

DGSAN: Discrete Generative Self-Adversarial Network

Ehsan Montahaei Danial Alihosseini Mahdiyeh Soleymani Baghshah

Sharif University of Technology, Tehran, Iran

ehsan.montahaei@gmail.com, dalihosseini@ce.sharif.edu, soleymani@sharif.edu

Abstract

Although GAN-based methods have received many achievements in the last few years, they have not been entirely successful in generating discrete data. The most crucial challenge of these methods is the difficulty of passing the gradient from the discriminator to the generator when the generator outputs are discrete. Despite the fact that several attempts have been made to alleviate this problem, none of the existing GAN-based methods have improved the performance of text generation compared with the maximum likelihood approach in terms of both the quality and the diversity. In this paper, we proposed a new framework for generating discrete data by an adversarial approach in which there is no need to pass the gradient to the generator. The proposed method has an iterative manner in which each new generator is defined based on the last discriminator. It leverages the discreteness of data and the last discriminator to model the real data distribution implicitly. Moreover, the method is supported with theoretical guarantees, and experimental results generally show the superiority of the proposed DGSAN method compared to the other popular or recent methods in generating discrete sequential data.

Keywords: discrete data; generative model; self adversarial; sequential.

1. Introduction

The early deep generative models that were utilized to generate sequential discrete data such as natural language were Recurrent Neural Networks (RNNs). However, RNN-based methods for discrete sequence generation that employ *teacher forcing* approach for training suffer from the so-called *exposure bias* problem [1, 2]. Moreover, the Generative Adversarial Networks (GANs) [3] were not entirely successful in generating sequential discrete data [4, 5] despite their success in other domains, especially image generation.

In the last years, various attempts were made to apply GANs to discrete domains, but some difficulties exist in the training process of these networks on discrete data. More precisely, in the discrete domains, passing the gradient to the generator is infeasible (due to the sampling at the output of the generator) during the training process [4, 5]. To overcome this issue, [4] utilized Reinforcement Learning (RL) for the first time to this end. Although RL provides a way to train the generator, it encounters some problems such as vast action space, the sparsity of the reward, and the high variance of the estimation. Since just receiving a scalar as the reward signal from the discriminator is not such informative, many methods such as LeakGAN [6], RankGAN [7], and MaliGAN [8] were proposed to pass a more informative signal to the generator. Furthermore, some studies like Gumbel-Softmax GAN [9], TextGAN [10], FM-GAN [11], and RelGAN [12] do not utilize an RL-based approach, but instead, they apply the Gumbel-Softmax trick (that provides a continuous approximation of the discrete multinomial distribution) to pass

the gradient to the generator. Even though both of the above approaches try to alleviate the gradient passing problem and attempt to make the information flow from the discriminator to the generator more effective, they are failed to take full advantage of the discriminator.

In this paper, a new framework for adversarially training generative models of discrete data called DGSAN is proposed. In this framework, by considering an explicit distribution for the generator (due to the advantage of finite discrete domains) and finding a closed-form relation between the next generator, the current generator, and the current discriminator, the gradient passing problem is resolved. In the proposed method, the generator and the discriminator are unified in a single network. This network provides the probability distribution of data while also prepares the (conditional) probability of assigning the input to the class of real data. As a result of this integration, the gradient passing issue is bypassed, and the training stability is achieved.

Below, we describe the main contributions of the proposed method:

- In the proposed method, as opposed to the existing GAN-based methods for discrete sequence generation, neither the RL approach nor the Gumbel-Softmax trick is required. Instead, by using the closed-form solution of the new generator, the difficulty of training will be resolved. Among the existing methods, MaliGAN [8] is the most related work to ours. However, it has significant differences. MaliGAN updates the generator via an RL approach using a gradient estimator that is based on importance sam-

pling. In contrast, the proposed method uses the optimal solution of the discriminator for the current generator directly. MaliGAN uses two different networks for discriminator and generator. However, DGSAN uses a single network for both of them.

- In standard GANs [3], for each generator, the optimal discriminator can be described according to the generator and the real data distribution (even when the generator is not optimal). In the proposed method, by considering domains like discrete ones for which an explicit generative distribution is obtainable, it would be possible to approximate the real data distribution from the discriminator and the current generator.
- The proposed method is not limited to the discrete sequence generation. In those domains such as finite discrete ones for which an explicit generative model is considered, the generator can be trained by the proposed adversarial method instead of the maximum likelihood estimation.

2. Related Work

As mentioned above, the GANs in discrete domains encounter a problem in propagating the gradient into the generator. SeqGAN [4] has overcome this problem by using a REINFORCE-like algorithm in the training of the generator. It assumes the generator as an agent that receives more reward from the discriminator to generate more realistic sentences and uses the Monte Carlo tree search to estimate the expected reward. This method has some difficulties, such as reward sparsity and high variance of training. Multiple studies were carried on this approach and tried to improve the RL-based solution by transferring more information from the discriminator to the generator. RankGAN [7] trains a ranker instead of a discriminator which relatively assigns a higher score to the more realistic sequences. In other words, the score of a sentence shows how realistic it is compared to other sentences in the current batch of data. Therefore, the generator will receive a more informative gradient. LeakGAN [6] takes advantage of the feudal networks by considering the generator as a two-level module containing a manager and a worker. The feature layer of the discriminator is fed to the manager as leaked information. Boundary-Seeking GAN [13] has also changed the generator’s objective using the discriminator’s output to make the generator closer to the approximated real data distribution by minimizing $\mathcal{D}_f(\tilde{P}||Q)$ where Q shows the generator. This objective is optimized via an importance sampling approach. There exists a very similar method to Boundary-Seeking GAN called MaliGAN [8], which was mainly proposed for sequence generation. MaliGAN reduces the variance of training by employing various techniques such as Mixed MLE-MALI training and estimating reward with action-value function $Q(s, a)$. Although this

method may show similarities to ours, it still uses an RL-based approach and does not take the complete information from the discriminator. In the proposed method, when the previous generator’s distribution is available, the next generator estimating the real data distribution is straightforwardly reachable, and no additional KL optimization is needed. Hence, none of the stabilizing tricks utilized in these methods, are not required.

Studies like TextGAN [10], FM-GAN [11], and RELGAN [12] employ the Gumbel-softmax trick to prepare the gradient signal for the generator. TextGAN [10] proposes a new objective for the generator. Its generator tries to make the feature distribution of generated data closer to that of the real data according to the Maximum Mean Discrepancy (MMD) measure. TextGAN attempts to push the generator’s focus from the last layer of the discriminator to its last feature layer. FM-GAN [11] leverages a variant of earth mover’s distance (EMD) between the feature distributions of real and synthetic sentences, called feature mover’s distance (FMD), as a new generator’s objective. RELGAN [12] incorporates a relational memory as a new component for modeling the long-distance dependency. Furthermore, its discriminator utilizes multiple representations to prepare a more informative signal for the generator.

All recent GAN-based methods for the discrete sequence generation, attempt to strengthen the information flow from discriminator to the generator to provide a more informative training signal. Nonetheless, they were failed to take full advantage of the discriminator. Despite the previous works, in our method, each generator has a hybrid nature. From one perspective, it can be seen as a discriminator, while in another perspective, it is an explicit generative model. Hence, the generator and discriminator are combined in a single model, and all information is pushed to the generator accordingly.

In addition to the GAN-based methods, there are the latent-based generative models, and among them, the Variational Auto-Encoder (VAE) is the most popular one. Since the calculation of the likelihood of real samples or the posterior distribution requires the tractability of the partition function, several approaches have been proposed to train these models; e.g. the VAE optimizes the lower bound of the likelihood called *ELBO* [14]. However, VAE encounters a difficulty in the text generation problem called the *KL Collapse* problem [15]. As a consequence of this problem, the decoder ignores the latent space and operates independently. Various studies have been done to identify the main origin of this issue [16, 17, 18]. From these studies, the decoder’s universality in conjunction with the undesired global and local optimums has been considered as a source of this difficulty.

3. Preliminaries on GAN

In the standard GANs, we have some samples from the real data distribution P , which we need to learn. The gen-

Algorithm 1 DGSAN general algorithm

```

1: Input:  $B$ : Batch size,  $D$ : Number of DGSAN iterations
2: Initialize two random generator  $Q^{\text{old}}$ , and  $Q_\theta$ .
3: for  $D$  iterations do
4:   while not converged do
5:     Sample  $\{\mathbf{x}_r^{(i)}\}_{i=1}^B$  from real data.
6:     Generate  $\{\mathbf{x}_f^{(i)}\}_{i=1}^B$  samples from  $Q^{\text{old}}$ .
7:     Train  $Q_\theta$  by optimizing Eq. 5.
       Use  $\{\mathbf{x}_r^{(i)}\}_{i=1}^B$  for the first term, and  $\{\mathbf{x}_f^{(i)}\}_{i=1}^B$  for the
       second term to estimate the expectation.
8:   end while
9:    $Q^{\text{old}} = Q_\theta$ 
10: end for

```

erator attempts to learn a network Q , which can generate samples similar to the real ones, and the discriminator intends to determine whether a sample is real or synthetic. Formally, the objective is [3]:

$$V(Q, D) = \mathbb{E}_{x \sim P}[\log D(x)] + \mathbb{E}_{x \sim Q}[\log(1 - D(x))], \quad (1)$$

which is maximized with respect to D and minimized with respect to Q . It is shown that the optimal D for a fixed distribution Q is found as [3]:

$$D_Q^*(x) = \frac{p(x)}{p(x) + q(x)}. \quad (2)$$

Moreover, for the optimal discriminator D_Q^* , minimizing $V(Q, D_Q^*)$ w.r.t. Q leads to the minimization of Jensen-Shannon divergence between P and Q [3].

