



# Stochastic Curiosity Exploration for Dialogue Systems

Jen-Tzung Chien, Po-Chien Hsu

Department of Electrical and Computer Engineering, National Chiao Tung University, Taiwan

## Abstract

Traditionally, task-oriented dialogue system is built by an autonomous agent which can be trained by reinforcement learning where the reward from environment is maximized. The agent is learned by updating the policy when the goal state is observed. However, in real world, the extrinsic reward is usually sparse or missing. The training efficiency is bounded. The system performance is degraded. It is challenging to tackle the issue of sample efficiency in sparse reward scenario for spoken dialogues. Accordingly, a dialogue agent needs additional information to update its policy even in the period when reward is absent in the environment. This paper presents a new dialogue agent which is learned by incorporating the intrinsic reward based on the information-theoretic approach via stochastic curiosity exploration. This agent encourages the exploration for future diversity based on a latent dynamic architecture which consists of encoder network, curiosity network, information network and policy network. The latent states and actions are drawn to predict stochastic transition for future. The curiosity learning are implemented with intrinsic reward in a metric of mutual information and prediction error in the predicted states and actions. Experiments on dialogue management using PyDial demonstrate the benefit by using the stochastic curiosity exploration.

**Index Terms:** dialogue management, deep reinforcement learning, information-theoretic learning, variational autoencoder

## 1. Introduction

Reinforcement learning (RL) system aims to learn an agent to take actions and interact with environment via sequential decision making. During the learning procedure, the agent obtains the observations from environment, decides an action according to the policy and then receives the feedback known as the reward. The success of deep RL using deep neural networks has brought many useful applications in real environments including dialogue management, game playing, autonomous driving and robot control. Deep RL is applicable to implement a deep policy network to carry out different dialogues for specific tasks where the model-based solution is popular. Model-based RL aims to learn the state transition of the environment based on Markov decision process (MDP) by using the trajectories which are stored while interacting with environment. This paper addresses how MDP is implemented to carry out the model-based agent where the cumulative reward is maximized through understanding the state transition of environments. In practical circumstances, it is challenging to understand the environment and predict the future when facing the unseen scenarios. We would like to learn a dialogue agent with self learning through exploration of useful states based on a latent dynamic system.

This study proposes a new latent dynamic representation to implement the model-based RL for dialogue system where the stochastic curiosity exploration is developed and performed. Variational autoencoder (VAE) [1, 2, 3, 4, 5] is introduced to represent the latent dynamics of states and actions. As we know,

VAE is powerful as a generative model which comprises an inference model as encoder and a generative model as decoder. The encoder compresses the states or actions into latent representations while the decoder generates the synthesized samples from latent states or actions. The variational future is therefore planned according to this variational RL where the prediction is made by using latent states and actions. If the future latent state differs from the actual latent state, it means that the agent still has some regions which have not been explored. Accordingly, the agent measures the stochastic curiosity in a form of mutual information between the predicted latent state and latent action. The predicted state with sufficient mutual information requires further exploration. The intrinsic reward based on this mutual information is maximized to fulfill the stochastic curiosity exploration for dialogue management. In the experiments, we illustrate the performance of stochastic curiosity exploration in deep RL based on deep Q network [6, 7]. An open-source end-to-end statistical spoken dialogue system toolkit using PyDial [8, 9] is evaluated. The dialogue management module is examined with a number of environments under different conditions. The learning objective is assessed to show the merit of stochastic curiosity learning for the success of spoken dialogues.

## 2. Reinforcement Learning for Dialogues

Statistical spoken dialogue system (SDS) is seen as an autonomous process where deep reinforcement learning is applicable to explore task-oriented goal. Basically, SDS consists of different modules in presence of various uncertainties. A user interacts with dialogue system using speech input and speech output via automatic speech recognition (ASR) and text-to-speech (TTS) synthesis, respectively. The module of spoken language understanding (SLU) is used to understand the semantics of sentence hypotheses from ASR. SLU provides a type of meaning for input utterance based on a slot-value pair as the state. A dialogue management (DM) module is developed to use the state observations to implement reinforcement learning for sequential decision making and sequential belief tracking. DM is seen as the policy of an agent for choosing actions in dialogue system. Natural language generation (NLG) is a module to generate the sentence based on the actions chosen by DM. TTS then provides the response speech to interact with user by using the text from NLG. Deep RL is implemented to train a desirable policy in DM module to fulfill the task-oriented goal.

