Mobile Networks and Applications
https://doi.org/10.1007/s11036-021-01795-5

AI for Online Customer Service: Intent Recognition and Slot Filling Based on Deep Learning Technology

Yirui Wu[1] · **Wenqin Mao**[1] · **Jun Feng**[1]

Accepted: 1 December 2020

Abstract
Cloud/edge computing and deep learning greatly improve performance of semantic understanding systems, where cloud/edge computing provides flexible, pervasive computation and storage capabilities to support variant applications, and deep learning models could comprehend text inputs by consuming computing and storage resource. Therefore, we propose to implement an intelligent online custom service system with power of both technologies. Essentially, task of semantic understanding consists of two subtasks, i.e., intent recognition and slot filling. To prevent error accumulation caused by modeling two subtasks independently, we propose to jointly model both subtasks in an end-to-end neural network. Specifically, the proposed method firstly extracts distinctive features with a dual structure to take full advantage of interactive and level information between two sub-tasks. Afterwards, we introduce attention scheme to enhance feature representation by involving sentence-level context information. With the support of cloud/edge computing infrastructure, we deploy the proposed network to work as an intelligent dialogue system for electrical customer service. During experiments, we test the proposed method and several comparative studies on public ATIS and our collected PSCF dataset. Experiment results prove the effectiveness of the proposed method by obtaining accurate and promising results.

Keywords Cloud/edge computing for deep learning · Combing AI and cloud/edge for custom service · Semantic understanding · Intent recognition · Slot filling

1 Introduction

With significant progress achieved by cloud/edge computing and deep learning, how to properly combine strength of both technologies becomes a hot topic [1, 2]. Essentially, deep learning models generally require larger computation and storage resource to construct deeper structure, in order to better optimize performance of big data driven tasks. Being capable to meet requirements of high amount resource, cloud/edge computing helps build a solid foundation for big data applications indeed [3]. Under guidance of involving both technologies, quantity of successful big data applications [4, 5] in domains like smart cities, internet of things, e-commerce, driverless cars, and etc, have emerged. Following such idea, we pay special attention to develop an intelligent online customer service system on the basis of big data technologies.

As an effective supplement to manual customer service, automatical online customer service dialogue systems have been invented for a relatively long time. Several impressive implements include the psychological consultation dialogue system ELIZA [6] developed by Massachusetts Institute of Technology, the kinship dialogue system SAD-SAM [7] developed by Carnegie Mellon University, and so on. As shown in Fig. 1, a customer service dialogue system usually consists of five functional modules: speech recognition, semantic understanding, dialogue management, text generation and speech synthesis.

Task of speech recognition module is to convert a continuous temporal signal representing user's speech into an instance of text. With help of grammatical and semantic analysis, semantic understanding module analyze generated text instances of speech recognition module,

✉ Jun Feng
fengjun@hhu.edu.cn

Yirui Wu
wuyirui@hhu.edu.cn

Wenqin Mao
hhuwqmao@gmail.com

[1] College of Computer and Information, Hohai University, Nanjing, China

Published online: 23 July 2021

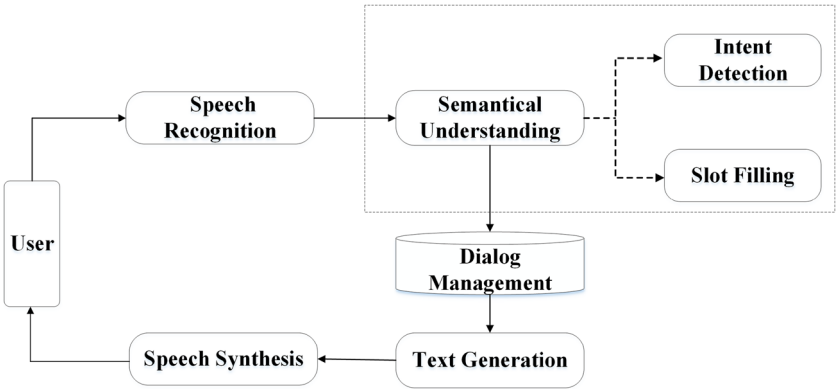

Fig. 1 General structure design of a custom service dialogue system with five separated functional modules

thus expressing user's intention in the way that can be easily processed by the following dialogue management module. Due to diversity and ambiguity of natural language, semantic understanding still faces many challenges and can be divided as two subtask, i.e., intention recognition and slot filling. Through analyzing results generated by semantic understanding module with additional dialogue context, history information and other useful information, dialogue management module automatically determine proper strategies and actions for answering. Afterwards, text generation module is responsible to generate text instances under guidance of dialogue management module. Finally, speech synthesis module speaks generated text out to send proper feedbacks to users. It's noted that speech synthesis module is not necessary for online custom service system. Meanwhile, simultaneously performing multiple online talks still remains an open question, due to low capability of semantic understanding model to comprehend and answer complicated questions, and low resource of computing infrastructure for efficient parallel running.

To solve such problem, researchers and engineers have tried to apply deep neural networks and cloud/edge computing infrastructure, resulting in several excellent projects, such as GoogleBrain, Siri of Apple, Jimi of JD and so on. These successful practices allow people to view the feasibility of applying deep learning and could/edge computing technologies [8–10] to construct dialogue systems. Essentially, core procedure of these systems is semantic understanding, which helps computers to comprehend meanings and intentions of users' questions via natural language. Intent recognition aims to classify a given question into a certain semantic category based on context information embedded in the question. Standard classifiers such as support vector machines (SVM) and boosting have been proven effective in task of intent recognition. Slot filling could be regarded as a sequence labeling task to find the maximum probability for slot value assignment based on the given sentence, where Conditional Random Field (CRF) and its variants have been widely used for task of slot filling.

Facing challenges brought by processing big data and requests from users for smarter and faster online customer service, we intend to further develop powerful deep learning structure as a appropriate solution. Specifically, former methods prefer independent modeling of two subtasks, since separated subtasks could be easily stepwise trained and optimized for local extremum. However, such modeling idea ignores fusion information of both tasks, leading to incomplete or ineffective modeling of features. This could be easily explained by an example. Supposing the purpose of a sentence is to search for a proper flight recognized by subtask of intent recognition, it's mostly common to answer information about the departure city, arrival city, and vice versa for subtask of slot filling. Therefore, both tasks will benefit from sufficient information exchange, if we could perform two subtasks simultaneously. Furthermore, ene-to-end modeling via a single neural network could prevent error accumulation during training, which fits with current developing trend of deep learning domain, i.e., all in one network.

