Rethinking Exposure Bias In Language Modeling

Yifan Xu *, Kening Zhang *, Haoyu Dong, Yuezhou Sun, Wenlong Zhao & Zhuowen Tu
University of California San Diego
{yix081,kez040, had002, yus174, wez094,ztu}@ucsd.edu

Abstract
Exposure bias describes the phenomenon that a language model trained under the teacher forcing schema may perform poorly at the inference stage when its predictions are conditioned on its previous predictions unseen from the training corpus. Recently, several generative adversarial networks (GANs) and reinforcement learning (RL) methods have been introduced to alleviate this problem. Nonetheless, a common issue in RL and GANs training is the sparsity of reward signals. In this paper, we adopt two simple strategies, multi-range reinforcing and multi-entropy sampling, to amplify and denoise the reward signal. Our model produces an improvement over competing models with regards to BLEU scores and road exam, a new metric we designed to measure the robustness against exposure bias in language models.

1 Introduction
Likelihood-based language models with deep neural networks have been widely adopted to tackle language tasks (Graves et al., 2013; Karpathy and Fei-Fei, 2015; Bahdanau et al., 2014; Devlin et al., 2018). By far, one of the most popular training strategies is teacher forcing, which derives from the general maximum likelihood estimation (MLE) principle (Williams and Zipser, 1989). Under the teacher forcing schema, a model is trained to make predictions conditioned on ground-truth inputs. Although this strategy enables effective training of large neural networks, it is susceptible to aggravate exposure bias: a model may perform poorly at the inference stage, once its self-generated prefix diverges from the previously learned ground-truth data (Bengio et al., 2015).

A common approach to mitigate this problem is to impose supervision upon the model’s own exploration. To this objective, existing literature have introduced REINFORCE (Williams, 1992) and actor-critic (AC) methods (Konda and Tsitsiklis, 2000) (including language GANs (Yu et al., 2017)), which offer direct feedback on a model’s self-generated sequences, so the model can later, at the inference stage, deal with previously unseen exploratory paths. However, due to the well-known issue of reward sparseness and the potential noises in the critic’s feedback, these methods are reported to risk compromising the generation quality, specifically in terms of precision.

In this paper, we adopt two simple strategies, multi-range reinforcing and multi-entropy sampling to overcome the reward sparseness during training. With the tricks applied, our model demonstrates an improvement over competing models. In addition, we propose road exam as a new metric to reveal a model’s robustness against exposure bias.

2 Related Works
As an early work to address exposure bias, Bengio et al. (2015) proposed a curriculum learning approach called scheduled sampling, which gradually replaces the ground-truth tokens with the model’s own predictions while training. Later, Huszár (2015) criticized this approach for pushing the model towards overfitting onto the corpus distribution based on the position of each token in the sequence, instead of learning about the prefix.

In recent RL-inspired works, Ranzato et al. (2015) built on the REINFORCE algorithm to directly optimize the test-time evaluation metric score. Bahdanau et al. (2016) employed a similar approach by training a critic network to predict the metric score that the actor’s generated sequence of tokens would obtain. In both cases, the reliance on a metric to accurately reflect the quality of gen-
erated samples becomes a major limitation. Such metrics are often unavailable and difficult to design by nature.

In parallel, adversarial training was introduced into language modeling by SeqGAN (Yu et al., 2017). This model consists of a generator pre-trained under MLE and a discriminator pre-trained to discern the generator’s distribution from the real data. Follow-up works based on SeqGAN alter their training objectives or model architectures to enhance the guidance signal’s informativeness. RankGAN replaces the absolute binary reward with a relative ranking score (Lin et al., 2017). LeakGAN allows the discriminator to leak its internal states to the generator at intermediate steps (Guo et al., 2017). Shi et al. (2018) models a reward function using inverse reinforcement learning (IRL). While much progress have been made, we surprisingly observed that SeqGAN (Yu et al., 2017) shows more stable results in road exam in Section 5.3. Therefore, we aim to amplify and denoise the reward signal in a direct and simple fashion.

