


## Article

# Multi-Sensor Context-Aware Based Chatbot Model: An Application of Humanoid Companion Robot

Ping-Huan Kuo <sup>1,2</sup> , Ssu-Ting Lin <sup>3</sup>, Jun Hu <sup>3</sup> and Chiou-Jye Huang <sup>4,\*</sup>

<sup>1</sup> Department of Mechanical Engineering, National Chung Cheng University, Chiayi 62102, Taiwan; phkuo@ccu.edu.tw

<sup>2</sup> Advanced Institute of Manufacturing with High-Tech Innovations (AIM-HI), National Chung Cheng University, Chiayi 62102, Taiwan

<sup>3</sup> Department of Intelligent Robotics, National Pingtung University, Pingtung 90004, Taiwan; cbc105010@gmail.com (S.-T.L.); wuorsut@gmail.com (J.H.)

<sup>4</sup> Department of Data Science and Big Data Analytics, Providence University, Taichung 43301, Taiwan

\* Correspondence: cjh1007@gm.pu.edu.tw; Tel.: +886-4-2632-8001 (ext. 15124)

**Abstract:** In aspect of the natural language processing field, previous studies have generally analyzed sound signals and provided related responses. However, in various conversation scenarios, image information is still vital. Without the image information, misunderstanding may occur, and lead to wrong responses. In order to address this problem, this study proposes a recurrent neural network (RNNs) based multi-sensor context-aware chatbot technology. The proposed chatbot model incorporates image information with sound signals and gives appropriate responses to the user. In order to improve the performance of the proposed model, the long short-term memory (LSTM) structure is replaced by gated recurrent unit (GRU). Moreover, a VGG16 model is also chosen for a feature extractor for the image information. The experimental results demonstrate that the integrative technology of sound and image information, which are obtained by the image sensor and sound sensor in a companion robot, is helpful for the chatbot model proposed in this study. The feasibility of the proposed technology was also confirmed in the experiment.

**Keywords:** multi-sensor fusion; natural language; processing; context-aware computing; chatbot; companion robot



**Citation:** Kuo, P.-H.; Lin, S.-T.; Hu, J.; Huang, C.-J. Multi-Sensor Context-Aware Based Chatbot Model: An Application of Humanoid Companion Robot. *Sensors* **2021**, *21*, 5132. <https://doi.org/10.3390/s21155132>

Academic Editor: Eui Chul Lee

Received: 15 June 2021  
Accepted: 27 July 2021  
Published: 29 July 2021

**Publisher's Note:** MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.



**Copyright:** © 2021 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

## 1. Introduction

Many studies have proposed robots that can chat with people. Opportunities to use chatbots will continue to increase. In addition to accompanying older adults living alone, chatbots are also applied to medical treatment. For example, individuals may be afraid of public spaces due to mental illnesses. Under these circumstances, chatbots can be used to provide them with companion and health care. Numerous studies of chatbots have been reported. Traditional chatbots only use the before and after sentences for training. However, such training results in low diversity, and accuracy is also low. This traditional system also renders chatbots unable to flexibly answer questions. For example, when being asked “What do you see?” traditional chatbots can only answer “I don’t know” due to the lack of image information. The aforementioned problem is only one example; numerous others are similar.

Numerous studies regarding chatbots have been published. Ref. [1] proposed and applied a hybrid chatbot to communication software. This technology combines a retrieval-based model and a QANet model and can answer user questions through an E-Learning system. In addition to supporting English, this technology can also converse with chatbots in Chinese and has performed well in experiments. Ref. [2] also applied chatbots in digital learning. In this study, the role of the proposed chatbot is to resolve Visual, Auditory, Read, Write, and Kinesthetic (VARK) information throughout the process of

responding to the learner. Furthermore, the proposed chatbot also records the learner's beta brainwaves and uses machine learning algorithms to classify learning orientations. Ref. [3] explored many studies on chatbots used in Facebook Messenger, mainly discussing how teachers use chatbots to supplement their teaching. The article also stated that the use of chatbots to assist in teaching is not yet mature and, therefore, chatbots as a field of research offer considerable room for development. Ref. [4] proposed a neural network-based chatbot that integrates personal data for analysis and estimation and uses the useful data to return appropriate responses. The main architecture for this technology is LSTM [5], and experiment results have substantiated the feasibility of this technology. In Ref. [6], bidirectional recurrent neural network (BiRNN) technology was used as the basis for a chatbot, and a GRU [7] was used as the core of the technology. The system architecture was exceedingly large and required substantial computing resources, but was verified favorably in experiments. Ref. [8] proposed a chatbot called Xatkit and stated that although it requires improvement, it currently can be applied in some real-world scenarios, and could be applied in many venues in the future.

Chatbots have diverse applications. To address the massive demand for manpower in customer service, some companies have proposed automated customer service chatbots. However, automated customer service chatbots sometimes misunderstand sentences. Ref. [9] presented research into automated customer service chatbots and stated that although chatbot technology is not yet fully mature, demand for this technology is notably high. Moreover, cognitive behavioral therapy (CBT) is a common method of therapy for patients with panic disorder. Ref. [10] explored the feasibility of using chatbots on mobile devices for CBT. The study found that using chatbots for CBT was both feasible and effective, pioneering the application of chatbots in medicine.

Deep learning as a field of research has burgeoned in the early 21st century. Its technological developments are diverse and have been applied to many research topics [11–15]. The present paper features deep learning techniques to resolve questions that chatbots cannot answer due to environmental contexts. The first step toward this type of practical learning is to enable bots to understand the present environment through visual images and to describe the environment through text. The most direct method is to design bots that use image caption technologies to understand their environments. The original purpose of image captions was, and is, to automatically generate annotations for an image. For example, if an image is entered as input, the image caption model can write a sentence about the image based on the image status. Therefore, in this article, the described method enables the bot to view its present environment clearly and to know about any object in its environment. Conventional chatbot models typically are only able to perform model training based on context; this type of model training is unable to consider image data. Even if the model can understand its present environment, it cannot incorporate the understood image data into dialogue responses. In this study, the proposed method can deepen the chatbot model's considerations of the results of its previous conversations, and together with the current dialogue, can train the chatbot. A statement generated by image caption technology can also be an input for the chatbot and can be incorporated into the next dialogue. This method enables the chatbot to answer questions more accurately; the highly diverse responses vary based on the present environment.

The four major contributions of this work are: (1) integrating image and sound data to achieve further precision and practicality in chatbot and human communications; (2) adjusting and optimizing a framework based on a commonly used LSTM model to train the image caption and chatbot model efficiently, thus improving scenario recognition abilities; (3) comparing the original LSTM-based model with the adjusted, GRU-based version and verifying that the training efficiency was improved by the adjustments; and (4) testing the trained model on a humanoid robot in a series of experiments to verify that the other model proposed in this paper and the applications of both models are effective and feasible.

The remaining sections of this paper are as follows. Section 2 details the operational concepts and process of this paper. Section 3 introduces the models and algorithmic framework being proposed. Section 4 presents and compares the experiment results. The discussions are addressed in Section 5. Section 6 is the conclusion.

## 2. System Architecture

In the past, chatting with bots was typically only possible through direct voice or text conversations using conventional chatbot models. A typical bot was unable to consider its surrounding environment through its vision when responding. The present research relies on chat training with images that the bot can see. To allow the chatbot models to integrate image data into input sets, this paper proposes an improved model that integrates image captions, natural language processing, and similar technologies; this model enables the bot to chat and respond according to the environment it sees. The feasibility and practicality of the model were substantiated by the experiment results. In the future, this method can be applied to chatbots and home care, as well as to increase the precision of typical chatbot's verbal expressions and to enable typical chatbots more aware of reality when conversing with humans.

The overview of the proposed framework is illustrated in Figure 1. Prior to undergoing any training, the chatbot can be said to be like a baby—incapable of anything. But after going through correct learning, the chatbot slowly becomes more human-like and gains the ability to communicate with people. The training process resembles the process by which adults teach children to recognize objects, typically holding a picture and telling the child, “This is a dog”. Although the ability to express oneself in words may not be entirely present during the initial training, the rate of failure can be reduced through continuous training. Therefore, we can use this process to enable robots to develop cognitive abilities regarding their surroundings. This process is the same as teaching children to speak, that is, developing their ability to speak through continuous conversations and correcting their mistakes. Teaching robots is the same—although the bot may answer incorrectly in the beginning, through continuous learning, they can learn to have normal conversations with people. The concept map of this paper is presented in Figure 2; the chatbot can perform analyses based on current image data. When the user asks questions, not only is the chatbot able to make considerations based on the environmental data it currently detects, the bot is also able to generate a corresponding response based on the conversation context. The benefit of this method is in the integration of multiple pieces of data, allowing the chatbot to better make responses focusing on relevant questions. In this manner, the bot's responses become more precise and more in line with reality.