### 2.1. Reinforcement Learning

Deep RL using deep Q network (DQN) [6, 7, 10] is introduced. DQN builds a value-based agent where a deep neural network (DNN) is incorporated to calculate the state-action value functions  $Q_{\theta}(s_t, a_t)$  in network outputs at each time  $t$  for individual actions  $a_t$  where state  $s_t$  is used as input. This calculation is to estimate the expected return  $Q(s_t, a_t) = \mathbb{E}[R_t | s_t, a_t]$  where  $R_t = \sum_{k=0}^{\infty} \gamma^k r_{t+k}$  as the state-action value function. Here,

$r_t$  is the extrinsic reward from environment and  $\gamma \in (0, 1]$  is a discount factor. DNN parameters  $\theta$  in value network  $Q_\theta(s_t, a_t)$  is updated by minimizing a square error loss function  $J(\theta) = (r_t + \gamma \max_a Q_\theta(s_{t+1}, a) - Q_\theta(s_t, a_t))^2$ . DQN aims to predict the Q value  $Q_\theta(s_t, a_t)$  which gets close to the temporal-difference (TD) [11, 12] target  $r_t + \gamma \max_a Q_\theta(s_{t+1}, a)$ . TD error is minimized. In addition to this value network, DQN implements Q learning by using replay buffer as well as target network. Replay buffer memories the transitions  $\{s_t, a_t, r_t, s_{t+1}\}$  which are sampled in minibatches to calculate the gradient of  $J(\theta)$  with respect to  $\theta$  where the slow convergence in Q learning due to too close consecutive states is improved. Besides, there is no correct target in Q learning. An additional target network is merged to calculate  $r_t + \gamma \max_a \hat{Q}_\theta(s_{t+1}, a)$  as TD target. It is not suitable to use the same network to calculate target value  $\hat{Q}_\theta(s_{t+1}, a)$  and Q value  $Q_\theta(s_t, a_t)$ . Parameter of target network  $\hat{\theta}$  is replaced by that of value network  $\theta$  periodically every a number of time steps. The exploration for environment dynamics in DQN is based on  $\epsilon$ -greedy algorithm.

## 2.2. Environment Exploration

RL aims to balance the trade-off between exploration and exploitation in real environment [13]. To improve the learning efficiency, the strategy based on variational information exploration (VIE) [14] was proposed. VIE used the entropy search, a popular Bayesian optimization method, to encourage agent to explore efficiently. The maximization of information gain was implemented to realize the agent’s belief in environment dynamics. Correspondingly, the sum of entropy reduction due to new state  $s_{t+1}$  along a trajectory  $\sum_t (H(\theta|\xi_t, a_t) - H(\theta|s_{t+1}, \xi_t, a_t))$  given with an experience of history is calculated.  $H(\cdot)$  is the entropy function. The intrinsic reward at each time  $t$  is expressed as  $r_t^i = \text{KL}(p(\theta|s_{t+1}, \xi_t, a_t) \| p(\theta|\xi_t))$  where  $\text{KL}(\cdot \| \cdot)$  denotes the Kullback-Leibler (KL) divergence,  $\theta$  denotes the parameter of transition model  $p(s_{t+1}|s_t, a_t, \theta)$  and  $\xi_t = \{s_1, a_1, \dots, s_t\}$  denotes the history of an agent who experiences until time step  $t$ . In practice, the KL divergence in  $r_t^i$  is implemented by  $\text{KL}(q(\theta|\phi_{t+1}) \| q(\theta|\phi_t))$  using variational distribution  $q(\theta|\phi_t)$  with parameter  $\phi_t$  updated by variational inference at each time  $t$ . Agent is trained by maximizing extrinsic reward  $r_t^e$  as well as intrinsic reward  $r_t^i$ , i.e.  $r_t = r_t^e + r_t^i$ . VIE follows the Bayesian perspective by maximizing the uncertainty reduction for an agent who explores the environment dynamics through state transitions. In [14], the Bayes-by-backprop network (BBN) [15] was used to learn the state transition of an environment. Weights of BBN were sampled from Gaussian distribution which implemented a robust network against the mode collapse. BBN calculated the entropy reduction or KL divergence between the posteriors of weight distributions at consecutive time steps, and used this intrinsic reward to encourage agent to explore. An efficient exploration for environment was performed by maximizing the expected sum of reduction of uncertainty in environment dynamics. Maximum entropy reduction was assured. In [16], the curiosity exploration (CE) was carried out as a model-based module which was called the intrinsic curiosity module (ICM). ICM extracted the transition information from input tuple  $\{s_t, a_t, s_{t+1}\}$  and used it as the intrinsic reward  $r_t^i$  for agent learning. CE trained a forward dynamics model  $f(\cdot)$  with parameter  $\theta_f$  that predicted the feature representation of next state  $\hat{\phi}(s_{t+1}) = f(\phi(s_t), a_t, \theta_f)$  where  $\phi(\cdot)$  was a feature extractor. The difference between the features of predicted state  $\hat{\phi}(s_{t+1})$  and actual state  $\phi(s_{t+1})$  was

