IMPROVE DIVERSE TEXT GENERATION BY SELF LABELING CONDITIONAL VARIATIONAL AUTO ENCODER

Yuchi Zhang, Yongliang Wang, Liping Zhang, Zhiqiang Zhang, Kun Gai

Alibaba Group, Beijing, China
{yuchi.zyc, yongliang.wyl, shanrui.zlp, zhang.zhiqiang}@alibaba-inc.com, jingshi.gk@taobao.com

ABSTRACT

Diversity plays a vital role in many text generating applications. In recent years, Conditional Variational Auto Encoders (CVAE) have shown promising performances for this task. However, they often encounter the so called KL-Vanishing problem. Pervious works use heuristic methods to avoid KL-vanishing, but it is hard to find an appropriate degree to which these methods should be applied. In this paper, we propose an explicit optimizing objective function to guide the encoder towards the "best encoder" and directly pull the CVAE away from KL-vanishing. A labeling network is introduced to estimate the "best encoder". It provides a continuous label in the latent space of CVAE to help build a close connection between latent variables and targets. The whole proposed method is named Self Labeling CVAE (SLCVAE). To boost the research of diverse text generation, we also propose a large native one-to-many dataset. Extensive experiments are conducted on two tasks, which show that our method largely improves the generating diversity while achieving comparable accuracy compared with state-of-the-art algorithms.

Index Terms— Self Labeling, CVAE, KL-vanishing, text generation, diversity

1. INTRODUCTION

Text generating techniques are widely used in various tasks, such as dialogue generation [1, 2], image caption [3, 4] and question-answer systems [5, 6], etc. Encoder-decoder models such as SEQ2SEQ[7] have been widely adopted in text generating tasks due to its accuracy. However, as conventional encoder-decoder models encode same input patterns to same unique representative vectors without any variation, their ability of generating different sentences from one input (also known as the "one-to-many" problem [2]) are limited. Thus they are not good at handling text generating tasks which further require results with diversity besides accuracy. Such as open-domain dialogue systems and selling point generation in e-commerce systems.

In the early periods, methods are proposed to interfere the inference stage of a well-trained encoder-decoder model to encourage abundant outputs. Such as MMI-AntiLM [8] and diverse beam search [9]. The drawback of such methods is that they do not optimize the encoder-decoder models to fit multi-target data and the quality of their generating results is limited by the trade-off between accuracy and diversity.

Recently, variational encoder-decoders such as Variational Auto Encoder (VAE) [10, 11] and Conditional VAE (CVAE) [12, 13] have shown great potentials in solving the "one-to-many" problems. These methods introduced an intermediate latent variable and assume that each configuration of the latent variable corresponds to a feasible response. Thus diverse responses can be generated by sampling the variable. However, both VAE and CVAE have encountered the KL-vanishing problem that the decoder tends to model the targets without making use of the latent variables. To solve such problem, various methods have been proposed. Such as *KL annealing* (KLA), and *word-dropout* operation (WD) proposed in [14], and *bag-of-word* (BOW) loss proposed in [2]. These approaches, in essence, weaken the decoder or strengthen the encoder to make compensation to the objective function of VAE/CVAE and mitigate the KL vanishing problem. However, it is hard to determine how weak/strong the decoder/encoder should be.

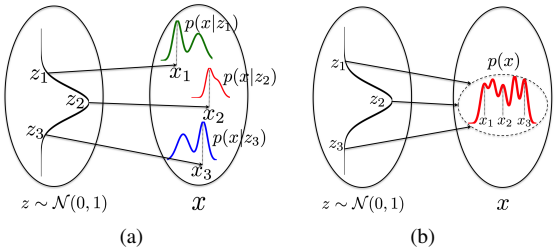

Fig. 1. Illustration of the generation process and KL-vanishing. (a) In assumption, each configuration of the latent variable is mapped by the decoder into a different decoding distribution $p(x|z)$. Benefiting from the latent distribution of z, all $p(x|z)$ with different z can have a good coverage of the entire target space of x, while $p(x|z)$ itself can be simple and only responsible for decoding one target. (b) When KL-vanishing takes place, zs lose the expressiveness of x and collapse to a same decoding distribution $p(x)$. Thus $p(x)$ trying to fit the entire space of x alone has a very complex structure and might only have a poor coverage of the space and lack the diversity.

