

PIN: A Novel Parallel Interactive Network for Spoken Language Understanding

1st Peilin Zhou
ADSPLAB
Peking University
Shenzhen, China
zhoupl@pku.edu.cn

2nd Zhiqi Huang
ADSPLAB
Peking University
Shenzhen, China
zhiqihuang@pku.edu.cn

3rd Fenglin Liu
ADSPLAB
Peking University
Shenzhen, China
fenglinliu98@pku.edu.cn

4th Yuexian Zou*
ADSPLAB
Peking University
Shenzhen, China
zouyx@pku.edu.cn

Abstract—Spoken Language Understanding (SLU) is an essential part of the spoken dialogue system, which typically consists of intent detection (ID) and slot filling (SF) tasks. Recently, recurrent neural networks (RNNs) based methods achieved the state-of-the-art for SLU. It is noted that, in the existing RNN-based approaches, ID and SF tasks are often jointly modeled to utilize the correlation information between them. However, we noted that, so far, the efforts to obtain better performance by supporting bidirectional and explicit information exchange between ID and SF are not well studied. In addition, few studies attempt to capture the local context information to enhance the performance of SF. Motivated by these findings, in this paper, Parallel Interactive Network (PIN) is proposed to model the mutual guidance between ID and SF. Specifically, given an utterance, a Gaussian self-attentive encoder is introduced to generate the context-aware feature embedding of the utterance which is able to capture local context information. Taking the feature embedding of the utterance, Slot2Intent module and Intent2Slot module are developed to capture the bidirectional information flow for ID and SF tasks. Finally, a cooperation mechanism is constructed to fuse the information obtained from Slot2Intent and Intent2Slot modules to further reduce the prediction bias. The experiments on two benchmark datasets, i.e., SNIPS and ATIS, demonstrate the effectiveness of our approach, which achieves a competitive result with state-of-the-art models. More encouragingly, by using the feature embedding of the utterance generated by the pre-trained language model BERT, our method achieves the state-of-the-art among all comparison approaches.

I. INTRODUCTION

Spoken Language Understanding (SLU) technology plays a crucial part in goal-oriented dialogue systems. It typically involves intent detection (ID) and slot filling (SF) tasks. As the names imply, intent detection aims to identify users' intents, while slot filling focuses on capturing semantic constituents from user utterances [1]. As shown in Fig. 1, given a user query 'Book a restaurant on next fall for 5', which is sampled from the SNIPS dataset [2], intent BookRestaurant is assigned to the whole sentence, and each token in the sentence corresponds to one specific slot type. Due to the process interdependence between SLU and subsequent dialogue components, such as the dialogue manager and the natural language generator, performance on these two tasks, i.e., ID and SF, determines the upper limit for the utility of such dialogue system [3].

* is corresponding author.

Utterance	Book	a	restaurant	on	next	fall	for	5
Slots	O	O	B-restaurant_type	O	B-timeRange	I-timeRange	O	B-party_size_number
Intent	BookRestaurant							

Fig. 1: An example of an utterance with BIO format annotations for slot filling, which indicates the slot of restaurant type, time range, and party size number from an utterance with an intent BookRestaurant.

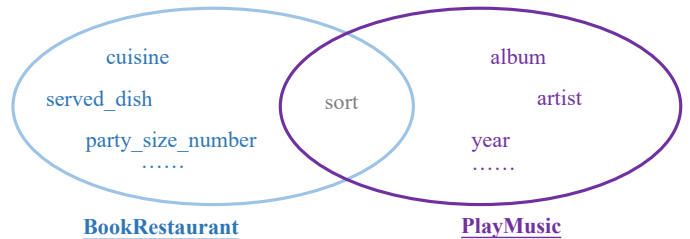


Fig. 2: An example of the co-occurrence characteristic between slot tags and intent labels. The underlined text represents the intent label, and texts inside the circle represent the slot tags correspond to that intent.

Intuitively, intent detection and slot filling are associated with each other [4], which can be observed in Fig. 2. For instance, when the intent of an utterance is PlayMusic, the slots of the utterance are more likely to be artist rather than cuisine, and vice versa (i.e., The more topic-related slots appear in an utterance, the more likely the intent of the utterance corresponds to that topic.). As the accumulation of annotated queries, the co-occurrence characteristic between slot tags and intent labels can become more prominent and perceptible, providing hints about the mutual dependence of ID and SF. Hence, it is promising to achieve a complementary effect by modeling the two tasks in a joint fashion and sharing knowledge between them. Some works [5]–[7] proposed to model intent-slot relation by sharing parameters, outperforming previous separated models by a large margin. More recently, gate mechanism and attention mechanism were also introduced to the RNN-based models [4] [8], which provides a new perspective for joint ID and SF modeling.

However, these methods still suffer from various limitations.

For one thing, local context information is not fully exploited in their models, ignoring the intuition that local context is a useful architectural inductive prior for SF. For another thing, most methods fail to take full advantage of the supervised signals due to their implicit or unidirectional modeling style of the intent-slot relations. Those limitations will hinder the further improvement of SLU systems, especially the overall accuracy, which highly depends on the joint performance of ID and SF.