3 Model Description

Problem Re-Formulation: Actor-Critic methods (ACs) consider language modeling as a generalized Markov Decision Process (MDP) problem, where the actor learns to optimize its policy guided by the critic, while the critic learns to optimize its value function based on the actor’s output and external reward information.

As Pfau and Vinyals (2016) points out, GAN methods can be seen as a special case of AC where the critic aims to distinguish the actor’s generation from real data and the actor is optimized in an opposite direction to the critic.

Actor-Critic Training: In this work, we use a standard single-layer LSTM as the actor network. The training objective is to maximize the model’s expected end rewards with policy gradient (Sutton et al., 2000):

\[
\mathcal{L}(\theta) = - \mathbb{E}_{X_{1:T} \sim \pi_{\theta}} \sum_{t=1}^{T} Q_{\phi}(x_{t}, h_{t}) \log \pi_{\theta}(x_{t} | h_{t})
\]

Then, we use a CNN as the critic to predict the expected rewards for current generated prefix:

\[
\mathcal{L}(\phi) = - \mathbb{E}_{X_{1:T} \sim \pi_{\theta}} (r(X_{1:T}) - Q_{\phi}(X_{1:T}))^2
\]

In practice, we perform a Monte-Carlo (MC) search with roll-out policy following Yu et al. (2017) to sample complete sentences starting from each location in a predicted sequence and compute their end rewards. Empirically, we found out that the maximum, instead of average, of rewards in the MC search better represents each token’s actor value and yields better results during training. Therefore, we compute the action value by:

\[
Q_{\phi}(x_{t}, h_{t}) = \max_{X_{1:T} \in MC^{\theta}(X_{1:t}, T)} Q_{\phi}(X_{1:T})
\]

In RL and GANs training, two major factors behind the unstable performance are the large variance and the update correlation during the sampling process (Mnih et al., 2016; Volodymyr et al., 2013). We address these problems using the following strategies:

Multi-Range Reinforcing: Our idea of multi-range supervision takes inspiration from deeply-supervised nets (DSNs) (Lee et al., 2015). Under deep supervision, intermediate layers of a deep neural network have their own training objectives and receive direct supervision simultaneously with the final decision layer. By design, lower layers in a CNN have smaller receptive fields, allowing them to make better use of local patterns. Our “multi-range” modification enables the critic to focus on local n-gram information in the lower layers while attending to global structural information in the higher layers. This is a solution to the high variance problem, as the actor can receive amplified reward with more local information compared to Yu et al. (2017).

Multi-Entropy Sampling: Language GANs can be seen an online RL methods, where the actor is updated from data generated by its own policy with strong correlation. Inspired by Anonymous (2020), we empirically find that altering the entropy of the actor’s sample distribution during training is beneficial to the AC network’s robust performance. In specific, we alternate the temperature \(\tau\) to generate samples under different behavior policies. During the critic’s training, the ground-truth sequences are assigned a perfect target value of 1. The samples obtained with \(\tau < 1\) are supposed to contain lower entropy and to converge less from the real data, that they receive a higher target value close to 1. Those obtained with \(\tau > 1\) contain higher entropy and more errors that their target values are lower and closer to 0. This mechanism decorrelates updates during sequential sampling by sampling multiple diverse entropy distributions from actor synchronously.
3.1 Effectiveness of Multi-Range Reinforcing and Multi-Entropy Sampling

Table 1 demonstrates an ablation study on the effectiveness of multi-range reinforcing (MR) and multi-entropy sampling (ME). We observe that ME improves BLEU$_{F5}$ (precision) significantly while MR further enhances BLEU$_{F5}$ (precision) and BLEU$_{F5}$ (recall). Detailed explanations of these metrics can be found in Section 4.

| Architecture | BLEU$_{F5}$ | BLEU$_{B5}$ |
|--------------|-------------|-------------|
| TF           | 15.4 ± 0.17 | 30.5 ± 0.08 |
| AC           | 13.8 ± 0.16 | 30.3 ± 0.13 |
| AC (with ME) | 22.4 ± 0.25 | 30.0 ± 0.09 |
| AC (with ME & MR) | 24.5 ± 0.14 | 31.6 ± 0.10 |

Table 1: Performance of alternative architectures on EMNLP2017 WMT News Dataset. Higher is better.