measured as the prediction error or a negative intrinsic reward which was minimized to encourage the curiosity and pursue the exploration for the best state prediction. The value of intrinsic reward in CE with a scaling factor  $\eta > 0$  was computed as  $r_t^i = \frac{\eta}{2} \|\hat{\phi}(s_{t+1}) - \phi(s_{t+1})\|_2^2$ . A self-supervised prediction was performed. In addition, an inverse neural network model with parameter  $\theta_g$  was merged to predict an action  $\hat{a}_t = g(s_t, s_{t+1}, \theta_g)$  where the discrepancy between the predicted  $\hat{a}_t$  and actual actions  $a_t$  was minimized. Using CE, the tuples  $\{s_t, a_t, s_{t+1}\}$  were collected to jointly train the policy network  $\pi(\cdot)$ , the feature encoder  $\phi(\cdot)$ , the forward model  $f(\cdot)$  and the inverse model  $g(\cdot)$ . Environment dynamics were explored but the intrinsic reward was insufficient to meet the situation of sparse reward in dialogue system.

## 3. Stochastic Curiosity Exploration

VIE promotes the exploration by maximizing the information gain using Bayesian neural network while CE pursues the self-supervised exploration by maximizing the curiosity or minimizing the prediction error of the compressed states. This paper proposes a new dialogue manager where the information-driven curiosity is maximized and regularized for deep reinforcement learning using the stochastic curiosity exploration (SCE) [17, 18, 19]. In addition to policy network  $\pi_\theta(a_t|s_t, r_t)$ , the encoder network, curiosity network and information network are constructed.

### 3.1. Latent Curiosity Representation

The dialogue agent is trained by using the trajectories of state  $s_t$ , action  $a_t$  and reward  $r_t$  at different time steps  $t$ . It is crucial to build a dynamic system for environment dynamics which characterizes the relations among current state  $s_t$ , action  $a_t$  and next state  $s_{t+1}$ . Basically, states and actions are high-dimensional. To facilitate stochastic modeling, the latent dynamic representation is presented to characterize the unknown environment dynamics where the uncertainties of state  $s_t$  and action  $a_t$  are represented for information-theoretic exploration. This latent dynamic system also promotes the learning efficiency where the redundancy in information extraction can be reduced during exploration in high-dimensional state space. This paper refers the variational autoencoder (VAE) [1] and builds a new latent dynamic system where the compressed representations of high-dimensional states and actions are extracted by the learned encoders. In the implementation, low-dimensional random state  $z_{s_t}$  and action  $z_{a_t}$  are calculated by using individual *encoder networks* where high-dimensional state  $s_t$  and action  $a_t$  are used as inputs, respectively. The same encoder is adopted for current state  $s_t$  and next state  $s_{t+1}$ . Stochastic modeling is embedded in a latent dynamic representation based on  $z_{s_t}$ ,  $z_{a_t}$  and  $z_{s_{t+1}}$ , which are learned by maximizing the evidence lower bounds (ELBO) [1, 20] for state encoder  $\mathcal{L}(s; \theta_s, \phi_s) = \mathcal{L}_{\theta_s} - \mathcal{L}_{\phi_s}$  and action encoder  $\mathcal{L}(a; \theta_a, \phi_a) = \mathcal{L}_{\theta_a} - \mathcal{L}_{\phi_a}$ . There are two terms in ELBOs. One is the log likelihoods  $\mathcal{L}_{\theta_s} = \mathbb{E}_{q_{\phi_s}(z_s|s)}[\log p_{\theta_s}(s|z_s)]$  and  $\mathcal{L}_{\theta_a} = \mathbb{E}_{q_{\phi_a}(z_a|a)}[\log p_{\theta_a}(a|z_a)]$  with the latent variables  $z_s$  and  $z_a$  sampled by the variational distributions  $q_{\phi_s}(z_s|s)$  and  $q_{\phi_a}(z_a|a)$ , respectively. The other is the KL terms  $\mathcal{L}_{\phi_s} = \text{KL}(q_{\phi_s}(z_s|s) \| p(z_s))$  and  $\mathcal{L}_{\phi_a} = \text{KL}(q_{\phi_a}(z_a|a) \| p(z_a))$  for regularizing variational distributions to get close to their priors  $p(z_s)$  and  $p(z_a)$ , respectively, which are standard Gaussians.

