Towards Explainable Joint Models via Information Theory for Multiple Intent Detection and Slot Filling

Xianwei Zhuang†, Xuxin Cheng*, Yuexian Zou†
School of ECE, Peking University, China
{xwzhuang, chengxx}@stu.pku.edu.cn, zouyx@pku.edu.cn

Abstract

Recent joint models for multi-intent detection and slot filling have obtained promising results through modeling the unidirectional or bidirectional guidance between intent and slot. However, existing works design joint models heuristically and lack some theoretical exploration, including (1) theoretical measurement of the joint-interaction quality; (2) explainability of design and optimization methods of joint models, which may limit the performance and efficiency of designs. In this paper, we mathematically define the cross-task information gain (CIG) to measure the quality of joint processes from an information-theoretic perspective and discover an implicit optimization of CIG in previous models. Based on this, we propose a novel multi-stage iterative framework with theoretical effectiveness, explainability, and convergence, which can explicitly optimize information for cross-task interactions. Further, we devise an information-based joint model (InfoJoint) that conforms to this theoretical framework to gradually reduce the cross-task propagation of erroneous semantics through CIG iterative maximization. Extensive experiment results on two public datasets show that InfoJoint outperforms the state-of-the-art models by a large margin.

Introduction

Spoken language understanding (SLU) is a critical task in dialog systems (Young et al. 2013), which generally includes two subtasks: intent detection (ID) and slot filling (SF) (Tur and De Mori 2011). Recently, joint models (Goo et al. 2018; Qin et al. 2019; Cheng et al. 2023c; Zhu et al. 2023a,b) for ID and SF have achieved impressive performance, and have proved that there exists a strong correlation between this two tasks (Weld et al. 2022). The recent studies (Gangadharanah and Narayanaswamy 2019) recognize that a single utterance often contains multiple intentions in real-world scenarios, i.e., Multi-Intent SLU. Thus, joint multi-intent SLU gradually attracting increasing attention. One of the mainstream insights works on devising attention mechanisms against joint tasks. AGIF (Qin et al. 2020, 2021b) explore unidirectional joint models from ID to SF based on graph attention (GAT) for multi-intent SLU. Cheng, Yang, and Jia (2023) proposes a variant of attention for reducing error propagation. Co-guiding Net (Xing and Tsang 2022) proposes heterogeneous attention (HGAT) to achieve bidirectional interaction. These works heuristically design cross-task interaction modules to improve performance.

Although promising progress has been made in previous works, these studies regrettably lack exploration of some core issues at the theoretical level, including:

(1) Why does the joint SLU model work better and how to quantitatively measure the enhancement quality of joint models in theory? Our work innovatively defines the cross-information gain (CIG) to measure the interaction quality of the beneficial information between ID and SF, and explains that the process of joint optimization is the process of maximizing CIG and minimizing the information gap between dual-task branches. Based on this perspective, we discover that previous works implicitly improve CIG although they do not explicitly mention CIG (i.e., there is a positive correlation between CIG and performance, as shown in Fig. 1).

(2) How to design an explainable model for the joint processes? Existing joint models are essentially single-stage models (Fig. 2a and b) which perform single interaction, and lack theoretical guidance. We construct the model as a
multi-stage Markov model (Fig. 2c) and claim the method can iteratively optimize in the positive direction and outperform single-stage models under some theoretical constraints. We further analyze the theoretical constraints for its effectiveness and convergence.

Furthermore, we devise a novel information-based joint model (InfoJoint) which satisfies the above theoretical constraints. Specifically, InfoJoint utilizes multi-level cross-task contrastive learning to maximize CIG. In addition, directional constraints and entropy weighting strategies are developed to facilitate positive optimization and balance learning in each batch, respectively. Finally, InfoJoint optimizes CIG through iterative training to achieve deep interaction and gradually eliminate error propagation between the joint tasks. In summary, our framework has several appealing facets: (1) **Explainability:** Joint models are guided by theory to fully explore relevance through iterative enhancement. (2) **Convergence:** The convergence of joint optimization is theoretically guaranteed. (3) **Universality:** This framework is architecture-independent (component-independent) and compatible with previous works. The main contributions of this paper are presented as follows:

1. **(Theory)** We propose a novel explainable multi-stage joint framework for multi-intent SLU with theoretical quantifiability, effectiveness, and convergence.
2. **(Methodology)** Based on the information-based principled framework, we devise an iterative-enhancement model termed InfoJoint, which adopts multi-level cross-task contrastive learning, directional constraints, and entropy weighting to achieve the effective interaction and performance improvement.
3. **(Experiments)** Extensive experiments on two public multi-intent SLU datasets MixATIS and MixSNIPS demonstrate that InfoJoint significantly outperforms the best baselines in terms of all evaluation metrics.