To describe interactive information between intent recognition and slot filling, the proposed method jointly model both subtasks in an end-to-end manner. Specifically, we firstly propose a dual network structure based on Bi-GRU (Bidirectional Gated Recurrent Unit), CNN (Convolutional Neural Network), and CRF to complete semantic understanding under the guidance of joint modeling. Inspired by successful implementations of attention scheme in domains of natural language processing and computer vision, we involve context information extracted by attention scheme for feature enhancement, resulting in more distinguish feature designs for both subtasks. Finally, we apply the proposed method to work as an intelligent dialogue system for online electrical customer service, which proves the efficiency of the proposed methods to deal with challenges in a specifical service domain.

The contribution of this paper is two-fold:

- Instead of stepwise modeling, we establish a dual structure of Bi-GRU, CNN and CRF to jointly model subtasks, i.e., intention recognition and slot filling, which could make full use of interactive information between two subtasks for enhanced and distinctive feature designs. Moreover, the proposed method is capable to fuse advantages of different deep learning structures for performance improvement.
- Following training requests of the proposed dual structure, we design an asynchronous training strategy with respect to different loss functions for two subtasks.
- We accelerate the running speed of the proposed model by implementing the proposed system on cloud/edge computing infrastructure. We believe the proposed online custom service system is a promising trial to involve both deep leaning and cloud/edge computing technologies for big data driven applications.

The rest of the paper is organized as follows. In Section 2, a brief overview of existing methods for semantic understanding and attention modeling is introduced. In Section 3, the proposed dual structure and attention scheme are illustrated in detail. In the fourth section, comparative experiments are conducted on one English and one Chinese datasets, and performance analysis will be discussed. Finally, the paper will be summarized.

2 Related work

Existing methods related to our work can be categorized into the following two categories: semantic understanding and attention scheme.

2.1 Introduction to semantic understanding

Semantic understanding is the ability of a machine to process the meaning and context behind real-world information, which generally consists of two subtasks, i.e., intent recognition and slot filling. The task of recognizing the intentions of an agent by analyzing some or all of their actions and/or analyzing the changes in the state resulting from their action, meanwhile the goal of slot filling is to identify from a running dialog different slots, which correspond to different parameters of the user's query. Intention recognition is generally regarded as a classification problem for meanings of question. Therefore, traditional solution for intention recognition is bag of words model to perform information retrieval, thus searching and locating the exact category label for input question. Meanwhile, slot filling is formulated as a task of semantic parsing to fill out most appropriate slot value with help of probability model.

With significant development of deep learning technologies, researchers try to recognize users' intentions with deep structures. Early, Ravuri et al. [11] firstly construct a combination network of GRU and LSTM units for feature extraction, and then build feature vectors for words to perform classification on input sentences. Then, Xia et al. [12] adopt a novel weighting mechanism to dynamically allocate appropriate weights for each semantic word. After aggregating weighted semantic values into feature, they take advantage of capsule model to perform hierarchical classification task on texts, which successfully completes task of intent recognition with remarkable accuracy. Drawing from literature linking gaze and visual attention, Singh et al. [13] combine gaze and model-based AI planning to build probability distributions over a set of possible intentions, which not only realizes application of online human intention recognition, but also provides ideas for intent recognition in text domain.

Deep learning technology is introduced by researchers to perform slot filling as well. Inspired by word and character level features extracted by a two-way LSTM-CNNs structure [14], Ma et al. [15] further propose a BiLSTM-CNNNs-CRF network to optimize the output tag sequence with help an additional CRF module. Afterwards, Liu et al. [16] propose LM-LSTM-CRF-aware model, which successfully constructs character-aware neural network to extract character-level feature representations under the guidance of multi-task framework. Recently, Liu et al. [17] propose a Coarse-to-fine approach (Coach) for cross-domain slot filling, which first learns the general pattern of slot entities by detecting whether the tokens are slot entities or not, and then predicts the specific types for the slot entities. Since traditional slot filling predicts a one-hot vector for each word and lacks semantic correlation modelling, Zhu et al. [18] construct three categories of distributed label embedding for each slot using different kinds of prior knowledge, i.e., atomic concepts, slot descriptions, and slot exemplars. After experiments, their proposed label embeddings are proved to share text patterns and reuses data with different slot labels.

Similar with idea of the proposed method, i.e., joint modeling semantic understanding, researchers try to construct a joint deep structure to model interactive information between intent recognition and slot filling. For example, Zhang et al. [19] adopt GRU to learn feature representation for each time step and predict each slot value. Meanwhile, their proposed model is capable to capture global features of sentence intent through functions of pooling layers. Afterwards, Liu et al. [20] introduce attention mechanism into the alignment-based RNN model for tasks of intention recognition and slot filling, where additional text context features

are extracted by attention mechanism for intent classification and slot label prediction. Since word level information is not well modeled in previous works, Chen et al. [21] propose WAIS, word attention for joint intent recognition and slot filling. Considering that intent recognition and slot filling have a strong relationship, they further propose a fusion gate that integrates the word level information and semantic level information together for jointly training the two tasks.

Through quantity of experiments, Yin et al. [22] draw a conclusion that CNN has certain advantages in identifying target tasks, while RNN has advantages in sequence recognition modeling due to its unique memory design. Compared with RNN, CRF model is capable to calculate the joint probability after labeling the entire sequence, instead of splicing the optimal labels at each moment. Therefore, it's essential to combine advantages of these three different models for joint modeling, which is core idea of the proposed dual structure with Bi-GRU, CNN and CRF models.

2.2 Attention scheme

Due to high flexibility, attention scheme is widely used in deep learning domain, where attention scheme is generally formed as a dimension of interpretability into internal representations. Generally speaking, we could construct attention scheme by two steps. During the first step to calculate weights based on similarity between input signal and pre-trained weights, Multi-Layer Perception (MLP) is utilized to calculate similarity as

$$sim(Q, W_i) = MLP(Q, W_i) \tag{1}$$

where Q is input signal and W_i refers to one of the pre-trained weights. Afterwards, Softmax function is adopted to perform normalization on calculated similarity and emphasize on informative parts:

$$\alpha_i = softmax(sim(Q, W_i)) = \frac{e^{sim(Q,W_i)}}{\sum_{j=1}^{L} e^{sim(Q,W_i)}}; \tag{2}$$

where L is the number of pre-trained weights.

In the second step to re-weight original values based on calculated weights, attention values $Atten$ can be obtained by summing weighted original values with:

$$Atten = \sum_{i=1}^{L} \alpha_i \cdot v_i \tag{3}$$

where v_i refers to original values and operation $\cdot$ means element-wise operation. By calculating with two stages above, we can get the attention value $Atten$ for original vector v with the input signal Q.