This latent dynamic representation has twofold considerations. One is to reflect the stochastic features from heteroge-

neous states and actions in dialogue system. The other is to facilitate the stochastic curiosity exploration which tackles the sparse reward as well as the model uncertainty. The agent learns under latent dynamic space based on the intrinsic reward calculated from current random variables  $z_{s_t}$  and  $z_{a_t}$  according to the *curiosity network* with output probability  $p(\hat{z}_{s_{t+1}}|z_{a_t}, z_{s_t})$ . This network provides high-level abstraction to predict future state  $\hat{z}_{s_{t+1}}$ . The agent of SCE learns to explore those unseen regions in environment by optimizing the stochastic curiosity. This curiosity-driven learning is performed by maximizing the difference between the predictive information of next state  $\hat{z}_{s_{t+1}}$  and the information of next state  $z_{s_{t+1}}$  measured in real environment. The stochastic curiosity is calculated by KL divergence between the distribution of predicted state  $p(\hat{z}_{s_{t+1}}|z_{s_t}, z_{a_t})$  via curiosity network and the variational distribution of actual state  $q(z_{s_{t+1}}|s_t)$  via state encoder. The objective is to explore the future by maximizing

$$\mathcal{L}_{\theta_{\text{cur}}} = \text{KL}(p(\hat{z}_{s_{t+1}}|z_{s_t}, z_{a_t})||q(z_{s_{t+1}}|s_{t+1})) \quad (1)$$

which is implemented through sampling the latent variables of predicted state  $\hat{z}_{s_{t+1}}$  from predictive distribution as well as actual state  $z_{s_{t+1}}$  from variational distribution. The intrinsic reward is therefore measured by  $r_t^i = \|\hat{z}_{s_{t+1}} - z_{s_{t+1}}\|_2^2$ . In addition, this study boosts the information-theoretic learning in SCE by merging a regularization in curiosity-based intrinsic reward.

### 3.2. Information-Theoretic Regularization

This study learns an agent by interacting with environment by enhancing the mutual information [21, 22, 23, 24] between latent variables of the predicted state  $\hat{z}_{s_{t+1}}$  and the selected action  $z_{a_t}$  from policy network. An *information network* is incorporated to calculate

$$\mathcal{I}(\hat{z}_{s_{t+1}}, z_{a_t}) = \mathbb{E}_{p(\hat{z}_{s_{t+1}}, z_{a_t})} \left[ \log \frac{p(\hat{z}_{s_{t+1}}, z_{a_t})}{p(\hat{z}_{s_{t+1}})p(z_{a_t})} \right] \quad (2)$$

which is maximized to promote the dependencies between the outputs of curiosity network  $\hat{z}_{s_{t+1}}$  and action encoder  $z_{a_t}$ . The selected action sufficiently supports the exploration based on stochastic curiosity. Intrinsic reward is then formed by

$$\tilde{r}_t^i = \|\hat{z}_{s_{t+1}} - z_{s_{t+1}}\|_2^2 + \mathcal{I}(\hat{z}_{s_{t+1}}, z_{a_t}). \quad (3)$$

The information-theoretic regularization is imposed for curiosity-driven exploration. However, it is challenging to construct a neural estimation for mutual information. An analytical neural network solution to deep RL based on SCE is required.

Here, we refer [25] to derive the variational lower bound of mutual information, parameterized by neural network, as a tractable and scalable objective in the implementation. This problem is to select a family of functions  $M_\theta : \mathcal{A} \times \mathcal{B} \rightarrow \mathbb{R}$  which are parametrized by a neural network model with parameters  $\theta \in \Theta$ . The lower bound of mutual information is derived as a neural information in a form of [26]

$$\mathcal{I}_\Theta(A, B) = \sup_{\theta \in \Theta} \mathbb{E}_{\mathbb{P}_{AB}}[M_\theta] - \log(\mathbb{E}_{\mathbb{P}_A \otimes \mathbb{P}_B}[e^{M_\theta}]) \quad (4)$$

where  $\mathbb{P}_{AB}$  and  $\mathbb{P}_A \otimes \mathbb{P}_B$  denote the joint and product of marginals in probability space, respectively. In the implementation, the parameter of information network  $\theta_{\text{inf}}$  is merged in the lower bound of mutual information between latent state  $\hat{z}_{s_{t+1}}$  and latent action  $z_{a_t}$ . The neural network parameter  $\theta_{\text{inf}}$  is used to calculate the function  $M_{\theta_{\text{inf}}}$  for the objective of mutual information (MI)  $\mathcal{L}_{\theta_{\text{inf}}}$ . The bound  $\mathcal{L}_{\theta_{\text{inf}}}$  is calculated by the joint distribution  $p(\hat{z}_{s_{t+1}}, z_{a_t})$  and the marginal distributions  $p(\hat{z}_{s_{t+1}})$