**Related Work**

**Joint model for intent detection and slot filling.** Early studies (Yao et al. 2014; Kurata et al. 2016; Zhang and Wang 2016) recognize that there is a close connection between ID and SF. Motivated by this, some studies (Goo et al. 2018; E et al. 2019; Liu et al. 2019a; Qin et al. 2019; Zhang et al. 2019; Wu et al. 2020; Qin et al. 2021a; Ni et al. 2023) start to model the relationship between ID and SF in a multi-task manner to improve the performance. However, these works ignore the multi-intent context in real-world scenarios, which are more challenging.

**Joint model for multi-intent SLU.** Kim, Ryu, and Lee (2017) start to explore multi-intent scene recognition and Gangadharaiah and Narayanaswamy (2019) propose the first joint model for multi-intent SLU. Qin et al. (2020, 2021b) propose GAT-based joint models to improve interaction performance, and Xing and Tsang (2022) propose a HGAT-based model to achieve the bidirectional interaction. GISCO (Song et al. 2022) considers the global correlation between ID and SF. SSRAN (Cheng, Yang, and Jia 2023) proposes a scope-sensitive attention network to model the dual-task interaction. Different from the above methods, we emphasize maximizing CIG iteratively to achieve the explainable cross-task interaction, which gives a significant insight on improving the performance of multi-intent SLU.

**Theoretical Analysis**

**Problem Definition**

We design a joint-learning network $\Psi(\cdot; \theta)$ that directly consumes a input utterance $U = \{u_i\}_{i=1}^n$ for multi-intent SLU. The model obtains predictions for multi-label intent classification $O^I = \{o^I_i\}_{i=1}^m$ and slot labels $O^S = \{o^S_i\}_{i=1}^n$ that map the utterance $U$, where $m$ denotes the number of intents in a given utterance and $n$ denotes the utterance length. We further define the training set of utterances as $D^U = \{U\}$.

**Theoretical Formulation**

**Markov Modeling.** As shown in Fig. 2a and Fig. 2b, to go beyond the single-stage heuristic modeling of dual-task interaction, we model the joint enhancement process as a multi-stage Markov process which is illustrated in Fig. 2c. This means that the hidden state random variables in the next stage $t + 1$ are only related to the current state $t$ for $t \in N$. 
Formally, we obtain the context-sensitive hidden states $X^0_I = \{x^0_{s,t}\}_{t=1}^n$ and $X^0_O = \{x^0_{s,t}\}_{t=1}^n$ through the shared encoder, where $I$ and $S$ represent ID and SF respectively. Note that the two tasks share the same semantics features in the initial stage, i.e., $X^0_I = X^0_O$. We further define the task-specific features obtained in the $t$-th iteration stage of the intent-detection branch and slot-filling branch as random variables $X^t_I = \{x^t_{s,t}\}_{t=1}^n$ and $X^t_S = \{x^t_{s,t}\}_{t=1}^n$.

**Cross-Task Information Gain.** To explicitly express the relationship between the two tasks, we consider quantifying the degree of joint enhancement from an information-theoretic perspective, and mathematically defining the CIG of the two tasks (implicit random variable $X_i$ and $X_j$):

$$CTIG(X_i; X_j) = H_o(X_i) - H_o(X_i | X_j),$$

where $H_o(\cdot)$ denotes information entropy under the model $o$, and $i = I, j = S$ or $i = S, j = I$. CIG is mathematically symmetric and non-negative, which shows a reduction of the uncertainty of the variable (task) $X_i$ given another variable (task) $X_j$. Based on this, the assumption that the two tasks $X_i$ and $X_j$ are mutually reinforcing can be expressed as:

$$CTIG(X_i; X_j) = CTIG(X_j; X_i) > 0.$$

From this perspective, we explain that the designs of previous joint models essentially strive to improve CIG to obtain task-specific bidirectional information and thereby improve prediction performance.