4 Model Evaluation

4.1 Modeling Capacity & Sentence Quality

We adopt three variations of BLEU metric from Shi et al. (2018) to reflect precision and recall. BLEU$_F$, or forward BLEU, is a metric for precision. It uses the real test dataset as references to calculate how many n-grams in the generated samples can be found in the real data. BLEU$_B$, or backward BLEU, is a metric for recall. This metric takes both diversity and quality into computation. A model with severe mode collapse or diverse but incorrect outputs will receive poor scores in BLEU$_B$. BLEU$_{HA}$ is the harmonic mean of BLEU$_F$ and BLEU$_B$, given by:

$$\text{BLEU}_{HA} = \frac{2 \cdot \text{BLEU}_F \cdot \text{BLEU}_B}{\text{BLEU}_F + \text{BLEU}_B}$$

4.2 Exposure Bias Attacks

Road Exam is a novel test we propose as a direct evaluation of exposure bias. In this test, a sentence prefix of length $K$, either taken from the training or testing dataset, is fed into the model under assessment to perform a sentence completion task. Thereby, the model is directed onto either a seen or an unseen “road” to begin its generation. Because precision is the primary concern, we set $\tau = 0.5$ to sample high-confidence sentences from each model’s distribution. We compare BLEU$_F$ of each model on both seen and unseen completion tasks and over a range of prefix lengths. By definition, a model with exposure bias should perform worse in completing sentences with unfamiliar prefix. The sentence completion quality should decay more drastically as the the unfamiliar prefix grows longer.

5 Experiment

5.1 Datasets

We evaluate on two datasets: EMNLP2017 WMT News and Google-small, a subset of Google One Billion Words.

- EMNLP2017 WMT News is provided in (Zhu et al., 2018), a benchmarking platform for text generation models. We split the entire dataset into a training set of 195,010 sentences, a validation set of 83,576 sentences, and a test set of 10,000 sentences. The vocabulary size is 5,254 and the average sentence length is 27.
- Google-small is sampled and pre-processed from its the Google One Billion Words. It contains a training set of 699,967 sentences, a validation set of 200,000 sentences, and a test set of 99,985 sentences. The vocabulary size is 61,458 and the average sentence length is 29.

5.2 Implementation Details

Network Architecture: We implement a standard single-layer LSTM as the generator (actor) and an eight-layer CNN as the discriminator (critic). The LSTM has embedding dimension 32 and hidden dimension 256. The CNN consists of 8 layers with filter size 3, where the 3rd, 5th, and 8th layers are directly connected to the output layer for multi-range supervision. Other parameters are consistent with Zhu et al. (2018).

Training Settings: Adam optimizer is deployed for both critic and actor with learning rate $10^{-4}$ and $5 \cdot 10^{-3}$ respectively. The target values for the critic network are set to $[0, 0.2, 0.4, 0.6, 0.8]$ for samples generated by the RNN with softmax temperatures $[0.5, 0.75, 1.0, 1.25, 1.5]$.

