Differentiated Distribution Recovery for Neural Text Generation

Jianing Li†, Yanyan Lan†*, Jiafeng Guo†, Jun Xu†, Xueqi Cheng†

† CAS Key Laboratory of Network Data Science and Technology,

Institute of Computing Technology, Chinese Academy of Sciences, Beijing, China

† University of Chinese Academy of Sciences, Beijing, China

** Department of Statistics, University of California, Berkeley*

{lijianing, lanyanyan, guojiafeng, junxu, cxq}@ict.ac.cn

BACKGROUND

Problem:

The primary goal of text generation models is to achieve high generation quality, featuring by less grammatical and logical errors, which is essential to pass the Turing Test. We focus on the unconditional setting, where the model generates text from scratch without receiving any conditions. The most commonly used method is RNN-based language models (RNNLM)[1] trained with Maximum Likelihood Estimation (MLE). However, RNNLM tends to generate flawed text, showing unsatisfactory Turing Test pass rate. Recent works try to improve the generation quality by incorporating adversarial training and reinforcement learning[2,3]. While leading to higher quality, generation diversity decreases relatively, and it is hard to control the balance. What is worse, such methods usually pose heavy training overhead.

Motivation:

We find out that the deficiency of RNNLM in generating high-quality text lie in the objective of MLE. MLE minimizes the Kullback-Leibler Distance between the model distribution and the real distribution, making the training sensitive to rare patterns and errors in the training data. Instead of a precise distribution recovery as in MLE, we propose Differentiated Distribution Recovery (DDR), which improves generation quality by focusing more on significant patterns and neglecting bad training samples.

DIFFERENTIATED DISTRIBUTION RECOVERY

Main Idea: Make the generation probability be proportional to the β-th power of real probability

$$\boxed{Q(x)} \propto \boxed{P(x)}^{\beta}, \qquad \beta > 1$$

Generation Probability   Real Probability

Implementation: Change the objective function for RNNLM

$$\max_{Q} \mathbb{E}_{x\sim P}\, f[Q(x)],$$
$$f(Q(x);\alpha) = \alpha \cdot Q(x)^{\frac{1}{\alpha}} - \alpha,$$
$$\alpha = \frac{\beta}{\beta-1} > 1$$

$$\begin{aligned}\mathcal{L}(\mathcal{D};\alpha) &= -\frac{1}{N}\sum_{i=1}^{N} \alpha \cdot Q(Y_{1:T}^{i})^{\frac{1}{\alpha}} \\ &= -\frac{\alpha}{N}\sum_{i=1}^{N}\prod_{t=1}^{T} Q(y_t^i|Y_{1:t-1}^i)^{\frac{1}{\alpha}} \\ &= -\frac{\alpha}{N}\sum_{i=1}^{N}\exp\left\{\frac{1}{\alpha}\sum_{t=1}^{T}\log Q(y_t^i|Y_{1:t-1}^i)\right\}\end{aligned}$$

Theorem

Let P and Q be two discrete distributions. With an objective defined as

$$\max_{Q} \mathbb{E}_{x\sim P}\, f[Q(x)],$$
$$f(Q(x);\alpha) = \alpha \cdot Q(x)^{\frac{1}{\alpha}} - \alpha, \qquad \alpha > 1,$$

The optimal Q with respect to the objective can be written as:

$$Q^*(x) = \frac{P(x)^{\beta}}{\sum_x P(x)^{\beta}}, \qquad \beta = \frac{\alpha}{\alpha - 1}$$

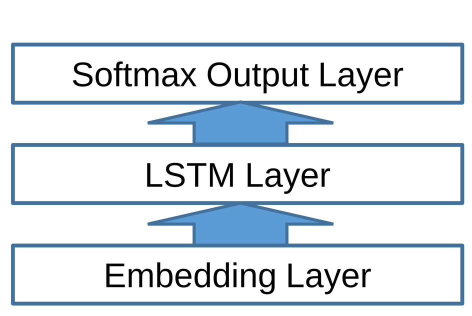
Advantages of DDR:

- Generation quality can be greatly improved
- Easy to implement
- No extra training overhead than a standard RNNLM
- The tradeoff between generation quality and diversity can be easily controlled by a hyper-parameter α

EXPERIMENTS

Settings

General model architecture



Datasets:

- Synthetic data
- MSCOCO Image Caption dataset
- EMNLP2017 WMT News dataset

Baselines:

- RNNLM [1]
- SeqGAN [2]
- LeakGAN [3]

Evaluations metrics

Quality metric:

- NLL[2] for synthetic dataset
- BLEU-N[4] for real datasets

Diversity metric:

- Distinct-N[5] for all datasets

$$NLL = -\mathbb{E}_{Y_{1:T}\sim Q}\left[\sum_{t=1}^{T}\log G_{oracle}(y_t|Y_{1:t-1})\right]$$

$$Distinct\_N = \frac{\#\,Unique\ n\_grams}{\#\,Total\ n\_grams}$$

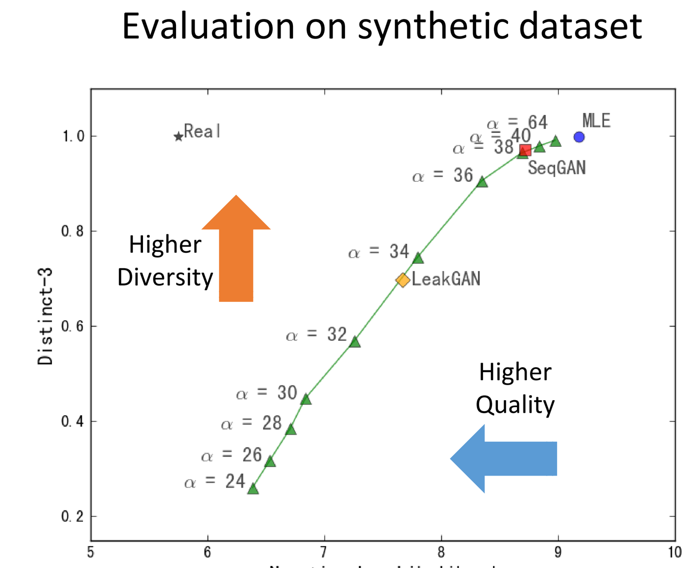
Turing Test Score:
Human evaluation on MSCOCO dataset, showing percentage of generated samples which are regarded as real.
Bad Samples%:
Add 10% noise in MSCOCO dataset and report percentage of generated bad samples.

Method	Turing Test Score	Bad Samples %
Ground Truth	0.772	-
MLE	0.490	8.0
SeqGAN	0.706	2.2
LeakGAN	0.758	**0.0**
DDR($\alpha=64$)	0.586	0.4
DDR($\alpha=32$)	-	0.1
DDR($\alpha=16$)	-	**0.0**
DDR($\alpha=8$)	0.692	**0.0**
DDR($\alpha=2$)	**0.932**	**0.0**

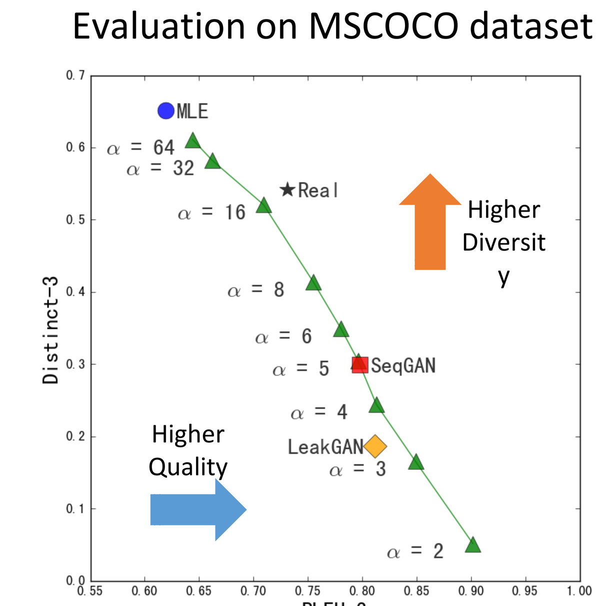
Conclusion:

- DDR achieves higher Turing Test pass rate with lower α, which is consistent with NLL and BLEU-N metrics
- DDR achieves higher Turing Test pass rate than MLE consistently, and surpass SeqGAN/LeakGAN with $\alpha = 2$, showing significantly higher generation quality
- DDR makes the model more robust against noises in training data than MLE.