**Optimization Constraints.** Our model exploits CIG to improve performance on the premise that the following three constraints are satisfied during the optimization process:

**Proposition 1 (Intrinsic Constraint)** Suppose $t$ is the number of steps for the iterative optimization process, \(\forall t \in N^+\),

$$CTIG(X^t_i; X^{t-1}_j) > 0,$$

where $i = I, j = S$ or $i = S, j = I$.

Proposition 1 is the foundation of the joint model, which mathematically describes that the amount of beneficial information provided by one task to another is always positive.

**Proposition 2 (Directional Constraint)** Suppose $t_x$ and $t_y$ are the number of steps for the iterative optimization process, \(\forall t_x, t_y \in N^+\), if $t_x > t_y$, then

$$CTIG(X^{t_x}_i; X^{t_x-1}_j) > CTIG(X^{t_y}_i; X^{t_y-1}_j),$$

where $i = I, j = S$ or $i = S, j = I$.

Proposition 2 constrains the function $CTIG(X^t_i; X^{t-1}_j)$ to monotonically increase w.r.t. the number of iterations $t$. Intuitively, a better expression of the current task will provide more information gain for another. This proposition ensures that the model is positively optimized in each iteration.

**Proposition 3 (Upper-bound Constraint)** The non-continuous set function $CTIG(X^t_i; X^{t-1}_j)$ w.r.t the number of iterative steps $t$ has a supremum $\mathcal{M}T(O^I, O^S)$, i.e.,

$$\sup \{CTIG(X^t_i; X^{t-1}_j) : t \in N^+\} = \mathcal{M}T(O^I, O^S),$$

where $\mathcal{M}T(O^I, O^S)$ is defined as the mutual information between intent-detection labels $O^I$ and slot-filling labels $O^S$, and $i = I, j = S$ or $i = S, j = I$.

Proposition 3 is a boundary condition that describes the existence of an upper bound on the information gain provided by the two tasks to each other. Further, this upper bound is determined by the optimal predictions (i.e., labels $O^I$ and $O^S$) of the two tasks.

**Convergence Analysis.** We next show the convergence ability of our joint model for universal optimization during the iterative process. By the convergence of functions, intuitively, a model should approach the limit value, i.e., the optimal value, after sufficient limited optimization steps.

**Theorem 1** (Convergence Property) Suppose $T = \{t : t \in N^+\}, CTIG : T \rightarrow \mathbb{R}^+$ is a non-continuous set function w.r.t the number of iterative steps $t$ and simultaneously satisfies Proposition 1-3, then, $\forall e > 0, \exists \delta > 0, \text{for } t \in T, \text{if } t > \delta,$

$$|CTIG(X^t_i; X^{t-1}_j) - \mathcal{M}T(O^I; O^S)| < e.$$ (6)

**Proof.** Through Proposition 1-3, we can derive that the non-continuous set function $CTIG(X^t_i; X^{t-1}_j)$ is a monotonically increasing bounded function. According to Monotone Convergence Theory (Yeh 2006), it can be further derived that the function converges to the supremum $\mathcal{M}T(O^I; O^S)$.

Theorem 6 theoretically demonstrates the convergence of our framework. To further measure convergence, the following energy function $E(\cdot)$ can be defined to estimate the iteration situation as $E(t) = \frac{CTIG(X^t_i; X^t_j)}{\mathcal{M}T(O^I; O^S)}$. We note that there is no explicit analytical expression of $CTIG$ about $t$, and therefore the energy function $E(t)$ cannot be directly calculated. Following previous works (Bugliarello et al. 2020; Ji et al. 2022a), we introduce a Monte Carlo estimator in the training set $D^t$ to approximate $E(t)$.

**Method**

**Architecture Overview**

In this section, we describe the overview of InfoJoint as shown in Fig. 3a and introduce the relationship between theory and design. Firstly, to avoid overparameterization caused by a large number of decoders, we encode the time $t$ and then concatenate it with the input embedding of decoders to achieve parameter reuse. Secondly, $X^t_i$ and $X^{t-1}_j$ are obtained asynchronously. We realize that $CTIG(X^t_i; X^{t-1}_j) \rightarrow CTIG(X^t_i; X^t_j)$ when the model is sufficiently iterated. Therefore, we utilize $CTIG(X^t_i; X^t_j)$ to approximate $CTIG(X^t_i; X^{t-1}_j)$ to achieve the synchronous optimization. Thirdly, inspired by Ji et al. (2022b); Cheng et al. (2023a), we utilize contrastive learning to minimize the InfoNCE loss to maximize a lower bound on CIG.

