The Thirty-Second AAAI Conference
on Artificial Intelligence (AAAI-18)

Towards Better Variational Encoder-Decoders in Seq2Seq Tasks

Xiaoyu Shen,[1] **Hui Su**[2]
[1]Saarland University, Saarbrücken, Germany
[2]Institute of Software, Chinese Academy of Science, China

Abstract

Variational encoder-decoders have shown promising results in seq2seq tasks. However, the training process is known difficult to be controlled because latent variables tend to be ignored while decoding. In this paper, we thoroughly analyze the reason behind this training difficulty, compare different ways of alleviating it and propose a new framework that helps significantly improve the overall performance.

Introduction

Conditional variational autoencoder (CVAE) with RNN encoder-decoders has been applied in several seq2seq tasks. However, it usually runs into the KL-vanishing problem that the RNN part ends up explaining everything without making use of the latent representation. In this paper, we take dialogue generation task as example, analyze why the KL-vanishing problem arises and compare different current strategies to tackle this problem. Instead of directly modeling the discrete dialogue distribution with latent variables, we propose a new framework that first extracts continuous vectors from the dialogue data which follow a simpler distribution, then establishes the link between them. Combined with a scheduled sampling trick, it can significantly outperform previous approaches. We hope the analysis and proposed framework can facilitate the research of CVAE seq2seq models.

Analysis

CVAE in dialogue generation CVAE generates dialogues as follows: z, which stands for a high-level latent representation, is sampled from the prior distribution $p_\theta(z|c)$, then response x is generated from $p_\theta(x|z,c)$. Though calculating the exact log-likelihood is intractable, it can be efficiently trained by optimizing the evidence lower bound (ELBO):

$$\begin{aligned} &-\log p_\theta(x|c) = -\log \int_z p_\theta(z|c)p_\theta(x|c,z)dz \\ &\leq -\mathbb{E}_{q_\phi(z|c,x)}[\log p_\theta(x|c,z)] + \mathrm{KL}(q_\phi(z|c,x)||p_\theta(z|c)) \end{aligned} \tag{1}$$

$q_\phi(z|c,x)$ is the approximated posterior distribution and c is the dialogue context. $q_\phi(z|x,c)$ and $p_\theta(z|c)$ are usually set as Gaussian distributions with diagonal covariance matrix.

KL-Vanishing Problem Straightforwardly optimizing with Eq. 1 suffers from the KL-vanishing problem because the RNN decoder $p_\theta(x|c,z)$ is a universal function approximator and tends to represent the density distribution without referring to the latent variable. At the beginning of the training process, when the approximate posterior $q_\phi(z|x,c)$ carries little useful information, it is natural for the model to blindly set $q_\phi(z|x,c)$ closer to the Gaussian prior $p_\theta(z|c)$ so that the extra cost from the KL divergence can be avoided.

We show there are two ways of alleviating this problem: weakening the power of decoders or improving the expressiveness of approximated posterior. Details can be found in the supplementary material.

Current Approaches Keeping the ELBO objective, we can weaken the decoder or improve the approximated posterior. For the former, word drop-out (Bowman et al. 2016) or Bag-of-word (BOW) loss (Zhao, Zhao, and Eskenazi 2017) are two popular ways. For the latter, (Serban et al. 2017) applies a piecewise distribution to replace the Gaussian prior distribution. (Kingma et al. 2016) used a normalizing flow to approximate the posterior distributions. All have their own advantages and disadvantages.

Some other ways modify the ELBO objective like KL-annealing, free bits (Kingma et al. 2016) or adversarial encoding (Makhzani et al. 2016). Detailed analysis is in the supplementary material.

Model

Our method does not aim at directly modeling the discrete dialogue data, which is difficult to match the distribution family of continuous latent variables. Instead, we first extract continuous variables from dialogue data that follow a simpler distribution and build a conditional variational encoder-decoder based on them. These continuous variables can reflect certain attributes of the dialogue data and can help recover the original dialogues. Specifically, we divide the training step into two phases: a denoising autoencoder (DAE) phase which aims at extracting continuous representations from dialogue data and a CVAE phase which builds a normal CVAE based on the extracted continuous representations.

In the CVAE phase, A sample h is obtained from the DAE by transforming dialogue texts into a continuous embedding and is used as a target for the maximum likelihood training

of the CVAE. We assume the generative model $p_\theta(\tilde{h}|z,c) = \mathcal{N}(h, I)$, the loss function is:

$$\min_{\phi} \mathrm{KL}(q_\phi(z|h,c)\|p_\phi(z|c)) + \frac{1}{2}\mathbb{E}_{q_\phi(z|h,c)}||g_\phi(z) - h||_2^2; \quad h = f_\theta(x,c) \tag{2}$$

The second squared loss item is the Gaussian likelihood, θ is fixed as part of DAE during training.