Attention scheme is commonly used in visual tasks to perform feature enhancement. Early, Xu et al. [23] propose a memory network with spatial attention embedded, which could use input question to choose relevant regions for answering. Afterwards, Yang et al. [24] present a multi-layer stacked attention network to infer the answer progressively, which achieves question-guided attention on every region in the image. Regarding former works as top-down attention, Anderson et al. [25] firstly construct bottom-up attention by Faster R-CNN to detect visual objects, and then assign weights to visual objects through question-guided top-down attention. Their proposed model successfully highlights key objects of input images, thus achieving accurate answers.

Most recently, Woo et al. [26] build Convolutional Block Attention Module (CBAM) as a lightweight attention scheme, which sequentially computes attention values along spatial and channel dimensions. Afterwards, they fuse attention and image feature map for automatical feature enhancement. Compared with CBAM, Cao et al. [27] propose Global context network (Gcnet), which simplifies non-local neural network and acts to be more efficient and faster in run-time. Afterwards, Zhao et al. [28] propose end-to-end Recurrent Attention (RA) model for pedestrian recognition, which highlight the spatial property of generated feature map by involving strength of Recurrent Learning and Attention scheme. Experiments show that they successfully extract context relationship among attribute categories of pedestrians to achieve more accurate recognition results. To obtain more accurate detection results with multiple attention modules, Zhao et al. [29] propose PFAN (Pyramid Feature Attention Network), which adopt spatial attention mechanism for low-level network structures and channel attention mechanism for high-level network.

With the idea of dynamic setting on receptive field, Liu et al. [30] propose RFB (Receptive Field Block) network, which uses dilated convolutions to obtain more noteworthy information on size of receptive field. Further emphasizing on the importance of receptive fields, Li et al. [31] design SKNet (Selective Kernel Network) to assign weights for both channel related information and size of convolution kernel. Afterwards, Xiao et al. [32] take receptive field, spatial and channel attention information into account, and propose TCAM (Text-Context-Aware Module) to solve the problem of multi-oriented and multi-language, which proves the effectiveness of applying different kinds of context information on task of text detection.

3 The proposed method

Inspired by successful applications of RNN-based or encoder-decoder based sequential processing models for semantic understanding, the proposed method involves

advantages of multiple types of deep learning structures for improvement, where we propose a novel dual structure to describe interactive information parts between two subtasks. Afterwards, the proposed method involve an attention scheme to describe context information of the input question, thereby improving performance. Finally, a novel asynchronous training method with design of loss function is proposed to train the proposed dual structure.

3.1 Design of dual structure for semantic understanding

Text can be viewed as a special category of temporal data, if we regard each word as a state for a sentence. By modifying aspect for data category, quantity of temporal data processing algorithm can be applied to solve text-related problems. We pay special attention on LSTM (Long Short-Term Memory) network, where its unique gate mechanism and memory design make it suitable to process long time-series data with features of updating passing-by information by only remembering important parts. With its unique designs and properties, text processing can be highly efficient to deal with relatively long sentences by acquiring meaningful and informative parts. However, the original unidirectional LSTM network can only encode information of sentence from front to back. Essentially, the exacted meaning of a single word can greatly be affected by its context information described by neighboring words in either front or back positions. To extract as much as information from sentences, we propose to encode information from both directions with Bi-LSTM model, which could capture two-way semantic dependence among words, thus better describing context information of the whole sentence.

Since the proposed method is originally designed to apply in cloud/edge based online custom service systems for parallel running, a light design in structure should be emphasized to pursue for high working efficiency with relatively small computing resource. Guided by this principle for algorithm design, we further modify LSTM cell to GRU cell, where GRU structure applies two forms of gates, i.e., Reset gate and Update gate, to not only save much storage and computing power with less parameters, but also keep similar working performance with LSTM network.

Besides the analysis view of regarding text as temporal data, traditional way is to process text as spatial data, where researchers usually adopt CNN to extract local features by encoding dependency relationship among words. In other words, words can be regarded as neighbour units with spatial relationship, which is suitable for CNN to model and describe.

Based on the above discussion, we propose to involve both Bi-GRU and CNN models for text processing, which is capable to extract sufficient information from input question sentences from spatial and temporal views. Moreover, we design a dual structure to model interactive information between two subtasks, i.e., intent recognition and slot filling, where two relatively independent modules are designed to carry out two subtasks. Meanwhile, a novel information exchange design formed by several hidden layers is proposed to promote feature fusion between two kinds of features extracted for two separated tasks. We believe such design could help make full use of interactive information between two subtasks for feature enhancement, thus greatly improving performance for both subtasks.

We show the detailed designs of the proposed dual structure in Fig. 2, where two Bi-GRU models are adopted to extract distinctive feature representations for intent classification and slot filling tasks respectively, a CNN model is placed in the upper part to encode local property of texts for intent classification, and a CRF model is introduced in lower part to model sequential information for slot filling. It's noted that feature information exchange formed by several layers severs for interactive feature fusion between two subtasks as well as two Bi-GRU models.

Specifically, a question, such as "Flight from Denver to Greece" as shown in Fig. 2, is asked by a user and regarded as input for the proposed method. During processing, the input question sentence can be represented as $X = \{x_1, x_2, ..., x_T\}$, where x_t refers to the tth word in sentence, and T is the total number of words inside the sentence. Afterwards, two Bi-GRU models extract temporal feature representations h_t^i and h_t^s for intent recognition and slot filling respectively, where i could be comprehended as either index of word or state index. It's noted that both feature representations are extracted from final hidden layer of Bi-GRU models.

Since task of intention recognition doesn't have strict requirement for alignment of input features and output intention category label, we propose to perform feature information exchange by concatenating h_t^i and h_t^s, which could be represented as

$$\tilde{h}_t^i = concate(h_t^i, h_t^s) \tag{4}$$

where all feature expression and concatenating processes can be regarded as operations for the encoding structure.