In this work, we propose a novel Parallel Interactive Network (PIN) to address above issues. For the first issue, a Gaussian self-attentive encoder [9] is introduced to better capture local structure and contextual information at each token, which incorporates valuable inductive prior knowledge for SF. For the second issue, we design a Intent2Slot module and a Slot2Intent module to model the bidirectional information flow between SF and ID. Specifically, inspired by the Dual Process Theory (DPT) [10] in neurocognitive science, we divide the information processing in these modules into two stages: the implicit interaction stage and the explicit interaction stage. These two stages correspond to two different processing styles in which the human brain operates: implicit (intuitive), unconscious learning and explicit (rational), conscious learning. In the implicit interaction stage, the relationships between intents and slots are implicitly captured in the parameters of the shared encoder, which is then utilized by the intuitive decoders to obtain token-level intent distribution and slot label distribution. In the explicit interaction stage, those distribution information obtained in former stage is explicitly utilized by rational decoders to reduce the solution space. Finally, a cooperation mechanism, which comprehensively considers information from above two stages, is performed to reduce the prediction bias and thereby improve the precision and accuracy of model predictions.

To verify the effectiveness of our proposed method, we conduct experiments on two real-world datasets, i.e., ATIS [11] and SNIPS [2], which are popularly used as benchmarks in recent works. Empirical results show that our method achieves competent performance on intent error rate, slot F1-score, and sentence-level semantic frame accuracy compared with other baselines. In addition, Bidirectional Encoder Representation from Transformer (BERT) [12] is explored to further improve the performance of our model.

In summary, the key contributions are as follows:

- We propose a novel parallel interactive network (PIN), which divides the mutual guidance between ID and SF into two interaction stages, i.e., implicit interaction stage and explicit interaction stage, to improve the performance and interpretability of our approach.
- We propose a novel cooperation mechanism within the PIN model in order to effectively combine and balance the information provided by the two interaction stages. It can further refine the prediction results of the proposed model and alleviate the error propagation problem.
- We validate our approach on two benchmark datasets. The experimental results demonstrate the effectiveness of

our approach, which outperforms all comparison methods in terms of most metrics on the two publicly benchmark datasets.

II. RELATED WORK

A. Intent Detection

In recent years, most researchers treat intent detection as a Semantic Utterance Classification (SUC) problem [13]. The early traditional method is to use rule-based templates [14] to match the most appropriate intent, which is extremely inflexible and effort-consuming if intent categories changed. Some statistical-based methods, such as Naive Bayes [15], Adaboost [16], Support Vector Machine (SVM) [17] and logistic regression [18], have also been explored to improve the performance of intent detection. But these approaches still have difficulty in understanding the deep semantics of user utterances due to the ambiguity and irregularity of user expressions. With the rise of deep learning techniques, many neural network based models are proposed to better solve this classification problem. [19] attempted to use Deep belief networks (DBNs) in call routing classification. [20]–[22] proposed using CNN to extract features of sentences and has achieved excellent results. More recently, [23] adopted GRU and LSTM to capture the long-range dependency between words, which showed excellent performance on the intent detection problem. Besides, [24] demonstrated that capsule network could also be applied to this task.

B. Slot Filling

Typically, slot filling is often regarded as a sequence labelling task. Compared with sentence-level intent detection, word-level slot filling heavily relies on fine-grained semantic features. Traditional rule-based approaches directly extract semantic concepts from the user utterance based on hand-crafted rules [25]. They are in the throes of poor generalization and bad transferability. Another line of works centered on feature-based supervised learning algorithms [26], such as Maximum Entropy Models [27], Hidden Markov Models (HMM) [28], and Conditional Random Fields (CRF) [29]. Though achieving better performance than rule-based models, they still suffered from designing elaborate features, which is obviously time-consuming and tedious. Recently, DL-based models for SF have taken a leading role and achieved state-of-the-art results. Bi-directional and hybrid RNN [30], deep LSTM [31], RNN-EM [32], attention-based encoder-decoder [33], CNN [34], and joint pointer and attention [35] are some typical works of this research direction.

C. Joint Modelling

Intuitively, intents and slots represent the semantics of user actions from different granularity perspectives. Hence, there have been some works attempting to jointly model intent detection and slot filling. Hence, how to better model the interactions between slots and intents is the crux of the matter. [5] proposed using an attention-based neural network to jointly model the two tasks. Without directly exchanging