Following the above rules, the general framework of our InfoJoint is shown in Fig. 3, which consists of four core components: a shared encoder, a time encoding module, a multi-level cross-task contrastive learning (MCCL) module and bidirectional information extraction modules. We utilize MCCL and bidirectional information extraction modules to gradually extract and improve the CIG through $t$ iterations. Instead of treating all input utterances indiscriminately, we propose an entropy-based weighting strategy to balance utterance information in each batch.
Shared Encoder and Time Encoding

Following Qin et al. (2020, 2021b); Xing and Tsang (2022), we adopt a bidirectional LSTM (BiLSTM) to produce a series of context-sensitive hidden states \( H = \{h_t\}_{t=1}^n \) over the word embeddings \( \hat{U} = \{\hat{u}_t\}_{t=1}^n \), and a self-attention mechanism to capture context-aware features \( A \in \mathbb{R}^{n \times d_k} \):

\[
h_t = \text{BiLSTM}(\hat{u}_t, h_{t-1}, h_{t+1}),
A = \text{softmax}(\frac{QK^T}{\sqrt{d_k}})V,
\]

where \( Q, K \) and \( V \) are matrices obtained by mapping the input word vector through different linear projections, and \( d_k \) denotes the dimension of keys \( K \).

We finally fuse these two representations as the encoding features: \( E_t = E_t^0 = H||A \), where \( || \) is a concatenation operation. To distinguish the decoder states at different stages, we utilize sine position encoding (Vaswani et al. 2017) to encode the iteration time \( t \) to participate in iterative optimization, before it exceeds \( t_{\text{max}} \) (set to 15).

Bidirectional Information Extraction

These two modules (shown in Fig. 3b) are mainly proposed to extract the valuable information which is conducive to obtaining better expression from the two tasks, namely extracting \( CIG(X_i; X_{i-1}^{t-1}, X_j^{t-1}) \). We employ the Heterogeneous Graph Attention Network (HGAT) (Velickovic et al. 2018; Wang et al. 2019) to implement this bidirectional guidance.

Specifically, to explicitly leverage slot information to guide multi-intent detection, we obtain the estimated slot labels sequence \( S_t = \{s_t^i\}_{i=1}^n \) by a pre-decoder, and construct a slot-to-intent semantic graph \( G_{S2I} = (V_{S2I}, E_{S2I}) \) similar to Xing and Tsang (2022). Further, we feed the node semantics \( H_t^i = \{h_t^{i,j}\}_{j=1}^n = \{x_t^i, \phi^{\text{emb}}(s_t)|_{j=1}^n\} \) into HGAT with \( K \) attention heads and finally produce more effective representation \( E_t^i = \{E^t_{[i,j]}\}_{j=1}^n \):

\[
\mathcal{F} \left( h_{[i,j]}^{t-1}; h_{[i,j]}^t \right) = a_{[k,r]}^T \left[ W_{j}^{[k,r]} h_{[i,j]}^t || W_{j}^{[k,r]} h_{[i,j]}^t \right],
\]

\[
\alpha_{[k,r]} = \frac{\exp \left( \sigma (\mathcal{F} (h_{[i,j]}^t, h_{[i,j]}^t)) \right)}{\sum_{j' \in N_i} \exp \left( \sigma (\mathcal{F} (h_{[i,j]}^t, h_{[i,j']}) \right)},
\]

\[
E_{[i,j]}^t = K \sigma \left( \sum_{k=1}^{K} \alpha_{[k,r]} W_{j}^{[k,r]} h_{[i,j]}^t \right),
\]

where \( \phi^{\text{emb}}(\cdot) \) is a projector; \( r \) denotes the type of edge from node \( j \) to node \( i \); \( \sigma \) represents the activation function; set
$N_i$ is the first-order neighbors of node $i$ on graph $G_{ZL}$; and $a^{[k,r]}$ and $V_f^{[k,r]}$ are trainable matrix of $r$ on the $k$-th head.