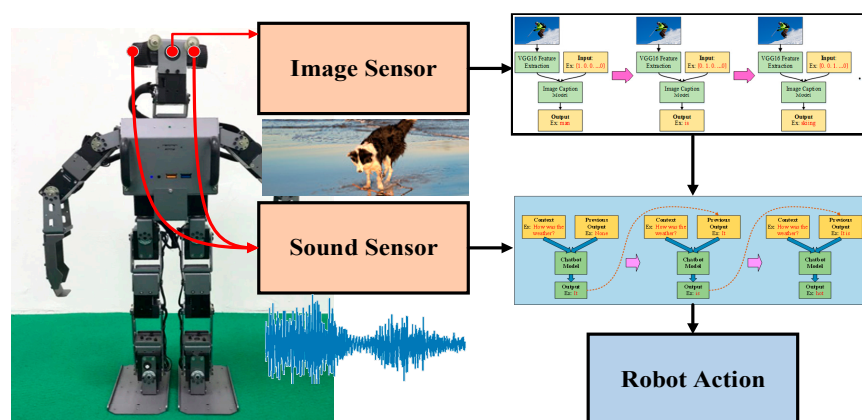


Figure 1. An overview of the proposed framework.

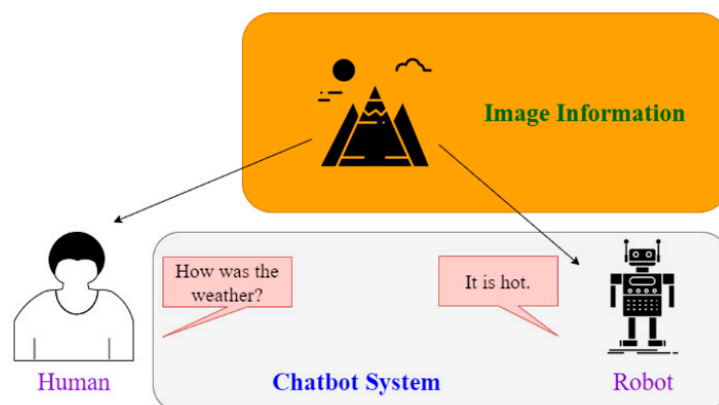


Figure 2. The concept of the proposed framework.

This paper can be split into two major areas. The first is picture scenario analysis, and the second is natural language processing. The system framework is presented in Figure 3. When the entire system is operating, we consider two inputs: image data and voice inputs. For the image inputs, first, we perform feature extraction through the VGG model. This reduces the dimensionality of the image data, avoiding training problems resulting from overly large input dimensions. Following the feature extraction, we then use the image caption model to generate a paragraph of text that matches the image data. For the voice inputs, we must first convert the sound signals into text messages through voice-to-text conversion technologies. Then the chatbot model must generate appropriate responses based on text messages converted from the user's voice and text information about the pictures. In this context, the output data type is text. Therefore, for the user to be able to hear the chatbot's response, we use text-to-voice conversion technologies to convert the text message into sounds, either to send to the user or to enable the bot to perform a corresponding action. In this manner, the whole system framework not only considers signals from sound, it also incorporates considerations of images. Therefore, the proposed method improves chatbot affinity by enabling chatbots to generate responses to be more in line with reality and to interact better with humans.

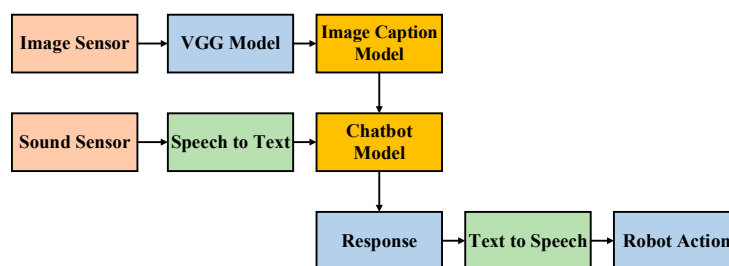
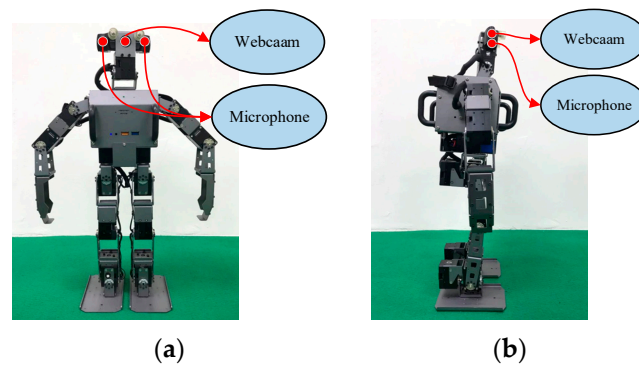


Figure 3. The architecture of the system.

To validate the method proposed in this paper, we applied the trained models with a humanoid robot, then verified the efficacy of the method through a series of experiments. Humanoid robots are the closest robot type to humans and are therefore suitable for chatting and companionship purposes; therefore, a humanoid robot was chosen for the experiments in this paper. The humanoid robot used in this paper is presented in Figure 4; the humanoid robot is 51 cm tall and weighs approximately 3.5 kg. It has 20 degrees of freedom, and has an inbuilt Logitech C920 HD Pro Webcam and microphone. The robot is powered by a XM430 motor and installed with an Intel Core i3 processor and OpenCR controller. Its Inertial Measurement Unit involves a three-axis gyroscope, a three-axis acceleration sensor, and a three-axis magnetometer. The operating system is Ubuntu.



**Figure 4.** The architecture of the humanoid robot: (a) the front view; and (b) the side view.

### 3. The Proposed Model

This section introduces the framework of the proposed model in detail. Because most of the data processed in this work was time-series data, RNNs constituted an appropriate framework. The most well-known RNN framework is the commonly used LSTM. The literature currently features many LSTM variations, such as the GRU framework. However, the GRU framework is simpler than the LSTM framework, but in actual operations the GRU can achieve results that are equal to (if not better than) the results of LSTM. Also, because the GRU framework is simpler, under normal conditions GRU has better training efficiency than LSTM. To compare the two RNN models, we must modify and enhance the original LSTM RNN in the hopes of achieving better convergence efficiency than the original neural network.

Many studies have performed in-depth explorations of chatbots, and this paper takes the framework from reference [16] as the basis for improvements. In the original version, an adversarial learning method for generative conversational agents (GCA) is applied for the chatbot model. Such a learning process originally also belonged to one of the sequence-to-sequence training regimes. However, such an approach can also transfer to an end-to-end training process by applying the specific structure. The chatbot model can generate an appropriate sentence by considering the dialogue history. The main core of GCA is LSTM, and the model also considers the current answer and the previous conversation history. The concept for this paper also was implemented with favorable results in the Keras framework [17], as presented in [18]. The Keras framework is a deep learning library which includes many Application Programming Interfaces (APIs) for building the corresponding models. Keras is also one of the most used deep learning frameworks. In [18], Keras is used for the implementation of the chatbot model. Based on the previous work in [18], this study improves the model structure and gives better performance than the original version. The detailed framework of the chatbot is exhibited in Figure 5; Figure 5a presents the original LSTM version [18]; Figure 5b displays the GRU version that was optimized in this study. In addition to the LSTM-to-GRU transformation, the differences between the two versions include the addition of two dense (fully connected) layers to the new version. The parameters are listed in Table 1. We also added a dropout to prevent overfitting and used Adam for the optimizer and categorical cross entropy for the generated loss function. A more detailed comparison of the two versions is laid out in the next section.