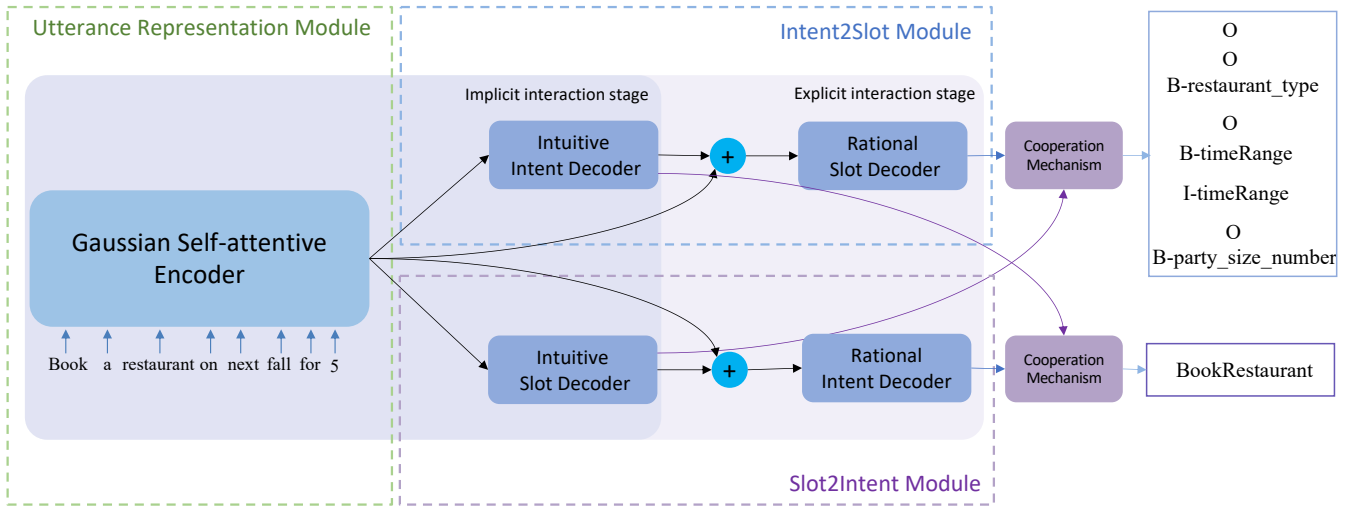


Fig. 3: Illustration of our Parallel Interactive Network (PIN) for joint intent detection and slot filling. The PIN consists of the Utterance Representation Module, the Intent2Slot Module and the Slot2Intent Module. For a given utterance, the Utterance Representation module will first read and encode it as context-aware text representation, which is then fed to the Intent2Slot Module and Slot2Intent Module to model the interaction between ID and SF in both implicit and explicit manners. Finally, a cooperation mechanism is constructed to fuse the information obtained from Slot2Intent and Intent2Slot modules to further reduce the prediction bias.

information between intents and slots, it simply relied on a joint loss function to “implicitly” consider both cues. [4] introduced a slot-gated mechanism to improve slot filling by conditioning on the learned intent results. [8] proposed using intent-augmented embedding as a special gate function to guide the process of slot filling. [36] proposed a Stack-Propagation framework to explicitly incorporate token-level intent information to slot filling, which can further reduce the error propagation. However, these unidirectional interaction approaches ignored the fact that slot filling results are also instructive to the intent detection task. Thus it is necessary to consider the cross-feeding interactions between intents and slots for these two tasks. For instance, [37] proposed a bi-directional interrelated model to achieve the effects of mutual enhancement. [38] adopted a hierarchical capsule network to leverage the hierarchical relationships among words, slots, and intents in an utterance. In contrast to these lines of work, we propose to explicitly model the mutual support in a parallel manner, which is more straightforward and does better in capturing the complicated interdependence among slots and intents. Moreover, empirical results on two benchmark datasets demonstrate the effectiveness of our method, which can further improve the performance of SLU systems.

III. APPROACH

A. Utterance Representation Module

In the utterance representation module, we use BiLSTM with Gaussian self-attention mechanism to leverage both advantages of local structure and contextual information for a given utterance, which are useful for ID and SF tasks.

The BiLSTM [39], which consists of a forward LSTM and a backward LSTM, transforms each input utterance $\mathbf{X} =$

(x_1, x_2, \dots, x_T) into hidden states $\mathbf{H} = (h_1, h_2, \dots, h_T)$ as follows:

$$\vec{h}_t = \overrightarrow{LSTM}(\phi^{\text{emb}}(x_i), \vec{h}_{t-1}) \quad (1)$$

$$\overleftarrow{h}_t = \overleftarrow{LSTM}(\phi^{\text{emb}}(x_i), \overleftarrow{h}_{t+1}) \quad (2)$$

$$h_t = \vec{h}_t \oplus \overleftarrow{h}_t \quad (3)$$

where $\phi^{\text{emb}}(x_i)$ is the embedding of word x_i and \oplus is concatenation operator.

Recently, self-attention mechanisms have found broad application in various NLP tasks such as machine translation [40], language understanding [12], [41], and extractive summarization [42]. However, the vanilla self-attention mechanism used in Transformer [40] treats the same words almost equally without considering their positions, which contradict with our intuition that adjacent words contribute more semantically to central words [9]. As a result, the vanilla self-attention mechanism fails to model the local structure of texts, which can provide useful prior knowledge to the SF task. To mitigate this problem, we apply a Gaussian prior to self-attention mechanisms to better capture both local and contextual information for each token. Formally, we follow the recently proposed Gaussian Transformer [9] to define the Gaussian self-attention function as:

$$c_i = \sum_j \text{Softmax}(-|wd_{i,j}^2 + b| + (x_i \cdot x_j)) x_j \quad (4)$$

$$\mathbf{C} = (c_1, c_2, \dots, c_T) \quad (5)$$

where c_i is the contextual representation (i.e., context vector) for i -th token, x_i and x_j represents the i -th token (i.e., central token) and j -th token from the sentence \mathbf{X} , $d_{i,j}$ represents the

distance between tokens, $|\cdot|$ represents the absolute value, $w > 0, b \leq 0$ are scalar parameters. For more details about Gaussian attention, please refer to [9].