Orthogonal to current approaches above, we propose an explicit optimization objective for the encoder to move towards the "best encoder" for better expressiveness to fit current decoder. Specifically, an additional module called "labeling network" is used to estimate the "best encoder" for the current decoder. Then a loss which measures the difference between the latent variable of CVAE and predicted variable from labeling network is added to the original objective function of the CVAE. Since this loss pulls the encoder towards the "best encoder" approximated by the labeling network and meanwhile original CVAE pulls encoder to the prior, an equilibrium will be reached where KL-vanishing can be avoided. A large scale dataset called EGOODS which contains native one-to-many text data of high quality is constructed to accelerate the research of diverse text generation. Experiments are conducted on a public dialogue dataset and the proposed EGOODS dataset, which demonstrate that our method called SLCVAE improves the diversity of text generation without losing accuracy compared to several state-of-art methods.

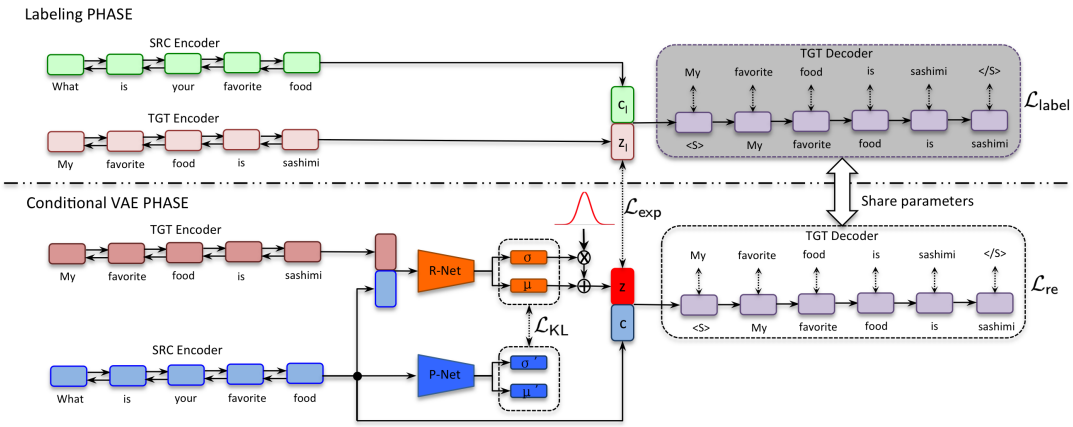

Fig. 2. Overview of the proposed method. The top part is the Labeling Phase and the bottom part is the CVAE Phase. The model optimized alternatively trained between the two phases. SRC and TGT are abbreviations of source and target. R-Net and P-Net are Recognition Network and Prior Network for the reparameterization trick [2]. $\mathcal{L}_{\text{re}}$ denotes a reconstruction loss in ELBO and $\mathcal{L}_{\text{KL}}$ denotes the KL divergence term.

2. SELF LABELING CVAE

Conventional VAE makes use of a latent variable z sampled from a prior distribution to generate data x. The logarithm likelihood of the data x is optimized by maximizing the evidence lower bound (ELBO) [10, 11] :

$$\log p(x) \geq \mathbb{E}_{q(z|x)}[\log p(x|z)] - KL(q(z|x)||p(z)) \quad (1)$$

where $q(z|x)$ and $p(x|z)$ are output distributions of the encoder and decoder respectively. And KL means the Kullback–Leibler divergence [15]. Note that our goal is to generate diverse x using different z. Two conditions should be satisfied: First, each z should correspond to a unique x through the decoder. Second, z should obey the prior distribution $p(z)$. Maximizing Equation 1 encourages the latter by pulling encoder's output distribution of z to $p(z)$. However, with $q(z|x)$ moving towards $p(z)$ during the optimizing procedure, z loses the discriminative information of different x and the decoder tends to fit the data even without the help of encoders. Such phenomenon is called KL-vanishing [14, 2]. As a consequence, the first condition is violated and multiple zs will collapse to a same averaged output distribution $p(x)$ as is shown in Fig. 1(b).