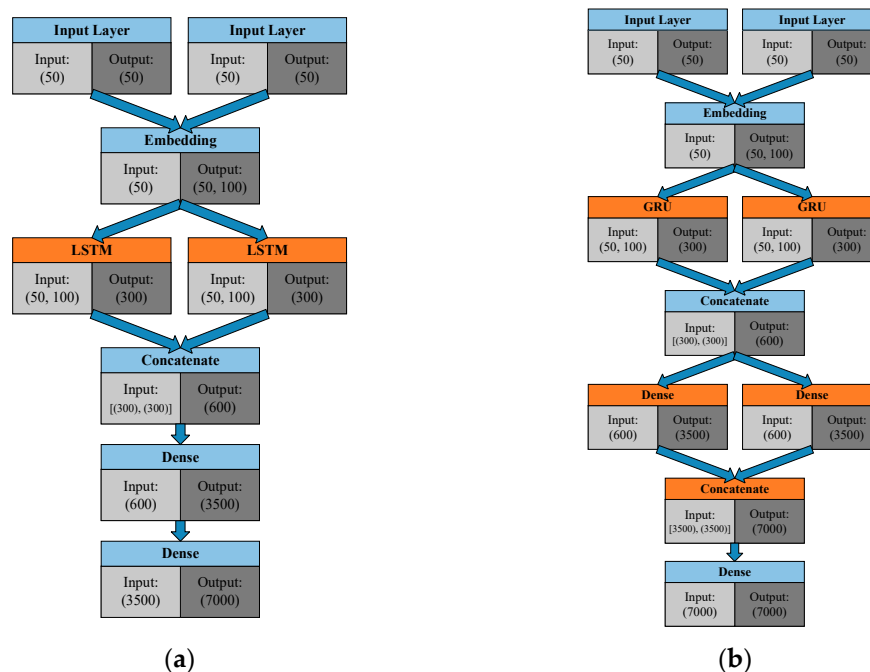


Figure 5. Architecture of the chatbot: (a) LSTM version; and (b) Enhanced GRU version.

Table 1. Parameters setting of the chatbot.

Parameter	Value
Dropout Rate	0.25
GRU Cell	300
Optimizer	Adam
Loss Function	Categorical Cross Entropy

The chatbot model proposed in this study was based mainly on the concept of a generative conversational agent (GCA) [16]. Records of a conversation with a chatbot can be stored using the  $x$  vector and one-hot encoding. Accordingly, the following equations are defined:

$$X = [\bar{x}_1, \bar{x}_2, \bar{x}_3 \dots \bar{x}_{S_s}] \quad (1)$$

$$Y = [\bar{y}_1, \bar{y}_2, \bar{y}_3 \dots \bar{y}_{S_s}] \quad (2)$$

$$E_c = W_e X \quad (3)$$

$$E_a = W_e Y \quad (4)$$

where  $W_e$  represents the embedding layer parameter,  $E_c$  represents the corresponding conversation record, and  $E_a$  represents the unfinished response. An RNN is then used to extract the embedding vectors of the conversation record and response, as illustrated in the following equation:

$$e_c = \Gamma_c(E_c; W_c) \quad (5)$$

$$e_a = \Gamma_a(E_a; W_a) \quad (6)$$

$$e = [e_c \ e_a] \quad (7)$$

$$y_h = \sigma(W_1 e + b_1) \quad (8)$$

$$P = \varphi(W_2 y_h + b_2) \quad (9)$$

where  $W_1$  and  $W_2$  are synaptic weights,  $b_1$  and  $b_2$  are biases,  $\sigma$  is the Rectified Linear Unit (ReLU), and  $\varphi$  is the Softmax activation function. Finally, the index with the largest

probability value is identified from  $P$  before the next iteration is started. Accordingly, the most suitable response can be output.

The operational method of the chatbot used in this paper is displayed in Figure 6. The figure displays the two input sources of the chatbot. The left-hand input source is the previous sentence spoken by the other party, and the right-hand input source is a collection of previous output results. Because of the unique advancements of the proposed training, the chatbot model can notably understand and analyze the contextual relationships of the sentences. Finally, we export the generated dialogue word-by-word in sequence, then assemble the words into a complete sentence. When the chatbot expresses the output results for the user to hear, the user may respond a second time based on the answer just received. Next, in addition to the second response from the user, the chatbot's own previous response is added to the chatbot's left-hand input. This additional input is exceedingly helpful to the chatbot's ability to understand the contextual meaning. Furthermore, this additional input can be substituted for the output text produced by the image data. This allows the chatbot to become familiar with the present image data, enabling the chatbot to assess the image content and generate the best response. Through repeated training, the chatbot model can fully grasp the relationship of the conversation with the user; in this manner, the chatbot can become more fluent in conversations with people, narrowing the gap between chatbots' and humans' communication skills.

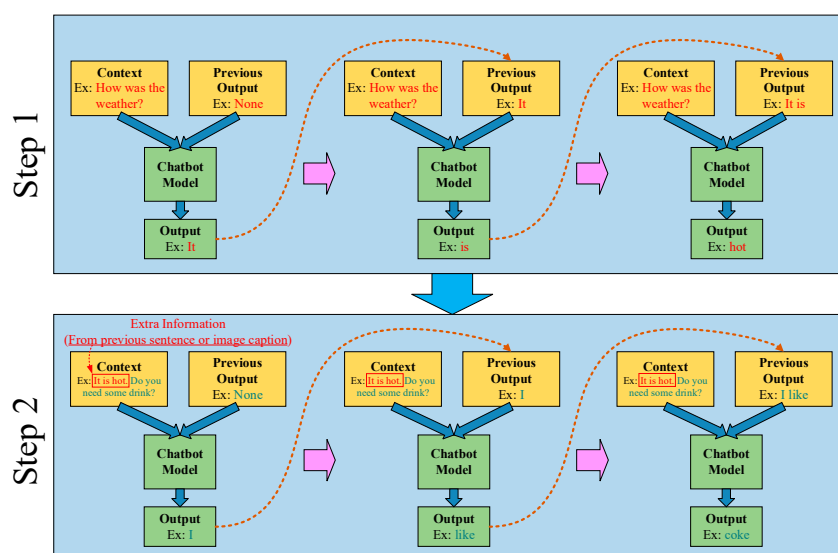
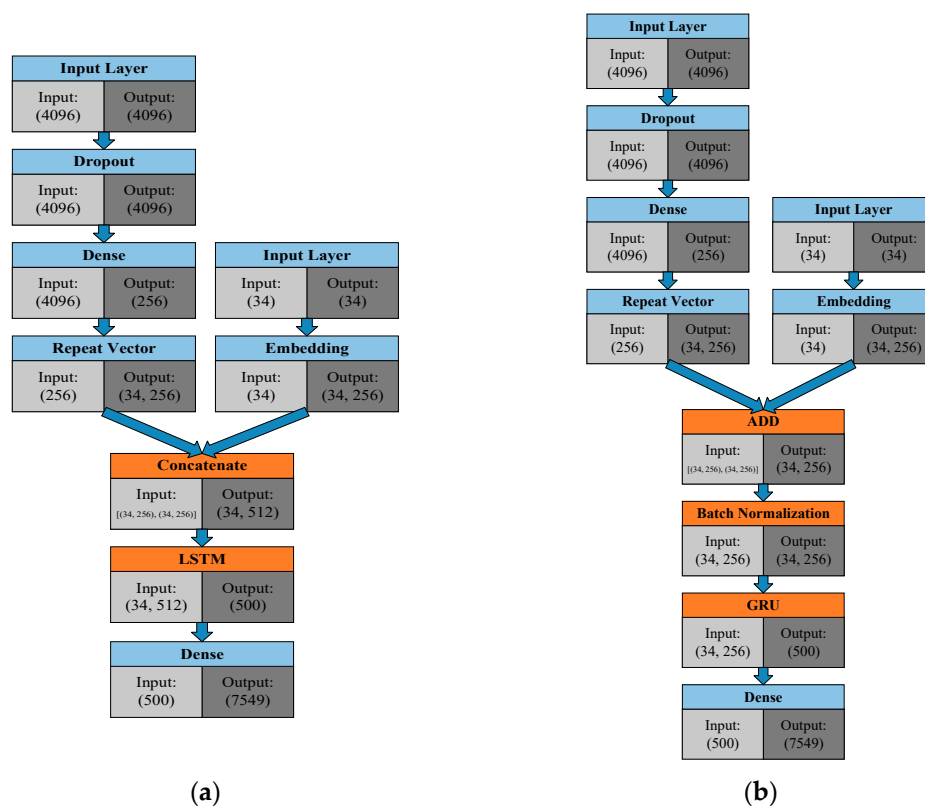


Figure 6. The workflow of the designed chatbot model.

Image caption technology has been a topic of in-depth exploration in many papers, and this paper builds on the framework in references [19,20]. The concept discussed in this paper was also implemented with favorable results in the Keras framework [17], as demonstrated in [21]. A detailed framework of our image caption model is presented in Figure 7, with Figure 7a as the original LSTM version [21] and Figure 7b as the optimized GRU version used in this paper. The original LSTM framework was exchanged for a GRU framework, to reduce the number of training parameters; the concatenation layer that had originally been in the middle was implemented through a tensor adding method. The improved approach, in addition to being able to consider input data from both sides, can also eliminate half of the parameters for this layer and improve training efficiency. Furthermore, batch normalization (BN) techniques were added to the model; BN can adaptively adjust the value range of the input data, reducing situations in which the activation function cannot be calculated as a result of the value of the previous layer being too great or too small. The parameters of the image caption model discussed in this paper can be found in Table 2. We also added dropout to prevent overfitting; we chose Adam for

the optimizer and categorical cross entropy for the output loss function. The next section compares these two versions in greater detail.