4. Proposed Method

In the following subsections, we first describe the general framework for adversarially training explicit generative models. Then, the theoretical analysis of the proposed training approach is presented. Finally, using the proposed framework, a method for sequence generation is presented.

4.1. Basic Model

We intend to learn an explicit generative model Q . To this end, we begin from an initial model and improve it iteratively. Consider that the generator found in the last step is called Q^{old} , and we want to find a new generator Q^{new} from it. A discriminator between the real distribution and Q^{old} is required to form the objective function for optimizing Q . The discriminator is modeled as:

$$D(x) = \frac{q^{\text{new}}(x)}{q^{\text{new}}(x) + q^{\text{old}}(x)}. \quad (3)$$

The intuition behind employing the above discriminator is based on the optimal discriminator of GAN. Indeed, if we want Q^{new} to be optimal (i.e. equal to the real distribution

P), the optimal discriminator can be formulated as in Eq. 3 (according to Eq. 2). For optimizing the new generator q_θ , the following objective is used (according to Eq. 1):

$$\max_{\theta} \mathbb{E}_{x \sim P} [\ln D(x)] + \mathbb{E}_{x \sim Q^{\text{old}}} [\ln(1 - D(x))], \quad (4)$$

which can be reformulated as below (according to Eq. 3):

$$\max_{\theta} \mathbb{E}_{x \sim P} \left[\ln \frac{q_\theta(x)}{q_\theta(x) + q^{\text{old}}(x)} \right] + \mathbb{E}_{x \sim Q^{\text{old}}} \left[\ln \frac{q^{\text{old}}(x)}{q_\theta(x) + q^{\text{old}}(x)} \right]. \quad (5)$$

Then, the generator Q_θ is found using samples of the real distribution P and the old generator. This procedure is iteratively done. In each DGSAN iteration, the Q_θ of the last iteration is considered as Q^{old} , and a new Q_θ is learned. According to Eq. 5, since samples come from Q^{old} and optimization is based on θ , the difficulty of gradient propagation through the discreteness of Q_θ 's output (new generator) is bypassed. In the implementation phase, we use an equal objective based on Softplus [19]:

$$\min_{\theta} \mathbb{E}_{x \sim P} [\text{Softplus}(\ln q^{\text{old}}(x) - \ln q_\theta(x))] + \mathbb{E}_{x \sim Q^{\text{old}}} [\text{Softplus}(\ln q_\theta(x) - \ln q^{\text{old}}(x))]. \quad (6)$$

An overview of the training procedure is presented in Alg. 1.

4.2. Theoretical Analysis

This section will discuss the convergence of the training algorithm, proposed in Section 4.1 (Algorithm 1). The proofs are provided in Appendix.

Theorem 4.1. *Let P, Q^{old} , and Q_θ denote three distributions. A relation between these distributions can be formulated as:*

$$\begin{aligned} \mathcal{D}_{JS}(P \| Q^{\text{old}}) &= \mathcal{L}(P, Q^{\text{old}}, Q_\theta) + \mathbb{E}_{x \sim Q^{\text{old}}} \left[\mathcal{B}_f \left(\frac{p(x)}{q^{\text{old}}(x)} \parallel \frac{q_\theta(x)}{q^{\text{old}}(x)} \right) \right], \\ &= \mathcal{L}(P, Q^{\text{old}}, Q_\theta) + \mathbb{E}_{x \sim P} \left[\mathcal{B}_f \left(\frac{q^{\text{old}}(x)}{p(x)} \parallel \frac{q^{\text{old}}(x)}{q_\theta(x)} \right) \right], \end{aligned} \quad (7)$$

where \mathcal{D}_{JS} denotes the Jensen–Shannon divergence, \mathcal{B}_f shows the Bregman divergence, and $f(u) = u \ln u - (u + 1) \ln(u + 1)$ is the function used in the Bregman divergence. Moreover, \mathcal{L} is defined as:

$$\begin{aligned} \mathcal{L}(P, Q^{\text{old}}, Q_\theta) &= \mathbb{E}_{x \sim P} \left[\ln \frac{q_\theta(x)}{q_\theta(x) + q^{\text{old}}(x)} \right] \\ &\quad + \mathbb{E}_{x \sim Q^{\text{old}}} \left[\ln \frac{q^{\text{old}}(x)}{q_\theta(x) + q^{\text{old}}(x)} \right]. \end{aligned} \quad (8)$$

Corollary 4.1.1. *When maximizing $\mathcal{L}(P, Q^{\text{old}}, Q_\theta)$ w.r.t. θ , $\frac{p(x)}{q^{\text{old}}(x)}$ is estimated by $\frac{q_\theta(x)}{q^{\text{old}}(x)}$.*

Algorithm 2 DGSAN sequential data algorithm

-
- 1: **Input:** B : Batch size, M : Maximum length of the given sequence, D : Number of DGSAN iterations per length, T : Sampling temperature
 - 2: Initialize two random generators Q^{old} , and Q_θ .
 - 3: $l = 1$
 - 4: **repeat**
 - 5: **for** D iterations **do**
 - 6: **while** not converged **do**
 - 7: Sample $\{\mathbf{s}^{(i)}\}_{i=1}^B$ from real data.
 - 8: Select the value of k randomly from $\{0, 1, \dots, M - l\}$, as the length of the prefix sequence.
 - 9: Split elements of $\{\mathbf{s}^{(i)}\}_{i=1}^B$ by considering tokens between k and $k + l$ as $\mathbf{x}_r^{(i)} = \mathbf{s}_{k:k+l}^{(i)}$, and the first k tokens as $\mathbf{c}^{(i)} = \mathbf{s}_{1:k}^{(i)}$.
 - 10: Generate $\{\mathbf{x}_f^{(i)}\}_{i=1}^B$ samples from Q^{old} conditioned on $\{\mathbf{c}^{(i)}\}_{i=1}^B$ with the temperature T .
 - 11: Train Q_θ by optimizing Eq. 11
 Use $\{\mathbf{x}_r^{(i)}\}_{i=1}^B$ in the first term and $\{\mathbf{x}_f^{(i)}\}_{i=1}^B$ in the second term to estimate the expectation in Eq. 11. All probabilities are conditioned on $\mathbf{c}^{(i)}$.
 - 12: **end while**
 - 13: $Q^{\text{old}} = Q_\theta$
 - 14: **end for**
 - 15: $l ++$
 - 16: **until** Not reached to max epoch
-

According to the above corollary, we can use Eq. 5 to find the best new generator for the current generator. Thus, if we reach the global optimum, Q_θ will be the real data distribution. Moreover, based on the latter corollary, adding noise to the distribution Q^{old} can help the training process in the points where Q^{old} is near to zero.

Since it is ideal but may be impossible to reach the global optimum of $\mathcal{L}(P, Q^{\text{old}}, Q_\theta)$, the next theorem is provided to support the convergence of the method in the case of reaching a non-optimal solution. By looking at each iteration from the discriminator training viewpoint, we will have the following theorem.

Theorem 4.2. *If $D_\theta(x) = \frac{q_\theta(x)}{q_\theta(x) + q^{\text{old}}(x)}$ is between a random and an optimal discriminator $D^*(x) = \frac{p(x)}{p(x) + q^{\text{old}}(x)}$ for Q^{old} , $\mathcal{D}_{JS}(P||Q_\theta)$ will be strictly less than $\mathcal{D}_{JS}(P||Q^{\text{old}})$.*

This theorem shows that when the discriminator is non-optimal, under the above condition, the Jensen-Shannon divergence decreases after each iteration of Algorithm 1, and thus, Q finally converges to P .

4.2.1. The \mathbf{f} Family

We can also extend the loss function and the above theorems to more general ones supporting all strictly convex f , instead of $f(u) = u \ln u - (u + 1) \ln(u + 1)$. More precisely, inspired by several works on the original GAN [20, 21, 22, 13, 8] that have extended the GAN's initial framework, we also extend the presented framework and theorems in Section 4.2 to f -divergence family that is more general than the JS divergence \mathcal{D}_{JS} .

Theorem 4.3. *For every P, Q^{old} , and Q_θ distributions, and f -divergence with strictly convex f ,*

$$\begin{aligned} \mathcal{D}_f(P||Q^{\text{old}}) &= \mathcal{L}_f(P, Q^{\text{old}}, \tau_\theta) + \mathbb{E}_{x \sim Q^{\text{old}}} \left[\mathcal{B}_f \left(\frac{p(x)}{q^{\text{old}}(x)} \parallel \frac{q_\theta(x)}{q^{\text{old}}(x)} \right) \right], \\ \mathcal{L}_f(P, Q^{\text{old}}, \tau_\theta) &= \mathbb{E}_{x \sim P} [\tau_\theta(x)] - \mathbb{E}_{x \sim Q^{\text{old}}} [f^*(\tau_\theta(x))], \end{aligned} \quad (9)$$

where $\tau_\theta(x) = f' \left(\frac{q_\theta(x)}{q^{\text{old}}(x)} \right)$, and f^* is the Fenchel conjugate of f .

It is worth mentioning that [20] also presents another variation of Theorem 4.3.

By maximizing $\mathcal{L}_f(P, Q^{\text{old}}, \tau_\theta)$ respect to the θ , the corresponding Bregman divergence will be minimized and the desired $p(x)$ estimation is achieved.