We can adopt a similar approach to extract intent-to-slot information and obtain a more abstract feature $E^S_k$ for Eq. 8.

**Maximize Cross-Task Information Gain**

Based on the existing theoretical discovery (Poole et al. 2019) that InfoNCE loss (Oord, Li, and Vinyals 2018) lower bound the mutual information, we propose MCCL (as shown in Fig. 3c) to estimate bidirectional information (i.e., CIG) in practice. Specifically, we adopt a memory bank strategy (He et al. 2020) to obtain more effective negative samples. At each iteration, the model samples a set of matching pairs $M^X = \{(X_{i,k}, X_{S,k})\}_{k=1}^{M^X}$ in the memory queue. Subsequently, we perform contrastive learning at the levels of utterance and token to fully estimate and optimize CIG:

**Utterance level.** We map the matching sample set $M^X$ to the $d_U$-dimension space and obtain a matching set in embedding space $S^U = \{(Z^U_k, Z^S_k)\}_{k=1}^{M^X}$ through the utterance-level projector, where $Z^U_k, Z^S_k \in \mathbb{R}^{1 \times d_U}$. The utterance-level contrastive loss $L_U$ can be formulated as:

$$L_U = -\frac{1}{|S^U|} \sum_{k=1}^{|S^U|} \log \frac{\exp(Z^U_k \cdot Z^S_k / \tau_1)}{\sum_{j=1, j \neq k}^{S^U|} \exp(Z^U_k \cdot Z^S_j / \tau_1)},$$

where $\tau_1$ denotes the temperature coefficient.

**Token level.** We map $M^X$ to the $d_T$-dimension space and obtain $K$ embeddings $Z^U_k = \{z^U_{k,i}\}_{i=1}^k$ and $Z^S_k = \{z^S_{k,i}\}_{i=1}^k$ through the projector, where $z^U_{k,i}, z^S_{k,i} \in \mathbb{R}^{1 \times d_T}$. We further integrate the tokens of all utterances to form a token-level matching set $S^T = \{\{z^U_{k,i}, z^S_{k,i}\}_{k,i}\}_{1}^{K \times n}$. The token-level contrastive loss $L_T$ can be formulated as:

$$L_T = -\frac{1}{|S^T|} \sum_{k=1}^{|S^T|} \log \frac{\exp(z^U_{k,i} \cdot z^S_{k,i} / \tau_2)}{\sum_{j=1, j \neq k}^{S^T|} \exp(z^U_{k,i} \cdot z^S_{j,i} / \tau_2)},$$

where $\tau_2$ denotes the temperature coefficient.

**Training and Inference**

**Entropy weighting.** We find that the information of utterances trained synchronously in a batch is imbalanced. To alleviate this issue, we propose a weighting strategy based on information entropy. Specifically, we define semantic uncertainty $\hat{H}$ as the indicator of utterance importance, and further obtain the weight $\alpha^U$ of the utterance $U$ in the batch $B$:

$$\hat{H}(U) = -\sum_{i=1}^n p_i^U \log p_i^U, \quad \alpha^U = \frac{\hat{H}(U)}{\sum_{U \in B} \hat{H}(U)},$$

where $p_i^U$ denotes the statistic frequency of $i$-th word $u_i$ in the $U$ under the training set $D^U$.

**Directional constraints.** To further ensure each iteration of our InfoJoint model is positively optimized (i.e., satisfying Proposition 2), we define an Exponential Decay Function $\varphi_{mg}(t)$ as a margin penalty for MCCL method, where, $\varphi_{mg}(t) = a \cdot e^{-kt}, a = 0.9$ and $k = 2$. Therefore, the total contrastive loss for stage $t$ is:

$$L^t_{mccl} = \varphi_{mg}(t) \cdot L_{elc} = \varphi_{mg}(t) \cdot (L_U + L_T). \quad (14)$$

For stage $t$, $E^U_i$ and $E^S_i$ are fed to intent and slot post-decoder, producing the intent and slot label distributions for each utterance: $Y^I_t$ and $Y^S_t$. Then, the loss of intent $Y^I_t$ and slot $Y^S_t$ predictions can be calculated by the binary cross-entropy loss and negative log-likelihood loss: $L^t_I$ and $L^t_S$. Intuitively, the predictions in the $t$ stage should be better than those in the $t-1$ stage. We further design another margin penalty $L^t_{mg}$ for this rule (i.e., Proposition 2):

$$L^t_{mg} = \max\{0, L^{t-1}_I - L^t_I\} + \max\{0, L^{t-1}_S - L^t_S\}. \quad (15)$$