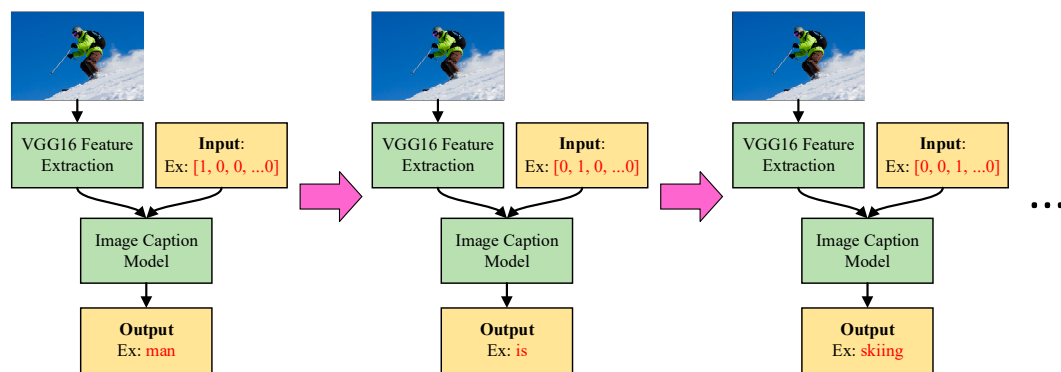
**Figure 7.** Architecture of the image caption: (a) LSTM version; and (b) Enhanced GRU version.

**Table 2.** Parameter settings of Image caption.

Parameter	Value
Dropout Rate	0.5
GRU Cell	500
Optimizer	Adam
Loss Function	Categorical Cross Entropy

Figure 8 presents the operations of the image caption model; the figure demonstrates that the image caption model can also be split into left-hand and right-hand inputs. The left-hand inputs are image-related; after an image is input into the model, we must first use VGG16 [22] to perform the task of feature extraction. The VGG16 model used in this paper is a complete model that has already been trained using a large volume of picture data. The final output layer of the original was not necessary for this work; but because this model had already been trained, the final output layer was eliminated, and the other parts were treated as excellent feature extraction tools. If we input an entire picture into the image caption model from the beginning, this may result in training issues from overly large input data. The use of the feature extraction method reduces the input dimensions and can present the picture's data in the most streamlined and adequate manner as vectors. The image caption model's right-hand inputs are a series of one-hot vectors that represent the current sequence of the output text. After repeated iterative processes, the image caption model describes the image in text by generating the text word-by-word to present a complete sentence.





**Figure 8.** Workflow of the designed Image caption model.

The model training process discussed in this paper is presented in Figure 9. The detailed pseudocode is listed in Algorithm 1. The details are addressed in Appendix A.

---

**Algorithm 1: The procedure of the proposed system.**

---

1. Data preprocessing (image caption)
  2. Image caption model initialization
  3. **for** all training epochs
  4.   **for** all image caption training samples
  5.     Training process
  6.   **end for**
  7. **end for**
  8. Data preprocessing (chatbot)
  9. Chatbot model initialization
  10. **for** all training epochs
  11.   **for** all chatbot training samples
  12.     Training process
  13.   **end for**
  14. **end for**
- 

The process can be split into two major sections: the first section is the image caption model, and the second part is the natural language processing part. In the image caption model, we first perform preprocessing on the data by consolidating and coding all the image data and the labeled text data. Then, after we initialize the model, we conduct model training. In the training process, we must observe whether the convergence conditions have been satisfied; if not, then training must be continued. If the text caption model training is complete, then the process enters its next stage. In the natural language processing part, the first and most crucial task is still the data preprocessing, followed by the initialization of the chatbot model. Next, training of the chatbot model must be performed using the processed data. Similarly, if the chatbot model does not satisfy the convergence conditions, the training must continue. Conversely, if the chatbot model has satisfied the convergence conditions, the training immediately ends. When the convergence conditions are met, the entire training process is complete. However, the image caption model must be trained first because this model plays a key role in incorporating image information in the chatbot input. If the image caption model is inadequately trained, the image information input into the chatbot becomes meaningless and may result in erroneous interpretation. After the training of the image caption model is completed, it can be used to convert image information to textual information, which is then fed to the input end of the chatbot. This conversion approach enables a chatbot model to be constructed without substantial modification. This also allows the model to interpret image information, which is a primary focus of this study. The problem of overfitting was considered during the training process, which was terminated immediately when overfitting occurred.

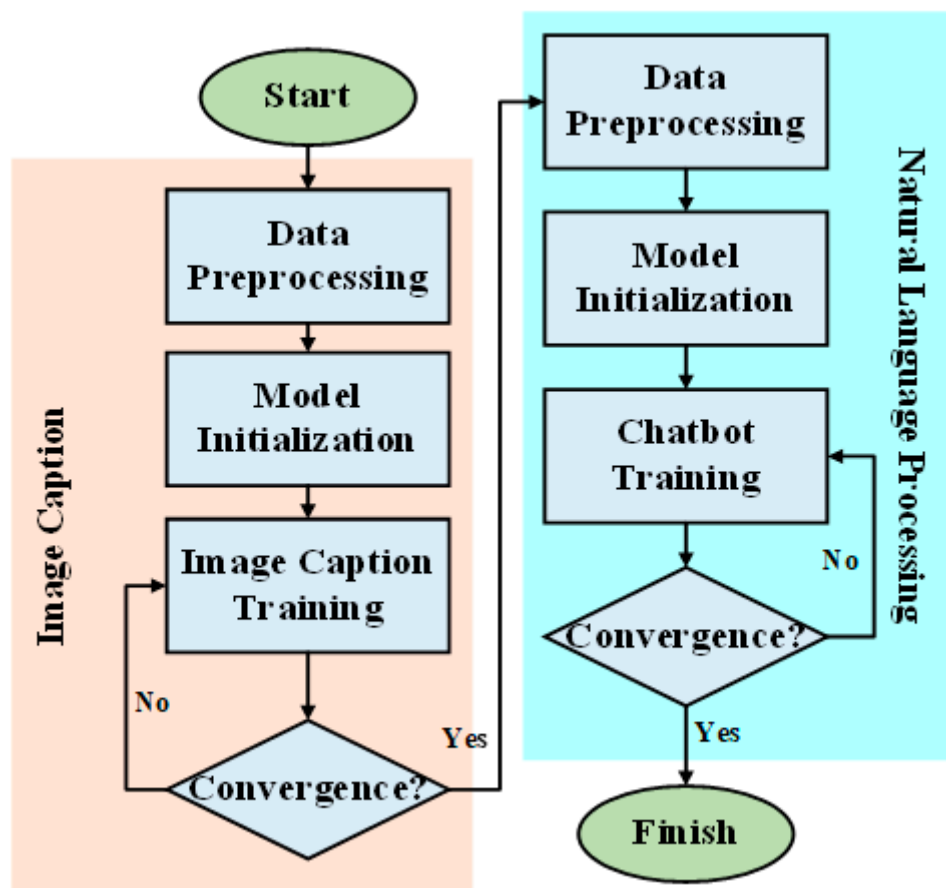


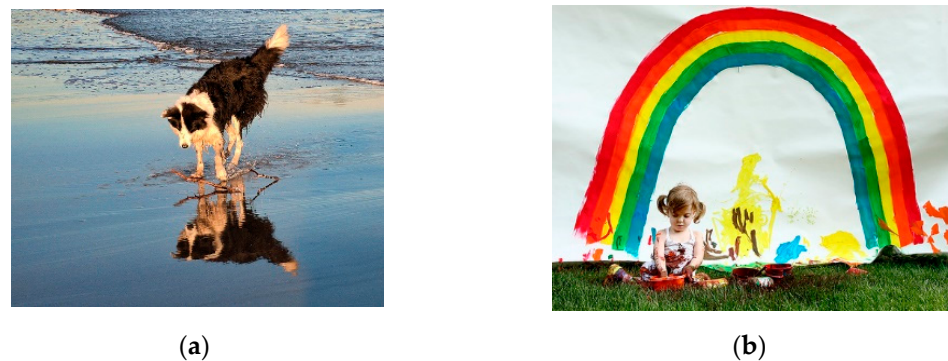
Figure 9. System flowchart of the training process.