Theorem 4.4. *If $D_\theta(x) = \frac{q_\theta(x)}{q_\theta(x) + q^{\text{old}}(x)}$ is between random and optimal discriminator $D^*(x) = \frac{p(x)}{p(x) + q^{\text{old}}(x)}$ for Q^{old} , we will have $\mathcal{D}_f(P||Q_\theta) < \mathcal{D}_f(P||Q^{\text{old}})$ for every f -divergence by strictly convex f .*

4.3. Method for sequence generation

In theory, By considering \mathbf{x} as x_1, x_2, \dots, x_M which they are tokens of the sequence, we can apply our method on the sequences to learn the joint distribution $P(x_1, x_2, \dots, x_M)$. However, learning such a high dimensional distribution is not feasible. Therefore, We train the single token distribution $P(x_i|x_{1:i-1})$ firstly, then the distribution on the sequences of two tokens $P(x_i, x_{i+1}|x_{1:i-1})$, and so forth (i.e. curriculum learning).

By considering all of the distributions in Eq. 5, conditioned on a prefix of real data, the generator's objective is obtained as follows:

$$\begin{aligned}
\max_{\theta} \mathbb{E}_{c \sim P_{prefix}} & \left[\mathbb{E}_{\mathbf{x} \sim P|c} \left[\ln \frac{q_{\theta}(\mathbf{x}|c)}{q_{\theta}(\mathbf{x}|c) + q^{\text{old}}(\mathbf{x}|c)} \right] \right. \\
& \left. + \mathbb{E}_{\mathbf{x} \sim Q^{\text{old}}|c} \left[\ln \frac{q^{\text{old}}(\mathbf{x}|c)}{q_{\theta}(\mathbf{x}|c) + q^{\text{old}}(\mathbf{x}|c)} \right] \right] \quad (10) \\
& = \mathbb{E}_{\mathbf{x}, c \sim P_{prefix}} \left[\ln \frac{q_{\theta}(\mathbf{x}|c)}{q_{\theta}(\mathbf{x}|c) + q^{\text{old}}(\mathbf{x}|c)} \right] \\
& \quad + \mathbb{E}_{c \sim P_{prefix} \mathbf{x} \sim Q^{\text{old}}|c} \left[\ln \frac{q^{\text{old}}(\mathbf{x}|c)}{q_{\theta}(\mathbf{x}|c) + q^{\text{old}}(\mathbf{x}|c)} \right], \quad (11)
\end{aligned}$$

where P_{prefix} denotes any prefix of sequences from real data and \mathbf{x} shows conditional samples of length l .

In the above objective, we want to learn $q_{\theta}(x_{k:k+l}|x_{1:k-1})$ distribution for different values of k and l . By considering $q_{\theta}(x_{k:k+l}|x_{1:k-1}) = \prod_{i=k}^{k+l} q_{\theta}(x_i|x_{1:i-1})$, we can rewrite all distributions based on the (conditional) distributions of single tokens.

The condition sequence x_1, \dots, x_k is first embedded as $\mathbf{e}_1, \dots, \mathbf{e}_k$ and then mapped into the sequence of hidden states $\mathbf{h}_1, \dots, \mathbf{h}_k$ using a recurrent unit $\mathbf{h}_i = g(\mathbf{h}_{i-1}, \mathbf{e}_{i-1})$ (in which x_0 shows the start token) and the conditional distribution is modeled as $q_{\theta}(x_i|x_{1:i-1}) = \text{softmax}(\mathbf{V}\mathbf{h}_i)$. Let θ denote the set of parameters of the generator network (containing the parameters of g and the matrix \mathbf{V}).

Usually sequence modeling is converted to learning $q(x_k|x_{1:k-1})$ whereas the joint distribution can be rewritten as $q(x_{1:M}) = \prod_{k=1}^M q(x_k|x_{1:k-1})$. Since the conditioned sequence always brings from real data during training while it is just the generated sequence by the model in the test time, there is a well-known problem called exposure bias. In our method, we incrementally raise the length l of the predicting sequence since increasing this length alleviates the exposure bias problem consequently. Increasing the predicting sequence’s length leads the model to see generated samples during the training phase. In the ideal case, if the length of the predicting sequence is equal to the target sequence length, exposure bias vanish entirely.

An overview of our training procedure for generating discrete sequences is provided in Algorithm 2.

5. Experiments and Results

In this section, we conduct experiments to examine the proposed method for the sequence generation problem and compare it to the state-of-the-art methods.

5.1. Evaluation Metrics

First of all, Negative Log-Likelihood (NLL) as a popular metric for evaluating generative models is presented, then three n-gram based measures for evaluating sequence generation are introduced. Finally, a metric that considers the semantic is described.

5.1.1. Negative Log Likelihood (NLL)

It shows the negative log-likelihood of real data in the model. The GAN-based methods tend to have poor NLL scores on the training/test data since this measure is much sensitive to the mode collapse phenomenon [23]. It should be noted that the NLL is closely related to another well-known metric called Perplexity (so, we just report NLL to avoid redundancy).

5.1.2. BLEU (BL)

BLEU is a metric designed to evaluate the machine translation models [24]. It measures n-gram similarities of a test sentence to a reference set and then takes the geometric mean of n-gram similarities to produce a score for the test sentence. As discussed in [23], a model repeatedly generating one high-quality sentence can get a high BLEU score while completely losing diversity. Therefore, BLEU evaluates the quality of the samples and is not sensitive to their diversity [25, 26, 27].

5.1.3. Backward BLEU (BBL)

The Backward BLEU was introduced by [28] to measure the diversity of generated samples. It considers generated samples as the reference and evaluates each test sample by BLEU.

5.1.4. MS-Jaccard (MSJ)

Either having a model, repeatedly generating just one high-quality sample or having a model generating a wide variety of low-quality samples, is disappointing. Thus, to consider this trade-off and have a metric jointly measuring the validity and coverage, the recently proposed MS-Jaccard metric is taken into account [27]. The n-grams of generated samples and real samples are considered as two multi-sets which preserve the repetition of n-grams. Then, the similarity of the resulted multi-sets is computed by the Jaccard similarity measure whose higher scores are more desired. If the generated sentences do not have diversity (e.g., when the mode collapse happens) or lose quality, the n-gram distribution of generated texts will be different from that of the real texts and this measure will be decreased.

5.1.5. Fréchet BERT Distance (FBD)

FBD [27] is inspired by FID [29] which is a famous metric in image generation task. Unlike other metrics, the FBD considers the semantic of sentences. The FBD considers the distance between the distributions of real and generated text in a feature space found by BERT (as a powerful language representation model) [30]. Indeed, the distance is the Fréchet distance between the two Gaussian distributions fitted to real and generated data in the feature space.

Table 1: Performance of models (using different measures) on *Amazon* dataset.

Method	NLL	FBD	MSJ3	MSJ5	MSJ7	BL3	BL5	BL7	BBL3	BBL5	BBL7
DGSAN	113.146	0.420	0.442	0.254	0.138	0.945	0.728	0.442	0.724	0.416	0.209
MLE	99.369	3.254	0.560	0.265	0.121	0.768	0.397	0.174	0.786	0.425	0.193
VAE	-	2.579	0.556	0.264	0.121	0.773	0.399	0.177	0.787	0.425	0.192
MaliGAN	123.069	8.951	0.313	0.147	0.069	0.740	0.407	0.188	0.686	0.311	0.136
RankGAN	127.364	6.511	0.307	0.141	0.064	0.773	0.363	0.158	0.669	0.303	0.129
SeqGAN	141.881	9.668	0.294	0.140	0.064	0.807	0.419	0.177	0.681	0.324	0.141

Table 2: Performance of models (using different measures) on *Yelp Restaurant* dataset.

Method	NLL	FBD	MSJ3	MSJ5	MSJ7	BL3	BL5	BL7	BBL3	BBL5	BBL7
DGSAN	56.398	0.382	0.334	0.152	0.067	0.756	0.417	0.245	0.538	0.259	0.157
MLE	50.201	2.276	0.322	0.125	0.045	0.542	0.246	0.149	0.551	0.245	0.147
VAE	-	2.297	0.344	0.133	0.048	0.554	0.242	0.145	0.559	0.251	0.150
MaliGAN	55.150	1.351	0.226	0.086	0.031	0.562	0.291	0.191	0.484	0.210	0.130
RankGAN	58.463	1.804	0.250	0.101	0.035	0.678	0.333	0.206	0.466	0.209	0.129
SeqGAN	53.135	1.022	0.276	0.114	0.042	0.617	0.307	0.198	0.504	0.226	0.138

Table 3: Performance of models (using different measures) on *Coco* dataset.

Method	NLL	FBD	MSJ3	MSJ5	MSJ7	BL3	BL5	BL7	BBL3	BBL5	BBL7
DGSAN	73.630	1.949	0.208	0.091	0.035	0.647	0.307	0.154	0.489	0.206	0.107
MLE	99.525	3.209	0.224	0.088	0.029	0.517	0.209	0.105	0.549	0.219	0.109
VAE	-	2.559	0.230	0.094	0.033	0.530	0.215	0.109	0.568	0.234	0.116
MaliGAN	93.243	3.931	0.176	0.065	0.021	0.503	0.200	0.100	0.504	0.192	0.098
RankGAN	103.071	3.198	0.118	0.038	0.010	0.474	0.186	0.098	0.381	0.135	0.074
SeqGAN	93.681	2.344	0.175	0.067	0.021	0.524	0.229	0.120	0.440	0.169	0.089

5.2. Datasets

We have used three real-world datasets: Image COCO captions¹, Yelp restaurant reviews²[31], and Amazon Review³ [32] to cover a wide range of linguistic datasets. Table 4 shows the statistics of datasets.

Table 4: Datasets Statistics

Dataset	Amazon Review	Image COCO captions	Yelp restaurant reviews
Train set size	174588	10000	40000
Validation set size	19436	10000	10000
Test set size	19436	10000	10000
Vocabulary size	6282	6488	7802
Max Length	41	38	30
Mean Length	28.87	11.62	9.45

5.3. Experiment Setup

Several experiments are conducted to compare our method with some popular GAN-based methods, MLE,

and VAE for sequence generation.