and  $p(z_{a_t})$  by using  $N$  minibatch samples. Joint distribution is computed from the transitions in replay buffer. Marginals are calculated by shuffling the individual samples of  $\hat{z}_{s_{t+1}} \in Z_S$  or  $z_{a_t} \in Z_A$  in the transitions. The bound of MI is derived by

$$\begin{aligned} \mathcal{I}(z_s, z_a) &= \sum_{z_s \in Z_S, z_a \in Z_A} p(z_s, z_a) \log \frac{p(z_s, z_a)}{p(z_s)p(z_a)} \\ &\geq \sup_{\theta_{\text{inf}} \in \Theta} \mathbb{E}_{p(z_s, z_a)}[M_{\theta_{\text{inf}}}] - \log(\mathbb{E}_{p(z_s)p(z_a)}[e^{M_{\theta_{\text{inf}}}}]) \\ &= \frac{1}{N} \sum_{n=1}^N M_{\theta_{\text{inf}}}(z_s^{(n)}, z_a^{(n)}) - \log\left(\frac{1}{N} \sum_{n=1}^N e^{M_{\theta_{\text{inf}}}(z_s^{(n)}, z_a^{(n)})}\right) \\ &\triangleq \mathcal{L}_{\theta_{\text{inf}}}. \end{aligned} \quad (5)$$

The variational RL [27] based on SCE is then implemented by training the state encoder  $\{\theta_s, \phi_s\}$ , action encoder  $\{\theta_a, \phi_a\}$ , curiosity network  $\theta_{\text{cur}}$ , information network  $\theta_{\text{inf}}$  and policy network  $\pi_\theta(a_t|s_t)$  by jointly maximizing the hybrid objective

$$\mathcal{L} = \mathcal{L}(s; \theta_s, \phi_s) + \mathcal{L}(a; \theta_a, \phi_a) + \lambda_c \mathcal{L}_{\theta_{\text{cur}}} + \lambda_i \mathcal{L}_{\theta_{\text{inf}}}. \quad (6)$$

where two tuning parameters  $\lambda_c$  and  $\lambda_i$  are used.

## 4. Experiments

Stochastic curiosity exploration was implemented in deep RL for spoken dialogue management using PyDial toolkit [8, 9, 28].

### 4.1. Spoken dialogue System

PyDial is an open-source end-to-end evaluation system for task-oriented dialogue where a number of benchmark environments with different dialogue modules were simulated [29]. The dialogue management module based on deep RL using DQN and other algorithms could be investigated. Different exploration methods were evaluated by 18 dialogue tasks which were built by 6 environments (with different semantic error rate (SER) (0%, 15% or 30%) [30, 31], action masking (on or off) and user model (standard or unfriendly)) and 3 domains (Cambridge (CR) and San Francisco (SFR) restaurants, and laptops (LAP)). SER reflected the semantic level where the true user act was corrupted by noise. Action masking was considered to test the learning capability of algorithms. Unfriendly user implied that users provided limited information. This work used the default setting of hyperparameters in DQN. DQN adopted the  $\epsilon$ -greedy exploration with a linear schedule starting from  $\epsilon = 0.3$  and then annealed to 0. Regularization parameters  $\lambda_c = 0.2$  and  $\lambda_i = 1.0$  were specified. The encoder, curiosity, information and policy networks were modeled by two to three fully-connected layers with the setting of activation functions provided by Pydial. Adam optimizer [32] with initial learning rate 0.001 was used. The replay buffer was set to be 6000. The maximum number of turns in dialogue was 25. The discount factor was 0.99. Every model was trained over ten different random seeds. The models were estimated with 10000 training dialogues and evaluated over 500 test dialogues. Dialogue performance was assessed by using the metrics of success rate and reward for policy model using different explorations. Success rate was defined as the percentage of dialogues which were completed successfully. Reward was defined as  $20 \cdot D - T$  where  $D$  was the success indicator (0 or 1) and  $T$  was the number of dialogue turns.

### 4.2. Performance Evaluation

The DQNs with  $\epsilon$ -greedy exploration (denoted by baseline) and other explorations based on VIE [14], CE [16] and the proposed

Table 1: Comparison of success rates and rewards by using different explorations in DQN under 18 benchmarking tasks.