#### 4. The Experimental Results

The first part of this section is a detailed explanation of the training data. Next, we present training and comparisons based on the model proposed in the previous section. Finally, we describe how the consolidated functions were integrated with the humanoid robot to test the effectiveness of the proposed model and analyze the feasibility.

##### 4.1. The Training Data

The databank for sentence training in this paper consisted of data sets of dialogue from online English classes [18]. Although image caption technology typically uses COCO data sets [23], which involve substantial amounts of data, as the training data, this experiment used the Flickr 8k Dataset [24], which has a smaller data volume, so that the model training results can be used with actual robots that may not have rapid computing abilities. The Flickr 8k Dataset has 8000 color images in total, and each image has five captions (see Figure 10 for the training data). From the images, we can see that each image has a chunk of text that describes the picture; these descriptions are man-made captions. Therefore, each image has various appropriate descriptions. Therefore, one of the key points of this paper is that image data must be described in text and incorporated into the chatbot's considerations. In this manner, conversations between humans and bots can become more precise and more in line with actual scenarios. Furthermore, in addition to the data mentioned, some new images were also added to the experiment in pursuit of greater realism and more accurate representation of the experimental scenario.

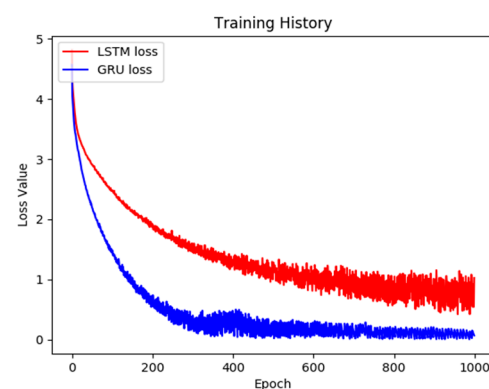


**Figure 10.** Training samples in Flickr 8 k [24]: (a) caption: “A dog at the beach”; and (b) caption: “A little girl covered in paint sits in front of a painted rainbow with her hands in a bowl”..

#### 4.2. The Experimental Results

To discuss the strengths and weakness of the proposed model and the original LSTM model in depth, we made the following comparisons. The chatbot training models are compared in Figure 11, whereas the image caption training models are compared in Figure 12. The figures demonstrate that during the chatbot model training process, the proposed method is faster and stabler and has a lower final training error rate. This also substantiates that the model proposed in this paper outperforms the original and conventional frameworks. However, during the training process of the image caption model, the performance of updated model proposed in this study was also slightly better. Although the error value sometimes increased slightly during the training process, overall it did not affect convergence. The final convergence results proved that the method proposed in this paper significantly outperformed conventional chatbot models. Conventional chatbot models, even in the later stages of training, exhibit trends of increasingly higher error values. This phenomenon once more substantiates that the stability and the validity of the method proposed in this paper are excellent.

To verify the practicality of the method proposed in this study, we incorporated a humanoid robot system into the following experiment and integrated the trained model into the humanoid robot. The experiment was split into a total of four scenarios; each scenario was matched with a picture to represent the present situation. The bot, based on the situation and user dialogue in the moment, made a corresponding response. Therefore, the bot’s inputs were not limited to the user’s voice data, but also included image data. The experimental results are depicted in Figures 13–16.



**Figure 11.** The learning curve of the chatbot model.

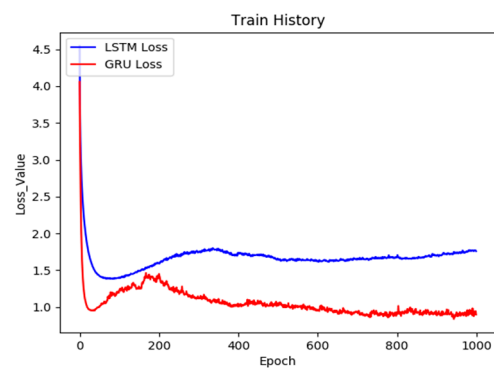


Figure 12. The learning curve of the image caption model.



Figure 13. Experimental result in scenario A.

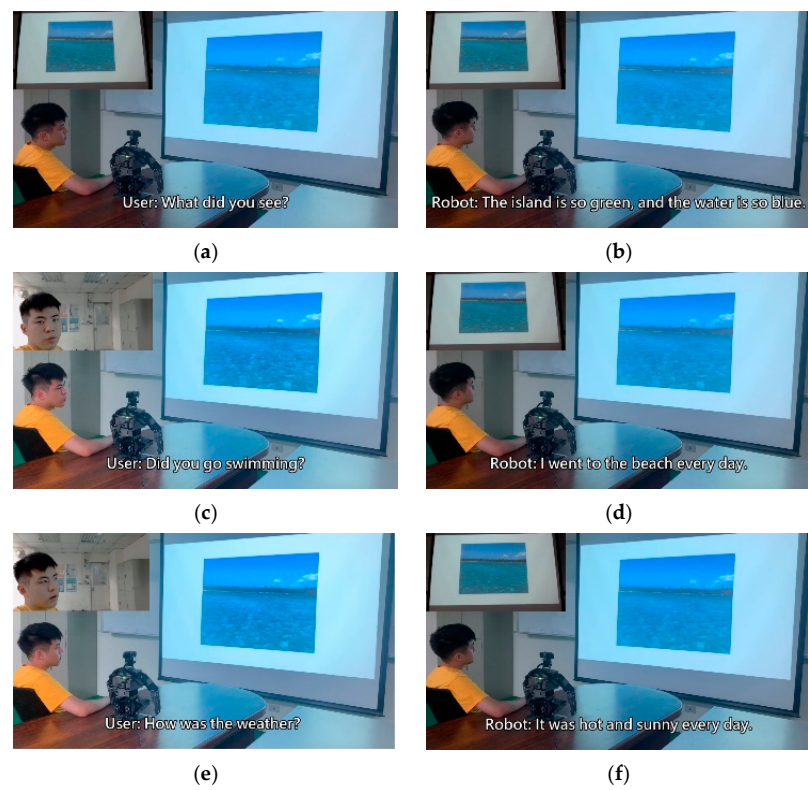


Figure 14. Experimental result in scenario B.

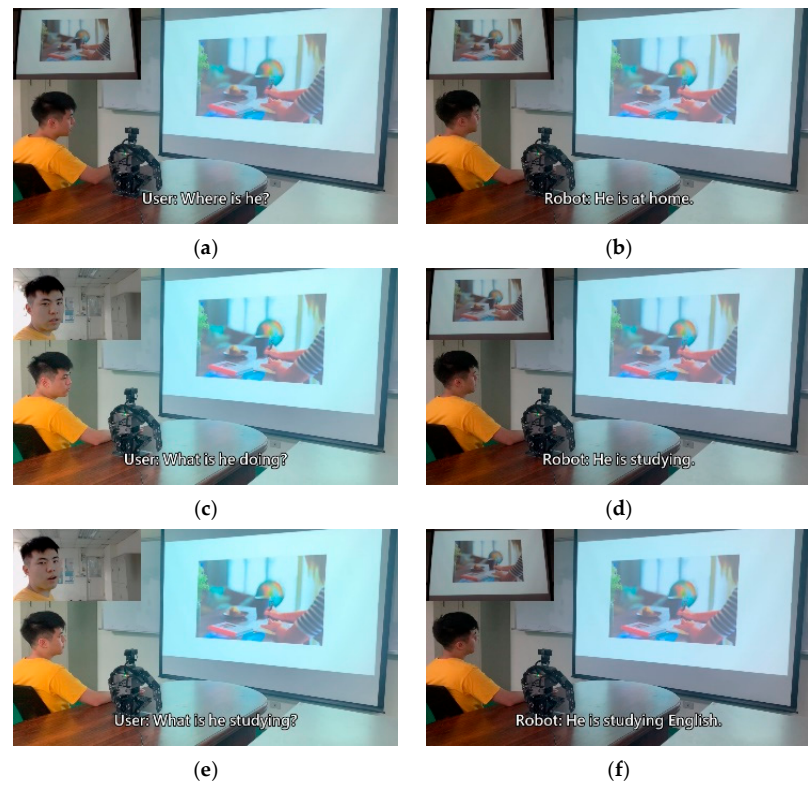
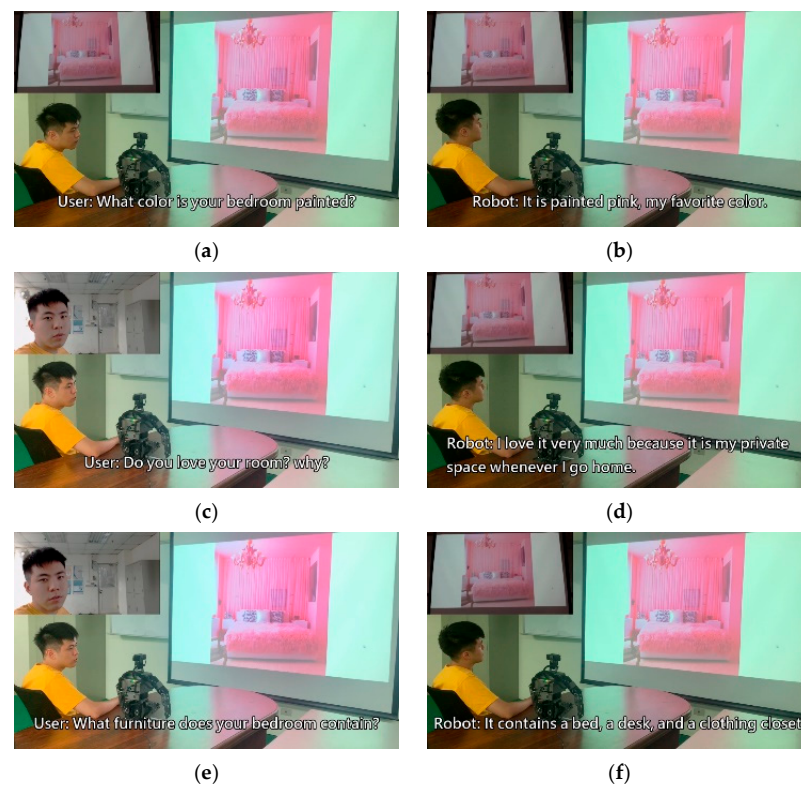