5.3 Discussion

Table 2 and Table 3 compare models on EMNLP2017 WMT News and Google-small. Our model outperforms the others in BLEU$_{F5}$, BLEU$_{B5}$, and BLEU$_{HAS}$, indicating a high diversity and quality in its sample distribution. It is

1. https://github.com/geek-ai/Texygen
2. http://www.statmt.org/lm-benchmark/
Table 2: Results on EMNLP2017 WMT News dataset. The 95 % confidence intervals from multiple trials are reported.

| Model                      | BLEU$_{F5}$ | BLEU$_{B5}$ | BLEU$_{HA5}$ |
|----------------------------|-------------|-------------|--------------|
| Teacher Forcing (TF)       | 15.4 ± 0.11 | 30.5 ± 0.05 | 20.5 ± 0.10  |
| Scheduled Sampling (SS) (Bengio et al., 2015) | 12.1 ± 0.14 | 30.3 ± 0.06 | 17.3 ± 0.14  |
| SeqGAN (Yu et al., 2017)   | 16.6 ± 0.09 | 28.7 ± 0.37 | 21.0 ± 0.11  |
| RankGAN (Lin et al., 2017) | 17.7 ± 0.14 | 30.1 ± 0.06 | 22.3 ± 0.11  |
| LeakGAN (Guo et al., 2017) | 19.8 ± 0.11 | 31.6 ± 0.04 | 24.4 ± 0.10  |
| MEMR                       | 24.5 ± 0.08 | 31.6 ± 0.06 | 27.9 ± 0.07  |

Table 3: Results on the Google-small dataset. The 95 % confidence intervals from multiple trials are reported. † This dataset was not tested in (Guo et al., 2017) and we are unable to train LeakGAN on this dataset using the official code due to its training complexity (taking 10+ hours per epoch).

| Model                      | BLEU$_{F5}$ | BLEU$_{B5}$ | BLEU$_{HA5}$ |
|----------------------------|-------------|-------------|--------------|
| Teacher Forcing (TF)       | 9.6 ± 0.03  | 12.9 ± 0.02 | 11.00 ± 0.02 |
| Scheduled Sampling (SS) (Bengio et al., 2015) | 6.2 ± 0.04  | 10.7 ± 0.02 | 7.8 ± 0.04   |
| SeqGAN (Yu et al., 2017)   | 20.7 ± 0.02 | 14.4 ± 0.02 | 17.0 ± 0.01  |
| RankGAN (Lin et al., 2017) | 21.4 ± 0.06 | 12.7 ± 0.02 | 15.9 ± 0.02  |
| LeakGAN† (Guo et al., 2017) | -           | -           | -            |
| MEMR                       | 22.0 ± 0.07 | 15.8 ± 0.02 | 18.4 ± 0.03  |

Figure 1 demonstrates the road exam results on EMNLP News. All models decrease in sampling precision (reflected via BLEU$_{F4}$) as the fed-in prefix length ($K$) increases, but the effect is stronger on the unseen test data, revealing the existence of exposure bias. Nonetheless, our model trained under ME and MR yields the best sentence quality and a relatively moderate performance decline. Although TF and SS demonstrate higher BLEU$_{F5}$ performance with shorter prefixes, their sentence qualities drop drastically on the test dataset with longer prefixes. On the other hand, GANs begin with lower BLEU$_{F4}$ precision scores but demonstrate less performance decay as the prefix grows longer and gradually out-perform TF. This robustness against unseen prefixes exhibits that supervision from a learned critic can boost a model’s stability in completing unseen sequences.

The better generative quality in TF and the stronger robustness against exposure bias in GANs are two different objectives in language modeling, but they can be pursued at the same time. Our model’s improvement in both perspectives exhibit one possibility to achieve the goal.

6 Conclusion

We have presented multi-range reinforcing and multi-entropy sampling as two training strategies built upon deeply supervised nets (Lee et al., 2015) and multi-entropy sampling (Anonymous, 2020). The two easy-to-implement strategies help alleviate the reward sparseness in RL training and tackle the exposure bias problem.
Acknowledgments

The authors are grateful for the supports by NSF IIS-1618477, NSF IIS-1717431, and a grant from Samsung Research America.

References

Anonymous. 2020. Neural program synthesis by self-learning. In Submitted to International Conference on Learning Representations. Under review.