Task		Baseline		VIE		CE		εSCE		SCE		SCE-MI	
		Suc.	Rew.	Suc.	Rew.	Suc.	Rew.	Suc.	Rew.	Suc.	Rew.	Suc.	Rew.
Env. 1	CR	92.5%	12.2	91.6%	12.2	94.4%	12.6	94.5%	12.6	94.8%	12.7	<b>95.6%</b>	<b>13.0</b>
	SFR	74.1%	7.6	81.5%	9.0	83.0%	9.3	83.7%	9.1	<b>84.7%</b>	9.6	84.6%	<b>9.7</b>
	LAP	73.0%	7.5	74.0%	7.5	78.3%	8.2	77.6%	8.1	76.8%	8.0	<b>78.5%</b>	<b>8.3</b>
Env. 2	CR	90.7%	11.7	94.8%	12.6	95.1%	12.2	94.9%	12.5	<b>95.7%</b>	<b>12.8</b>	95.1%	12.4
	SFR	90.1%	10.7	83.0%	8.9	87.4%	10.0	87.0%	10.2	87.5%	10.5	<b>92.5%</b>	<b>11.3</b>
	LAP	84.9%	9.1	79.7%	8.2	79.4%	7.7	83.4%	9.0	87.2%	9.4	<b>89.4%</b>	<b>10.5</b>
Env. 3	CR	93.6%	12.0	94.2%	12.0	94.5%	12.0	94.7%	12.3	95.2%	12.3	<b>96.1%</b>	<b>12.4</b>
	SFR	73.3%	6.1	71.7%	5.9	75.9%	6.8	76.3%	6.8	77.3%	7.0	<b>82.7%</b>	<b>8.0</b>
	LAP	69.0%	5.6	66.1%	5.0	69.0%	5.7	71.1%	5.8	73.0%	6.1	<b>73.8%</b>	<b>6.4</b>
Env. 4	CR	86.4%	9.7	91.1%	10.9	92.9%	11.3	89.5%	10.3	90.2%	10.8	<b>93.6%</b>	<b>11.4</b>
	SFR	80.5%	8.5	78.9%	8.0	79.0%	7.8	85.0%	8.3	84.7%	8.8	<b>87.1%</b>	<b>9.5</b>
	LAP	78.6%	7.5	75.9%	7.0	83.2%	8.0	82.5%	8.0	<b>83.3%</b>	<b>8.5</b>	80.2%	7.8
Env. 5	CR	90.2%	10.0	91.2%	10.3	93.5%	10.6	93.9%	10.5	93.5%	10.7	<b>94.8%</b>	<b>11.2</b>
	SFR	71.6%	4.3	74.7%	4.8	73.3%	4.7	78.3%	5.3	80.4%	6.2	<b>82.2%</b>	<b>6.5</b>
	LAP	51.8%	0.8	57.7%	1.6	53.3%	1.0	57.9%	1.4	59.4%	1.9	<b>61.1%</b>	<b>2.2</b>
Env. 6	CR	89.9%	10.2	89.3%	10.1	89.9%	10.2	89.6%	10.1	<b>90.2%</b>	<b>10.3</b>	89.7%	10.2
	SFR	64.0%	3.3	63.9%	3.2	63.5%	3.1	65.5%	3.9	66.6%	3.7	<b>70.4%</b>	<b>4.6</b>
	LAP	56.2%	2.1	59.8%	2.6	54.3%	2.1	58.0%	2.4	58.3%	2.7	<b>61.9%</b>	<b>3.3</b>

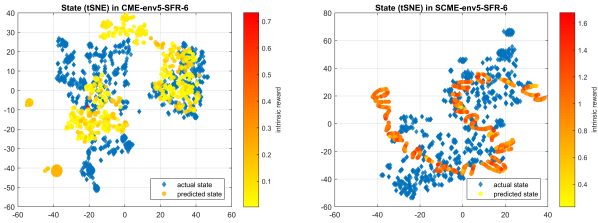


Figure 1: Predicted state and actual state after training with 6000 dialogues in Env. 5 of SFR domain. Color bar indicates the values of intrinsic rewards. **Left: CE. Right: SCE.**

SCE were implemented for comparison. Figure 1 demonstrates the latent variables of predicted states  $\hat{z}_{s_{t+1}}$  and actual states  $z_{s_{t+1}}$  by using CE and SCE. *t*-SNE [33] was applied to show two-dimensional samples. The values of intrinsic rewards in the prediction are shown by color. We can see that CE (left) does encourage the agent to explore the unknown environment, but not all the predicted regions have high intrinsic reward. The proposed SCE (right) also encourages the agent to explore the unknown regions where the state distribution sufficiently reflects the environment. The intrinsic rewards in the predicted states are high. SCE does maximize the exploration.