Figure 15. Experimental result in scenario C.



**Figure 16.** Experimental result in scenario D.

The experiment combined image caption with natural language processing technology. The robot was expected to give corresponding responses through related image and sound information. In the experimental procedure, the robot used a microphone to capture sound and converted sound into text through the speech-to-text (STT) technique. Then, the robot converted the obtained image information into text by using a deep artificial neural network. Sound and image information was converted into textual information for storage then inputted to the natural language processing model. The robot was able to give reasonable responses according to the information converted from image and sound. After the robot computed suitable sentences, it used the text-to-speech (TTS) technique to convert textual information into sound information. Thus, the text was read to allow the user to understand the robot's response.

In some instances, the robot's responses were not perfect. These imperfect responses may have arisen because the databank of image data and dialogue data was insufficient and, therefore, the robot was unable to fully consider all circumstances relevant to the required response. In the future, the model may process dialogue records that combine images; the model may then integrate larger and more comprehensive chatbot training databanks to increase the fluency of communications between humans and robots.

## 5. Discussion

In this experiment, the algorithm developed was integrated with a humanoid robot to facilitate conversation with humans. The robot used its built-in microphone to capture users' voice signals in addition to using the image sensor installed on its head to capture image information in the surrounding environment. Because the robot used in the experiment features multiple degrees of freedom, it is capable of aligning its sight with that of the user to converse. Moreover, it can turn its head to observe the surrounding environment while conversing. Accordingly, the robot uses the image-text conversion model to incorporate relevant textual information into the chatbot input. The content of the conversation revealed that the response provided by the robot was associated with the image information it received. This indicated that the robot does not merely consider the relationships between

voice signals when providing a response. This experiment verified that the proposed chatbot model is highly applicable to humanoid robots with companionship functions. The experimental video can be accessed in [25].

Two neural networks were used in this study, namely the chatbot model and image caption model. The main function of the image caption model is to convert implicit information in images and express it as text. The chatbot model interprets user conversations and previous responses to provide the user with the most suitable reply. The input of the chatbot model considers not only previous responses but also the text information output from the image caption model (e.g., the left input of the chatbot displayed in Figure 6). Therefore, the input end of the chatbot model retains both the information regarding the previous response and the text information output from the image caption model. This is equivalent to incorporating image information for model computation and interpretation. Accordingly, this not only resolves the problem concerning the lack of image and voice data when generating a response but also provides a convenient approach to integrate image and voice data for subsequent processing by the chatbot model. The feasibility of this approach was verified in the experiment.

In this study, dialogues from online English classes were used as the main conversation training data, whereas the Flickr 8k Dataset was used to train the image caption model. Because the conversation training data were in the form of textual information, image information first had to be converted to textual information before it could be input into the chatbot model. This conversion process was achieved using the image caption model, which was trained using the Flickr 8k Dataset before it became capable of image–text conversion. The converted textual information was then input into the chatbot model, allowing the model to consider such information when performing interpretation. This is equivalent to indirectly inputting image information into the chatbot model. This method converts all image information to textual information, facilitating information integration without the need for substantial modification to the model framework or parameter configuration. In addition, few training datasets currently exist for models that combine conversational text data and image data. Therefore, the proposed method was developed for effective training of the chatbot model through an existing conversational training dataset and image information. In the future, more training datasets can be included to improve the quality of conversations generated by the proposed chatbot model. This will further improve the applicability of the model to companion robots.

The main contribution of this study was the proposal of a chatbot model that integrates voice and image data to provide responses. This model can be applied in physical humanoid robots, allowing them to converse with humans. The experimental results revealed that, when no notable interferences existed in voice and image signals, the robot could engage in conversations fluently in most situations. Whether the responses provided by the humanoid robot can be considered accurate or fulfill user needs depends on the subjective evaluation of individuals. Consequently, because of the presence of various human-related factors, an objective quantitative evaluation of the chatbot model's accuracy was difficult. Nevertheless, the effectiveness of the model used in this study has been verified in other studies, such as [16], indicating that the public has mostly recognized the functionality of chatbots. The model used in this study was modified based on that in [16] and therefore exhibited more favorable performance and converged faster than did the original model. In addition, the modified model demonstrated more favorable learning efficiency. The experiment performed in this study also provided strong evidence to verify the effectiveness of the proposed model.

Advancements in medical technology have made aging societies inevitable. The home care for older adults living alone is a pressing social issue. Human beings generally live in groups, and people need to converse with others. If an elderly individual has no company at home, this has a large psychological influence. Therefore, building a chatbot model for the elders may solve the above problem. The developed technique in this study can also be one of the solutions of the above issue.

In this study, the proposed chatbot model is executed in the humanoid robot. However, this model can also be executed in any computer. The processing speed is related to the computing core level of the processor. In this study, because the user will communicate with the chatbot model in the experiments, we chose the humanoid robot for the platform. During the communication, the robot can also turn its head to look around and observe the environmental information. The humanoid robot can also execute some motions to react to the user. However, if the user has no humanoid robot for the platform, a simple computer is also feasible. Therefore, the requirements of this approach are not too high. Most people can access this chatbot model.

The validations of the proposed model are shown in the experiments. The functions of the proposed framework are completely demonstrated by applying the four scenarios. In the experiment, the robot can hear what the user says and use the SST technique to recognize the speech signal. The humanoid robot can also look around to observe the environmental information for the image input. Based on the previous sentences and the current information, the robot can give corresponding responses and reactions to the user. All the functions are intergraded in the humanoid robot and validated in the experiments.

In this paper, the output of the proposed chatbot model is a sentence for the response. It cannot be treated as a classification or normal regression problem. Therefore, defining the accuracy evaluation method for the proposed chatbot model is not applicable in these cases. For evaluating the performance of the proposed enhanced models, we use the learning curves for the comparisons. The learning curves not only indicate the convergence speed, but also show the training loss. The training loss of the proposed model can also prove that the error values are also lower than the original models. Besides, for presenting the actual performance of the trained model, performing the experiments with designed scenes may be better. As a result, our proposed model gives good results in the six experiments, as shown in the experimental video provided in this paper. Moreover, the experiments also show the feasibility and superiority of the proposed model.

There are two main cores in this proposed framework. The first one is the image caption model. The image caption model is used to extract the features and transfer these features to a one-sentence caption. Therefore, this model can help the humanoid robot to understand the content of the image. However, the image caption model should also be well trained. Hence, the Flickr 8k dataset is applied for training the image caption model. This dataset includes many pictures with the corresponding caption sentences. By using this dataset, the image caption model can be well trained for generating the corresponding sentence. The other important part of the proposed framework is the chatbot model. In this part, the generated sentence will also become one of the inputs of the chatbot model. Considering the chatting history and the generated sentence, the chatbot model can give the user the correct responses. However, the chatbot model also must be trained by the collected dataset of dialogue. By combining these two well-trained models, the whole proposed framework can talk with the users. This is also the reason why the robot can understand the meaning of the picture and give the correct response to the user.

