	82.03	82.35	86.84	86.51	84.84
GRU-Self-attention	88.08	86.35	83.03	87.66	88.70	86.74
BiGRU-Self-attention	89.64	89.03	86.31	88.60	90.81	88.88

shown in equation 4.1:

$$Accurary = \frac{TP + TN}{TP + FP + TN + FN} \tag{4.1}$$

Among them, TP represents the number of times the model identifies correct results as "correct"; FP represents the number of times the model identifies correct results as "errors". TN represents the number of times the model identifies incorrect results as "correct"; FN represents the number of times the model identifies incorrect results as "errors".

4.4. Parameter settings. This experiment sets the Epoch value of the BiGRU Self attention model to 10 and the Attention_size to 128. Using Glove for word vector initialization, with a word vector dimension of 200; Using BiGRU for feature extraction with 100 hidden units; Using adaptive Adam algorithm for training optimization, the learning rate and learning rate decay rate are set to 0.001 and 0.0001.

4.5. Analysis of experimental results.

4.5.1. Verification of intent recognition classification model. In order to verify the intention recognition performance of the BiGRUSelf attention model used in intelligent speech interaction robots, experiments were conducted to compare the classification accuracy of classification methods used in other speech recognition studies such as DCNN, CNN-LSTM, GRU Self attention, etc. on the dataset. The recognition results of each model on different categories of file texts are shown in Table 4.1.

As shown in the table, BiGRU Self attention consistently has the highest recognition accuracy on various categories of file texts, with an average recognition accuracy of 88.99%, better recognition and classification performance [17]. Compared to traditional recognition methods such as DCNN and CNN-LSTM, the recognition results of BiGRU Self attention have been significantly improved, with significant improvement effects; Compared to GRU Self attention using only unidirectional GRU for feature extraction, BiGRU Self attention using bidirectional GRU contains more information, has a more comprehensive understanding of the context, and further improves recognition accuracy by an average of about 2. 03%, resulting in superior model performance.

Feature extraction is an important step in intent recognition. In order to verify the improvement of the BiGRU Self attention model in terms of training speed, the same 1000 data points were selected from the dataset to train BiGRU and BiLSTM feature extraction from commonly used intention recognition classification models. The training time for both is shown in Figure 4.1.

As shown in the Figure, under the same training data conditions, BiGRU has a faster training speed, a training time shortened by about 4 times compared to BiLSTM, and a higher training efficiency. It can be inferred that compared to intention recognition models based on Glove2BiLSTM-Self attention, Glove2BiGRU-Self attention should also have faster training speed and more obvious advantages.

4.5.2. Verification of Speech Interaction System Based on Intent Recognition. In order to verify the effectiveness of the author's design intention recognition method in the intelligent voice interaction system of robots, 500 user interaction statements of different categories were selected from the dataset with 100 entries per category, and intention recognition tests were conducted on intelligent voice interaction robots equipped with intention recognition modules[18]. The test is divided into three groups, and the recognition results of the robot are counted according to different categories. The average of the three tests is taken as the final value of the test. The measurement indicators for the effectiveness of robot intent recognition include recognition accuracy and recognition time. The recognition accuracy of the three tests is shown in Table 4.2.

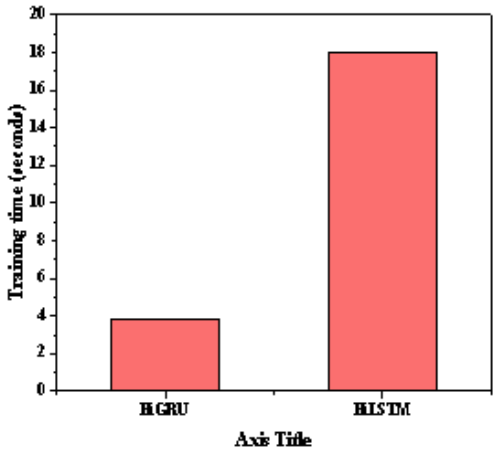

Fig. 4.1: Comparison of model training speed (Unit: seconds)

Table 4.2: Display of Test Results for Intelligent Voice Interaction Robot Intent Recognition Module (Unit:%)

Category	Test 1	Test 2	Test 3	average
Encyclopedia	91.55	93.15	88.15	91.05
Chatting type	88.14	86.51	93.17	89.23
Q&A category	86.57	89.84	88.11	88.17
Search class	91.52	86.54	85.01	87.65
News	91. 53	93.16	91.50	92.06

From Table 4.2, it can be seen that the robot's intention recognition accuracy for five types of interaction statements in three tests is above 88%, and the recognition results are stable and high. The average recognition accuracy is similar to the test results of Glove2BiGRU-Self attention, proving that the intention recognition model proposed by the author can be successfully applied to the intelligent speech interaction system of the robot.