Thus we propose to strengthen the connection between the latent z and target x via maintaining the expressiveness of the encoder. As illustrated in Fig. 1(a), considering that an expressive z has the ability to recover a unique target through the decoder, the decoder itself can then be used to find the most expressive z' given a certain target x. This in concept equivalents to finding the inverse image of x of the decoder. So the inverse image z' of x can be regarded as the effectiveness label x in the continuous latent space. And if z' has been obtained, then we are reasonably motivated to pull the encoder distribution $p(z|x)$ to be close to z' to maintain the expressiveness of the encoder.

However, finding the inverse image of the decoder exactly is not an easy task. To overcome this, we introduce an extra network to approximate z' which is the inverse image of x output by the decoder. This network, whose output is denoted as z_{label}, estimates the effectiveness label of z in essence and is therefore named as *labeling network*. It can also be considered as an approximation to the ideal encoder for the current decoder.

Specifically, the *labeling network* shares the same network structure with the original encoder of VAE, but it only outputs the variable z_{label} rather than the reparameterized distribution. As the output of the VAE encoder is a distribution $q(z|x)$, we put the expressive constraint on the expectation of the L_2 distance $||z - z_{label}||^2$ between encoded latent variable and z_{label} over the encoder distribution $q(z|x)$. Thus an expressiveness objective function is defined as follows:

$$\mathcal{L}_{\text{exp}} = \mathbb{E}_{q(z|x)}[||z - z_{label}||^2] \quad (2)$$

which is minimized to encourage the encoder to be more expressive. By using $g(x)$ to denote the labeling network i.e. $z_{label} = g(x)$, and adding $\mathcal{L}_{\text{exp}}$ as an additional term to the VAE's objective function, we get the total objective function in following:

$$\begin{aligned}\mathcal{L}_{\text{SLVAE}} = &-\mathbb{E}_{q(z|x)}[\log p(x|z)] + KL(q(z|x)||p(z)) \\ &+\lambda \mathbb{E}_{q(z|x)}[||z - g(x)||^2]\end{aligned} \quad (3)$$

From this formulation, we can see that, $q(z|x)$ is not only pulled to $p(z)$ like before, but also pulled to the estimated "best encoder" for the decoder. The hyper-parameter λ is used to control the importance of the expressiveness objective. The "best encoder" can expand a comprehensive coverage of the target space through the current decoder. Thus they will reach an equilibrium at which the $p(z|x)$ is close to $p(z)$ and also remains the expressiveness. As we incorporate a labeling network into original VAE to estimate the most expressive latent label given the decoder and strengthen the connection between the latent z and target x through the decoder itself, we call this method Self Labeling VAE (SLVAE).

When it comes to CVAE, things remain the same except that everything is conditioned on c. And the objective function becomes:

$$\begin{aligned}\mathcal{L}_{\text{SLCVAE}} = &-\mathbb{E}_{q(z|x,c)}[\log p(x|z,c)] + KL(q(z|x,c)||p(z|c)) \\ &+\lambda \mathbb{E}_{q(z|x,c)}[||z - g(x)||^2]\end{aligned} \quad (4)$$

Similarly, we call this model SLCVAE.

As we discussed before, $g(x)$ should be the "best encoder" for the decoder to recover x. Thus we should optimize $g(x)$ by maximizing the following objective function:

$$\log p(x|z_{label}, c) = \log p(x|g(x), c) \quad (5)$$

with the decoder fixed.

Fig. 2 shows the overview of the whole proposed method. To optimize Equation. 4 and Equation. 5, we parameterize all the three modules: the encoder $q_\phi(z|x,c)$ and decoder $p_\theta(x|z,c)$ of the

CVAE, and the labeling network $g_\gamma(x, c)$. An alternative training schedule is used with two phases: the CVAE phase and the Labeling phase.