In this paper, the humanoid robot is just one of the platforms of the proposed model. The proposed model can also be executed on just a computer. However, because the humanoid robot has many degrees of freedom, the robot can interact with the user by the motions. Furthermore, the robot can also track the human face while the user is talking (as shown in the experiments). The conversation between the robot and the human sounds relatively reasonable, because the training process of the proposed model is based on the human dialogue histories. The robot can learn the “rules” of the human conversations. Besides, the prediction results also adopt the concept of the probability for selecting the best response to the user. It is hard to say that using a humanoid robot is necessary in this example. However, we think that using a humanoid robot for this application might be a good choice.

There are many researches about companion robots for helping the elderly. In [26], a companion robot is designed for the elderly person. Such a robot can operate in assistive



environments in a smart home. The experiments indicate that such a system is both useful and useable. Literature [27] discusses assistive social robots in elderly care. This article mentioned that social robots are useful in eldercare. Moreover, the robot can also increase the quality of life. Literature [28] proposes a smart companion robot for elderly people. This robot applies the cloud computing technologies and fuzzy logic for the proposed robot. The experiments shows that the proposed robot system is feasible for human users. Literature [29] proposes a companion robot interaction system for supporting elderly people. Such a robot can also recognize the emotions of the users by using image processing techniques. Literature [30] proposes a home robot for elderly people with cognitive impairment. Such a robot can interact with an 89-year-old lady, and also give good results in the experiments. Literature [31] discusses the influence of social presence on acceptance of a companion robot by older people. The robot named “iCat” is applied for the experiments. The robot iCat with touch screen can help the elderly by the interactions, and the results are also validated in the experiments. Literature [32] proposes an assistive robot for older adults. Such a robot applies speech recognition techniques in the experiments. The user can also play games with the proposed robot. Based on the above researches, companion robots are very important and helpful for the elderly. In this paper, our humanoid can also be one of the companion robots to help elderly people.

However, in this paper, the humanoid robot is just a platform of the chatbot model. The proposed approach might be used for elderly people in the future. The proposed model can be executed in just one computer. It is unnecessary to perform the model in only humanoid robots. However, if the chatbot model is integrated in a humanoid robot, the humanoid robot can also perform some expressive motions during the chatting process. This might increase willingness to use this system. Therefore, we believe that the user somehow benefits by performing the chatbot model in a humanoid robot.

Although there are currently many studies that allow robots to use audio and visual data simultaneously to analyze information and make appropriate responses, in this study, the audio and visual data are analyzed in advance using different models, and the results of the analysis are input into the chatbot model. In order for the chatbot model to consider the visual data, we converted the image into text so that the robot can recognize the content of the image and use it as one of the points of reference for its next reply. This approach is still rare in current research and it is also the novelty of this research.

The goal of this research is to allow humanoid robots to have the ability to converse with humans. Through the aid of visual images the robot can perceive the external environment for it to make more correct judgments. Therefore, the output of the model proposed in this study is a string of words, which is difficult to define through comparative metrics. Consequently, to measure the effectiveness of the model, we drew a convergence curve and compared it with the model architectures proposed by other researches. According to the experiment results, the convergence effect of our proposed model is better, and this is also one of the main contributions of this paper.

In this research we use general STT and TTS methods for the audio signal processing model and have not optimized or modified the methods. The main targets for model optimization in this research are the image caption model and chatbot model used in image processing. Experiment results show that the performance of the modified model in this paper is better than that of the original model.

The image used in the experiment video is one of the images used during training in the image caption model. The light and shadow changes of the image do have a very big impact on the robot’s analysis process, therefore during the experiment, we tried to keep the ambient light in a more stable state. As for how to reduce the impact of ambient light on image analysis, it is one of the goals that we urgently need to overcome in the future.

Although in most cases the model proposed in this study can communicate with humans and generate acceptable results, greater distraction in the environment (such as a lot of background noise or an environment with large changes in light and shadow) may affect the results of model analysis and give unreasonable responses. Another possible

reason for experiment failure lies in insufficient data training. The more complete the data training the better the model processing capabilities. Therefore, if we wish to enhance the effectiveness of the proposed model, the above two points will be very critical.

The model performance comparison proposed in this paper is mainly focused on the convergence of the model. Using different databases to train the same model will also produce completely different training results. During the interaction process between humans and the model proposed in this paper, human feelings can be very subjective, therefore human feelings are not taken into consideration in the effective analysis of the model.

This research is mainly aimed at optimizing the training efficiency of the model. In the actual application phase in the future, if it is needed to customize the model for elderly users, the first item for consideration is the establishment of a database. It is an indispensable procedure to collect and organize data on phrases and conversational nuances commonly used by elderly users. In addition, we must further study and explore the habits of using technology products by elderly users. In the future it may be possible to incorporate medical related knowledge to improve the model, thereby bridging the gap between elderly users and robots.

## 6. Conclusions

This study proposed a chatbot model that combines image and sound information and applied it to a companion chatbot. Image caption and natural language processing technologies were adopted together to design a chatbot system. The experimental results indicate that the model, combining image and sound information, is beneficial in particular chatting scenarios. Overall, the improved GRU model demonstrated superior efficacy in this application than did LSTM. The convergence in model training was also faster and more stable. In addition, the experiment verified the effectiveness and feasibility of the proposed method. In the future, this technology can be extended into the field of older adult care. It may make older people's lives less lonely and thus improve their quality of life. The direction of this study is also helpful to aging societies.

**Author Contributions:** Conceptualization, P.-H.K.; Data curation: P.-H.K.; Formal analysis: P.-H.K.; Funding acquisition: P.-H.K.; Investigation: P.-H.K.; Methodology: P.-H.K.; Project administration: P.-H.K. and C.-J.H.; Resources: P.-H.K. and C.-J.H.; Software: P.-H.K., S.-T.L., J.H. and C.-J.H.; Supervision: P.-H.K. and C.-J.H.; Validation: P.-H.K., S.-T.L., J.H. and C.-J.H.; Visualization: P.-H.K. and C.-J.H.; Writing—original draft: P.-H.K. and C.-J.H.; Writing—review and editing: P.-H.K. and C.-J.H. All authors have read and agreed to the published version of the manuscript.

**Funding:** This work was supported by the Ministry of Science and Technology, Taiwan, under Grants MOST 109-2221-E-194-053-MY3.

**Institutional Review Board Statement:** Not applicable.

**Informed Consent Statement:** Not applicable.

**Conflicts of Interest:** The authors declare no conflict of interest.

## Appendix A

The VGG16 model: as presented in Figure A1, is a picture classifier with great performance and was often used when early deep neural networks first became popular.

The described methods are TTS and STT technologies, and their operational methods are presented in Figure A2. The TTS and STT technologies currently in development are already rather comprehensive. In this paper, these two technologies were used and integrated into the humanoid robot to make the robot's responses more natural and more fluent when answering humans.



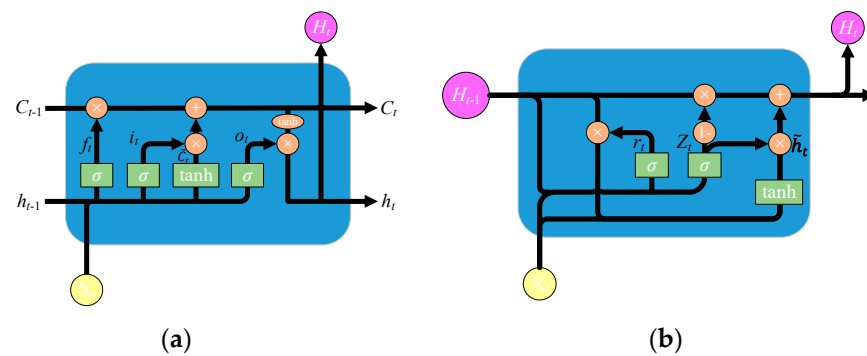


Figure A3. The cell structures of (a) LSTM; and (b) GRU.

$$f_t = \sigma(W_f \cdot [h_{t-1}, x_t] + b_f) \quad (\text{A1})$$

$$i_t = \sigma(W_i \cdot [h_{t-1}, x_t] + b_i) \quad (\text{A2})$$

$$\tilde{C}_t = \tanh(W_c \cdot [h_{t-1}, x_t] + b_c) \quad (\text{A3})$$

$$C_t = f_t * C_{t-1} + i_t * \tilde{C}_t \quad (\text{A4})$$

$$o_t = \sigma(W_o \cdot [h_{t-1}, x_t] + b_o) \quad (\text{A5})$$

$$h_t = o_t * \tanh(C_t) \quad (\text{A6})$$

GRU is a variation on LSTM; the major difference is that it combines the forget gate and input gate into a single update gate. Other than some minor changes, the operational concepts are markedly similar. The final model is more streamlined than the original LSTM model. The calculations are presented in (A7)–(A10).