The generator networks (i.e. the decoder in VAE and the generator in MLE, GANs, and DGSAN) are the same and implemented with one-layer LSTM, having 128 embedding and 64 hidden dimension.

All GANs including SeqGAN, MaliGAN, and RankGAN are implemented in Texygen framework⁴[25] and have a discriminator with two CNN layers including the filter size 2 and 3 and the number of filters 100 and 200, respectively.

The VAE’s encoder is an LSTM with 128 embedding and 128 hidden dimension, and the latent dimension is 64. To mitigate the *KL Collapse* problem, it is trained with *KL term annealing* trick which is implemented with *Tanh* function. Also, word dropout is taken into account. Its implementation is based on [this Github repository](https://github.com/geek-ai/TextGAN-PyTorch)⁵.

All evaluations are based on the best checkpoints of models concerning the FBD score which is evaluated every 5 epochs on validation data during training. In our experiments, the number of DGSAN iterations per length (D) was set to 5, and the temperature in DGSAN training (T) was chosen as 2.0.

MLE, VAE, and DGSAN are trained for 1000 epochs

¹<https://github.com/pclucas14/GansFallingShort>

²<https://github.com/shentianxiao/language-style-transfer>

³<https://github.com/williamSYSU/TextGAN-PyTorch>

⁴<https://github.com/geek-ai/Textygen>

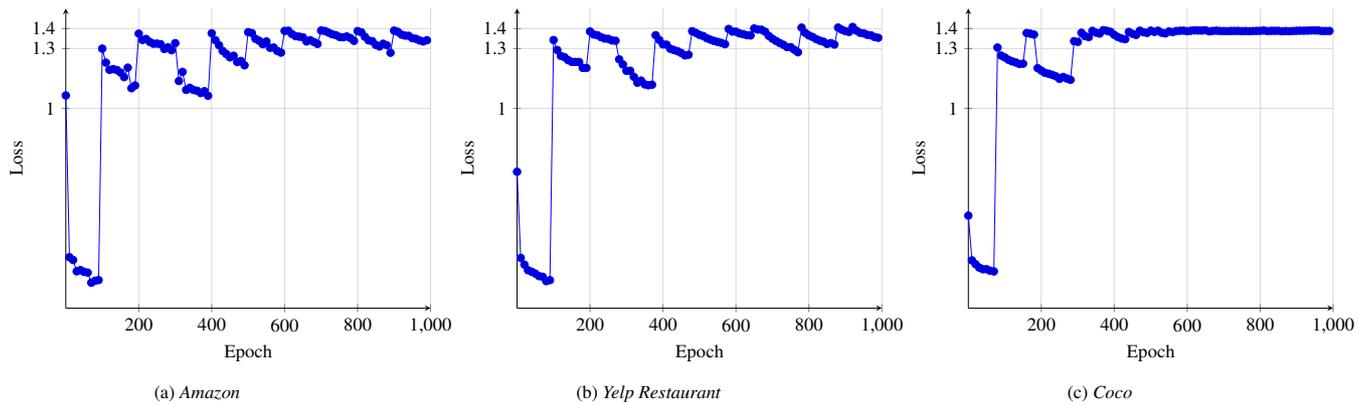


Fig. 1: DGSAN training loss of different datasets

on all datasets. Since the GANs include the discriminator training phase, each epoch of their training lasts much more. Therefore, to have a fair training time for all models, fewer epochs are considered for GANs. They are trained for 280 epochs including 80 epochs of the MLE pre-training. Moreover, training of the GANs usually ends in the mode collapsed state, and the best FBD based checkpoint is not near the end. This is approved with our experiments; e.g., the RankGAN achieves the best FBD score around the 115th epoch which is far from the last epoch.

The baseline and evaluation codes is available at [this Github repository](#)⁶. Also, the implementation code of DGSAN is available in [this Github repository](#)⁷.

5.4. Results

As discussed in Section 5.3, all models are selected based on the best FBD score, since it not only evaluates the semantic and syntax validity of samples, but also evaluates the diversity of them simultaneously [27]. Moreover, as the FBD considers the whole sentence, it is independent of any length parameter selection, such as n in the N-gram metrics.

Tables 1, 2, and 3 show the results of different methods on the Amazon, Yelp Restaurant, and the Coco datasets, respectively. In addition to the FBD, three N-gram based metrics, including MS-Jaccard (MSJ), BLEU (BL), and Backward BLEU (BBL), are reported with different n of 3, 5, 7. The number of generated samples is same as the test set size.

Recent studies [23, 27] show that the regular training of GANs typically leads them to collapse in generating high-quality samples while other baselines such as MLE will perform better when considering both quality and diversity. Although, our experiments are based on the best FBD checkpoints to alleviate this problem for GAN-based methods; nevertheless, their results are not competitive, yet; even with the MLE. It is noticeable that the DGSAN method has outperformed other methods according to the FBD and BLEU on all of the datasets. There is also a correlation between the improvement of the DGSAN results and n in the N-gram based metrics. Moreover, as the

FBD considers the whole of the sentences, the improvement of DGSAN in sequence generation with longer coherency can be inferred from the results.

Figure 1 shows the loss values of DGSAN during the training. The loss value approximately converged to the value of 1.38 which is consistent with the theory (indicating that the discriminator loss when it can not distinguish between real and fake, is $2 \ln 2 \approx 1.38$). It is worth mentioning that difficulty in reaching to this equilibrium is usually a big challenge for GAN-based methods specially on discrete data.

Finally, some samples of generated sentences by each of the models are provided in the Appendix. For each of the models and datasets, the provided samples are selected in two ways, the random selection and high-quality selection based on BLEU5 among generated samples.

6. Conclusion

In this paper, we proposed a generative adversarial model for domains in which we can consider an explicit distribution and sample from that distribution, including discrete domains. As opposed to the existing approaches, we do not need an RL-based approach for the generator’s training. As a consequence of finding the generator via a closed-form solution in each iteration, we removed the gradient passing to the generator. Moreover, the GAN stability issues during training caused by seeking the saddle point of the objective are bypassed in the proposed method.

Since generating a sequence of discrete data is an essential task in natural language generation, we examined the proposed method in generating natural language texts. Experiments demonstrate that the proposed method can model a more similar distribution to the real distribution (than those generated by the compared methods) according to the measures that approximate the similarity of the distributions.