After feature concatenating operation, we introduce a CNN model for further processing, which consists of one convolution layer and a global pooling layer. After processing of CNN model, another layer of GRU network is added at the top of upper structure, which works together with CNN model as decoding structure. Specifically, output of GRU network s_t^i can be represented as

$$g_t^i = \phi_i(g_{t-1}^i, f_C(\tilde{h}_t^i)) \tag{5}$$

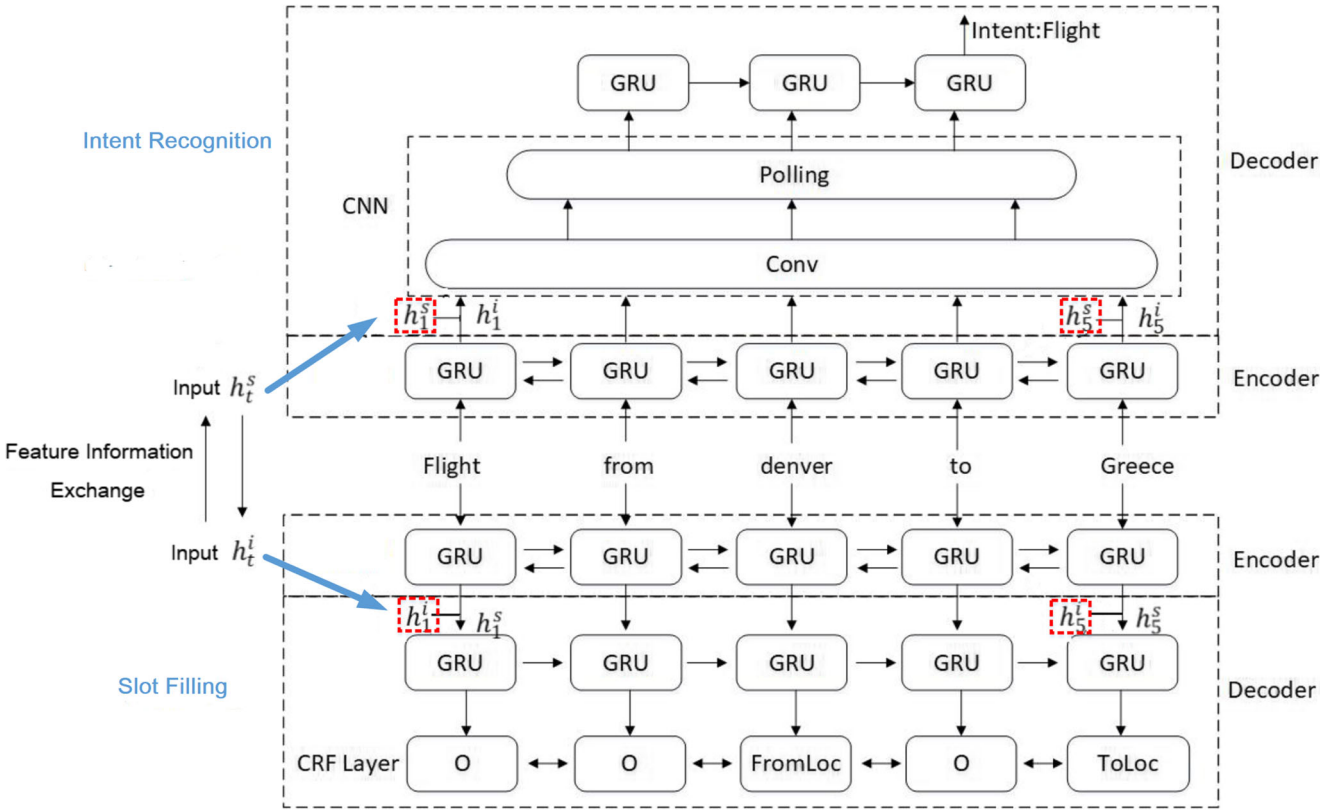

Fig. 2 Design of the proposed dual structure based on Bi-GRU, CNN and CRF models, where the upper part performs subtask of intent recognition by Bi-GRU and CNN models, the lower part is responsible for slot filling by Bi-GRU and CRF models, and we put an example of input question in the middle part. It's noted that both upper and lower parts are designed in an encoder and decoder manner, and a novel information exchange design is used for feature fusion between two subtask

where function $\phi_i()$ represents the operations of GRU cells, g_t^i and g_{t-1}^i refer to output of GRU network at current state t and the last state $t-1$, and function $f_C()$ refers to operations in the adopted CNN model. Since we suppose the input question have T words or T states, we can calculate g_T^i for final feature output, which can be further used to output intention category label via classier of logical regression. It's noted all these operations of CNN and top GRU network can be viewed as decoding structure for semantical meanings of input questions.

In the designed structure for slot filling subtask, we introduce a decoding layer of GRU network for feature alignment task after feature extraction in Bi-GRU encoding layer, which could be represented as

$$g_t^s = \phi_s(g_{t-1}^s, h_t^s, h_t^i) \tag{6}$$

where function $\phi_s()$ represents the operations of GRU cells, g_t^s and g_{t-1}^s refer to output of GRU network at current state t and the last state $t-1$.

After feature alignment, a CRF model is introduced for further processing, where state transition matrix in CRF model could effectively utilize former and later semantic labels to determine meaning of current word. In fact, introducing CRF for jointly decoding could improve accuracy of slot filling by modeling relationship among neighboring labels, thus accurately defining a proper label chain for the entire sentence in global manner. For example, label of time can't be determined to follow label of location, due to the general order of persons' talking habit.

Specifically, we firstly construct initial score matrix G based on output of the decoding GRU network g_t^s:

$$G_{t,y} = \varphi(g_t^s, y), where\ t = 1, ..., T, y = 1, ..., K \tag{7}$$

where t and y are the indexes for row and column of matrix G respectively, N and K are total numbers of words and label categories respectively, and function $\varphi()$ is designed to transform feature information into initial classifying category score by a layer of fully-connected network.

Afterwards, we consider start and end of sentences as two categories of labels with index 0 and $K+1$, where the total number of label set is increased to $K+2$. We thus define label sequence for the input sentence X as $Y = \{y_1, y_2, ..., y_T\}$, where each label could vary from 0 to $K+1$.