On the basis of the above intent recognition results, remove the parts that were identified incorrectly, and then conduct a recognition time test on the parts that were identified correctly. Define the end time of user command issuance as the initial time, and record the completion time of robot intent recognition one by one. The number of tests is also divided into three, and the data categories involved in the test include five categories. The average time of three tests for each category of data is used as the test result. The time required for robots to complete correct intent recognition under different types of data is shown in Figure 4.2.

From Figure 4.2, it can be seen that the speech interaction robot loaded with Glove2BiGRU-Self attention not only has a high recognition accuracy but also has a fast recognition speed for the intention recognition of five categories of user interaction statements, with recognition time all within 0.34 seconds. Glove2BiGRU-Self attention still has high performance on robot humans, ensuring that the robot can accurately and quickly interact with user speech intelligently.

4.5.3. Practical application verification of the proposed method.

(1) Speech instruction recognition. In order to verify the implementation of an intelligent speech interaction system based on intent recognition in robots, the operation of the intent recognition module in the speech interaction system was tested using the voice command of "The weather in Yibin today" as an example. After the voice recognition node recognizes the voice command "The weather in Yibintoday" sent by the user and

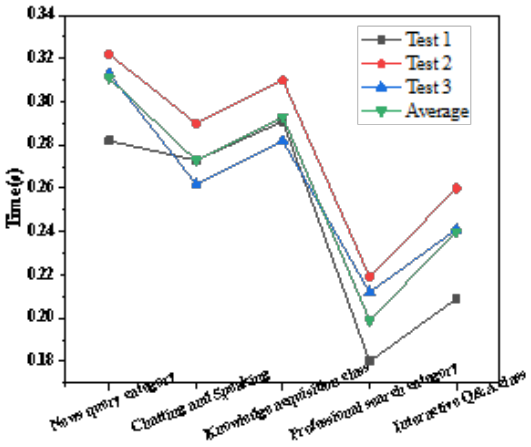

Fig. 4.2: The time required for intelligent voice robots to correctly complete intent recognition (Unit: seconds)

Table 4.3: Intelligent Voice Interaction Results of Robots under Five Categories of Communication Conversations

Category	Interrogative sentence	Answer	Result determination
Chatting type	What's the scenery like in Yibin	I heard Yibin is a very nice place	Meet expectations
Q&A category	It's bargain time	March 29, 2023 at 5:10 p.m	Meet expectations
News	What is the score of the Warriors today	109∼120	Meet expectations

publishes it as text, it automatically subscribes to the command text, searches for answers through the Internet, and publishes the search results as text of "Yibin: Wednesday, March 29, cloud to light rain southwest wind, the lowest temperature 12 degrees", and then performs text to speech conversion by the voice synthesis node, and finally plays them through robot voice. According to the test results, it can be concluded that intelligent voice robots can accurately recognize the intentions of user voice commands. In order to further verify the recognition performance of robots in human-machine voice interaction for user intentions, a multi category topic dialogue and communication were conducted with robots through voice input. In the five categories of dialogue communication, each representative is selected, and the results of robot intelligent voice interaction are shown in Table 4.3.

From the Table 4.3, it can be seen that the robot's response content is consistent with the user's voice input intention, which proves that the robot can perform intention recognition through an intelligent voice interaction system equipped with the Glove2BiGRU-Self attention intention recognition module, and search and output results based on the user's instruction intention.

(2) Mobile instruction recognition. In order to further verify the application effect of the design intent recognition method in intelligent voice interactive robots, the robot is controlled for movement through voice commands. Custom dialogue construction is carried out in the intent recognition module, and the constructed mobile command information includes the following content:

Instruction 1:"move_go_forward"
Instruction 2:"move_rotate_left"
Instruction 3:"move_rotate_right"
Instruction 4:"move_go_back"

The program matches keywords based on the text output of voice commands. When the flag is 0, it

indicates the inclusion of any keyword, indicating that the program matches successfully. The robot will move based on the successfully matched instructions. According to the command "forward left right backward", the little turtle made corresponding movements, proving that the intelligent voice interaction system built using the proposed Glove2BiGRU-Selfattention intention recognition method can control the robot's actions by matching the command content, and the movement results meet the requirements of voice commands.

5. Conclusion. In summary, the designed intelligent speech interaction robot based on intention recognition is constructed by introducing the Glove2BiGRU-Self attention classification prediction model, completing the construction of the system intention recognition function module, and integrating various functional modules of the system using ROS distributed architecture to achieve intelligent speech interaction between users and robots. The results show that compared to the intent recognition methods based on DCNN model, CNN-LSTM model, and unidirectional GRU Self attention model, the proposed intent recognition method based on GloveBibGRU-Self attention model significantly improves the recognition accuracy, reaching 88.88%. In addition to accuracy, the recognition efficiency of the model has also improved. In terms of feature extraction, compared to traditional BiLSTM based feature extraction methods, the BiGRU feature extraction method reduces training time by 4 times and has higher training efficiency. When applied to actual robot voice interaction systems, it still has high performance, achieving rapid interaction between human-machine speech. Through research, using GRU instead of traditional LSTM for speech feature classification simplifies the model structure, improves model training efficiency, and through the bidirectional modeling of GRU, the feature nodes contain richer information, resulting in higher generalization of extracted features. This enables the model to learn features more effectively, improve classification accuracy, and enable robots to accurately and quickly respond to user speech instructions. However, there are still shortcomings in this study. With the increase of training data, the classification and recognition time of the model also increases, and the recognition accuracy correspondingly decreases. Further research and improvement are needed to ensure that the model still performs well in more diverse and flexible speech instructions in reality.