References

- [1] F. Huszar, CoRR abs/1511.05101 (2015). [arXiv:1511.05101](https://arxiv.org/abs/1511.05101).
- [2] S. Bengio, O. Vinyals, N. Jaitly, N. Shazeer, in: C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, R. Garnett (Eds.), *Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems 2015*, December 7-12, 2015, Montreal, Quebec, Canada, pp. 1171–1179.
- [3] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. C. Courville, Y. Bengio, CoRR abs/1406.2661 (2014). [arXiv:1406.2661](https://arxiv.org/abs/1406.2661).
- [4] L. Yu, W. Zhang, J. Wang, Y. Yu, in: S. P. Singh, S. Markovitch (Eds.), *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence*, February 4-9, 2017, San Francisco, California, USA., AAAI Press, 2017, pp. 2852–2858.
- [5] I. J. Goodfellow, CoRR abs/1701.00160 (2017). [arXiv:1701.00160](https://arxiv.org/abs/1701.00160).
- [6] J. Guo, S. Lu, H. Cai, W. Zhang, Y. Yu, J. Wang, in: S. A. McIlraith, K. Q. Weinberger (Eds.), *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence*, New Orleans, Louisiana, USA, February 2-7, 2018, AAAI Press, 2018.
- [7] K. Lin, D. Li, X. He, M. Sun, Z. Zhang, in: [33], pp. 3158–3168.
- [8] T. Che, Y. Li, R. Zhang, R. D. Hjelm, W. Li, Y. Song, Y. Bengio, CoRR abs/1702.07983 (2017). [arXiv:1702.07983](https://arxiv.org/abs/1702.07983).
- [9] M. J. Kusner, J. M. Hernández-Lobato (2016). URL: <https://arxiv.org/abs/1611.04051>.
- [10] Y. Zhang, Z. Gan, K. Fan, Z. Chen, R. Henao, D. Shen, L. Carin, in: D. Precup, Y. W. Teh (Eds.), *Proceedings of the 34th International Conference on Machine Learning, ICML 2017*, Sydney, NSW, Australia, 6-11 August 2017, volume 70 of *Proceedings of Machine Learning Research*, PMLR, 2017, pp. 4006–4015.
- [11] L. Chen, S. Dai, C. Tao, D. Shen, Z. Gan, H. Zhang, Y. Zhang, L. Carin (2018). URL: [DBLP:conf/nips/2018](https://arxiv.org/abs/1801.03978).
- [12] W. Nie, N. Narodytska, A. Patel, *International Conference on Learning Representations (ICLR)* (2019).
- [13] R. D. Hjelm, A. P. Jacob, A. Trischler, G. Che, K. Cho, Y. Bengio, in: *International Conference on Learning Representations*.
- [14] D. P. Kingma, M. Welling, in: Y. Bengio, Y. LeCun (Eds.), *2nd International Conference on Learning Representations, ICLR 2014*, Banff, AB, Canada, April 14-16, 2014, Conference Track Proceedings. URL: [http://arxiv.org/abs/1312.6114](https://arxiv.org/abs/1312.6114).
- [15] S. R. Bowman, L. Vilnis, O. Vinyals, A. M. Dai, R. Józefowicz, S. Bengio, in: Y. Goldberg, S. Riezler (Eds.), *Proceedings of the 20th SIGNLL Conference on Computational Natural Language Learning, CoNLL 2016*, Berlin, Germany, August 11-12, 2016, ACL, 2016, pp. 10–21. URL: <https://doi.org/10.18653/v1/k16-1002>. doi:10.18653/v1/k16-1002.
- [16] X. Chen, D. P. Kingma, T. Salimans, Y. Duan, P. Dhariwal, J. Schulman, I. Sutskever, P. Abbeel, in: *5th International Conference on Learning Representations, ICLR 2017*, Toulon, France, April 24-26, 2017, Conference Track Proceedings, OpenReview.net, 2017. URL: <https://openreview.net/forum?id=BysvGP5ee>.
- [17] S. Zhao, J. Song, S. Ermon, in: *The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019*, *The Thirty-First Innovative Applications of Artificial Intelligence Conference, IAAI 2019*, *The Ninth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019*, Honolulu, Hawaii, USA, January 27 - February 1, 2019, AAAI Press, 2019, pp. 5885–5892. URL: <https://doi.org/10.1609/aaai.v33i01.33015885>. doi:10.1609/aaai.v33i01.33015885.
- [18] J. He, D. Spokoyny, G. Neubig, T. Berg-Kirkpatrick, in: *7th International Conference on Learning Representations, ICLR 2019*, New Orleans, LA, USA, May 6-9, 2019, OpenReview.net, 2019. URL: <https://openreview.net/forum?id=rylDfnCqF7>.
- [19] C. Dugas, Y. Bengio, F. Bélisle, C. Nadeau, R. Garcia, in: T. K. Leen, T. G. Dietterich, V. Tresp (Eds.), *Advances in Neural Information Processing Systems 13*, *Papers from Neural Information Processing Systems (NIPS) 2000*, Denver, CO, USA, MIT Press, 2000, pp. 472–478. URL: <http://papers.nips.cc/paper/1920-incorporating-second-order-functional-knowledge-for-better-option-pricing>.
- [20] M. Sugiyama, T. Suzuki, T. Kanamori, *Annals of the Institute of Statistical Mathematics* 64 (2012) 1009–1044.
- [21] S. Nowozin, B. Cseke, R. Tomioka, in: D. D. Lee, M. Sugiyama, U. von Luxburg, I. Guyon, R. Garnett (Eds.), *Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016*, December 5-10, 2016, Barcelona, Spain, pp. 271–279.
- [22] B. Poole, A. A. Alemi, J. Sohl-Dickstein, A. Angelova, CoRR abs/1612.02780 (2016). [arXiv:1612.02780](https://arxiv.org/abs/1612.02780).
- [23] M. Caccia, L. Caccia, W. Fedus, H. Larochelle, J. Pineau, L. Charlin, in: *International Conference on Learning Representations (ICLR)*. URL: <https://openreview.net/forum?id=BJgza6VtPB>.
- [24] K. Papineni, S. Roukos, T. Ward, W. Zhu, in: *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, July 6-12, 2002, Philadelphia, PA, USA., ACL, 2002, pp. 311–318.
- [25] Y. Zhu, S. Lu, L. Zheng, J. Guo, W. Zhang, J. Wang, Y. Yu, in: K. Collins-Thompson, Q. Mei, B. D. Davison, Y. Liu, E. Yilmaz (Eds.), *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval, SIGIR 2018*, Ann Arbor, MI, USA, July 08-12, 2018, ACM, 2018, pp. 1097–1100. doi:10.1145/3209978.3210080.
- [26] T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger, Y. Artzi, CoRR abs/1904.09675 (2019). [arXiv:1904.09675](https://arxiv.org/abs/1904.09675).
- [27] D. Alihosseini, E. Montahaei, M. Soleymani Baghshah, in: *Proceedings of the Workshop on Methods for Optimizing and Evaluating Neural Language Generation*, Association for Computational Linguistics, Minneapolis, Minnesota, 2019, pp. 90–98. URL: <https://www.aclweb.org/anthology/W19-2311>. doi:10.18653/v1/W19-2311.
- [28] Z. Shi, X. Chen, X. Qiu, X. Huang, *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence* (2018). URL: <http://dx.doi.org/10.24963/ijcai.2018/606>. doi:10.24963/ijcai.2018/606.
- [29] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, S. Hochreiter, in: [33], pp. 6629–6640.
- [30] J. Devlin, M. Chang, K. Lee, K. Toutanova, CoRR abs/1810.04805 (2018). [arXiv:1810.04805](https://arxiv.org/abs/1810.04805).
- [31] T. Shen, T. Lei, R. Barzilay, T. S. Jaakkola, in: I. Guyon, U. von Luxburg, S. Bengio, H. M. Wallach, R. Fergus, S. V. N. Vishwanathan, R. Garnett (Eds.), *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017*, 4-9 December 2017, Long Beach, CA, USA, pp. 6830–6841. URL: <http://papers.nips.cc/paper/7259-style-transfer-from-non-pa>.
- [32] J. J. McAuley, J. Leskovec, in: Q. Yang, I. King, Q. Li, P. Pu, G. Karypis (Eds.), *Seventh ACM Conference on Recommender Systems, RecSys '13*, Hong Kong, China, October 12-16, 2013, ACM, 2013, pp. 165–172. URL: <https://doi.org/10.1145/2507157.2507163>. doi:10.1145/2507157.2507163.
- [33] I. Guyon, U. von Luxburg, S. Bengio, H. M. Wallach, R. Fergus, S. V. N. Vishwanathan, R. Garnett (Eds.), *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017*, 4-9 December 2017, Long Beach, CA, USA, 2017.

7. Appendix

Proofs of the theorems presented in Section 4.2 are described here.

Lemma 7.1. *If f is convex then*

$$f^*(f'(x)) = f'(x)x - f(x), \quad (12)$$

where f^* is the Fenchel conjugate of f .

Proof. For each convex function f :

$$\begin{aligned} f(u) &\geq f'(x)(u-x) + f(x) \\ \Rightarrow xf'(x) - f(x) &\geq uf'(x) - f(u) \\ \Rightarrow xf'(x) - f(x) &\geq \sup_{u \in \text{dom}_f} \{uf'(x) - f(u)\}. \end{aligned} \quad (13)$$

On the other side, it is obvious that for every x , we have:

$$\sup_{u \in \text{dom}_f} \{uf'(x) - f(u)\} \geq xf'(x) - f(x). \quad (14)$$

Therefore:

$$\sup_{u \in \text{dom}_f} \{uf'(x) - f(u)\} = xf'(x) - f(x). \quad (15)$$

From the Fenchel conjugate definition, we obtain:

$$\begin{aligned} f^*(t) &= \sup_{u \in \text{dom}_f} \{ut - f(u)\} \\ \Rightarrow f^*(f'(x)) &= \sup_{u \in \text{dom}_f} \{uf'(x) - f(u)\} \\ \Rightarrow f^*(f'(x)) &= xf'(x) - f(x). \end{aligned} \quad (16)$$

□

Proof of Theorem 4.3

Proof. According to the Bregman divergence definition, we reach the following equation:

$$\begin{aligned} \mathcal{B}_f\left(\frac{p(x)}{q^{\text{old}}(x)} \parallel \frac{q_\theta(x)}{q^{\text{old}}(x)}\right) &= \\ f\left(\frac{p(x)}{q^{\text{old}}(x)}\right) - f\left(\frac{q_\theta(x)}{q^{\text{old}}(x)}\right) &- f'\left(\frac{q_\theta(x)}{q^{\text{old}}(x)}\right)\left(\frac{p(x)}{q^{\text{old}}(x)} - \frac{q_\theta(x)}{q^{\text{old}}(x)}\right) \\ = f\left(\frac{p(x)}{q^{\text{old}}(x)}\right) - \frac{p(x)}{q^{\text{old}}(x)}f'\left(\frac{q_\theta(x)}{q^{\text{old}}(x)}\right) &+ \left(f'\left(\frac{q_\theta(x)}{q^{\text{old}}(x)}\right)\frac{q_\theta(x)}{q^{\text{old}}(x)} - f\left(\frac{q_\theta(x)}{q^{\text{old}}(x)}\right)\right). \end{aligned} \quad (17)$$

By using Lemma 7.1, and defining $\tau_\theta = \frac{q_\theta(x)}{q^{\text{old}}(x)}$, the above Bregman divergence can be reformulated as:

$$\begin{aligned} \mathcal{B}_f\left(\frac{p(x)}{q^{\text{old}}(x)} \parallel \frac{q_\theta(x)}{q^{\text{old}}(x)}\right) &= \\ f\left(\frac{p(x)}{q^{\text{old}}(x)}\right) - \frac{p(x)}{q^{\text{old}}(x)}f'\left(\frac{q_\theta(x)}{q^{\text{old}}(x)}\right) &+ f^*\left(f'\left(\frac{q_\theta(x)}{q^{\text{old}}(x)}\right)\right) \\ = f\left(\frac{p(x)}{q^{\text{old}}(x)}\right) - \frac{p(x)}{q^{\text{old}}(x)}\tau_\theta(x) &+ f^*(\tau_\theta(x)). \end{aligned} \quad (18)$$

Hence, by getting expectation, we have:

$$\begin{aligned} \mathbb{E}_{x \sim \mathcal{Q}^{\text{old}}}[\mathcal{B}_f\left(\frac{p(x)}{q^{\text{old}}(x)} \parallel r_\theta(x)\right)] &= \mathbb{E}_{x \sim \mathcal{Q}^{\text{old}}}\left[f\left(\frac{p(x)}{q^{\text{old}}(x)}\right)\right] \\ &- \mathbb{E}_{x \sim \mathcal{Q}^{\text{old}}}\left[\frac{p(x)}{q^{\text{old}}(x)}\tau_\theta(x)\right] + \mathbb{E}_{x \sim \mathcal{Q}^{\text{old}}}[f^*(\tau_\theta(x))] \\ &= \mathbb{E}_{x \sim \mathcal{Q}^{\text{old}}}\left[f\left(\frac{p(x)}{q^{\text{old}}(x)}\right)\right] - \mathbb{E}_{x \sim P}[\tau_\theta(x)] + \mathbb{E}_{x \sim \mathcal{Q}^{\text{old}}}[f^*(\tau_\theta(x))] \\ &= \mathcal{D}_f(P \parallel \mathcal{Q}^{\text{old}}) - \mathcal{L}_f(P, \mathcal{Q}^{\text{old}}, \tau_\theta). \end{aligned} \quad (19)$$