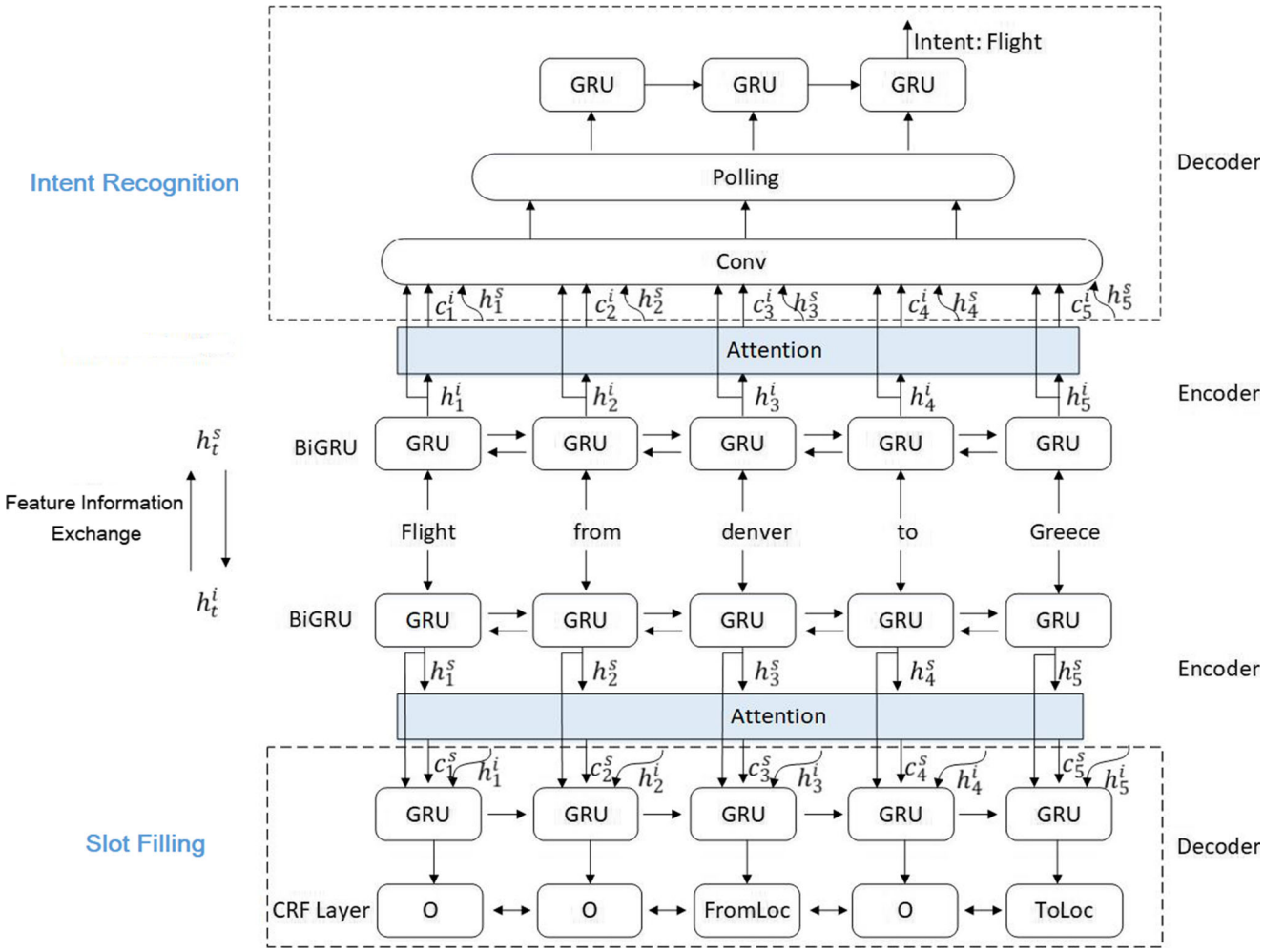

Fig. 3 Design of the proposed dual structure with two attention schemes, where attention schemes are designed to enhance feature representation ability by modeling text context information

The probability score for X to be classified as a specified label sequence $\tilde{Y}$ could be represented as:

$$S(X, \tilde{Y}) = \sum_{t=1}^{T} M(y_t, \tilde{y}_t) + \sum_{t=1}^{T} G_{t,\tilde{y}_t}, \tag{8}$$

where $\tilde{y}_t$ refers to settled category label for the tth word in $\tilde{Y}$, M is defined as transition score matrix, and $M_{y_t,\tilde{y}_t}$ refers to the corresponding score by modifying label of the tth word from y_t to $\tilde{y}_t$. It's noted matrix size for M is determined as $(K+2) \times (K+2)$ after increasing label set.

After calculating probability score for all possible label sequences, we output the corresponding label sequence as result of slot filling with the maximum score, which could be represented as:

$$Y^* = \arg\max_{Y^*} S(X, Y^*) \tag{9}$$

3.2 Design of dual structure with attention scheme

On the basis of the proposed dual structure, we propose a proper attention scheme, which could enhance representative ability of generated feature by involving additional context information.

We show the detailed design of dual structure with attention scheme in Fig. 3, where two attention schemes are introduced in the coding layer of intention recognition and slot filling parts, respectively. In fact, both attention schemes are designed with the same intention, i.e., add additional text context information in features for higher accuracy. We show structure design of the propose attention scheme in Fig. 4, where context information is involved for further processing by rewriting Eq. 5 and 6 as

$$g_t^i = \phi_i(g_{t-1}^i, f_C(\tilde{h}_t^i), c_t^i) \tag{10}$$

$$g_t^s = \phi_s(g_{t-1}^s, h_t^s, h_t^i, c_t^s) \tag{11}$$

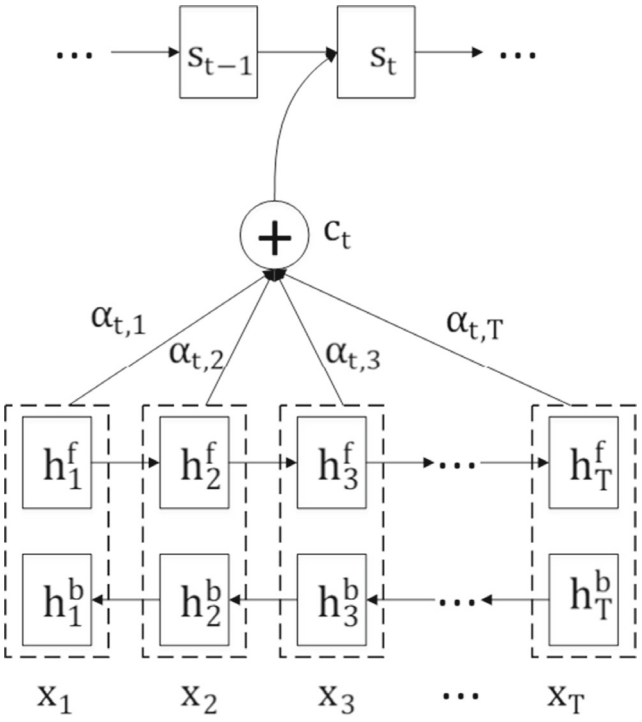

Fig. 4 Structure design of the proposed attention scheme to involve context information

where c_t^i and c_t^s represent the involving context information for intent recognition and slot filling parts, respectively.

Since both attention schemes are designed with similar idea, we focus on explaining the general procedures to compute c_t, where c_t could be regarded as weighted temporal feature representation:

$$c_t = \sum_{j=1}^{T} \alpha_{t,j} h_j \tag{12}$$

where j is the index for word, and $\alpha_{t,j}$ represent weights to estimate importance of the corresponding temporal features. We calculate $\alpha_{t,j}$ by the following equation:

$$\alpha_{t,j} = \frac{exp(e_{t,j})}{\sum_{k=1}^{T} exp(e_{t,k})} \tag{13}$$

$$e_{t,k} = f_g(s_t, h_k) \tag{14}$$

where $e_{t,j}$ calculates relationship between temporal feature of the jth word h_j in encoding layer and signal of tth word s_t in decoding layer, and function $f_g()$ refers to a feed-forward neural network to model relationship. Essentially, $\alpha_{t,j}$ tries to model relationship between the tth word at the decoding layer and the jth word at the encoding layer. In other words, we try to calculate the influence of the jth word in input sequence on the tth word of the target output sequence, which is regarded as text context information for the proposed dual structure.

