Adaptive Parameterization for Neural Dialogue Generation

Hengyi Cai[1,2], Hongshen Chen[3], Cheng Zhang[1], Yonghao Song[1], Xiaofang Zhao[1], Dawei Yin[3]
[1]Institute of Computing Technology, Chinese Academy of Sciences, Beijing, China
[2]University of Chinese Academy of Sciences, Beijing, China
[3]Data Science Lab, JD.com, Beijing, China

Overview

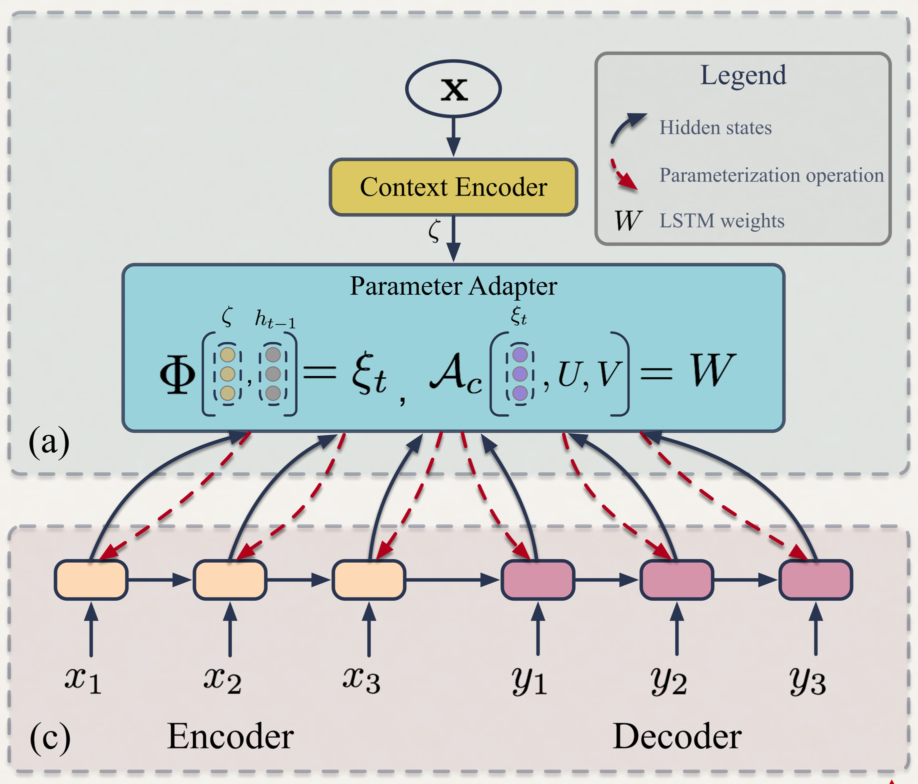


Figure 2: Context-aware parameterization. (a) Context-aware parameter adapter. (c) One layer of the dialogue generation model.

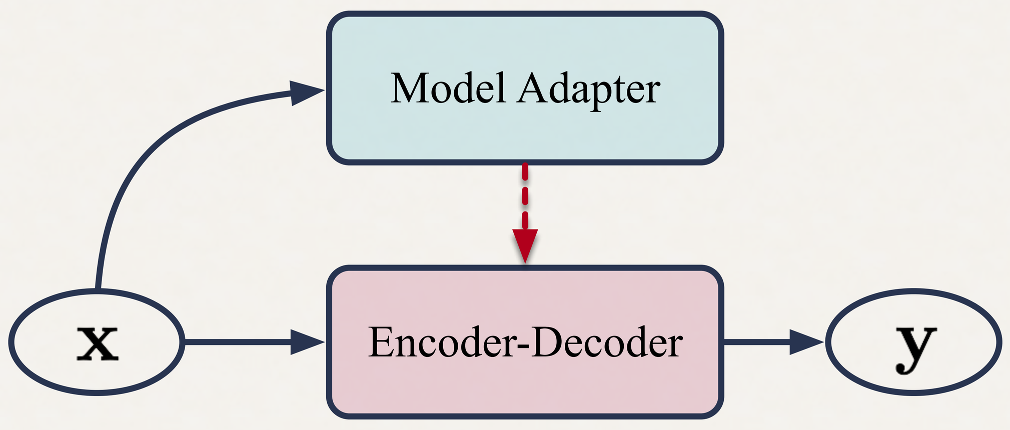


Figure 1: General model architecture. **Black solid lines** denote information flow, and the red dashed line indicates the adaptive parameterization operation.

