# PIN: A Novel Parallel Interactive Network for Spoken Language Understanding

Peilin Zhou, Zhiqi Huang, Fenglin Liu, Yuexian Zou*
ADSPLAB, School of ECE, Peking University

## 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.

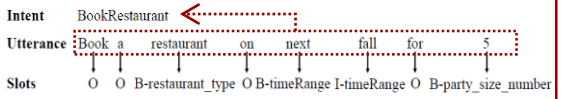


Fig. 1: An example of an utterance with BIO format annotations for slot filling, which indicates the slot of restaurant type, time range, and party size number from an utterance with an intent BookRestaurant.

Intuitively, intent detection and slot filling are associated with each other. Hence, it is promising to achieve a complementary effect by modeling the two tasks in a joint fashion and sharing knowledge between them.

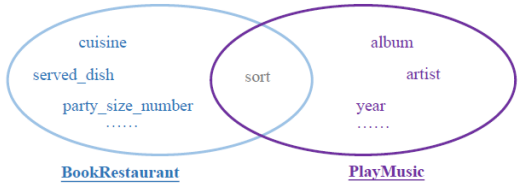


Fig.2: An example of the co-occurrence characteristic between slot tags and intent labels. The underlined text represents the intent label, and texts inside the circle represent the slot tags correspond to that intent.

**Limitation & Challenge:**

- Local context information is not fully exploited in their models, ignoring the intuition that local context is a useful architectural inductive prior for SF.
- Many methods fail to take full advantage of the supervised signals due to their implicit or unidirectional modeling style of the intent-slot relations.

**Solution:**

We propose a novel Parallel Interactive Network (PIN) to address above issues:

- A Gaussian self-attentive encoder is introduced to better capture local structure and contextual information at each token, which incorporates valuable inductive prior knowledge for SF.
- We design a Intent2Slot module and a Slot2Intent module to model the bidirectional information flow between SF and ID.

## Model

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.

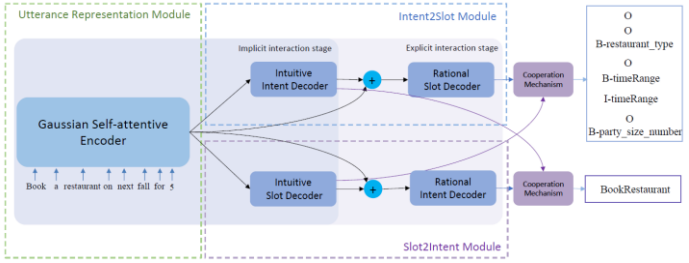


Fig. 3: Illustration of our Parallel Interactive Network (PIN) for joint intent detection and slot filling.

### 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.

$$\mathbf{E} = \mathbf{H} \oplus \mathbf{C}$$

$$\mathbf{H} = (h_1, h_2, \ldots, h_T) \quad \mathbf{C} = (c_1, c_2, \ldots, c_T)$$

$$\overrightarrow{h_i} = \overrightarrow{\text{LSTM}}\left(\phi^{\text{emb}}(x_i), \overrightarrow{h_{i-1}}\right) \quad \overleftarrow{h_i} = \overleftarrow{\text{LSTM}}(\phi^{\text{emb}}(x_i), \overleftarrow{h_{i+1}})$$

$$h_i = \overrightarrow{h_i} \oplus \overleftarrow{h_i}$$

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

### Slot2Intent Module

- Intuitive Slot Decoder

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

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

- Rational Intent Decoder

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

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

### Intent2Slot Module

The Intent2Slot Module has the similar structure as the Slot2Intent Module but switches the tasks for the two decoders.

- Intuitive Intent Decoder
- Rational Slot Decoder

### Cooperation Mechanism

$$r_t^S = \text{softmax}(\text{MLP}(\mathbf{h}_t^{RS})) \qquad r_t^I = \text{softmax}(\text{MLP}(\mathbf{h}_t^{RI}))$$

$$h_t^S = \mathbf{h}_t^{RS} \odot r_t^S + \mathbf{h}_t^{IS} \odot (1 - r_t^S) \qquad h^I = \Sigma_{t=1}^T \quad \mathbf{h}_t^{RI} \odot r_t^I + \mathbf{h}_t^{II} \odot (1 - r_t^I)$$

## Experiments

To evaluate the effectiveness of our proposed model, we conduct experiments on two benchmark datasets: the ATIS dataset and the SNIPS dataset. The results across all the models are presented in Table 1.

| 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** |

TABLE 1: Experiment results of our model and the baselines on two benchmark datasets.

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:

| 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** |

TABLE 2 Ablation experiments on two benchmarks to investigate the impacts of various components.

## Conclusions & Main Contribution

- 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.