3.3 Asynchronous training strategy for dual structure

Since the proposed method adopts a dual structure for two subtasks, we propose to train two subtask networks with asynchronous training strategy.

Under the idea of defining both loss functions with cross-entropy loss, we define the loss function for intention recognition network as

$$L_1 = -\sum_{j=1}^{K} y_j^i \log(y_j^i) \tag{15}$$

and loss function for slot filling subtask network as

$$L_2 = -\sum_{t=1}^{T}\sum_{j=1}^{J} y_{t,j}^s \log(y_{t,j}^s) \tag{16}$$

$$y_{t,j}^s = e^{S(X,\tilde{Y})}/Z \tag{17}$$

where superscript i and s indicate intent recognition and slot filling network, K, T and J refer to the total number of category labels, slot values, words in the text respectively, function $S(X, \tilde{Y})$ corresponds to Eq. 8 for calculation of matching score, and Z sums scores of all possible matches for normalization.

We propose to use asynchronous training strategy for jointly training, thus keeping two loss functions relatively separated. Such strategy could bring two major advantages:

(1) Compared with only using a joint model for both tasks, dual structure with asynchronous training strategy can not only capture more useful information, but also overcome structural limitations of adopting a single model.
(2) Interactive information between two tasks can be learned by sharing the hidden state of networks, i.e., temporal feature replantation, which coincides with the asynchronous training procedures for two loss functions.

Supposing the proposed method is processing the $t-1$th word for training, intent recognition and slot filling network would generate h_{t-1}^i and h_{t-1}^s at the $t-1$th iteration. In the tth iteration, intent recognition network would read a batch of input data x_t, exchanged temporal features for slot filling h_{t-1}^s. After processing of encoder and decoder layers, it would finally predict category label y_t^i for x_t, and calculates L_1 loss. Afterwards, the same batch data x_t and exchanged temporal features for slot filling h_t^i are regarded as input of slot filling network, which finally compute sequence $y_{t,j}^s$ and L_2 loss for time step t. When one of the networks is training, we should keep parameters of the other one fixed, thus achieving fast and converged asynchronous training.

4 Experiment

In this section, we show the effectiveness and efficiency of the proposed model for semantic understanding task. We first introduce dataset. Then, we describe implementation details. Finally, ablation and comparison experiments are conducted to prove the efficiency of the proposed, respectively.

4.1 Datasets

Two public dataset are used in this experiment, i.e., ATIS dataset and PCSF (Electrical Customer Service Dialogue) dataset. The ATIS dataset is an air travel information system dataset. Its corresponding dialogue content is the recording of the flight booker, where the labeling area is "Airline Travel" and it's widely used in the study of semantic understanding. ATIS dataset adopts the commonly used "I / O / B annotation method", which contains a total of 4978 training samples and 893 test samples. ATIS dataset is equipped with 26 intents, 128 slot values, and O tags.

PCSF dataset is a Chinese dataset, which mainly comes from the question-and-answer data pair between a customer service and the user. It's mainly about electric power, including dialogue data such as electric power repair, fault maintenance, electricity fee inquiry and recharge. It's noted that all cases are manually labeled by colleagues in the laboratory. There are 9860 samples in total, including 8381 training samples and 1479 test samples. Among these samples, PCSF dataset contains a total of 12 intent categories including electric power repair, electricity fee inquiry and etc for intent recognition. PCSF dataset contains 76 slot values, including time, location, user name, and etc for slot filling. In the example of power repair, we firstly determine what current intention for conversation, i.e., power repair business, and then we need determine slot values in the dialogue, including user name, user card number and other specific user data. Moreover, we show two examples of labels for questions of PCSF dataset in Fig. 5, where BOS represents beginning of the sentence with label 'O', and EOS represents end of the sentece.

4.2 Implementation details

We adopt the idea of cross-validation for all experiments. We test each model by 6 times and the average value of these total 6 experiments is used for final experimental result. In the experiment, the hidden layer vector size of the decoding layer network is set to 128, and the maximum input length accepted by the model is 60. When the input is greater than this maximum value, a truncation operation is taken in this experiment. Adam optimization algorithm is used to update network parameters. The initial learning rate is set to 0.005. The exponential decay rate β_1 of the first-order moment estimation is set to 0.9, meanwhile the exponential decay rate β_2 of the second-order moment estimation is set to 0.999. In order to prevent division by zero for updating parameters, the parameter ϵ_1 is set to a very small number as 10^{-8}, and batch size is settle as 128 for training data.

4.3 Ablation experiments

In order to prove the improvement of introducing the interactive information between slot filling and intent recognition subtasks, we conduct experiments with five different sets of settings, i.e., Dual(RNN), Dual, Dual+h^s, Dual+h_i, the proposed method, where we adopt RNN network for temporal feature extraction instead of Bi-GRU in setting of Dual(RNN).

Tables 1 and 2 give the ablation experimental results on ATIS and PSCF dataset, respectively. We could observe results achieved by Dual setting are lower than the baseline model in both intent recognition and slot filling tasks. In fact, GRU network greatly contributes to running speed rather than accuracy for both tasks, compared with the original RNN network. Considering the urgent request on computation efficiency to guarantee parallel and online running of custom service applications, we choose GRU network for our basis network to extract temporal features.