Make the parameters of encoder-decoder **adaptive to its input** by:

- **Context-aware parameterization** (Fig. 2)
- **Topic-aware parameterization** (Fig. 3)

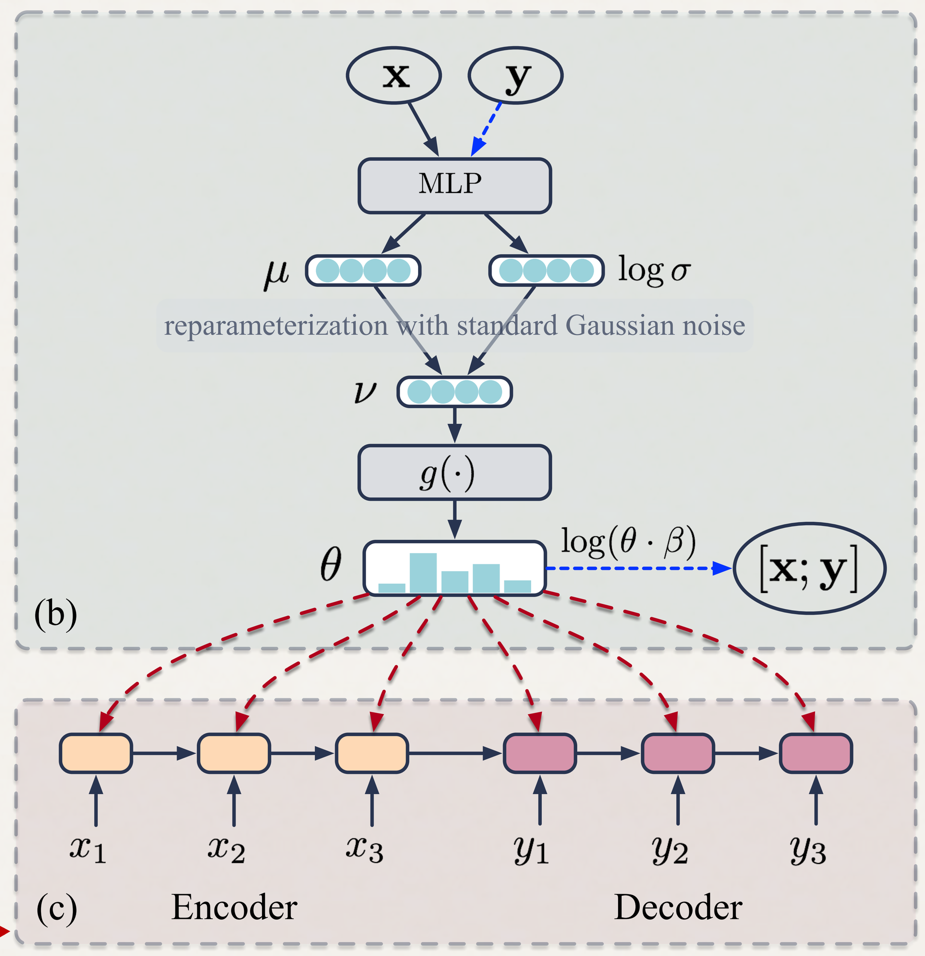
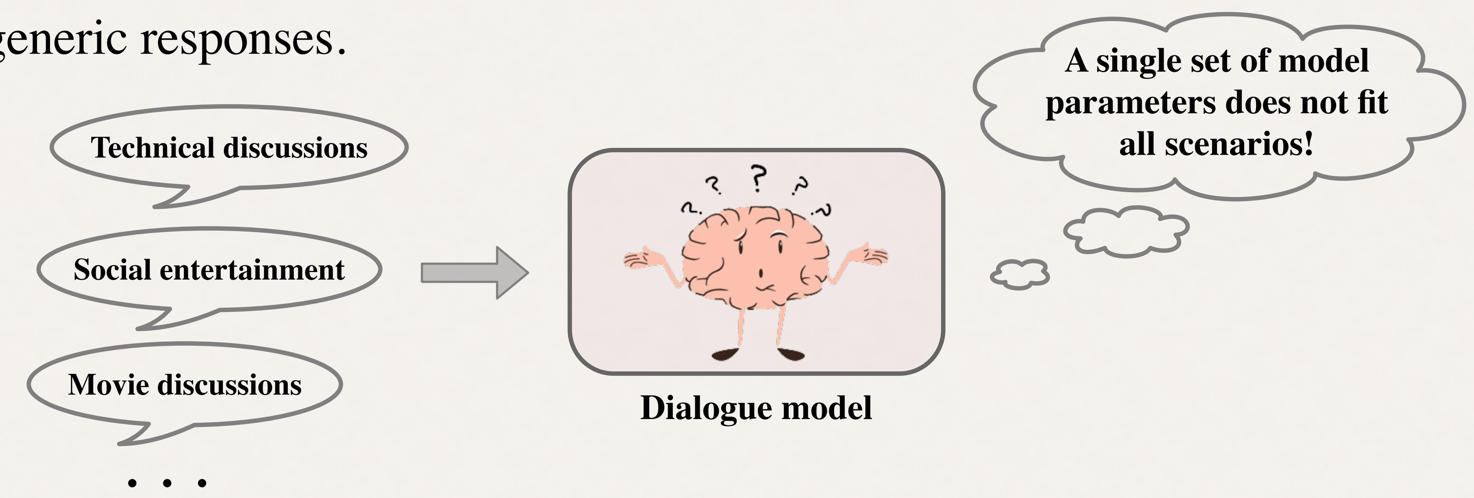