Therefore, the joint loss function of InfoJoint is:

$$L^t = \alpha L^t_I + (1 - \alpha) L^t_S + \beta L^t_{mg} + \lambda L^t_{mccl}, \quad (16)$$

where $\alpha, \beta$ and $\lambda$ are trade-off hyper-parameters.

**Details.** In the training stage, InfoJoint adopts Eq. 8 to perform iterative optimization from $t = 1$ to $t = T_{max}$ for each batch. For efficiency, we manually define $T_{max}$ instead of using the Energy function to determine convergence. In the inference stage, the iteration process and MCCL can be discarded without affecting the inference speed. We can directly obtain inference results by setting $t$ as the optimal $T_{max}$. And the final outputs $O^I_t$ and $O^S_t$ are obtained via applying the Top-K strategy over $Y^I_t$ and arg max over $Y^S_t$.

**Experiments**

**Experimental Settings**

**Datasets and Metrics.** Following previous works, we conduct our experiments on two public multi-intent SLU datasets\footnote{https://github.com/LooperXX/AGIF} to evaluate the effectiveness of InfoJoint, i.e., the cleaned version of MixATIS and MixSNIPS (Hemphill, Godfrey, and Doddington 1990; Coucke et al. 2018; Qin et al. 2020). MixATIS and MixSNIPS include 13162, 759, 828 utterances and 39776, 2198, 2199 ones for training, validation and testing respectively. For a fair comparison with previous works, we also adopt accuracy(Acc), F1 score and overall accuracy as metrics for multi-intent detection, SF and sentence-level semantic frame parsing respectively.

**Settings.** The hyper-parameters $\alpha, \beta$ and $\lambda$ of loss (Eq. 16) are set as 0.7, 0.2 and 0.4 on MixATIS, and 0.6, 0.2 and 0.4 on MixSNIPS. We adopt grid search to determine hyperparameters for optimal performance. The temperature $\tau_1$ and $\tau_2$ in Eq. 11 and 12 are empirically set as 0.05. We utilize Adam (Kingma and Ba 2015) with a learning rate of 0.001 and a weight decay of 1e-0 to train InfoJoint for both datasets. We train all models from scratch with 100 epochs. For batch size, we set 16 and 32 for MixATIS and MixSNIPS. All experiments are conducted on 4 RTX3090 GPUs.

NIPS. All experiments are conducted on 4 RTX3090 GPUs.
### Main Results and Analysis

The main experimental results are shown in Table 1. We can see that InfoJoint with $T_{\text{max}} = 10$ outperforms all baselines on both datasets. And we have more detailed observations:

1. Our multi-stage joint model is significantly superior to the baselines with single-stage unidirectional and bidirectional guidance in all metrics of both datasets. Compared with the unidirectional-guided state-of-the-art model SSRAN, InfoJoint achieves 2.7% improvement on Slot (F1), 2.0% improvement on Intent (Acc), 3.6% improvement on Overall (Acc) on the MixATIS dataset, and 0.3% improvement on Slot (F1), 1.1% improvement on Intent (Acc), 1.4% improvement on Overall (Acc) on the MixSNIPS dataset. Compared with the bidirectional-guided best model Co-guiding Net, InfoJoint also achieves significant and consistent performance improvements on all metrics.

2. InfoJoint achieves a significant improvement in terms of overall accuracy. We could observe that the bidirectional models can perform better in overall accuracy than the unidirectional ones. This suggests that the bidirectional guidance achieves calibration and alignment of dual tasks, thereby obtaining better semantic parsing results. And InfoJoint ensures a gradual increase of beneficial information in bidirectional interaction through iterative enhancement. This gradually reduces cross-task propagation of erroneous semantics, thereby further facilitating sentence-level semantic analysis.