The final utterance representation \mathbf{E} is the concatenation of the output of Gaussian self-attention and BiLSTM:

$$\mathbf{E} = \mathbf{H} \oplus \mathbf{C} \quad (6)$$

where $\mathbf{E} = (\mathbf{e}_1, \dots, \mathbf{e}_T) \in \mathbb{R}^{T \times 2d}$ and \mathbf{e}_i can grasp both local and contextual information for i th token by Gaussian self-attention mechanism. \mathbf{E} is then fed to subsequent Slot2Intent Module and Intent2Slot Module simultaneously.

B. Slot2Intent Module

In our work, the Slot2Intent module is designed to explicitly incorporate slot signals to intent prediction process. Specifically, a intuitive slot decoder is used to decode the utterance representation \mathbf{E} into slot label sequence. Then, these token-level slot features are fed to a rational intent decoder to guide the intent prediction process.

- **Intuitive Slot Decoder.** We use a unidirectional LSTM as the intuitive slot decoder. The decoder's hidden vector at each decoding time step t is calculated as:

$$\mathbf{h}_t^{IS} = LSTM(\mathbf{h}_{t-1}^{IS}, \mathbf{y}_{t-1}^{IS} \oplus \mathbf{e}_t) \quad (7)$$

$$\mathbf{y}_t^{IS} = \text{softmax}(\mathbf{W}_h^{IS} \mathbf{h}_t^{IS}) \quad (8)$$

where \mathbf{h}_{t-1}^{IS} is the previous decoder state, \mathbf{e}_t is the aligned encoder hidden state, \mathbf{W}_h^{IS} are trainable parameters of the model, \mathbf{y}_t^{IS} is the slot label distribution of the t -th token in the utterance.

- **Rational Intent Decoder.** To consider both the utterance representation and slot information, we use another unidirectional LSTM as the rational intent decoder. The encoder generates T hidden states by concatenating the slot output distribution \mathbf{y}_t^{IS} and the aligned encoder hidden state \mathbf{e}_t :

$$\mathbf{h}_t^{RI} = LSTM(\mathbf{h}_{t-1}^{RI}, \mathbf{y}_{t-1}^{RI} \oplus \mathbf{y}_t^{IS} \oplus \mathbf{e}_t) \quad (9)$$

$$\mathbf{y}_t^{RI} = \text{softmax}(\mathbf{W}_h^{RI} \mathbf{h}_t^{RI}) \quad (10)$$

where \mathbf{h}_{t-1}^{RI} is the previous decoder state, \mathbf{y}_t^{IS} is the output of intuitive slot decoder, \mathbf{W}_h^{RI} are trainable parameters of the model, and \mathbf{y}_{t-1}^{RI} is the previously predicted slot label distribution.

C. Intent2Slot Module

As shown in Fig. 3, the Intent2Slot Module has the similar structure as the Slot2Intent Module but switches the tasks for the two decoders. In this case, we introduce a intuitive intent decoder and a rational slot decoder to explicitly incorporate intent signals to slot filling process.

- **Intuitive Intent Decoder.** In practice, We use an unidirectional LSTM as the intuitive intent decoder. The decoder state \mathbf{h}_t^{II} of the t -th token is acquired by directly conditioning on the utterance representation and then utilized for producing intuitive intent features in the same way as intuitive slot decoder.

- **Rational Slot Decoder.** Similar to the rational intent decoder, the rational slot decoder takes the concatenation of the intent output distribution \mathbf{y}_t^{II} and the aligned encoder hidden state \mathbf{e}_t as input units. The rational slot feature \mathbf{h}_t^{RS} and the rational slot output distribution \mathbf{y}_t^{RS} of the t -th token are obtained in the same way as rational intent decoder.

D. Cooperation Mechanism

As described above, we have prepared the intuitive features \mathbf{h}_t^{IS} and \mathbf{h}_t^{II} and the rational features \mathbf{h}_t^{RS} and \mathbf{h}_t^{RI} . It is unreasonable to treat this two different information equally when predicting the intent and slot labels. For example, in the Slot2Intent Module, the wrongly predicted slot label might mislead the rational intent decoder, causing the rational feature \mathbf{h}_t^{RI} less reliable. In this case, the intuitive feature \mathbf{h}_t^{II} should matter more than the rational feature \mathbf{h}_t^{RI} , which could mitigate this error propagation problem. However, if the slot labels are predicted correctly and have strong correlation with some particular intent label, \mathbf{h}_t^{RI} is more important. Hence, we introduce a novel cooperation mechanism to make the model adaptively learn to adjust the balance:

$$r_t^S = \text{softmax}(\text{MLP}(\mathbf{h}_t^{RS})) \quad (11)$$

$$r_t^I = \text{softmax}(\text{MLP}(\mathbf{h}_t^{RI})) \quad (12)$$

$$h_t^S = \mathbf{h}_t^{RS} \odot r_t^S + \mathbf{h}_t^{IS} \odot (1 - r_t^S) \quad (13)$$

$$h^I = \sum_{t=1}^T \mathbf{h}_t^{RI} \odot r_t^I + \mathbf{h}_t^{II} \odot (1 - r_t^I) \quad (14)$$

where MLP represents Multilayer Perceptron, r_t^S and r_t^I are score vectors, h_t^S and h^I represent the fused slot and intent vectors, and \odot denotes element-wise multiplication of vectors.