Table 1 summarizes the success rates and rewards with 10000 training dialogues. SCE and SCE-MI denotes the proposed SCE without and with mutual information included in intrinsic reward for exploration, respectively. The SCE combined with  $\epsilon$ -greedy exploration (denoted by  $\epsilon$ SCE) is implemented for comparison. Basically, the average performances of all methods are not improved significantly in CR domain which is an easy domain in PyDial where the sophisticated exploration does not significantly help. In the SFR and LAP domains, the variants of SCE perform better than the other methods. This is because that the exploration is likely improved in the complicated environments with sparse reward. Particularly, SFR domain is known as a task with high-dimensional state space

where different SCEs work well. SCEs could build the latent dynamic space for deep RL which accommodates the informative features about high dimensional states. It is interesting that SCE combined with  $\epsilon$ -greedy exploration ( $\epsilon$ SCE) does not help clearly. The contribution from stand-alone stochastic curiosity exploration is assured. Importantly, SCE with mutual information likely obtains higher success rates and rewards than the method without mutual information. This implies that mutual information can regularize the agent and boost the learning towards exploring the informative states. We can see that SCE-MI has considerable improvement in environment 5 where an unfriendly user setting was considered. SCE-MI provides large curiosity for maximizing the exploration for unseen states in future. The proposed SCE can handle the heterogeneous environment with unfriendly user problem.

## 5. Conclusions

This paper presented the stochastic curiosity exploration as a general solution to model-based reinforcement learning. The exploration method was applied for dialogue management in an end-to-end task-oriented dialogue system. The proposed solution coped with the sparse reward task and maximized the exploration via information-theoretic learning. The variational inference was incorporated to learn the latent dynamics for heterogeneous environment in spoken dialogues. The high-dimensional state space was compactly represented. The curiosity network was trained to predict the latent future with diversity. The information network was learned to measure the mutual information as a regularized exploration for future. By using this embedding information as intrinsic reward, the agent learned by itself and explored for useful future. Experiments on variational reinforcement learning for dialogue system demonstrated the effectiveness of the proposed method in different environments. The reward was increased when the variational inference was executed for curiosity maximizing exploration. This study develops a new exploration scheme which can be extended to other reinforcement learning algorithms and tasks [34] for different information systems and applications.