3. We adopt a method similar to He and Garner (2023) to evaluate the performance of ChatGPT on these two datasets. As shown in Table 1, although ChatGPT has a strong ability for zero-shot learning in ID, it still lags far behind InfoJoint in overall accuracy. This difference suggests that ChatGPT may struggle to SF and comprehend the abstract connection between ID and SF. Hence, our work on joint multi-intent SLU remains of significant value to the community.

### Ablation Study

We conduct a set of ablation experiments (shown in Table 2) to verify the effectiveness of our theoretical framework.

**Effect of maximizing CIG.** We conduct experiments to study the impact of MCCL on InfoJoint as illustrated in Table 2 with groups (a)(b)(c)(g). It can be seen that the performance of InfoJoint significantly decreases without MCCL to improve the cross-task information gain. Moreover, the lack of any level of contrastive loss ($\mathcal{L}_U$ or $\mathcal{L}_T$) can also affect prediction performance. This verifies that (1) MCCL is capable of extracting CIG sufficiently and effectively. (2) Maximizing CIG can effectively model semantic-level and word-level interactions of dual tasks to improve performance.

**Effect of directional constraints.** To further examine the effectiveness of margin penalty strategies, we show the ablation study on groups (d)(e)(g) in Table 2. We observe that the absence of any strategy ($\varphi_{mg}(t)$ or $\mathcal{L}_{mg}$), especially $\varphi_{mg}(t)$, can result in a significant decrease in predictive performance. This indicates that the margin penalty $\varphi_{mg}(t)$ and $\mathcal{L}_{mg}$ can effectively constrain the model to optimize in the positive direction (i.e., satisfying Proposition 2).

---

**Table 1:** Quantitative comparison results on MixATIS and MixSNIPS. * denotes the improvement of InfoJoint over all baselines is statistically significant with $p < 0.05$ under t-test. The best results are in bold and the second best ones are underlined.

<table>
<thead>
<tr>
<th>Model</th>
<th>MixATIS</th>
<th>MixSNIPS</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Intent</td>
<td>Slot</td>
</tr>
<tr>
<td>Bi-Model (Wang, Shen, and Jin 2018)</td>
<td>70.3</td>
<td>83.9</td>
</tr>
<tr>
<td>Stack-Propagation (Qin et al. 2019)</td>
<td>72.1</td>
<td>87.8</td>
</tr>
<tr>
<td>Joint Multiple ID-SF (Gangadharaih and Narayanaswamy 2019)</td>
<td>73.4</td>
<td>84.6</td>
</tr>
<tr>
<td>AGIF (Qin et al. 2020)</td>
<td>74.4</td>
<td>86.7</td>
</tr>
<tr>
<td>GL-GIN (Qin et al. 2021b)</td>
<td>76.3</td>
<td>88.3</td>
</tr>
<tr>
<td>GISCO (Song et al. 2022)</td>
<td>75.0</td>
<td>88.5</td>
</tr>
<tr>
<td>Co-guiding Net (Xing and Tsang 2022)</td>
<td>79.1</td>
<td>89.8</td>
</tr>
<tr>
<td>SSRAN (Cheng, Yang, and Jia 2023)</td>
<td>77.9</td>
<td>89.4</td>
</tr>
<tr>
<td>ChatGPT (OpenAI 2023)</td>
<td>66.1</td>
<td>43.7</td>
</tr>
<tr>
<td>InfoJoint($T_{\text{max}}=5$)</td>
<td>79.8</td>
<td>90.6</td>
</tr>
<tr>
<td>InfoJoint($T_{\text{max}}=10$)</td>
<td>80.6*</td>
<td>91.4*</td>
</tr>
</tbody>
</table>

**Table 2:** Ablation study on both datasets for quantitatively evaluating the contribution of different components to InfoJoint. We repeat this experiment five times to obtain the statistical mean.
Effect of entropy weighting. By comparing the groups (f) and (g) of Table 2, we can further verify the effectiveness of the entropy weighting strategy in improving performance.

Method Analysis

Dynamic analysis of information gain. To further analyze the universality of the information-theoretic perspective and how InfoJoint works, we conduct an analysis experiment in MixATIS, which adopts a Monte Carlo estimator to obtain the CIG estimator of different models in each epoch (as shown in Fig. 4a). For AGIF and Co-guiding Net, we treat the output of the corresponding decoder as task random variables (i.e., $X_f$ and $X_g$), and discover that these two models implicitly optimize the CIG during training. And InfoJoint explicitly and significantly optimizes CIG to ensure deep alignment and interaction, resulting in better performance.