In the CVAE phase, we minimize the loss function of the SLCVAE:

$$\begin{aligned}\min_{\phi,\theta,\beta} \mathcal{L}_{\mathsf{SLCVAE}} = \min_{\phi,\theta,\beta}[&-\mathbb{E}_{q_\phi(z|x,c)}[p_\theta(x|z,c)] \\ &+KL(q_\phi(z|x,c)||p_\beta(z|c)) \\ &+\lambda\mathbb{E}_{q_\phi(z|x,c)}[||z-g_\gamma(x)||^2]]\end{aligned} \tag{6}$$

where β are parameters of the prior network. In this phase, the labelling network $g_\gamma(x, c)$ is fixed to provide a z_{label} corresponding to each x.

In the Labeling phase, we minimize the loss function of the labeling network:

$$\min_{\gamma} \mathcal{L}_{\mathsf{label}} = \min_{\gamma}[-p_\theta(x|g_\gamma(x), c)] \tag{7}$$

The decoder is fixed at this time to get the good expressive label for current decoder.

3. THE EGOODS DATASET

The "one-to-many" text generating problem is an active research topic and plays important roles in many tasks. However, there still lacks real one-to-many datasets to improve and evaluate the algorithms for this problem. Most current datasets that come from dialogue system are essentially one-to-one corpora. Although there may exist various underlying responses for a certain question, these datasets only contain one answer for each dialogue context due to data source limitations.

To fulfill the gap between the demand and status quo for one-to-many dataset, we collect a large scale item description corpus from a Chinese e-commerce website to construct the native one-to-many dataset. In this corpus, each item has one description provided by their sellers and multiple recommendation sentences written by third-party who is payed to make these sentences more attractive to customers. The descriptions provided by sellers are usually texts stacking many keywords of the item properties. On the contrary, the recommendation sentences are written according to item descriptions but read more smoothly. For the text generation task, we naturally use the sellers' descriptions as the source to generate multiple recommendation sentences mimicking humans. This corpus originates from a real business in which texts are of high quality and coherent with sources. We call this very large and native one-to-many dataset EGOODS.

After simple cleaning and formatting, EGOODS dataset contains 3001140 source and target pairs from 789582 items in total. So each source item description has 3.8 target recommendation sentences on average. The dataset is split into training/validation/testing parts with respect to items, each of which contains 2961317/19536/20287 pairs.

4. EXPERIMENTS

4.1. Experimental Setups

Our experiments are conducted on two text generating tasks: open-domain dialogue generation and recommendation sentence generation. For the first task, the public dialogue dataset Daily Dialog (DD) [16] is used. DD dataset is collected from different websites under 10 topics. It contains 13118 multi-turn dialogue sessions in English, and is split into training/validation/testing set of 11118/1000/1000 sessions. For each full speaker turn, we use all utterances but the last one as the dialogue context to predict the last one. Need to note that though there may exist various responses for a question, DD dataset essentially only contains one-to-one data. To better model and evaluate the diversity, the newly constructed one-to-many dataset EGOODS is adopted in the second task.

We compared our SLCVAE (SL) to 4 strong baselines: SEQ2SEQ [7], MMI-AntiLM [8], CVAE and CVAE with *bag-of-word loss* (BOW) [2]. Several training skills, such as *KL-annealing*(KLA) and *word dropout*(WD) [14], are used in combination with baselines and our method to improve the performance. All methods are required to generate 10 responses for each given input. Note that although the SEQ2SEQ model uses deterministic encoding vectors, the widely adopted beam search strategy can be applied during inference procedure to generate 10-best decoding results which corresponds to 10 responses (denoted as SEQ2SEQ+BS).

The whole structure of SLCVAE is implemented with the famous open source library PyTorch[17]. Encoders are two-layer bidirectional RNNs [18] with Gated Recurrent Units (GRU) [19] and the decoders are two-layer unidirectional RNN with GRUs throughout all experiments. For DD and EGOODS dataset respectively, the word embedding sizes and hidden dimensions of RNN are set to 32 and 128 according to the size of each dataset. And in all CVAE-based methods, the latent variable dimensions are set to 8 and 16 for two datasets separately. The coefficients of the labeling network (λ) are set to be 0.5 and 0.1. The Adam optimizer [20] with a learning rate of 0.0001 is used to train all models with batch sizes of 64 and 128 for two datasets.Training skills of KLA and WD are also used to get further better performance.