E. Task Learning

The final slot probability distribution of the t -th token and the intent probability distribution of the whole utterance are predicted by:

$$\mathbf{y}_t^S = \text{softmax}(\mathbf{W}_h^S \mathbf{h}_t^S) \quad (15)$$

$$\mathbf{y}^I = \text{softmax}(\mathbf{W}_h^I \mathbf{h}^I) \quad (16)$$

where \mathbf{W}_h^S and \mathbf{W}_h^I are the parameters to be learned.

TABLE I: Dataset Statistics

Dataset	ATIS	SNIPS
Vocab Size	722	11,241
Average Sentence Length	11.28	9.05
#Slots	120	72
#Intents	21	7
#Training Samples	4,478	13,084
#Development Samples	500	700
#Test Samples	893	700

TABLE II: Experiment results of our model and the baselines on two benchmark datasets.

Model	SNIPS			ATIS		
	Intent (Err)	Slot (F1)	Overall (Acc)	Intent (Err)	Slot (F1)	Overall (Acc)
Recursive NN [43]	2.7	88.3	-	4.6	94.0	-
Dilated CNN, Label-Recurrent [44]	1.7	93.1	-	1.9	95.5	-
Attention Bi-RNN [5]	3.3	87.8	74.1	8.9	94.2	78.9
Joint Seq2Seq [7]	3.1	87.3	73.2	7.4	94.2	80.7
Slot-Gated Model [4]	3.0	88.8	75.5	6.4	94.8	82.2
Stack-Propagation [36]	2.0	94.2	86.9	3.1	95.9	86.5
SF-ID,SF first [38]	2.6	91.4	80.6	2.2	95.8	86.8
SF-ID,ID first [38]	2.7	92.2	80.4	2.9	95.8	86.9
Graph LSTM [45]	2.3	93.8	85.6	3.6	95.8	86.2
PIN (our model)	0.9	94.5	88.0	2.8	95.9	87.1
Joint BERT [46]	1.4	97.0	92.8	2.5	96.1	88.2
Graph LSTM + ELMo [45]	1.7	95.3	89.7	2.8	95.9	87.6
Stack-Propagation + BERT [36]	1.0	97.0	92.9	2.5	96.1	88.6
PIN(our model) + BERT	0.8	97.1	93.2	2.2	96.3	88.8

Therefore, the cross entropy loss for SF and ID can be calculated as:

$$\mathcal{L}_{slot} = - \sum_{j=1}^N \sum_{t=1}^M l_j^{t,S} \log y_j^{t,S} \quad (17)$$

$$\mathcal{L}_{intent} = - \sum_{j=1}^N l_j^I \log y_j^I \quad (18)$$

where $l_j^{t,S}$ and l_j^I denote the ground truth label of slot and intent respectively, N is the number of training samples of each dataset, M is the number of slot labels, which depends on the length of each input utterance.

The training goal of this network is to minimize a joint loss function:

$$\mathcal{L} = \lambda \mathcal{L}_{slot} + (1 - \lambda) \mathcal{L}_{intent} \quad (19)$$

where λ is hyperparameter.

IV. EXPERIMENTS AND ANALYSIS

A. Datasets

To evaluate the effectiveness of our proposed model, we conduct experiments on two benchmark datasets: the ATIS (Airline Travel Information Systems) dataset [11] and the SNIPS dataset [2]. Table I shows the statistics of all datasets.

- **The ATIS dataset** contains transcribed audio recordings of people making flight reservations. It is annotated with 21 intent types and 120 slot types. Following the same data division as in [4], we use 4,478 utterances for training set, 500 utterances for validation set and another 893 utterances for test set.
- **The SNIPS dataset** is a balanced real-world dataset with about 2400 utterances per each of 7 intents, which is collected from the Snips personal voice assistant. There are 72 slot types in the dataset. In this paper, we use the same SNIPS corpus setting as previous related works [4], [38]. The training set contains 13,084 utterances, the validation set and the test set each contain 700 utterances respectively.

B. Evaluation Metrics

Three evaluation metrics are used as our choice to measure the quality of our proposed model. The performance of intent detection is measured by intent error rate, while slot filling is evaluated with F1-score. Furthermore, the sentence accuracy is utilized to indicate the overall performance of the SLU system, which depends on the coordination between the two tasks.

C. Training Details

The model was implemented in PyTorch and trained on a single NVIDIA GeForce GTX 2080TI GPU. We preprocess the datasets following [4]. The dimension of the word embedding is 512 for SNIPS dataset and 256 for ATIS dataset. The hidden size of the encoder is set to 256. We train our model with an Adam optimizer [47], with a learning rate of 0.001. In order to reduce overfit, we apply L2 regularization to our model, with a decay rate of 1×10^{-6} . We set the batch size to 16, the teacher forcing rate to 0.9, the dropout [48] rate to 0.4, and loss weight λ to 0.5. During training process, early-stop strategy is performed on both datasets.