Analysis of iteration and decay function. We further analyze the effects of $\varphi_{mg}(t)$ and $T_{max}$ as shown in Fig. 4b. We analyze three different types of functions: the exponential decay function $(a \cdot e^{-kt})$, the power function $(\frac{1}{1+kt})$, and the step function with an initial value of 0.5 and a decrease of 0.05 each time, where $a, k, \gamma = 0.9, 2, 3$. We find that the exponential function converges the fastest, possibly because the step function with an initial value of 0.5 and a decrease of 0.05 each time, where $a, k, \gamma = 0.9, 2, 3$. We find that the step function with an initial value of 0.5 and a decrease of 0.05 each time, where $a, k, \gamma = 0.9, 2, 3$. We find that the exponential function converges the fastest, possibly because it has the highest decay rate (gradient). We further observe that the larger $T_{max}$, the more sufficient cross-task interaction, and the higher the prediction performance.

Analysis of hyperparameter. As shown in Fig. 4c, we perform grid search on $\lambda$ and $\beta$ in the MixATIS dataset. We first fix $\beta = 0.2$ to balance MCCL and SLU tasks and observe that $\lambda = 0.4$ achieves the optimal balance between the main task and the information enhancement task. We then fix $\lambda = 0.4$ and find that $\beta = 0.2$ is the optimal trade-off.

Qualitative analysis. Following Zhu et al. (2023c), to better understand what the iterative interaction learns, we visualize the attention weight of InfoJoint and Co-guiding Net for comparison, as shown in Fig. 5. We observe that after sufficient iterations ($T_{max} = 10$), InfoJoint properly aggregates the intent “AddToPlaylist” at slots “add,” “song,” “to,” and “siesta.” This demonstrates InfoJoint successfully focuses weight on the correct slots during optimization and has a better interactive ability compared to prior methods.

Experiments with Pre-training Model

<table>
<thead>
<tr>
<th>Model</th>
<th>MixATIS</th>
<th>MixSNIPS</th>
</tr>
</thead>
<tbody>
<tr>
<td>RoBERTa</td>
<td>49.7</td>
<td>80.2</td>
</tr>
<tr>
<td>AGIF+RoBERTa</td>
<td>50.0</td>
<td>80.7</td>
</tr>
<tr>
<td>SSRAN+RoBERTa</td>
<td>54.4</td>
<td>83.1</td>
</tr>
<tr>
<td>Co-guiding Net+RoBERTa</td>
<td>57.5</td>
<td>85.3</td>
</tr>
<tr>
<td>InfoJoint$T_{max} = 10$+RoBERTa</td>
<td><strong>58.6</strong></td>
<td><strong>86.1</strong></td>
</tr>
<tr>
<td>BERT</td>
<td>51.6</td>
<td>83.0</td>
</tr>
<tr>
<td>SSRAN+BERT</td>
<td>55.3</td>
<td>85.6</td>
</tr>
<tr>
<td>InfoJoint$T_{max} = 10$+BERT</td>
<td><strong>58.9</strong></td>
<td><strong>86.4</strong></td>
</tr>
</tbody>
</table>

Table 3: Overall (Acc) performance for pre-trained models with different architectures.

Following Qin et al. (2020); Cheng et al. (2023b), we use the pre-trained RoBERTa (Liu et al. 2019b) and BERT (Devlin et al. 2019) encoders to replace the original shared encoder. As shown in Table 3, InfoJoint outperforms all baselines when utilizing RoBERTa and BERT as encoders by a large margin, which further shows the universality and effectiveness of our iterative joint method on multi-intent SLU.

Conclusion

In this paper, we quantify the quality of joint interaction in multi-intent SLU from the perspective of information theory, and propose a principled framework with explainability and convergence. Based on this, we devise a novel joint model termed InfoJoint to model multi-stage dynamic interaction. Extensive experiments on two public datasets and analyses verify the effectiveness of InfoJoint.
Acknowledgments
We thank all anonymous reviewers for their constructive and insightful comments. This paper was partially supported by NSFC (No: 62176008) and Shenzhen Science & Technology Research Program (No: GXWD20201231165807007-20200814115301001).

References
OpenAI. 2023. ChatGPT.