Accuracy and diversity are two sides of the generations we need to concern. Automatic quantitative measures for these purposes are still an open research challenge [21, 22]. [2] proposed BLEU-precision and BLEU-recall metrics for discourse-level accuracy and diversity respectively as following:

$$\begin{aligned}\mathsf{precision(c)} &= \frac{\sum_{i=1}^{N} \max_{j\in[1,M_c]} d(r_j, h_i)}{N} \\ \mathsf{recall(c)} &= \frac{\sum_{j=1}^{M_c} \max_{i\in[1,N]} d(r_j, h_i)}{M_c}\end{aligned} \tag{8}$$

where $d(r, h)$ means a similarity metric between a generated sentence h and a reference r. BLEU-1, BLEU-2 and BLEU-3 are adopted as such metric in our experiment and their average result is calculated as the final quantitative measure. However, BLEU-recall is defined based on lexical similarity, which might penalize a reasonable but not same prediction. Following [8], we also use the number of *distinct n-gram* to measure the word-level diversity. The *distinct* is normalized to [0, 1] by dividing the total number of generated tokens. In summary, BLEU-precision is reported as the accuracy measure, and BLEU-recall, *distinct-1* and *distinct-2* are reported as diversity measures.

We also conduct human evaluations on the EGOODS dataset. 7 human experts are employed to measure the fluency of generated sentences, coherence of each sentence to source and diversity. For fluency and coherence, experts are asked to vote to each sentence. Sentences which yield more than 4 votes are good sentences. The ratio of good sentences are reported. For diversity, 5 level of diverse scores are introduced. The higher the score, the more diverse the sentence is. The final diversity score of each sentence is the average

Table 1. Results on Daily Dialog (DD). The bottom 3 lines are CVAE based methods.

Methods	BLEU-prec	BLEU-recall	*distinct-1*	*distinct-2*
SEQ2SEQ+BS	0.164	0.282	0.002	0.007
MMI-AntiLM	0.153	0.275	0.002	0.012
KLA+WD	0.212	0.345	0.010	0.041
KLA+WD+BOW	0.210	0.344	0.013	0.066
KLA+WD+SL	**0.214**	**0.354**	**0.014**	**0.078**

Table 2. Results on EGOODS. The bottom 3 lines are CVAE based methods.

Methods	BLEU-prec	BLEU-recall	*distinct-1*	*distinct-2*
SEQ2SEQ+BS	**0.379**	0.388	0.0012	0.0042
MMI-AntiLM	0.356	0.374	0.0021	0.0146
KLA+WD	0.373	**0.405**	0.0039	0.0216
KLA+WD+BOW	0.374	0.404	0.0039	0.0231
KLA+WD+SL	0.373	**0.405**	**0.0049**	**0.0270**

score of all experts.

4.2. Results

4.2.1. Automatic Quantitative Measurement

Table. 1 shows the evaluation results of all methods on Daily Dialog dataset. Training skills of KLA and WD are used for all CVAE based methods. We can see that our proposed method outperforms all baselines in terms of all the 4 metrics on this task. This confirms our insight of the generating process that our labeling objective can lead to an equilibrium at which the KL-vanishing problem is significantly relieved and so result in better diversity. Remind that Daily Dialog is actually a *one-to-one* dataset. The better performance in diversity on DD demonstrates that our model can better exploit such training data without explicit *one-to-many* annotations.