D. Baselines

We compare our model against the following baselines:

- **Recursive NN** [43] introduced recursive neural networks to perform ID and SF together in one model, using purely lexical features and the parse tree.
- **Dilated CNN, Label-Recurrent** [44] proposed a label-recurrent and dilated convolution-based model to incorporate long-distance context hierarchically for joint SF and ID.
- **Attention Bi-RNN** [5] proposed an attention-based bi-directional RNN model to jointly model SF and ID. They explored various strategies to leverage explicit alignment information in the encoder-decoder models.
- **Joint Seq2Seq** [7] adopted multi-task learning method to jointly model SF and ID in a single bi-directional RNN with LSTM cells.

TABLE III: Ablation experiments on two benchmarks to investigate the impacts of various components.

Model	SNIPS			ATIS		
	Intent (Err)	Slot (F1)	Overall (Acc)	Intent (Err)	Slot (F1)	Overall (Acc)
w/o Slot2Intent module	3.1	95.8	86.5	3.1	95.7	86.5
w/o Intent2Slot module	2.0	94.3	87.0	3.0	95.7	86.7
w/o Gaussian self-attention	2.3	92.9	84.4	3.1	94.9	85.0
w/o cooperation mechanism	1.4	94.3	87.4	3.4	95.9	87.0
Full PIN model	0.9	94.5	88.0	2.8	95.9	87.1

- **Slot-Gated Model** [4] introduced a slot-gated mechanism into an attention-based model to improve slot filling by the learned intent context vector.
- **Stack-Propagation** [36] introduced Stack-Propagation framework to better incorporate the intent information for SF task.
- **SF-ID Network** [37] proposed a SF-ID network which introduced a interrelated mechanism and iteration mechanism to establish direct connections for ID and SF to help them promote each other mutually. They designed two modes for their model: SF-First and ID-First.
- **Graph LSTM** [45] was the first to introduce a graph-based model to utilize the semantic correlation between slot and intent.
- **Joint BERT** [46] proposed a joint intent detection and slot filling model based on BERT.

The results of baselines are taken from [43], [37] and [36] because we use the same datasets and evaluation metrics.

E. Results

The results across all the models are presented in Table II, where the proposed PIN model outperforms all the comparison methods for both datasets and almost all metrics obtain the improvement, except for the intent error rate on the ATIS.

In the SNIPS dataset, we can see that our model outperforms previous best result by 0.8% in terms of error rate on intent detection, by 0.3% in terms of F1-score on slot filling. It is also noteworthy that our proposed model performs better especially on sentence-level overall accuracy, where the relative improvement is around 1.1% for SNIPS. In ATIS dataset, although we achieve top 1 performance in terms of slot F1-score and sentence-level accuracy, the relative improvement is not as obvious as the SNIPS. We believe the reason lies in the different complexity of the two datasets. It should be noted that, all the intents in ATIS are limited in a single domain where flight information is requested, while those in SNIPS cover multiple topics such as music and weather, adding more diversity to the SNIPS dataset. As a result, SNIPS could provide more hints about the co-occurrence characteristic between intent and slot, which could be better utilized by our proposed model to boost the performances. Hence, compared with ATIS, the experimental results on SNIPS could better reflect the effectiveness of our PIN model. Furthermore, these results can also verify our assumption that intent information and slot information can be simultaneously utilized for better promoting both two tasks.

What is more, we also investigate the effect of incorporating the pre-trained language model to our proposed method. Specifically, we substitute the PIN encoder with the BERT pre-trained model and fine-tune it to further boost the performance. From the last block of Table II, PIN + BERT outperforms all comparison approaches and establishes the new state-of-the-art performances in terms of slot F1-score and overall accuracy. We attribute this to the powerful feature extraction capabilities of BERT pre-trained model. And our model can take advantage of these enriched features to enhance the mutual guidance between ID and SF, thereby further improving performance.

F. Ablation Study

The ablation study is performed to investigate the contribution of each component in our proposed model. We remove some important components used in our model and all the variants are described as follows:

- W/O Slot2Intent module, where no slot information is explicitly incorporated to intent prediction process. But we still use the intuitive intent decoder from the Intent2Slot module to predict the intent of the utterance.
- W/O Intent2Slot module, where no slot information is explicitly provided for slot filling process. Instead, we use the intuitive slot decoder from Slot2Intent module to perform SF task.
- W/O Gaussian self-attention, where no gaussian self-attention is performed in the utterance encoder. The utterance is encoded only by BiLSTM.
- W/O cooperation mechanism, where only the rational features are utilized for the two tasks. The other components keep the same as our model.

Table III presents our results on ATIS and SNIPS. We observe that all these components contribute to the superior performance of our proposed model.

If we remove the Gaussian self-attention from the holistic model, the performance on both tasks drops to some extent, and a more obvious decline can be found for the slot F1-score. We attribute it to the fact that the Gaussian self-attention is able to capture both local and contextual representation for each token, which is beneficial for slot filling task.

Besides, we also analyze the condition where either Slot2Intent Module or the Intent2Slot Module is removed. In other words, only unidirectional assistance is allowed. We can see from the table that the full model, which is a bidirectional one, performs better on most metrics compared with unidirectional variants. We believe the reason is that

the full model does better in exploiting the “synergy effect” between the two tasks with its parallel structure. Specifically, our proposed model can elaborately capture the patterns of co-occurrence characteristic between slots and intents, and simultaneously provide useful guidance for the two tasks, which is conducive to the overall accuracy.