By introducing h^s for intent recognition, we can view a large increase in performance of Intent(Acc), when comparing Dual+h^s and Dual. However, h^s doesn't promote performance of slot filling. This phenomenon can be

Question1:	BOS	我	想	投诉	一	下	电工	EOS	Intent
	|	|	|	|	|	|	|		|
Label1:	O	O	O	O	O	O	B_name.complain		**Complain**
Question2:	BOS	为	什么	平桥区	洋河乡	停电	了	EOS	Intent
	|	|	|	|	|	|	|		|
Label2:	O	O	O	B-loc.repair	I-loc.repair	O	O		**Repair**

Fig. 5 Two Examples of labels for questions in PCSF dataset. It's noted that both question are asked in Chinese

Table 1 Results of ablation experiments with different structure settings on ATIS dataset

Method	Slot(P)	Slot(R)	Slot(F1)	Intent(Acc)	Overall(ACC)
Dual(RNN)	95.6	92.8	94.2	91.1	78.9
Dual	94.1	93.3	93.7	90.8	78.6
Dual+h^s	94.3	93.3	93.8	97.6	83.9
Dual+h^i	98.7	97.5	98.1	91.0	84.1
The proposed	**98.8**	**97.6**	**98.2**	**97.7**	**89.6**

Bold values indicate the best performance among all comparative methods

Table 2 Results of ablation experiments with different structure settings on PSCF dataset

Method	Slot(P)	Slot(R)	Slot(F1)	Intent(Acc)	Overall(ACC)
Dual(RNN)	80.2	76.5	78.3	90.2	75.2
Dual	78.4	77.8	78.1	89.7	74.8
Dual+h^s	78.5	77.9	78.2	92.3	82.3
Dual+h^i	88.9	85.4	87.1	89.9	82.9
The proposed	**89.1**	**85.4**	**87.2**	**92.5**	**85.6**

Bold values indicate the best performance among all comparative methods

Table 3 Comparison performance among the proposed method and different comparative studies on ATIS dataset

Method	Slot(F1)	Intent(Acc)	Overall(ACC)
Joint Seq [33]	94.3	92.6	80.7
Attention BiRNN [34]	94.2	91.1	78.9
Slot-Gated Intent [35]	94.8	93.6	82.2
Self-Attentive [36]	95.1	96.8	82.2
Bi-Model [37]	95.5	96.4	85.7
SF-ID Network [38]	95.6	96.6	86.0
The proposed without Atten	97.3	96.9	86.4
The proposed	**98.2**	**97.7**	**89.6**

Bold values indicate the best performance among all comparative methods

Table 4 Comparison performance among the proposed method and different comparative studies on PSCF dataset

Method	Slot(F1)	Intent(Acc)	Overall(ACC)
Joint Seq [33]	80.4	90.3	79.4
Attention BiRNN [34]	78.3	90.2	75.2
Slot-Gated Intent [35]	81.3	91.7	81.2
Self-Attentive [36]	86.6	91.5	83.6
Bi-Model[37]	86.5	90.4	79.3
SF-ID Network [38]	84.6	90.6	84.1
The proposed without Atten	85.7	91.6	84.1
The proposed	**87.2**	**92.5**	**85.6**

Bold values indicate the best performance among all comparative methods

observed as well, when comparing between Dual+h^i and Dual. Both comparisons prove the effectiveness of introduce interactive information for processing. Last but not least, the proposed method introduce h^s and h^i for further improvement, which achieves the best performance in all measurements. Therefore, we believe it's quite beneficial to describe multiple categories of interactive information, i.e., h^s and h^i, for task of semantic understanding.

4.4 Comparison experiments

Tables 3 and 4 give the comparative results of the proposed method and comparative studies on ATIS and PSCF dataset. By comparing between The proposed without Atten and The proposed, we can find great improvement in performance by introducing attention scheme. In fact, text context information is highly effective in recognizing semantic meanings of sentences, which has been proved and applied in many natural language processing applications. The proposed method offers an appropriate attention scheme to enhance feature representation by modeling context information, which works well in dual structure and both subtasks.

By comparing results achieved the proposed method on PSCF and ATIS datasets, we can observe inconsistent performance, where the proposed method achieves worse results on PSCF dataset. Such phenomenon can be explained by the fact that Chinese dataset, i.e., PSCF dataset, generally requires quantity of preprocessing work. Due to its unique language property, Chinese must be segmented from paragraph to words, which can't be well completed with current popular preprocessing methods. Therefore, failure in segmentation leads to the problem of error accumulation in further processing, which greatly affects performance achieved by the proposed method.

Comparing with comparative studies, we can find the proposed method significantly improve performance on all three measurements. Even being not good at dealing with Chinese dataset, the proposed method still achieves better performance than other methods on PSCF dataset. In fact, most former methods prefer to independently model two subtasks or only focus to well performing only one subtask, which leads to the ignorance of iterative information. By introducing and well describing interactive information, the proposed dual structure has proved its effectiveness by experimental results.

5 Conclusion

This paper focuses on joint modeling of intent recognition and slot filling subtasks with technologies of deep learning and Cloud/Edge computing. In order to make full use of interactive information between two subtasks, a dual structure based on Bi-GRU, CNN and CRF models is proposed to describe interactive information. Afterwards, we introduce attention scheme on the dual structure to involve text context information for feature enhancement. Finally, an asynchronous training strategy is proposed to train two networks. Both ablation and comparative experimental results on an English and a Chinese dataset show the efficiency of the proposed model, comparing with several promising methods.

In the future, we would improve the proposed network with two novel features, i.e., ability to deal with complex sentences for desirable user experience and light-weight version to work on low-resource systems. We would like to encode context information from dialogue or other helpful information for better performance. Facing complexity of users' variant questions, the proposed network should offer specific and accurate answer after analyzing sentences on cloud based architecture. Moreover, the constructed online custom service should suggest users with currently accessible solutions to deal with questions which can't be understood automatically. With such workflow, we believe the proposed network could not only largely decreases manual burden of custom service, but also provides a complete solution with features of real-time feedback and high intelligence for solving users' problems. Current version of the proposed network requires large computation resource, which couldn't work well in low-resource platforms. Facing difficulty that some small companies couldn't afford enough equipments, it's essential to develop a light version of the proposed method, thus being fit with commercial usage on specific domains, such as electricity, telephone, and TV custom services.

Funding This work was supported by National Key R&D Program of China under Grant 2018YFC0407901, the Fundamental Research Funds for the Central Universities under Grant B200202177, the Natural Science Foundation of China under Grant 61702160, the Natural Science Foundation of Jiangsu Province under Grant BK20170892.

Data Availability The data used to support the findings of this study were supplied by Wenqin Mao under license and so cannot be made freely available. Requests for access to these data should be made to Yirui Wu (wuyirui@hhu.edu.cn).

Declarations

Conflicts of Interest The authors declare that they have no conflicts of interest.