□

Lemma 7.2. *Let α be a constant, if f can be written as $f(u) = uf(\frac{1}{u}) + \alpha(u-1)$, then:*

$$\frac{1}{x}\mathcal{B}_f(x \parallel y) = \mathcal{B}_f\left(\frac{1}{x} \parallel \frac{1}{y}\right). \quad (20)$$

Proof. By taking the derivative of f :

$$\begin{aligned} f(u) &= uf\left(\frac{1}{u}\right) + \alpha(u-1) \\ \Rightarrow f'(u) &= f\left(\frac{1}{u}\right) - \frac{f'\left(\frac{1}{u}\right)}{u} + \alpha, \end{aligned} \quad (21)$$

and using the definition of Bregman divergence, we have:

$$\begin{aligned} \mathcal{B}_f\left(\frac{1}{x} \parallel \frac{1}{y}\right) &= f\left(\frac{1}{x}\right) - f\left(\frac{1}{y}\right) - f'\left(\frac{1}{y}\right)\left(\frac{1}{x} - \frac{1}{y}\right) \\ &= \left(\frac{f(x)}{x} + \frac{\alpha}{x} - \alpha\right) - \left(\frac{f(y)}{y} + \frac{\alpha}{y} - \alpha\right) \\ &- \left((f(y) - yf'(y) + \alpha)\left(\frac{1}{x} - \frac{1}{y}\right)\right) \\ &= \left(\frac{f(x)}{x} + \frac{\alpha}{x} - \frac{f(y)}{y} - \frac{\alpha}{y}\right) \\ &- \left(\frac{f(y)}{x} - \frac{yf'(y)}{x} + \frac{\alpha}{x} - \frac{f(y)}{y} + f'(y) - \frac{\alpha}{y}\right) \\ &= \frac{f(x)}{x} - \frac{f(y)}{x} + \frac{yf'(y)}{x} - f'(y) \\ &= \frac{1}{x}\left(f(x) - f(y) - f'(y)(x-y)\right) \\ &= \frac{1}{x}\mathcal{B}_f(x \parallel y). \end{aligned} \quad (22)$$

□

Lemma 7.3. *Let α be a constant, if f satisfies $f(u) = uf(\frac{1}{u}) + \alpha(u-1)$, then:*

$$\mathbb{E}_{x \sim \mathcal{Q}}[\mathcal{B}_f\left(\frac{p(x)}{q(x)} \parallel r(x)\right)] = \mathbb{E}_{x \sim P}[\mathcal{B}_f\left(\frac{q(x)}{p(x)} \parallel \frac{1}{r(x)}\right)]. \quad (23)$$

Proof. Based on Lemma 7.2, for each x , we have:

$$q(x)\mathcal{B}_f\left(\frac{p(x)}{q(x)} \parallel r(x)\right) = p(x)\mathcal{B}_f\left(\frac{q(x)}{p(x)} \parallel \frac{1}{r(x)}\right). \quad (24)$$

Finally, using a summation on all x in Eq. 24, the proof is completed. □

Proof of Theorem 4.1

Proof. Let $f(u) = u \ln u - (u + 1) \ln(u + 1)$. The Fenchel conjugate of f is obtained as below:

$$f^*(t) = -\ln(1 - \exp(t)) \quad \text{s.t. } t \in \mathbb{R}_-. \quad (25)$$

Based on Theorem 4.3, we have:

$$\begin{aligned} \tau_\theta(x) &= \ln \frac{q_\theta(x)}{q_\theta(x) + q^{\text{old}}(x)} \\ \Rightarrow f^*(\tau_\theta(x)) &= -\ln\left(\frac{q^{\text{old}}(x)}{q_\theta(x) + q^{\text{old}}(x)}\right), \end{aligned} \quad (26)$$

and thus:

$$\begin{aligned} \mathcal{D}_{JS}(P||Q^{\text{old}}) &= \mathcal{L}_f(P, Q^{\text{old}}, \tau_\theta) \\ &+ \mathbb{E}_{x \sim Q^{\text{old}}} \left[\mathcal{B}_f\left(\frac{p(x)}{q^{\text{old}}(x)} \parallel \frac{q_\theta(x)}{q^{\text{old}}(x)}\right) \right], \\ \mathcal{L}_f(P, Q^{\text{old}}, \tau_\theta) &= \mathbb{E}_{x \sim P} \left[\ln \frac{q_\theta(x)}{q_\theta(x) + q^{\text{old}}(x)} \right] \\ &+ \mathbb{E}_{x \sim Q^{\text{old}}} \left[\ln \frac{q^{\text{old}}(x)}{q_\theta(x) + q^{\text{old}}(x)} \right], \end{aligned} \quad (27)$$

where \mathcal{D}_{JS} is Jensen-Shannon divergence.

On other side, f satisfies the assumptions of Lemma 7.3 by selecting $\alpha = 0$, Thus:

$$\mathbb{E}_{x \sim Q^{\text{old}}} \left[\mathcal{B}_f\left(\frac{p(x)}{q^{\text{old}}(x)} \parallel \frac{q_\theta(x)}{q^{\text{old}}(x)}\right) \right] = \mathbb{E}_{x \sim P} \left[\mathcal{B}_f\left(\frac{q^{\text{old}}(x)}{p(x)} \parallel \frac{q^{\text{old}}(x)}{q_\theta(x)}\right) \right], \quad (28)$$

and proof will be complete. \square

Proof of Corollary 4.1.1

Proof. Since $\mathcal{D}_{JS}(P||Q^{\text{old}})$ is constant with respect to θ , maximizing $\mathcal{L}(P, Q^{\text{old}}, Q_\theta)$ is equivalent to minimizing $\mathbb{E}_{x \sim P} \mathcal{B}_f\left(\frac{q^{\text{old}}(x)}{p(x)} \parallel \frac{q^{\text{old}}(x)}{q_\theta(x)}\right)$ and $\mathbb{E}_{x \sim Q^{\text{old}}} \mathcal{B}_f\left(\frac{p(x)}{q^{\text{old}}(x)} \parallel \frac{q_\theta(x)}{q^{\text{old}}(x)}\right)$ (according to Eq. 7). \square

Lemma 7.4. For each f -divergence with function f , there exists another f -divergence with function g defined as below where $\mathcal{D}_f(P||Q) = \mathcal{D}_g(P||Q)$ and $g'(1) = 0$.

$$g(u) = f'(1) - f'(1)u + f(u). \quad (29)$$

Proof.

$$\begin{aligned} \mathcal{D}_g(P||Q) &= \sum_x q(x)g\left(\frac{p(x)}{q(x)}\right) \\ &= \sum_x q(x) \left[f'(1) - f'(1)\frac{p(x)}{q(x)} + f\left(\frac{p(x)}{q(x)}\right) \right] \\ &= f'(1) - f'(1) + \sum_x q(x) \left[f\left(\frac{p(x)}{q(x)}\right) \right] \\ &= f'(1) - f'(1) + \mathcal{D}_f(P||Q) \\ &= \mathcal{D}_f(P||Q). \end{aligned} \quad (30)$$

As $f(1) = 0$ for a valid f -divergence, we will have:

$$g'(1) = f'(1) - f'(1) + f(1) = f(1) = 0. \quad (31)$$

\square

Lemma 7.5. Let P, Q be two distributions; if there exists a distribution $\phi(x)$ that $\min\{p(x), q(x)\} < \phi(x) < \max\{p(x), q(x)\}$ for all x , we will have:

$$\mathcal{D}_f(P||\Phi) < \mathcal{D}_f(P||Q), \quad (32)$$

where f is strictly convex.

Proof. by using $f(1) = 0$ property:

$$\begin{aligned} \mathcal{D}_f(P||\Phi) - \mathcal{D}_f(P||Q) &= \\ &= \sum_x p(x) \left[f\left(\frac{p(x)}{\phi(x)}\right) - f\left(\frac{p(x)}{q(x)}\right) \right] \\ &= \sum_{x:p(x)>q(x)} p(x) \left[f\left(\frac{p(x)}{\phi(x)}\right) - f\left(\frac{p(x)}{q(x)}\right) \right] \\ &+ \sum_{x:p(x)<q(x)} p(x) \left[f\left(\frac{p(x)}{\phi(x)}\right) - f\left(\frac{p(x)}{q(x)}\right) \right] \end{aligned} \quad (33)$$

and by using lemma assumption ($\min\{p(x), q(x)\} < \phi(x) < \max\{p(x), q(x)\}$),

$$\begin{aligned} \mathcal{D}_f(P||\Phi) - \mathcal{D}_f(P||Q) &= \\ &= \sum_{x:p(x)>\phi(x)>q(x)} p(x) \left[f\left(\frac{p(x)}{\phi(x)}\right) - f\left(\frac{p(x)}{q(x)}\right) \right] \\ &+ \sum_{x:p(x)<\phi(x)<q(x)} p(x) \left[f\left(\frac{p(x)}{\phi(x)}\right) - f\left(\frac{p(x)}{q(x)}\right) \right]. \end{aligned} \quad (34)$$