We also examine the effect of the cooperation mechanism in Table III. The results show that adding the cooperation mechanism on top of the decoders can further improve SLU performance. We attribute this to the fact that by adaptively fusing intuitive features and rational features, our model can better control the proportion of contributions at each stage, which is similar to the ensemble method.

V. CONCLUSION AND FUTURE WORK

In this paper, we propose a novel Parallel Interactive Network (PIN) for jointly modeling intent detection and slot filling. In our model, a Gaussian self-attentive encoder is first introduced to better capture the local context information of utterance, then two modules are introduced to model the mutual guidance between ID and SF. Finally, a cooperation mechanism is proposed to further improve the performance and robustness of our proposed PIN. Experiment results on two benchmark datasets show that the proposed PIN achieves competent performance compared with other baselines, demonstrating the effectiveness of our proposed PIN. In addition, by incorporating the pre-trained language model BERT, our method achieves the state-of-the-art among all comparison approaches.

For our future work, we will extend our model to handle cold start problem where few data samples are provided for training process.

ACKNOWLEDGMENT

Special acknowledgements are given to AOTO-PKUSZ Joint Research Center for Artificial Intelligence on Scene Cognition Technology Innovation for its support.

REFERENCES

- [1] G. Tur and R. De Mori, *Spoken language understanding: Systems for extracting semantic information from speech*. John Wiley & Sons, 2011.
- [2] A. Coucke, A. Saade, A. Ball, T. Bluche, A. Caulier, D. Leroy, *et al.*, “Snips voice platform: an embedded spoken language understanding system for private-by-design voice interfaces,” *CoRR*, vol. abs/1805.10190, 2018.
- [3] H. Chen, X. Liu, D. Yin, and J. Tang, “A survey on dialogue systems: Recent advances and new frontiers,” *Acm Sigkdd Explorations Newsletter*, vol. 19, no. 2, pp. 25–35, 2017.
- [4] C.-W. Goo, G. Gao, Y.-K. Hsu, C.-L. Huo, T.-C. Chen, K.-W. Hsu, and Y.-N. Chen, “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, 2018.
- [5] B. Liu and I. Lane, “Attention-based recurrent neural network models for joint intent detection and slot filling,” in *Interspeech 2016, 17th Annual Conference of the International Speech Communication Association, San Francisco, CA, USA, September 8-12, 2016* (N. Morgan, ed.), pp. 685–689, ISCA, 2016.
- [6] X. Zhang and H. Wang, “A joint model of intent determination and slot filling for spoken language understanding,” in *IJCAI*, vol. 16, pp. 2993–2999, 2016.
- [7] D. Hakkani-Tür, G. Tür, A. Celikyilmaz, Y.-N. Chen, J. Gao, L. Deng, and Y.-Y. Wang, “Multi-domain joint semantic frame parsing using bi-directional rnn-lstm,” in *Interspeech*, pp. 715–719, 2016.
- [8] C. Li, L. Li, and J. Qi, “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, 2018.
- [9] M. Guo, Y. Zhang, and T. Liu, “Gaussian transformer: a lightweight approach for natural language inference,” in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, pp. 6489–6496, 2019.
- [10] J. T. Wixted, “Dual-process theory and signal-detection theory of recognition memory,” *Psychological review*, vol. 114, no. 1, p. 152, 2007.
- [11] C. T. Hemphill, J. J. Godfrey, and G. R. Doddington, “The atlas spoken language systems pilot corpus,” in *Speech and Natural Language: Proceedings of a Workshop Held at Hidden Valley, Pennsylvania, June 24-27, 1990*, 1990.
- [12] J. Devlin, M. Chang, K. Lee, and K. Toutanova, “BERT: pre-training of deep bidirectional transformers for language understanding,” in *NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pp. 4171–4186.
- [13] Y. N. Dauphin, G. Tür, D. Hakkani-Tür, and L. P. Heck, “Zero-shot learning and clustering for semantic utterance classification,” in *ICLR (Poster)*, 2014.
- [14] J. Dowding, J. M. Gawron, D. E. Appelt, J. Bear, L. Cherny, R. C. Moore, and D. B. Moran, “GEMINI: A natural language system for spoken-language understanding,” *CoRR*, vol. abs/cmp-lg/9407007, 1994.
- [15] A. McCallum, K. Nigam, *et al.*, “A comparison of event models for naive bayes text classification,” in *AAAI-98 workshop on learning for text categorization*, vol. 752, pp. 41–48, Citeseer, 1998.
- [16] R. E. Schapire and Y. Singer, “Boostexter: A boosting-based system for text categorization,” *Machine learning*, vol. 39, no. 2-3, pp. 135–168, 2000.
- [17] P. Haffner, G. Tur, and J. H. Wright, “Optimizing svms for complex call classification,” in *2003 IEEE International Conference on Acoustics, Speech, and Signal Processing, 2003. Proceedings (ICASSP'03)*, vol. 1, pp. I–I, IEEE, 2003.
- [18] A. Genkin, D. D. Lewis, and D. Madigan, “Large-scale bayesian logistic regression for text categorization,” *technometrics*, vol. 49, no. 3, pp. 291–304, 2007.
- [19] A. Deoras and R. Sarikaya, “Deep belief network based semantic taggers for spoken language understanding,” in *Interspeech*, pp. 2713–2717, 2013.
- [20] Y. Kim, “Convolutional neural networks for sentence classification,” in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP 2014, October 25-29, 2014, Doha, Qatar*, pp. 1746–1751, 2014.
- [21] Y. Zhang and B. C. Wallace, “A sensitivity analysis of (and practitioners’ guide to) convolutional neural networks for sentence classification,” in *IJCNLP(1)*, pp. 253–263, Asian Federation of Natural Language Processing, 2017.
- [22] H. B. Hashemi, A. Asiaee, and R. Kraft, “Query intent detection using convolutional neural networks,” in *International Conference on Web Search and Data Mining, Workshop on Query Understanding*, 2016.
- [23] S. Ravuri and A. Stolcke, “Recurrent neural network and lstm models for lexical utterance classification,” in *Sixteenth Annual Conference of the International Speech Communication Association*, 2015.
- [24] C. Xia, C. Zhang, X. Yan, Y. Chang, and P. S. Yu, “Zero-shot user intent detection via capsule neural networks,” in *EMNLP*, pp. 3090–3099, Association for Computational Linguistics, 2018.
- [25] J.-H. Kim and P. C. Woodland, “A rule-based named entity recognition system for speech input,” in *Sixth International Conference on Spoken Language Processing*, 2000.
- [26] J. Li, A. Sun, J. Han, and C. Li, “A survey on deep learning for named entity recognition,” *arXiv preprint arXiv:1812.09449*, 2018.
- [27] J. N. Kapur, *Maximum-entropy models in science and engineering*. John Wiley & Sons, 1989.
- [28] S. R. Eddy, “Hidden markov models,” *Current opinion in structural biology*, vol. 6, no. 3, pp. 361–365, 1996.
- [29] J. Lafferty, A. McCallum, and F. C. Pereira, “Conditional random fields: Probabilistic models for segmenting and labeling sequence data,” 2001.
- [30] G. Mesnil, X. He, L. Deng, and Y. Bengio, “Investigation of recurrent-neural-network architectures and learning methods for spoken language understanding,” in *Interspeech*, pp. 3771–3775, 2013.