Figure 3: Topic-aware parameterization. (b) Latent topic inferrer.

Motivation

Neural conversation systems generate responses based on the sequence-to-sequence (SEQ2SEQ) paradigm. Typically, the model is equipped with **a single set of learned parameters** to generate responses for given input contexts. When confronting **diverse conversation scenarios**, its adaptability is rather limited and the model is hence prone to generate generic responses.



Method

Context-aware Parameterization (Fig. 2)

We parameterize the LSTM weights $W \in \mathbb{R}^{N_r \times N_c}$ by factorizing it as:

$$W = \mathcal{A}_c(\xi_t, U, V)$$
$$\xi_t = \Phi(\zeta, h_{t-1}),$$

where Φ is implemented as another LSTM unit, h_{t-1} is the previous encoder or decoder hidden states, ζ stands for the context representation output by the context encoder, and $\xi_t \in \mathbb{R}^{N_\zeta}$ is the context representation at timestep t. $\mathcal{A}_c$ denotes the context-aware parameterization function, defined as:

$$\mathcal{A}_c(\xi_t, U, V) = U\, \xi_t V^T,$$

Where $U \in \mathbb{R}^{N_r \times N_\zeta}$ and $V \in \mathbb{R}^{N_c \times N_\zeta}$ are learnable weights.

Topic-aware Parameterization (Fig. 3)

To enable parameter sharing among similar topics, we first distill the dialog topic distribution θ using variational topic inference and then construct the weights W upon θ.