As shown in Lemma 7.4, we can consider an equivalent divergence having the property $f'(1) = 0$. Thus, this lemma and the strict convexity of f conclude that f is strictly increasing function for $x > 1$ and strictly decreasing one for $x < 1$. As a result, each summation term on Eq. 34 has a negative value and the proof is completed. \square

Proof of Theorem 4.4

Proof. The function $f(x) = \frac{x}{1-x}$ is strictly increasing between zero and one, so we have:

$$\begin{aligned} &\min\{0.5, D^*(x)\} < D_\theta(x) < \max\{0.5, D^*(x)\} \\ \Rightarrow &\frac{\min\{0.5, D^*(x)\}}{1 - \min\{0.5, D^*(x)\}} < \frac{D_\theta(x)}{1 - D_\theta(x)} \\ &< \frac{\max\{0.5, D^*(x)\}}{1 - \max\{0.5, D^*(x)\}} \\ \Rightarrow &\min\left\{\frac{0.5}{1 - 0.5}, \frac{D^*(x)}{1 - D^*(x)}\right\} < \frac{D_\theta(x)}{1 - D_\theta(x)} \\ &< \max\left\{\frac{0.5}{1 - 0.5}, \frac{D^*(x)}{1 - D^*(x)}\right\} \\ \Rightarrow &\min\left\{1, \frac{p(x)}{q^{\text{old}}(x)}\right\} < \frac{D_\theta(x)}{1 - D_\theta(x)} \\ &< \max\left\{1, \frac{p(x)}{q^{\text{old}}(x)}\right\} \\ \Rightarrow &\min\{q^{\text{old}}(x), p(x)\} < \frac{D_\theta(x)}{1 - D_\theta(x)} q^{\text{old}}(x) \\ &< \max\{q^{\text{old}}(x), p(x)\} \\ \Rightarrow &\min\{q^{\text{old}}(x), p(x)\} < q_\theta(x) < \max\{q^{\text{old}}(x), p(x)\} \end{aligned} \quad (35)$$

according to Lemma 7.5, the proof is completed. \square

Proof of Theorem 4.2

Proof. In Theorem 4.4, a general case for every f -divergence is shown and Theorem 3.2 is a special case in which $f(u) = u \ln u - (u + 1) \ln(u + 1)$. \square

8. Samples of generated sentences

Here are the generated samples of different models on the Image Coco captions, Yelp Restaurant reviews, and Amazon datasets. In addition to reporting the randomly generated samples of all models, we sorted samples of different models based on BLEU5 (as quality metric) and selected top scoring ones to have a fairer comparison.

The best samples based on BLEU5 from DGSAN method trained on Coco dataset.

- a man and woman are sitting on a plate .
- a man and a woman standing in front of a large building .
- a man standing in front of a black dog .
- a picture of a person riding a horse in the street .
- a woman riding a horse in front of a fence .
- man and a woman standing in front of a bicycle .
- group of people are skiing on a snow covered field .
- a crowd of people standing next to each other .
- a man is riding a horse in the grass .
- a man and a woman standing in a kitchen .

Random samples from DGSAN method trained on Coco dataset.

- a pregnant woman with bananas on her head holding a black cat .
- a brown horse and a white horse running in sand .
- a kitchen with a stove and a breakfast bar .
- two dogs sitting in the side of the bathroom mirror .
- four skiers standing together in the snow in front of a mountain .
- the young boy is holding an apple with bites in it .
- a cutting board with chopped scallions and broccoli on it .
- two stuffed animals are on top of a plate .
- two people riding bikes on the back of a man's back of a group of people .
- a partially eaten apple by a verizon device .

The best samples based on BLEU5 from MLE method trained on Coco dataset.

- a group of people skiing down a ski slope .
- a person skiing down a snow covered slope .
- a person riding a skateboard on a city street .
- a man standing in front of a crowd of people .
- a horse in a field with trees .
- an orange sitting on a white plate .
- a group of men standing next to each other on a snow covered street .
- a man is sitting on a wooden bench .
- a couple is sitting on a wooden bench .
- a person that is on the side of a road .

Random samples from MLE method trained on Coco dataset.

- a fancy car propped up .
- three people on horses riding on the beach .
- the horses seem eager to walk down the path .
- a kitchen with a white stove and a black doorway .
- people are looking through birds in the process snowy living room .
- two red and yellow checkers .
- an image with boats waiting across the bridge into a sky .
- a picture of a woman using a cat .
- a pitcher preparing to throw a base ball .
- several food items at a vegetable stand including apples and citrus .

The best samples based on BLEU5 from MaliGAN method trained on Coco dataset.

- a man riding skis down a ski slope .
- a person standing on top of a plate .
- a dog is standing in front of a car .
- a couple sitting on a white counter .
- a woman sitting on a wooden bench .
- a cat sitting on a kitchen counter .
- a group of people sitting on a blue plate .
- a man sitting on a bed in a red .
- a group of people sitting on top of a table .
- a group of people sitting on a bed .

Random samples from *MaliGAN* method trained on *Coco* dataset.

- a woman with eye glasses in the dog is standing next to a large ornate tree .
- kittens playing with pictures of a plant pot .
- two kites shaped animals sitting on a light fence , with graffiti on the upper part of a bathroom .
- a black and white photograph of a street and buildings by a tower .
- a baseball player throwing a ball from the pitchers mound .
- the cyclists are waiting for a turn and a black dog sitting .
- a bathroom sink some beige and a mouse also in the shape of a toilet .
- a group of skateboarders gathered with their skateboards .
- a man riding a motorcycle on a dirt road .
- plantains on a stalk sitting on a porch on a sunny day .

The best samples based on BLEU5 from *RankGAN* method trained on *Coco* dataset.

- a person standing in front of a truck .
- motorcycles parked in front of a building .
- motorcycles parked in front of a building .
- motorcycles parked in front of a building .
- motorcycles parked in front of a building .
- motorcycles parked in front of a building .
- motorcycles parked in front of a building .
- motorcycles parked in front of a building .
- motorcycles parked in front of a building .
- motorcycles parked in front of a building .

Random samples from *RankGAN* method trained on *Coco* dataset.

- an image of a basket of fruit in a bowl .
- man preparing simple appliances .
- parking lit up inside of water that .
- a little kid that is doing a skateboard trick in the air .
- man in living room with a baby bowl .
- multiple banana bunches growing on a leafy tree .
- a woman in a black helmet jumping a hurdle while riding a horse .
- a man in brown shirt playing with a red frisbee .
- a kitchen with wooden counters and a stove with windows .
- some pots and one's and there are is on benches .

The best samples based on BLEU5 from *SeqGAN* method trained on *Coco* dataset.

- a close up of a bunch of people .
- a boy stands on a sidewalk in front of a store .
- two people riding on the side of a road .
- a woman standing next to a woman on a sidewalk .
- two giraffe standing next to each other on a shelf in a sink .
- a man is standing in front of a man riding a motorcycle .
- a close up of a bunch of people standing around .
- a close up of a cat sitting on a table .
- a close up of a cat sitting on a table .
- a woman is sitting on top of a motorcycle .

Random samples from *SeqGAN* method trained on *Coco* dataset.

- a green and orange garden in front of the green .
- a large passenger jet taking off from an airport runway .
- a team of baseball players stand in a field and visit and wait .
- a woman holds a bowl with bananas on her head .
- a variety of fruit placed in a wicker basket .
- a white bowl with shrimp , broccoli and rice .
- a boy with a black cap and eyeglasses has a black baseball mitt on his hand .
- a man riding across a snow covered slope .
- a person riding a bike down on a city street .
- a skater is just about to pull off a trick .

The best samples based on BLEU5 from *VAE* method trained on *Coco* dataset.

- a man riding a brown horse through a field .
- a person standing in front of a bicycle .
- a horse is standing in the middle of a road .
- a lot of people that are on the beach .
- a close up of a piece of chocolate cake .
- a person standing in front of a truck .
- a man is standing in front of a group of people .
- a woman riding skis on a snow covered ski slope .
- black and white photograph of a man in a park .
- a person cross country skiing on a snow slope .

Random samples from VAE method trained on Coco dataset.

- a person is standing alongside a deep pathway through a meadow with trees .
- many different types of leafy vegetables on a table .
- competitor in a down hill ski race during the olympics .
- two people on skis on slope with trees next to them .
- the view from a horse being led down a path with others .
- a kitchen with cabinets and a toilet with toiletries in it .
- a toilet with a wind vane in it's reflection .
- a boy jumping in mid air on a skateboard .
- three woman on skis in the snow smiling .
- a close up of a person sitting at a table eating a carrot .

The best samples based on BLEU5 from DGSAN method trained on Yelp Restaurant dataset.

- i love this place .
- i was not impressed at all .
- do n't stay here again .
- this place is great !
- the staff is friendly and helpful .
- do not stay here .
- i 'm a fan .
- the worst i 've ever had .
- i love this place .
- the food was amazing .

Random samples from DGSAN method trained on Yelp Restaurant dataset.

- this place is good not the best like the other expensive place .
- fries usually have way too much salt on them and/or are old .
- they lost our business permanently .
- the interior treats is excellent .
- big mistake .
- they have good thai iced tea too .
- on the 4th level they have a race car perfect for pictures and restrooms .
- this is one of the best joint i 've ever had .
- quality prices are very reasonable .
- it was good and fresh .

The best samples based on BLEU5 from MLE method trained on Yelp Restaurant dataset.

- the service was good .
- we will definitely be back .
- i love this place .
- i highly recommend it !
- the service was great and the food was delicious .
- i will be back .
- i will not be going back .
- the food is amazing !
- i love this place !
- this place is awesome !

Random samples from MLE method trained on Yelp Restaurant dataset.