Dzmitry Bahdanau, Philemon Brakel, Kelvin Xu, Anirudh Goyal, Ryan Lowe, Joelle Pineau, Aaron Courville, and Yoshua Bengio. 2016. An actor-critic algorithm for sequence prediction. arXiv preprint arXiv:1607.07086.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. arXiv preprint arXiv:1409.0473.

Samy Bengio, Oriol Vinyals, Navdeep Jaitly, and Noam Shazeer. 2015. Scheduled sampling for sequence prediction with recurrent neural networks. In Advances in Neural Information Processing Systems, pages 1171–1179.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805.

Alex Graves, Abdel-rahman Mohamed, and Geoffrey Hinton. 2013. Speech recognition with deep recurrent neural networks. In Acoustics, speech and signal processing (icassp), 2013 ieee international conference on, pages 6645–6649. IEEE.

Jiaxian Guo, Sidi Lu, Han Cai, Weinan Zhang, Yong Yu, and Jun Wang. 2017. Long text generation via adversarial training with leaked information. arXiv preprint arXiv:1709.08624.

Ferenc Huszár. 2015. How (not) to train your generative model: Scheduled sampling, likelihood, adversary? arXiv preprint arXiv:1511.05101.

Andrzej Karpathy and Li Fei-Fei. 2015. Deep visual-semantic alignments for generating image descriptions. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 3128–3137.

Vijay R Konda and John N Tsitsiklis. 2000. Actor-critic algorithms. In Advances in neural information processing systems, pages 1008–1014.

Chen-Yu Lee, Saining Xie, Patrick Gallagher, Zhengyou Zhang, and Zhuowen Tu. 2015. Deeply-supervised nets. In Artificial Intelligence and Statistics, pages 562–570.

Kevin Lin, Dianqi Li, Xiaodong He, Zhengyou Zhang, and Ming-Ting Sun. 2017. Adversarial ranking for language generation. In Advances in Neural Information Processing Systems, pages 3155–3165.

Volodymyr Mnih, Adria Puigdomenech Badia, Mehdi Mirza, Alex Graves, Timothy Lillicrap, Tim Harley, David Silver, and Koray Kavukcuoglu. 2016. Asynchronous methods for deep reinforcement learning. In International conference on machine learning, pages 1928–1937.

David Pfau and Oriol Vinyals. 2016. Connecting generative adversarial networks and actor-critic methods. arXiv preprint arXiv:1610.01945.

Marc’Aurelio Ranzato, Sumit Chopra, Michael Auli, and Wojciech Zaremba. 2015. Sequence level training with recurrent neural networks. arXiv preprint arXiv:1511.06732.

Zhan Shi, Xinchu Chen, Xipeng Qiu, and Xuanjing Huang. 2018. Towards diverse text generation with inverse reinforcement learning. arXiv preprint arXiv:1804.11258.

Richard S Sutton, David A McAllester, Satinder P Singh, and Yishay Mansour. 2000. Policy gradient methods for reinforcement learning with function approximation. In Advances in neural information processing systems, pages 1057–1063.

Mnih Volodymyr, Koray Kavukcuoglu, David Silver, Alex Graves, and Ioannis Antonoglou. 2013. Playing atari with deep reinforcement learning. In NIPS Deep Learning Workshop.

Ronald J Williams. 1992. Simple statistical gradient-following algorithms for connectionist reinforcement learning. Machine learning, 8(3-4):229–256.

Ronald J Williams and David Zipser. 1989. A learning algorithm for continually running fully recurrent neural networks. Neural computation, 1(2):270–280.

Lantao Yu, Weinan Zhang, Jun Wang, and Yong Yu. 2017. Seqgan: Sequence generative adversarial nets with policy gradient. In AAAI, pages 2852–2858.

Yaoming Zhu, Sidi Lu, Lei Zheng, Jiaxian Guo, Weinan Zhang, Jun Wang, and Yong Yu. 2018. Texygen: A benchmarking platform for text generation models. arXiv preprint arXiv:1802.01886.