- [31] K. Yao, B. Peng, Y. Zhang, D. Yu, G. Zweig, and Y. Shi, "Spoken language understanding using long short-term memory neural networks," in *2014 IEEE Spoken Language Technology Workshop (SLT)*, pp. 189–194, IEEE, 2014.
- [32] B. Peng, K. Yao, L. Jing, and K.-F. Wong, "Recurrent neural networks with external memory for spoken language understanding," in *Natural Language Processing and Chinese Computing*, pp. 25–35, Springer, 2015.
- [33] E. Simonnet, N. Camelin, P. Deléglise, and Y. Esteve, "Exploring the use of attention-based recurrent neural networks for spoken language understanding," 2015.
- [34] N. T. Vu, "Sequential convolutional neural networks for slot filling in spoken language understanding," in *INTERSPEECH*, pp. 3250–3254, ISCA, 2016.
- [35] L. Zhao and Z. Feng, "Improving slot filling in spoken language understanding with joint pointer and attention," in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 426–431, 2018.
- [36] L. Qin, W. Che, Y. Li, H. Wen, and T. Liu, "A stack-propagation framework with token-level intent detection for spoken language understanding," in *EMNLP/IJCNLP (1)*, pp. 2078–2087, Association for Computational Linguistics, 2019.
- [37] E. Haihong, P. Niu, Z. Chen, and M. Song, "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, 2019.
- [38] C. Zhang, Y. Li, N. Du, W. Fan, and P. S. Yu, "Joint slot filling and intent detection via capsule neural networks," in *ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pp. 5259–5267, 2019.
- [39] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.
- [40] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in neural information processing systems*, pp. 5998–6008, 2017.
- [41] A. Radford, K. Narasimhan, T. Salimans, and I. Sutskever, "Improving language understanding by generative pre-training," URL https://s3-us-west-2.amazonaws.com/openai-assets/researchcovers/languageunsupervised/language_understanding_paper.pdf, 2018.
- [42] Y. Liu, "Fine-tune BERT for extractive summarization," *CoRR*, vol. abs/1903.10318, 2019.
- [43] D. Guo, G. Tür, W. Yih, and G. Zweig, "Joint semantic utterance classification and slot filling with recursive neural networks," in *SLT*, pp. 554–559, IEEE, 2014.
- [44] A. Gupta, J. Hewitt, and K. Kirchhoff, "Simple, fast, accurate intent classification and slot labeling," *CoRR*, vol. abs/1903.08268, 2019.
- [45] L. Zhang, D. Ma, X. Zhang, X. Yan, and H. Wang, "Graph lstm with context-gated mechanism for spoken language understanding," in *AAAI*, 2020.
- [46] Q. Chen, Z. Zhuo, and W. Wang, "BERT for joint intent classification and slot filling," *CoRR*, vol. abs/1902.10909, 2019.
- [47] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015.
- [48] W. Zaremba, I. Sutskever, and O. Vinyals, "Recurrent neural network regularization," *CoRR*, vol. abs/1409.2329, 2014.