- great vibe , sad if you are going on .
- i had the eggplant and bottle ” as “ lobster ”
- that and also the line to check out was ridiculous .
- first class service and food .
- has great breakfast - fabulous businesses service .
- the customer service crap .
- do n't listen to me , you might like it .
- over the years the customer service has always been blah .
- the hallway was falling apart .
- we had our anniversary dinner at the steakhouse and it was delicious !

The best samples based on BLEU5 from MaliGAN method trained on Yelp Restaurant dataset.

- the food is always delicious .
- i love this place .
- i love this place !
- i love this place !
- they do n't care less .
- this place was amazing .
- i love this place !
- this place is great !
- this place is amazing !
- i will definitely go back .

Random samples from MaliGAN method trained on Yelp Restaurant dataset.

- it was great years ago and still is .
- an amazing evening without fault .
- what an amazingly horrible experience .
- the breaded cover was loose and thick .
- shrimp tempura tasted terrible .
- but i love their sushi !
- we just got delivery of our food and it was delicious !
- not recommended again .
- more elegant , and have you seen their golf course !
- i 'm a vegetarian and love eating indian food .

The best samples based on BLEU5 from RankGAN method trained on Yelp Restaurant dataset.

- i guess they do n't seem to care .
- i love this place .
- i love this place !
- it was very very good .
- this place is terrible .
- i 'm not going back .
- avoid at all costs .
- i love this place !
- i love this place .
- you do n't seem to care .

Random samples from RankGAN method trained on Yelp Restaurant dataset.

- rude bartenders and rude management !
- the staff has always been nice to me .
- i could n't hold a conversation with my date .
- the dinner was ok , nothing special and was small in portion .
- room was clean , but bad fit and finish .
- great too .
- our waitress carrie was disgusting .
- do n't go the time .
- they said it was a mistake and he was removed the next day .
- awesome !

The best samples based on BLEU5 from SeqGAN method trained on Yelp Restaurant dataset.

- very friendly and fast .
- i love this place .
- i love this place .
- i will be back .
- service was really good .
- the food is always delicious .
- the food was delicious .
- i love this place .
- this place is great .
- this place is good .

Random samples from SeqGAN method trained on Yelp Restaurant dataset.

- one of my favorite places to eat !
- it 's bland .
- wonderful vegan raw foods !
- incredible !
- i will be going again !
- great moroccan food and service .
- i did n't want to kill .
- drinks are fabulous perfectly priced .
- huge portions and the food was delicious as always .
- this is a very nice hotel .

The best samples based on BLEU5 from VAE method trained on Yelp Restaurant dataset.

- this place is awesome .
- the staff is very friendly .
- the food was great .
- just do n't go back .
- this place is great .
- the food is excellent .
- i love this place .
- i love this place .
- i highly recommend it !
- the food was great .

Random samples from VAE method trained on Yelp Restaurant dataset.

- expensive .
- they are rude and have the most disgusting attitude towards people .
- the bar menu was very good but the lobster lacks flavor quality .
- tonight , the last time .
- i like the food but i am concerned about the cleanliness of the place .
- just get me my chocolate milk dude !
- i had the spicy chicken and i was satisfied with it .
- , i trust you 'll enjoy .
- every time that quick , great attention to match on the service .
- love love .

The best samples based on BLEU5 from DGSAN method trained on Amazon dataset.

- my daughter and i love this game , it 's a fun game to play to pass the time . i love this game .
- i love this game . love it ! ! ! ! ! i love this game and the graphics are great . i would recommend this game to anyone .
- this was a good book that i could n't wait to read the next book in the series .
- could not put this book down . i ca n't wait to read the next one in the series . it 's great ! !
- i really liked this book . i would like to see more from this author . i am looking forward to the next book in the series !
- it was a good read . i am looking forward to reading the rest of the series .
- i really enjoyed reading this book . i would recommend this book . it is very entertaining .
- can not wait to read the next book in the series . i ca n't wait to read the next one .
- i really enjoyed this book , i could not put it down . ca n't wait to read the next one ! ! ! ! !
- when i have to wait for the next book . i loved it . i ca n't wait to read more in this series .

The best samples based on BLEU5 from *MaliGAN* method trained on *Amazon* dataset.

- worth the time . i recommend this book . i ca n't wait to read the next one .
- i really enjoyed this book , ca n't wait to read the second book in the series
- i really enjoyed this book . i am looking forward to the next book . i would love more .
- i loved this book . it kept me interested . i will be purchasing more
- i enjoyed reading this book and it was a good book . looking forward to the series . it
- i fell in love with this book . it was hard to follow . i loved it . i can not wait to read something
- this book held my interest . i would love to read more of this book .
- this was a great story . i would highly recommend this book . i really enjoyed the book and look forward to the rest of the bride .
- this book could n't put it down . the plot could have been longer . it was hard to put down .
- but i liked it and ca n't wait to read the second book in this series . i did n't know what i finished .

Random samples from *MaliGAN* method trained on *Amazon* dataset.

- set of cards and clive cussler book 's think ! science fiction is excellently written . this was a little more interesting . worth it ... doubt again
- it repeats of the plot the end . if you want this to test ur life in this series . you feel peaceful . it is good and highly recommend
- i enjoyed and his flaws and a happily ever after i progressed and the story was taking
- hard to put it down once you star ... i had to though because at times i laughed so hard . i am looking forward for the next titles .
- fun and challenging . my 9 and 11 year olds love it too , good for both adults and children , had a blast !
- love this author . intriguing story . great lead characters . plot moved along quickly . have read and/or purchased other books by this author . glad i started with this one .
- another such been purchasing this book . it was very enjoyable . i do . opinion by the electronic dollar . i liked the chapters . i really liked the characters but the questions were written with excitement .
- i just started playing this game & amp ; i 'm enjoying it thoroughly . i love the vibrant colors & amp ; graphics . i love it !!!
- the google keyboard does everything this does and is included on most devices . also , it looks better in my opinion .
- i love this game ! ! its so cute and addicting ! ! ! i 'm not into scary zombie games , this is fun to play and kid friendly

The best samples based on BLEU5 from *RankGAN* method trained on *Amazon* dataset.

- this book was very interesting . i do n't recommend this book - could not put it down .
- this book could n't put it down . very good ! i hope the next one .
- could n't put it down . i just could n't put down .
- well written . i will read more from this author in the future . a very enjoyable book .
- i could n't put the book down ! i could n't wait for the next one in the whole family
- i could n't wait for the next book in this series and i enjoyed the entire book .
- short book really enjoyed the book , could n't put it down , i ca n't wait for the next one .
- funny . could n't put it down , this book was great !
- i enjoyed the book cant wait for the next one . this is one of the best i 've ever read
- not sure that i could n't put the book down . very good . i could not wait to see my attention

Random samples from *RankGAN* method trained on *Amazon* dataset.

- good story not the best dean koontz story i 've ever read but it was interesting all the same i enjoyed it .
- history you will never finish it . not much to that written book from the beginning earlier ! i enjoyed the characters to make the characters real . they were very honest at the end .
- not much to a child . some bad things i did n't want to have little money in the most part . good read !
- if you have a word word of problems in the & # 34 ; anecdotes and learning more please and the end .
- not my favorite of a author . it had some twists and written not how the book ended . will recommend this one .
- i love reading sequels and reading each characters story . this one is especially sweet . ca n't wait for the next book !
- funny and great imagination of the real of the friends and the history of the reacher books . a best that i have ever read . very close . just keep on them
- i loved this book from the beginning to the end . such a good story of faith and love and how god will be with us always . i would recommend this book to everyone ! ! ! !
- i was not impressed with this app despite the fact i love hidden object games . would not recommend this app
- i am enjoying the book but it seems to be starting to drag just a bit for me .

The best samples based on BLEU5 from *SeqGAN* method trained on *Amazon* dataset.

- i liked this book but i could n't put it down . i loved this book . i love this . i am looking forward to the next in the series and i will definitely be reading more of
- if i did n't know how it did n't put it down .
- i just could n't put it down . it was a good read and i am glad i did n't see why . very recommend it to others
- i had a hard time putting it down , but i enjoyed the book and continued to be . i am looking forward to reading more of this author .
- did n't get better , but when i could n't put it down . i loved this book .
- this story was very good . i enjoyed the story so i loved it and i am hooked and i am looking forward to next in this series . i could n't put it down , and do not
- this was a great surprise and i could n't put the book down and read and had a hard time putting it down ! i am so glad i read it ! ! ! !
- why i do n't write a novel with a lot of twist .
- this one was a book i could n't put it down . i really enjoyed this book and could n't have read so far , but i did n't have to read but this has
- this is one of those books that i could n't put the book down . i could n't live it . i like it . it was a little predictable but it was written with a few

Random samples from *SeqGAN* method trained on *Amazon* dataset.

- the movie was was more from this as i do n't follow if we went into a short thing it if i read in this book but i found it very interesting if it is free , i do
- i love to read and them all i got really easy to put each to the reviews . i enjoyed this book ! i really love her books for us all on to reading out things . on the basis that
- it 's a nice entertainment to read this story and the idea to talk through flowers is new but the characters are not that intricate .
- this game is another fun dash game i really enjoyed it . kind of repetitive but it was enjoyable as well .
- keeps you in read some good ghost novels in this book , i i really have my therapist and writing by this and a good read ! i am really impressed with good info for anyone who likes to the kids
- i thought that this was a wonderful story ! ! i actually cried a few times , this book is fantastic & i loved it ! ! !
- michael connelly is an excellent mystery writer . this book is not his best but is good .
- a great story . not the less book will do , i enjoy this story . it deleted and kept me though there are a friend it has a great story and i do n't know what this story
- love these books , i loved the humor and got to know the way to the end . i found this book just an interesting little ones . but not that i would recommend it . can all the whistles of
- i have always wanted a horse and this game is so cute ! ! !