In the DAE phase, An observation x is sampled from the training data and fed into the transform function get a continuous vector representation $h = f_\theta(x)$. The corresponding latent variable z is sampled from the posterior distribution $q_\phi(z|h,c)$ provided by the CVAE part. The sampled latent variable z, together with x, forms a target for training the DAE. The objective function is:

$$\begin{aligned} \min_{\theta}\ & \max(\epsilon, \mathrm{KL}(q_\phi(z|h,c)\|p_\phi(z|c))) \\ & - \mathbb{E}_{q_\phi(z|h,c)}[log(p_\theta(x|\tilde{h},c))]; \\ & h = f_\theta(x), \tilde{h} = (1-p)g_\phi(z) + ph \end{aligned} \tag{3}$$

The first item is used to control KL divergence in a reasonable range such that the transformed h can be close to our Gaussian assumption. ϵ can be used to adjust the leverage between the reconstruction loss and KL divergence, where a lower ϵ value will lead to a lower KL divergence in the end. While the parameter ϕ is fixed, we can change the transform function f_θ to optimize with the KL divergence. ϵ acts as the reserved space as in free bits (Kingma et al. 2016), but we apply it on the whole dimension. p is the keeping rate defined in Eq. **??** in the supplementary material. Basically, we improve the variational encoder-decoder framework with a co-training process and a scheduled sampling strategy.

These two phases are trained alternatively until an equilibrium is achieved. When testing, a response can be generated by first sampling a latent variable from $p_\phi(z|c)$, then getting the noisy $\tilde{h}$ from $g_\phi(z)$ and decoding by $p_\theta(x|\tilde{h},c)$. There is no extra cost and $g_\phi(z)$ can be seen as adding an additional feedforward layer before feeding z as the input.

Experiment

We conduct our experiments on two dialogue datasets: Dailydialog (Li et al. 2017) and Switchboard (Godfrey and Holliman). These two datasets are randomly separated into training/validation/test sets with the ratio of 10:1:1.

Measurement and Comparison We compare our model with the basic HRED and several current approaches including KL-annealing (KLA), word drop-out (DO), free-bits (FB) and bag-of-words loss (BOW). The details are summarized in Table 1 and **??**. We set the reserved space for every dimension as 0.02 in free bits (FB) and also try reserving 5 bits for the whole dimension space (FB-all). We use an ϵ value 5 for our model with co-training (CO) and set the scheduled sampling (SS) weight $k = 2500$ or 5000 for Dailydialog or Switchboard. We also experiment with jointly training the DAE and CVAE part in our model and report the results. We measured the perplexity (PPL), KL divergence (KL) , negative log-likelihood (NLL) and percentage of unique sentences in the generated responses (Unique). NLL is averaged over all the 80-word slices within every batch. For latent variable models, NLL is computed as the ELBO, which is the lower bound of the real NLL.

Table 1: Metric Results on Dailydialog Dataset

Model	PPL	KL	NLL	Unique(%)
HRED	46.7	0.00	232.8	25.1
KLA	33.9	4.10	230.0	76.5
KLA+DO	29.8	3.81	224.5	78.2
KLA+BOW	27.7	7.78	236.4	**91.1**
FB	40.0	3.44	240.6	59.1
FB-all	28.8	4.96	226.9	82.9
CO	24.9	5.00	219.9	83.8
CO+DO	24.0	5.01	217.7	85.1
CO+SS	**22.1**	**4.93**	**212.3**	86.7
CO+SS(joint)	26.4	5.19	224.3	79.6

Results As can be seen, our model CO+SS achieves the lowest NLL with a high diversity over both datasets. The Schedule Sampling (SS) strategy significantly helps brings down the NLL. Jointly training the model brings recession on both the perplexity and KL divergence. Results on Switchboard are in the supplementary material.

Conclusion

CVAE with RNN encoder-decoders are known difficult to be trained. In this work, we thoroughly analyze the reason of the training difficulty and compare different current approaches, then propose a new framework that helps significantly improve the performance.

References

Bowman, S. R.; Vilnis, L.; Vinyals, O.; Dai, A. M.; Jozefowicz, R.; and Bengio, S. 2016. Generating sentences from a continuous space. *CONLL*.

Godfrey, J., and Holliman, E. Switchboard-1 release 2, 1997. *Linguistic Data Consortium, Philadelphia*.

Kingma, D. P.; Salimans, T.; Jozefowicz, R.; Chen, X.; Sutskever, I.; and Welling, M. 2016. Improved variational inference with inverse autoregressive flow. In *Advances in Neural Information Processing Systems*, 4743–4751.

Li, Y.; Hui, S.; Xiaoyu, S.; Wenjie, L.; Ziqiang, C.; and Shuzi, N. 2017. Dailydialog: A manually labelled multi-turn dialogue dataset. *IJCNLP*.

Makhzani, A.; Shlens, J.; Jaitly, N.; Goodfellow, I.; and Frey, B. 2016. Adversarial autoencoders. *ICLR*.

Serban, I. V.; II, A. G. O.; Pineau, J.; and Courville, A. 2017. Piecewise latent variables for neural variational text processing. *EMNLP*.

Zhao, T.; Zhao, R.; and Eskenazi, M. 2017. Learning discourse-level diversity for neural dialog models using conditional variational autoencoders. *ACL*.