Evaluation on synthetic dataset



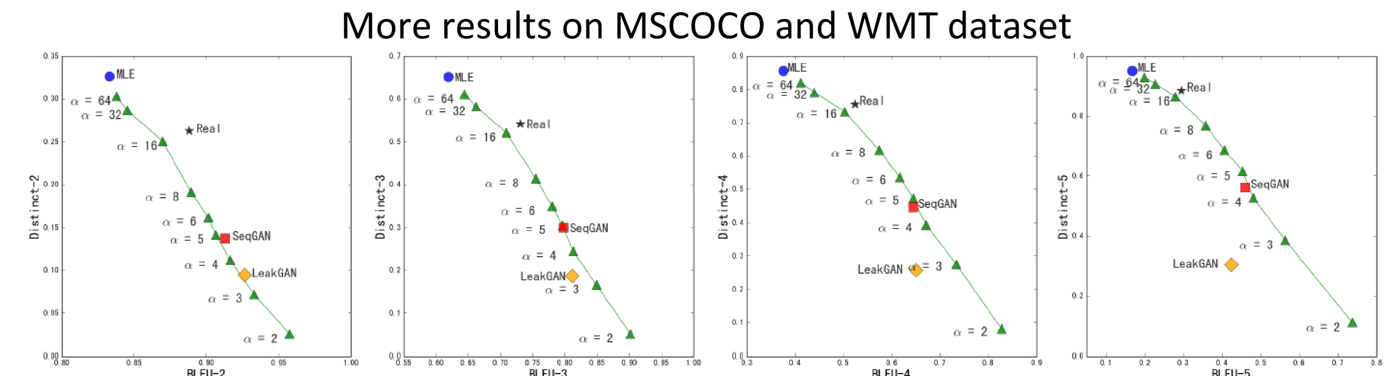
Evaluation on MSCOCO dataset



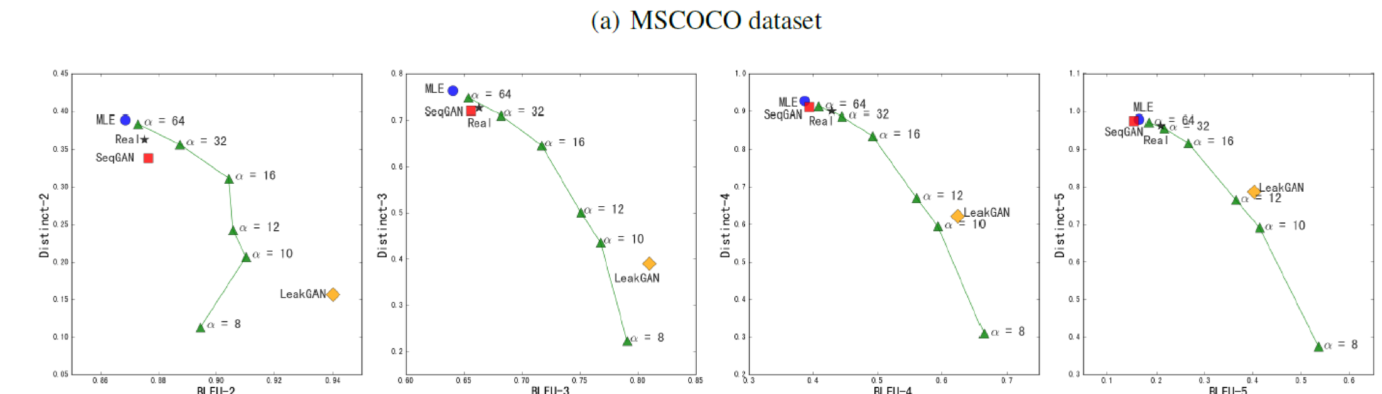
Conclusion:

- DDR achieves higher generation quality than MLE, and competitive generation quality as SeqGAN and LeakGAN on the same diversity level
- DDR with lower α achieves lower diversity but higher quality
- DDR achieve higher generation quality than SeqGAN and LeakGAN with low α values

More results on MSCOCO and WMT dataset



(a) MSCOCO dataset



(b) WMT dataset

REFERENCES

[1] Mikolov, T.; Karafia´t,M.; Burget, L.; Cˇ ernocky`, J.; and Khudanpur, S. 2010. Recurrent neural network based language model. In Eleventh Annual Conference of the International Speech Communication Association.
[2] Yu, L.; Zhang, W.; Wang, J.; and Yu, Y. 2017. Seqgan: Sequence generative adversarial nets with policy gradient. In AAAI, 2852–2858.
[3] Guo, J.; Lu, S.; Cai, H.; Zhang, W.; Yu, Y.; and Wang, J. 2017. Long text generation via adversarial training with leaked information. arXiv preprint arXiv:1709.08624.
[4] Papineni, K.; Roukos, S.; Ward, T.; and Zhu, W.-J. 2002. Bleu: a method for automatic evaluation of machine translation. In Proceedings of the 40th annual meeting on association for computational linguistics, 311–318. Association for Computational Linguistics.
[5] Li, J.; Galley, M.; Brockett, C.; Gao, J.; and Dolan, B. 2015. A diversity-promoting objective function for neural conversation models. arXiv preprint arXiv:1510.03055.