$$z_t = \sigma(W_z \cdot [h_{t-1}, x_t]) \quad (\text{A7})$$

$$r_t = \sigma(W_r \cdot [h_{t-1}, x_t]) \quad (\text{A8})$$

$$\tilde{h}_t = \tanh(W \cdot [r_t * h_{t-1}, x_t]) \quad (\text{A9})$$

$$h_t = (1 - z_t) * h_{t-1} + z_t * \tilde{h}_t \quad (\text{A10})$$

## References

1. Wu, E.H.-K.; Lin, C.-H.; Ou, Y.-Y.; Liu, C.-Z.; Wang, W.-K.; Chao, C.-Y. Advantages and Constraints of a Hybrid Model K-12 E-Learning Assistant Chatbot. *IEEE Access* **2020**, *8*, 77788–77801. [\[CrossRef\]](#)
2. Rajkumar, R.; Ganapathy, V. Bio-Inspiring Learning Style Chatbot Inventory Using Brain Computing Interface to Increase the Efficiency of E-Learning. *IEEE Access* **2020**, *8*, 67377–67395. [\[CrossRef\]](#)
3. Smutny, P.; Schreiberova, P. Chatbots for learning: A review of educational chatbots for the Facebook Messenger. *Comput. Educ.* **2020**, *151*, 103862. [\[CrossRef\]](#)
4. Liu, B.; Xu, Z.; Sun, C.; Wang, B.; Wang, X.; Wong, D.F.; Zhang, M. Content-Oriented User Modeling for Personalized Response Ranking in Chatbots. *IEEE/ACM Trans. Audio Speech Lang. Process.* **2018**, *26*, 122–133. [\[CrossRef\]](#)
5. Hochreiter, S.; Schmidhuber, J. Long Short-Term Memory. *Neural Comput.* **1997**, *9*, 1735–1780. [\[CrossRef\]](#)
6. Mao, G.; Su, J.; Yu, S.; Luo, D. Multi-Turn Response Selection for Chatbots with Hierarchical Aggregation Network of Multi-Representation. *IEEE Access* **2019**, *7*, 111736–111745. [\[CrossRef\]](#)

7. Chung, J.; Gulcehre, C.; Cho, K.; Bengio, Y. Empirical Evaluation of Gated Recurrent Neural Networks on Sequence Modeling. *arXiv* **2014**, arXiv:1412.3555.
8. Daniel, G.; Cabot, J.; Deruelle, L.; Derras, M. Xatkit: A Multimodal Low-Code Chatbot Development Framework. *IEEE Access* **2020**, *8*, 15332–15346. [[CrossRef](#)]
9. Sheehan, B.; Jin, H.S.; Gottlieb, U. Customer service chatbots: Anthropomorphism and adoption. *J. Bus. Res.* **2020**, *115*, 14–24. [[CrossRef](#)]
10. Oh, J.; Jang, S.; Kim, H.; Kim, J.J. Efficacy of mobile app-based interactive cognitive behavioral therapy using a chatbot for panic disorder. *Int. J. Med. Inform.* **2020**, *140*, 104171. [[CrossRef](#)]
11. Luong, M.; Pham, C. Incremental Learning for Autonomous Navigation of Mobile Robots based on Deep Reinforcement Learning. *J. Intell. Robot. Syst.* **2021**, *101*, 1. [[CrossRef](#)]
12. Pang, L.; Zhang, Y.; Coleman, S.; Cao, H. Efficient Hybrid-Supervised Deep Reinforcement Learning for Person Following Robot. *J. Intell. Robot. Syst.* **2020**, *97*, 299–312. [[CrossRef](#)]
13. Ma, J.; Lu, H.; Xiao, J.; Zeng, Z.; Zheng, Z. Multi-robot Target Encirclement Control with Collision Avoidance via Deep Reinforcement Learning. *J. Intell. Robot. Syst.* **2020**, *99*, 371–386. [[CrossRef](#)]
14. Yan, C.; Xiang, X.; Wang, C. Towards Real-Time Path Planning through Deep Reinforcement Learning for a UAV in Dynamic Environments. *J. Intell. Robot. Syst.* **2020**, *98*, 297–309. [[CrossRef](#)]
15. Azayev, T.; Zimmerman, K. Blind Hexapod Locomotion in Complex Terrain with Gait Adaptation Using Deep Reinforcement Learning and Classification. *J. Intell. Robot. Syst.* **2020**, *99*, 659–671. [[CrossRef](#)]
16. Ludwig, O. End-to-end Adversarial Learning for Generative Conversational Agents. *arXiv* **2017**, arXiv:1711.10122.
17. Keras. Available online: <https://keras.io/> (accessed on 25 July 2020).
18. Seq2seq Chatbot. Available online: <https://github.com/oswaldoludwig/Seq2seq-Chatbot-for-Keras> (accessed on 25 July 2020).
19. Vinyals, O.; Toshev, A.; Bengio, S.; Erhan, D. Show and Tell: A Neural Image Caption Generator. *arXiv* **2014**, arXiv:1411.4555.
20. Xu, K.; Ba, J.; Kiros, R.; Cho, K.; Courville, A.; Salakhutdinov, R.; Zemel, R.; Bengio, Y. Show, Attend and Tell: Neural Image Caption Generation with Visual Attention. *arXiv* **2015**, arXiv:1502.03044.
21. Image Captioning. Available online: <https://github.com/Div99/Image-Captioning> (accessed on 25 July 2020).
22. Simonyan, K.; Zisserman, A. Very Deep Convolutional Networks for Large-Scale Image Recognition. *arXiv* **2014**, arXiv:1409.1556.
23. COCO Dataset. Available online: <https://cocodataset.org/#home> (accessed on 25 July 2020).
24. Flickr 8k Dataset. Available online: <https://www.kaggle.com/adityajn105/flickr8k> (accessed on 7 May 2020).
25. Experimental Video. Available online: [https://youtu.be/y3v\\_cla6-Go](https://youtu.be/y3v_cla6-Go) (accessed on 25 July 2020).
26. Saunders, J.; Syrdal, D.S.; Koay, K.L.; Burke, N.; Dautenhahn, K. “Teach Me–Show Me”—End-User Personalization of a Smart Home and Companion Robot. *IEEE Trans. Hum. Mach. Syst.* **2015**, *46*, 27–40. [[CrossRef](#)]
27. Broekens, J.; Heerink, M.; Rosendal, H. Assistive social robots in elderly care: A review. *Gerontechnology* **2009**, *8*, 94–103. [[CrossRef](#)]
28. Pavón-Pulido, N.; López-Riquelme, J.A.; Pinuaga-Cascales, J.J.; Ferruz-Melero, J.; Dos Santos, R.M. Cybi: A smart companion robot for elderly people: Improving teleoperation and telepresence skills by combining cloud computing technologies and fuzzy logic. In Proceedings of the 2015 IEEE International Conference on Autonomous Robot Systems and Competitions, Vila Real, Portugal, 8–10 April 2015; pp. 198–203.
29. Lee, S.B.; Yoo, S.H. Design of the companion robot interaction for supporting major tasks of the elderly. In Proceedings of the 2017 14th International Conference on Ubiquitous Robots and Ambient Intelligence (URAI), Jeju, Korea, 28 June–1 July 2017; pp. 655–659.
30. Gross, H.-M.; Schroeter, C.; Mueller, S.; Volkhardt, M.; Einhorn, E.; Bley, A.; Langner, T.; Martin, C.; Merten, M. I’ll keep an eye on you: Home robot companion for elderly people with cognitive impairment. In Proceedings of the 2011 IEEE International Conference on Systems, Man, and Cybernetics, Anchorage, AK, USA, 9–12 October 2011; pp. 2481–2488.
31. Heerink, M.; Kröse, B.; Evers, V.; Wielinga, B. The influence of social presence on acceptance of a companion robot by older people. *J. Phys. Agents* **2008**, *2*, 33–40. [[CrossRef](#)]
32. Wu, Y.H.; Wrobel, J.; Cristancho-Lacroix, V.; Kamali, L.; Chetouani, M.; Duhaut, D.; Le Pevedic, B.; Jost, C.; Dupourque, V.; Ghriissi, M.; et al. Designing an assistive robot for older adults: The ROBADMOM project. *Irbm* **2013**, *34*, 119–123. [[CrossRef](#)]