References

1. Qi L, Wang X, Xu X, Dou W, Li S (2020) Privacy-aware cross-platform service recommendation based on enhanced locality-sensitive hashing. IEEE Trans Netw Sci Eng:1–1

2. Xu X, Shen B, Yin X, Khosravi MR, Wu H, Qi L, Wan S (2020) Edge server quantification and placement for offloading social media services in industrial cognitive iov. IEEE Trans Ind Inf 99:11. https://doi.org/10.1109/TII.2020.2987994
3. Qi L, He Q, Chen F, Zhang X, Dou W, Ni Q (2020) Data-driven web apis recommendation for building web applications. IEEE Trans Big Data:1–1
4. Qi L, He Q, Chen F, Dou W, Wan S, Zhang X, Xu X (2019) Finding all you need: Web apis recommendation in web of things through keywords search. IEEE Trans Comput Soc Syst 6(5):1063–1072
5. Xu X, Liu X, Xu Z, Dai F, Zhang X, Qi L (2020) Trust-oriented iot service placement for smart cities in edge computing. IEEE Internet Things J 7(5):4084–4091
6. Weizenbaum J (1966) Eliza-a computer program for the study of natural language communication between man and machine. Commun ACM 9(1):36–45
7. Simmons RF (1967) Answering english questions by computer. Autom Lang Process 8(1):253
8. Xu X, Mo R, Dai F, Lin W, Wan S, Dou W (2020) Dynamic resource provisioning with fault tolerance for data-intensive meteorological workflows in cloud. IEEE Trans Ind Inf 16(9):6172–6181
9. Xu X, Wu Q, Qi L, Dou W, Tsai S-B, Bhuiyan ZA (2020) Trust-aware service offloading for video surveillance in edge computing enabled internet of vehicles. IEEE Trans Intell Transp Syst:1–10. https://doi.org/10.1109/TITS.2020.2995622
10. Xu X, Zhang X, Liu X, Jiang J, Qi L, Bhuiyan MZA (2020) Adaptive computation offloading with edge for 5g-envisioned internet of connected vehicles. IEEE Trans Intell Transp Syst:1–10. https://doi.org/10.1109/TITS.2020.2982186
11. Ravuri SV, Stolcke A (2015) Recurrent neural network and LSTM models for lexical utterance classification. In: Proceedings of 16th Annual Conference of the International Speech Communication Association, pp 135–139
12. Xia C, Zhang C, Yan X, Chang Y, Yu PS (2018) Zero-shot user intent detection via capsule neural networks. In: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, pp 3090–3099
13. Singh RR, Miller T, Newn J, Velloso E, Vetere F, Sonenberg L (2020) Combining gaze and AI planning for online human intention recognition. Artif. Intell. 284:103275
14. Chiu JPC, Nichols E (2016) Named entity recognition with bidirectional lstm-cnns. Trans Assoc Comput Linguist 4:357–370
15. Ma X, Hovy EH (2016) End-to-end sequence labeling via bi-directional lstm-cnns-crf. In: Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics
16. Liu T, Yao J-G, Lin C-Y (2019) Towards improving neural named entity recognition with gazetteers. In: Proceedings of the 57th Conference of the Association for Computational Linguistics, pp 5301–5307
17. Liu Z, Winata GI, Xu P, Fung P (2020) Coach: A coarse-to-fine approach for cross-domain slot filling. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pp 19–25
18. Zhu S, Zhao Z, Ma R, Yu K (2020) Prior knowledge driven label embedding for slot filling in natural language understanding. IEEE ACM Trans Audio Speech Lang Process 28:1440–1451
19. Zhang X, Wang H (2016) A joint model of intent determination and slot filling for spoken language understanding. In: IJCAI, vol 16, pp 2993–2999
20. Liu B, Lane I (2016) Attention-based recurrent neural network models for joint intent detection and slot filling. In: Proceedings of 17th Annual Conference of the International Speech Communication Association, pp 685–689
21. Chen S, Yu S (2019) WAIS: word attention for joint intent detection and slot filling. In: Proceedings of The Thirty-Third AAAI Conference on Artificial Intelligence, pp 9927–9928
22. Yin W, Kann K, Yu M, Schütze H (2017) Comparative study of cnn and rnn for natural language processing. arXiv:1702.01923
23. Xu H, Saenko K (2016) Ask, attend and answer: Exploring question-guided spatial attention for visual question answering. In: Proceedings of European Conference on Computer Vision, vol 9911, pp 451–466
24. Yang Z, He X, Gao J, Deng L, Smola AJ (2016) Stacked attention networks for image question answering. In: Proceedings of Computer Vision and Pattern Recognition, pp 21–29
25. Anderson P, He X, Buehler C, Teney D, Johnson M, Gould S, Zhang L (2018) Bottom-up and top-down attention for image captioning and visual question answering. In: Proceedings of Computer Vision and Pattern Recognition, pp 6077–6086
26. Woo S, Park J, Lee J-Y, So Kweon I (2018) Cbam: Convolutional block attention module. In: Proceedings of European Conference on Computer Vision, pp 3–19
27. Cao Y, Xu J, Lin S, Wei F, Hu H (2019) Gcnet: Non-local networks meet squeeze-excitation networks and beyond. In: Proceedings of International Conference on Computer Vision Workshops, pp 1971–1980
28. Zhao X, Sang L, Ding G, Han J, Di N, Yan C (2019) Recurrent attention model for pedestrian attribute recognition. In: Proceedings of AAAI Conference on Artificial Intelligence, pp 9275–9282
29. Zhao T, Wu X (2019) Pyramid feature attention network for saliency detection. In: Proceedings of Computer Vision and Pattern Recognition, pp 3085–3094
30. Liu S, Huang D, Wang Y (2018) Receptive field block net for accurate and fast object detection. In: Proceedings of European Conference on Computer Vision, vol 11215, pp 404–419
31. Li X, Wang W, Hu X, Yang J (2019) Selective kernel networks. In: Proceedings of Computer Vision and Pattern Recognition, pp 510–519
32. Xiao Y, Xue M, Lu T, Wu Y, Palaiahnakote S (2019) A text-context-aware CNN network for multi-oriented and multi-language scene text detection. In: Proceedings of International Conference on Document Analysis and Recognition, pp 695–700
33. Hakkani-Tür D, Tür G, Celikyilmaz A, Chen Y-N, Gao J, Deng L, Wang Y-Y (2016) Multi-domain joint semantic frame parsing using bi-directional rnn-lstm. In: Proceedings of Annual Conference of the International Speech Communication Association, pp 715–719
34. Liu B, Lane I (2016) Attention-based recurrent neural network models for joint intent detection and slot filling. arXiv:1609.01454
35. Goo C-W, Gao G, Hsu Y-K, Huo C-L, Chen T-C, Hsu K-W, Chen Y-N (2018) Slot-gated modeling for joint slot filling and intent prediction. In: Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers), pp 753–757
36. Li C, Li L, Qi J (2018) A self-attentive model with gate mechanism for spoken language understanding. In: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, pp 3824–3833

37. Wang Y, Shen Y, Jin H (2018) A bi-model based rnn semantic frame parsing model for intent detection and slot filling. arXiv:1812.10235
38. Haihong E, Niu P, Chen Z, Song M (2019) A novel bi-directional interrelated model for joint intent detection and slot filling. In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pp 5467–5471