Variational topic inference. 1) θ is constructed from the latent variable ν: $\theta = \mathrm{softmax}(\nu)$. 2) Given θ, the marginal likelihood of dialog d is formulated as: $p(d) = \int_\theta p(\theta) \prod_{i=1}^{|d|} \sum_{z_i} p(w_i|\beta_{z_i}) p(z_i|\theta) d\theta$, where topic assignment z_i can be integrated out and the likelihood of a word $w_i \in d$ can be factorized as: $\log p(w_i|\beta, \theta) = \log \sum_{z_i} [p(w_i|\beta_{z_i}) p(z_i|\theta)] = \log(\theta \cdot \beta^T)$. β is the topic-word distribution.

Construct the weights W upon θ. $W = \mathcal{A}_\kappa(\theta, U, V) = U\theta V^T$.

Parameterization with Both Context and Topics

$$W = \Psi_t \mathcal{A}_c(\xi_t, U_c, V_c) + (1 - \Psi_t)\mathcal{A}_\kappa(\theta, U_k, V_k)$$
$$\Psi_t = sigmoid(\xi_t, \theta)$$

Experiments

Dataset

We construct an open-domain conversation corpus covering a broad range of resources including a movie discussions dataset collected from Reddit, an Ubuntu technical corpus, and a chitchat dataset.

Comparison

Models	BLEU	Ave.	Gre.	Ext.	Dist1	Dist2	Dist3
SEQ2SEQ	0.845	69.60	64.94	45.29	0.282	0.592	0.787
CVAE (Zhao et al., 2017)	1.546	71.23	66.67	47.14	0.546	1.716	2.731
LAED (Zhao et al., 2018)	0.7545	69.91	63.55	43.12	0.389	0.916	1.243
TA-S2S (Xing et al., 2017)	1.465	72.47	65.9	45.19	0.359	0.799	1.016
DOM-S2S (Choudhary et al., 2017)	1.189	74.42	66.6	48.47	0.497	1.294	1.814
This work (w/ context para.)	1.94	74.03	66.76	49.23	0.649	1.889	2.745
This work (w/ topic para.)	**2.051**	74.17	66.65	49.04	0.591	1.699	2.438
This work (w/ both)	1.90	**75.59**	**67.25**	**51.17**	**0.709**	**2.10**	**3.108**

Discussion and Analysis

The adaptive parameterization function $\mathcal{A}$ is reminiscent of the Singular Value Decomposition (SVD). We investigate the orthogonality of the learned U and V. We trained our model multiple times with different parameter initialization methods and observe that UU^T and VV^T approximate identity matrices. We conjecture that such SVD-alike parameterization implicitly enforces orthogonality during training, even though $\mathcal{A}$ does not perform matrix decomposition actually.

We visualize the learned topical word embedding in variational topic inferrer. The discernible clusters of the topical words in Fig. 4 demonstrate that the topic inferrer in topic-aware parameterization effectively distills the latent topic distribution of each conversation.

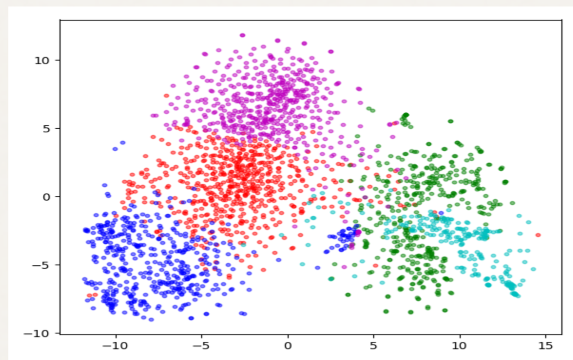


Figure 4: t-SNE projection of topical word embeddings.

Conclusion

We presents an adaptive neural dialogue generation model, which allows the dynamical parameterization of the model to each conversation and enables the generation of appropriate responses in diverse conversations. More specifically, we propose two adaptive parameterization approaches: context-aware parameterization which captures local semantics of the input context; and topic-aware parameterization which enables parameter sharing by first inferring the latent topics of the given context and then generating the parameters with the inferred latent topics. It should be noted that our approach is not isolated to only LSTMs. We would like to explore the effectiveness of the approach regarding other structures in future work.