## 6. References

- [1] D. P. Kingma and M. Welling, “Auto-encoding variational Bayes,” in *Proc. of International Conference on Learning Representations*, 2014.
- [2] T. Zhao, K. Lee, and M. Eskenazi, “Unsupervised discrete sentence representation learning for interpretable neural dialog generation,” in *Proc. of Annual Meeting of the Association for Computational Linguistics*, 2018, pp. 1098–1107.
- [3] J.-T. Chien, W.-L. Liao, and I. E. Naqa, “Exploring state transition uncertainty in variational reinforcement learning,” in *Proc. of European Signal Processing Conference*, 2020.
- [4] J.-T. Chien, “Deep Bayesian learning and understanding,” in *Proc. of International Conference on Computational Linguistics: Tutorial Abstracts*, 2018, pp. 13–18.
- [5] —, “Deep Bayesian natural language processing,” in *Proc. of Annual Meeting of the Association for Computational Linguistics: Tutorial Abstracts*, 2019, pp. 25–30.
- [6] V. Mnih, K. Kavukcuoglu, D. Silver, A. Graves, I. Antonoglou, D. Wierstra, and M. Riedmiller, “Playing Atari with deep reinforcement learning,” *arXiv preprint arXiv:1312.5602*, 2013.
- [7] V. Mnih, K. Kavukcuoglu, D. Silver, A. A. Rusu, J. Veness, M. G. Bellemare, A. Graves, M. Riedmiller, A. K. Fidjeland, G. Ostrovski *et al.*, “Human-level control through deep reinforcement learning,” *Nature*, vol. 518, no. 7540, pp. 529–533, 2015.
- [8] S. Ultes, L. M. R. Barahona, P.-H. Su, D. Vandyke, D. Kim, I. Casanueva, P. Budzianowski, N. Mrkšić, T.-H. Wen, M. Gasic *et al.*, “Pydial: A multi-domain statistical dialogue system toolkit,” in *Proceedings of ACL 2017, System Demonstrations*, 2017, pp. 73–78.
- [9] I. Casanueva, P. Budzianowski, P.-H. Su, N. Mrkšić, T.-H. Wen, S. Ultes, L. Rojas-Barahona, S. Young, and M. Gašić, “A benchmarking environment for reinforcement learning based task oriented dialogue management,” in *Proc. of NeurIPS Deep Reinforcement Learning Symposium*, 2017.
- [10] A. Barreto, W. Dabney, R. Munos, J. J. Hunt, T. Schaul, H. P. van Hasselt, and D. Silver, “Successor features for transfer in reinforcement learning,” in *Advances in Neural Information Processing Systems*, 2017, pp. 4055–4065.
- [11] R. S. Sutton, “Temporal credit assignment in reinforcement learning,” Ph.D. dissertation, University of Massachusetts Amherst, 1985.
- [12] —, “Learning to predict by the methods of temporal differences,” *Machine Learning*, vol. 3, no. 1, pp. 9–44, 1988.
- [13] N. Chentanez, A. G. Barto, and S. P. Singh, “Intrinsically motivated reinforcement learning,” in *Advances in Neural Information Processing Systems*, 2005, pp. 1281–1288.
- [14] R. Houthoofd, X. Chen, Y. Duan, J. Schulman, F. De Turck, and P. Abbeel, “VIME: Variational information maximizing exploration,” in *Neural Information Processing Systems*, 2016, pp. 1109–1117.
- [15] C. Blundell, J. Cornebise, K. Kavukcuoglu, and D. Wierstra, “Weight uncertainty in neural networks,” in *Proc. of International Conference on Machine Learning*, 2015, pp. 1613–1622.
- [16] D. Pathak, P. Agrawal, A. A. Efros, and T. Darrell, “Curiosity-driven exploration by self-supervised prediction,” in *Proc. of International Conference on Machine Learning*, 2017, pp. 2778–2787.
- [17] T. Hester and P. Stone, “Intrinsically motivated model learning for developing curious robots,” *Artificial Intelligence*, vol. 247, pp. 170–186, 2017.
- [18] J.-T. Chien and P.-C. Hsu, “Stochastic curiosity maximizing exploration,” in *Proc. of International Joint Conference on Neural Networks*, 2020.
- [19] M. Bellemare, S. Srinivasan, G. Ostrovski, T. Schaul, D. Saxton, and R. Munos, “Unifying count-based exploration and intrinsic motivation,” in *Advances in Neural Information Processing Systems*, 2016, pp. 1471–1479.
- [20] S. Watanabe and J.-T. Chien, *Bayesian Speech and Language Processing*. Cambridge University Press, 2015.
- [21] M. Gabrić, A. Manoel, C. Luneau, N. Macris, F. Krzakala, and L. Zdeborová, “Entropy and mutual information in models of deep neural networks,” in *Advances in Neural Information Processing Systems*, 2018, pp. 1821–1831.
- [22] N. Savinov, A. Raichuk, R. Marinier, D. Vincent, M. Pollefeys, T. Lillicrap, and S. Gelly, “Episodic curiosity through reachability,” in *Proc. of International Conference on Learning Representations*, 2018.
- [23] S. Mohamed and D. J. Rezende, “Variational information maximization for intrinsically motivated reinforcement learning,” in *Neural Information Processing Systems*, 2015, pp. 2125–2133.
- [24] Y. Tu, M.-W. Mak, and J.-T. Chien, “Variational domain adversarial learning with mutual information maximization for speaker verification,” *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 2013–2024, 2020.
- [25] M. I. Belghazi, A. Baratin, S. Rajeswar, S. Ozair, Y. Bengio, A. Courville, and R. D. Hjelm, “MINE: mutual information neural estimation,” *arXiv preprint arXiv:1801.04062*, 2018.
- [26] M. D. Donsker and S. S. Varadhan, “Asymptotic evaluation of certain Markov process expectations for large time, II,” *Communications on Pure and Applied Mathematics*, vol. 28, no. 2, pp. 279–301, 1975.
- [27] D. J. Rezende, S. Mohamed, and D. Wierstra, “Stochastic back-propagation and approximate inference in deep generative models,” *Proc. of International Conference on Machine Learning*, 2014.
- [28] J.-T. Chien and W. X. Lieow, “Meta learning for hyperparameter optimization in dialogue system,” in *Proc. of Annual Conference of International Speech Communication Association*, 2019, pp. 839–843.
- [29] Z. Lipton, X. Li, J. Gao, L. Li, F. Ahmed, and L. Deng, “BBQ-networks: Efficient exploration in deep reinforcement learning for task-oriented dialogue systems,” in *Proc. of AAAI Conference on Artificial Intelligence*, 2018.
- [30] J.-T. Chien, “Association pattern language modeling,” *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 5, pp. 1719–1728, 2006.
- [31] J.-T. Chien and C.-H. Chueh, “Joint acoustic and language modeling for speech recognition,” *Speech Communication*, vol. 52, no. 3, pp. 223–235, 2010.
- [32] D. P. Kingma and J. Ba, “Adam: A method for stochastic optimization,” in *Proc. of International Conference on Learning Representations*, 2015.
- [33] L. van der Maaten and G. E. Hinton, “Visualizing data using *t*-SNE,” *Journal of Machine Learning Research*, vol. 9, no. Nov, pp. 2579–2605, 2008.
- [34] H.-H. Tseng, Y. Luo, S. Cui, J.-T. Chien, R. K. Ten Haken, and I. El Naqa, “Deep reinforcement learning for automated radiation adaptation in lung cancer,” *Medical Physics*, vol. 44, no. 12, pp. 6690–6705, 2017.