
Sequence Generation with a Guider Network

Ruiyi Zhang¹ Changyou Chen² Zhe Gan³ Wenlin Wang¹ Zheng Wen⁴ Lawrence Carin¹

Abstract

Sequence generation with reinforcement learning (RL) has received significant attention recently. However, the sparse-reward issue remains to be a main challenge in the RL training process, where only a scalar guiding signal is available after an entire sequence has been generated. This type of sparse rewards tend to ignore global structural information of a sequence, causing generation of sequences semantically inconsistent. In this paper, we present a model-based RL approach to overcome this issue. Specifically, we propose a novel guider network to model the sequence-generation environment, which can assist next-word prediction and provide intermediate rewards for generator optimization. Extensive experiments demonstrate that the proposed method leads to improved performance for both unconditional and conditional sequence-generation tasks.

1. Introduction

Sequence generation is an important area of investigation within machine learning. Recent work has shown excellent performance on a number of tasks, by combining reinforcement learning (RL) and generative models. Example applications include image captioning (Ren et al., 2017; Rennie et al., 2016), text summarization (Li et al., 2018; Paulus et al., 2017; Rush et al., 2015), and adversarial text generation (Guo et al., 2017; Lin et al., 2017; Yu et al., 2017; Zhang et al., 2017; Zhu et al., 2018). The sequence-to-sequence framework (Seq2Seq) (Sutskever et al., 2014) is a popular technique for sequence generation. However, models from such a setup are typically trained to predict the next token given previous ground-truth tokens as input, causing what is termed *exposure bias* (Ranzato et al., 2016). By contrast, sequence-level training with RL provides an effective way to solve this challenge by treating sequence generation as

an RL problem. By directly optimizing an evaluation score (cumulative rewards) (Ranzato et al., 2016), state-of-the-art results have been obtained in many sequence-generation tasks (Paulus et al., 2017; Rennie et al., 2016). However, one problem in such a framework is that rewards in RL training are particularly sparse, since a scalar reward is typically only available after an entire sequence has been generated. For RL-based sequence generation, most existing works rely on a model-free framework, which has been criticized for its high variance and poor sample efficiency (Sutton & Barto, 1998). On the other hand, model-based RL methods do not suffer from these issues, but they are usually difficult to train in complex environments. Furthermore, a learned policy is usually restricted by the capacity of an environment model. Recent developments on model-based RL (Gu et al., 2016; Kurutach et al., 2018; Nagabandi et al., 2017) combine the advantages of these two approaches, and have achieved improved performance by learning a model-free policy, assisted by an environment model. In addition, model-based RL has been employed recently to solve problems with extremely sparse rewards, with curiosity-driven methods (Pathak et al., 2017).

Inspired by the ideas in (Gu et al., 2016; Kurutach et al., 2018; Nagabandi et al., 2017; Pathak et al., 2017), we propose a model-based RL method to overcome the sparse-reward problem in sequence-generation tasks. Our main idea is to employ a new guider network to model the generation environment in the feature space of sequence tokens, which is used to emit intermediate rewards by matching the predicted features from the guider network and features from generated sequences. The guider network is trained to encode global structural information of training sequences, useful to guide next-token generation in the generative process. Within the proposed framework, we also develop a new type of self-attention mechanism, to assist the guider network to provide high-level planning-ahead information. The intermediate rewards are combined with a final scalar reward, e.g., an evaluation score in a Seq2Seq generation model or the discriminator loss in the generative-adversarial-network (GAN) framework, to train a sequence generator with policy-gradient methods. Extensive experiments show improved performance of our method on sequence-generation tasks, relative to existing state-of-the-art methods.

¹Duke University ²SUNY at Buffalo ³Microsoft AI ⁴Adobe Research. Correspondence to: Ruiyi Zhang <ryzhang@cs.duke.edu>.

2. Background

2.1. Sequence-to-Sequence Model

A sequence-generation model learns to generate a sequence $Y = (y_1, \dots, y_T)$ conditioned on a possibly empty object X from a different feature space. Here $y_t \in \mathcal{A}$ with \mathcal{A} the alphabet set of output tokens. The pairs (X, Y) are used for training a sequence-generation model. We use T to denote the length of an output sequence, and $Y_{1,\dots,t}$ to indicate a subsequence of the form (y_1, \dots, y_t) . The output of a trained generator is denoted \hat{Y} . Since we focus on text generation in this paper, we will use *token* and *word* interchangeably to denote an element of a (text) sequence.