Performances of different methods on EGOODS are shown in Table. 2. Our method achieves comparable accuracy with baselines and best diversity among all methods. This demonstrates the effectiveness of SLCVAE on the *one-to-many* data. In detail, our method harvests the much better gains on word-level diversity while is only slightly better than CVAE on BLEU-recall. We explain this in two folds: First, strong baselines can benefit from the large scale and *one-to-many* nature of EGOODS to better fit the multiple targets. Another reason is that automatically evaluating the quality of generated texts is very challenging. BLEU-recall only measures the coverage of hypothesis for the annotated targets, and could not judge good algorithms precisely when the annotations are limited. In such situation, *distinct* measures the vocabulary a model actually uses and demonstrates its absolute lexical diversity. Furthermore, we observed that SEQ2SEQ+BS obtains the best BLEU-precision among all methods on EGOODS, but it performs much worse on Daily Dialog. Meanwhile, the BLEU-recall gap between SEQ2SEQ+BS and the best result on EGOODS is obviously small than that on DD. We point out that our dataset especially designed for "one-to-many" problem significantly improves the generation quality of SEQ2SEQ methods.

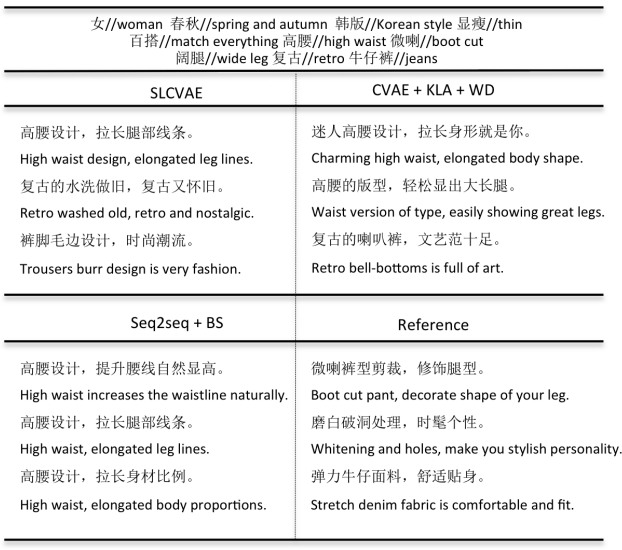

Fig. 3. Example of generated texts.

4.2.2. Human Evaluation

Human evaluation results on EGOODS are shown in Table 3. Such results show that our method achieves comparable fluency and coherence as baseline methods, but our diversity is much higher than other models. Although the SEQ2SEQ+BS method achieves the best fluency, it sacrifices too much diversity, which means the result is monotonous and dull.

Table 3. Human evaluation results.

Methods	Fluency(%)	Coherence(%)	Diversity
SEQ2SEQ+BS	**96**	65	1.55
KLA+WD	87	64	3.12
KLA+WD+BOW	83	**66**	3.18
KLA+WD+SL	91	**66**	**3.32**

4.2.3. Text Generating Examples

Fig. 3 shows an example of generated texts for EGOODS. All source and generated sentences are displayed with their English translations. 3 results are generated separately by SEQ2SEQ+BS, CVAE and our method SLCVAE. The results from all three methods are of good fluency and coherent to the input. But obviously SEQ2SEQ+BS fails to show different expressions thus gets poor diversity. Both CVAE model and our method tend to show stronger abilities in generating diversely than SEQ2SEQ+BS, since we can see that the generated results have better coverages for the references. This finding is consistent with the quantitative experiment results we have discussed above.

5. CONCLUSION

Recently CVAE based methods show great potentials for "one-to-many" text generating tasks. However CVAE working with RNNs tends to run into the KL-vanishing problem. In this paper, we propose the self labeling mechanism which connects the decoder with latent variable by an explicit optimization objective. It leads the encoder to reach an equilibrium at which the decoder can take full advantage of the latent variable. Experiments show that SLCAVE largely improves the generating diversity.

6. REFERENCES

[1] Jiwei Li, Will Monroe, Alan Ritter, Michel Galley, Jianfeng Gao, and Dan Jurafsky, "Deep reinforcement learning for dialogue generation," *arXiv preprint arXiv:1606.01541*, 2016.

[2] Tiancheng Zhao, Ran Zhao, and Maxine Eskenazi, "Learning discourse-level diversity for neural dialog models using conditional variational autoencoders," in *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2017, vol. 1, pp. 654–664.