REFERENCES

[1] Xing, Y., Lv, C., Wang, H., Cao, D., & Velenis, E. (2020). An ensemble deep learning approach for driver lane change intention inference. Transportation Research Part C: Emerging Technologies, 115, 102615.
[2] Wu, Z., Chen, Y., Zhao, B., Kang, X., & Ding, Y. (2021). Review of weed detection methods based on computer vision. Sensors, 21(11), 3647.
[3] Pérez-Hernández, F., Tabik, S., Lamas, A., Olmos, R., Fujita, H., & Herrera, F. (2020). Object detection binary classifiers methodology based on deep learning to identify small objects handled similarly: Application in video surveillance. Knowledge-Based Systems, 194, 105590.
[4] Xiong, D., Zhang, D., Zhao, X., & Zhao, Y. (2021). Deep learning for EMG-based human-machine interaction: A review. IEEE/CAA Journal of Automatica Sinica, 8(3), 512-533.
[5] Lou, H., Zhao, H., & Deng, W. . (2022). Research on civil aviation passenger question intention recognition based on text classification method of self-attention and deep neural network, 22(7), 4337-4347.
[6] Li, W., Shi, P., & Yu, H. (2021). Gesture recognition using surface electromyography and deep learning for prostheses hand: State-of-the-art, challenges, and future. Frontiers in neuroscience, 15, 621885.
[7] Wang, L., Kou, Q., Zeng, Q., Ji, Z., Zhou, L., & Zhou, S. . (2022). Substation switching device identification method based on deep learning. 2022 4th International Conference on Data-driven Optimization of Complex Systems (DOCS), 1-6.
[8] Li, G., Liu, F., Sharma, A., Khalaf, O. I., Alotaibi, Y., Alsufyani, A., & Alghamdi, S. (2021). Research on the natural language recognition method based on cluster analysis using neural network. Mathematical Problems in Engineering, 2021, 1-13.
[9] Qiu, S., Zhao, H., Jiang, N., Wang, Z., Liu, L., An, Y., ... & Fortino, G. (2022). Multi-sensor information fusion based on machine learning for real applications in human activity recognition: State-of-the-art and research challenges. Information Fusion, 80, 241-265.
[10] Zang, H., Cheng, L., Ding, T., Cheung, K. W., Wei, Z., & Sun, G. (2020). Day-ahead photovoltaic power forecasting approach based on deep convolutional neural networks and meta learning. International Journal of Electrical Power & Energy Systems, 118, 105790.
[11] Hassouneh, A., Mutawa, A. M., & Murugappan, M. (2020). Development of a real-time emotion recognition system using facial expressions and EEG based on machine learning and deep neural network methods. Informatics in Medicine Unlocked, 20, 100372.
[12] Liu, J., & Wang, X. (2021). Plant diseases and pests detection based on deep learning: a review. Plant Methods, 17, 1-18.
[13] Wan, S., Qi, L., Xu, X., Tong, C., & Gu, Z. (2020). Deep learning models for real-time human activity recognition with smartphones. Mobile Networks and Applications, 25(2), 743-755.

[14] Guo, L., Lu, Z., & Yao, L. (2021). Human-machine interaction sensing technology based on hand gesture recognition: A review. IEEE Transactions on Human-Machine Systems, 51(4), 300-309.
[15] Mujahid, A., Awan, M. J., Yasin, A., Mohammed, M. A., Damaševičius, R., Maskeliūnas, R., & Abdulkareem, K. H. (2021). Real-time hand gesture recognition based on deep learning YOLOv3 model. Applied Sciences, 11(9), 4164.
[16] Teng, F., Song, Y., & Guo, X. (2021). Attention-tcn-bigru: an air target combat intention recognition model. Mathematics, 9(19), 2412.
[17] Guo, L., Lu, Z., & Yao, L. (2021). Human-machine interaction sensing technology based on hand gesture recognition: A review. IEEE Transactions on Human-Machine Systems, 51(4), 300-309.
[18] Buerkle, A., Eaton, W., Lohse, N., Bamber, T., & Ferreira, P. (2021). EEG based arm movement intention recognition towards enhanced safety in symbiotic Human-Robot Collaboration. Robotics and Computer-Integrated Manufacturing, 70, 102137.

Edited by: Bradha Madhavan
Special issue on: High-performance Computing Algorithms for Material Sciences
Received: May 10, 2024
Accepted: Jun 25, 2024