[3] Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan, "Show and tell: A neural image caption generator," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 3156–3164.

[4] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio, "Show, attend and tell: Neural image caption generation with visual attention," in *International conference on machine learning*, 2015, pp. 2048–2057.

[5] Bridget B Beamon, Michael D Whitley, and Robert L Yates, "Justifying passage machine learning for question and answer systems," Apr. 4 2017, US Patent 9,613,317.

[6] Jonghoon Oh, Kentaro Torisawa, Chikara Hashimoto, Takuya Kawada, Stijn De Saeger, Junichi Kazama, and Yiou Wang, "Non-factoid question-answering system and computer program," July 4 2017, US Patent 9,697,477.

[7] Ilya Sutskever, Oriol Vinyals, and Quoc V Le, "Sequence to sequence learning with neural networks," in *Advances in neural information processing systems*, 2014, pp. 3104–3112.

[8] Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan, "A diversity-promoting objective function for neural conversation models," in *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2016, pp. 110–119.

[9] Ashwin K Vijayakumar, Michael Cogswell, Ramprasath R Selvaraju, Qing Sun, Stefan Lee, David Crandall, and Dhruv Batra, "Diverse beam search: Decoding diverse solutions from neural sequence models," *arXiv preprint arXiv:1610.02424*, 2016.

[10] Danilo Jimenez Rezende, Shakir Mohamed, and Daan Wierstra, "Stochastic backpropagation and approximate inference in deep generative models," in *International Conference on Machine Learning*, 2014, pp. 1278–1286.

[11] Diederik P Kingma and Max Welling, "Auto-encoding variational bayes," *arXiv preprint arXiv:1312.6114*, 2013.

[12] Xinchen Yan, Jimei Yang, Kihyuk Sohn, and Honglak Lee, "Attribute2image: Conditional image generation from visual attributes," in *European Conference on Computer Vision*. Springer, 2016, pp. 776–791.

[13] Kihyuk Sohn, Honglak Lee, and Xinchen Yan, "Learning structured output representation using deep conditional generative models," in *Advances in Neural Information Processing Systems*, 2015, pp. 3483–3491.

[14] Samuel R Bowman, Luke Vilnis, Oriol Vinyals, Andrew Dai, Rafal Jozefowicz, and Samy Bengio, "Generating sentences from a continuous space," in *Proceedings of The 20th SIGNLL Conference on Computational Natural Language Learning*, 2016, pp. 10–21.

[15] Solomon Kullback and Richard A Leibler, "On information and sufficiency," *The annals of mathematical statistics*, vol. 22, no. 1, pp. 79–86, 1951.

[16] Yanran Li, Hui Su, Xiaoyu Shen, Wenjie Li, Ziqiang Cao, and Shuzi Niu, "Dailydialog: A manually labelled multi-turn dialogue dataset," in *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, 2017, vol. 1, pp. 986–995.

[17] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer, "Automatic differentiation in pytorch," in *NIPS-W*, 2017.

[18] Mike Schuster and Kuldip K Paliwal, "Bidirectional recurrent neural networks," *IEEE Transactions on Signal Processing*, vol. 45, no. 11, pp. 2673–2681, 1997.

[19] Junyoung Chung, Caglar Gulcehre, Kyunghyun Cho, and Yoshua Bengio, "Empirical evaluation of gated recurrent neural networks on sequence modeling," in *NIPS 2014 Workshop on Deep Learning, December 2014*, 2014.

[20] DP Kingma, LJ Ba, et al., "Adam: A method for stochastic optimization," 2015.

[21] Chia-Wei Liu, Ryan Lowe, Iulian V Serban, Michael Noseworthy, Laurent Charlin, and Joelle Pineau, "How not to evaluate your dialogue system: An empirical study of unsupervised evaluation metrics for dialogue response generation," .

[22] Xiaowei Tong, Zhenxin Fu, Mingyue Shang, Dongyan Zhao, and Rui Yan, "One" ruler" for all languages: Multi-lingual dialogue evaluation with adversarial multi-task learning," *arXiv preprint arXiv:1805.02914*, 2018.