Starting from the initial hidden state s_0 , a recurrent neural network (RNN) produces a sequence of states (s_1, \dots, s_T) given a sequence-feature representation $(e(y_1), \dots, e(y_T))$, where $e(\cdot)$ denotes a function mapping a token to its feature representation. Let $e_t \triangleq e(y_t)$. The states are generated by applying a transition function $h : s_t = h(s_{t-1}, e_t)$ for T times. The transition function h is implemented by a cell of an RNN, with popular choices being the Long Short-Term Memory (LSTM) (Hochreiter & Schmidhuber, 1997) and the Gated Recurrent Unit (GRU) (Cho et al., 2014). We will use the LSTM for our model. To generate a token $\hat{y}_t \in \mathcal{A}$, a stochastic output layer is applied on the current state s_t :

$$\hat{y}_t \sim \text{Multi}(1, \text{softmax}(g(s_{t-1}))), \quad (1)$$

$$s_t = h(s_{t-1}, e(\hat{y}_t)), \quad (2)$$

where $\text{Multi}(1, \cdot)$ denotes one draw from a multinomial distribution, and $g(\cdot)$ represents a linear transformation. Since the generated sequence Y is conditioned on X , one can simply start with an initial state encoded from X : $s_0 = s_0(X)$ (Bahdanau et al., 2017; Cho et al., 2014). Such conditional RNN can be trained for sequence generation with gradient ascent by maximizing the log-likelihood of a generative model.

2.2. Model-Based Reinforcement Learning

Reinforcement learning aims to learn an optimal policy for an agent interacting with an unknown or highly complex environment. A policy is modeled as a conditional distribution $\pi(a|s)$, which specifies the probability of choosing an action $a \in \mathcal{A}$ at the state $s \in \mathcal{S}$. Formally, an RL problem is characterized by a Markov decision process (MDP), $\mathcal{M} = \langle \mathcal{S}, \mathcal{A}, P, r, \gamma \rangle$, where $\gamma \in (0, 1)$ is the discrete-time discount factor. If the agent chooses action $a \in \mathcal{A}$ at state $s \in \mathcal{S}$, then the agent will receive an immediate reward $r(s, a)$, and the state will transit to $s' \in \mathcal{S}$ with probability $P(s'|s, a)$. The agent knows \mathcal{S} , \mathcal{A} , and γ , but may not know P or r . The expected discounted total reward of a policy π is defined as:

$$J(\pi) = \sum_{t=1}^{\infty} \mathbb{E}_{P, \pi} [\gamma^{t-1} \cdot r(s_t, a_t)]. \quad (3)$$

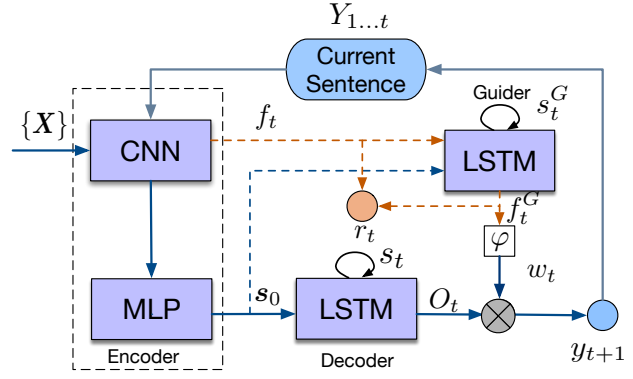


Figure 1. Model Overview: Sequence generation with a guider network. Solid lines mean gradients are backpropagated in training; dash lines mean gradients are not backpropagated.

The agent’s goal is to learn an optimal policy that maximizes $J(\pi)$.

In model-based RL, a model of environment dynamics is built to make predictions for future states conditioned on the current state, which can be used for action selection, e.g., next-token generation. This model is typically a discrete-time system, taking the current state-action pair (s_t, a_t) as input, and outputting an estimate of the next state $s_{t+\Delta t}$ at time $t + \Delta t$. At each step t , a_t is chosen based on the model, and the model will re-plan with the updated information from the dynamics. This control scheme is referred to as model-predictive control (MPC) (Nagabandi et al., 2017). Note that in our setting, the state s in RL typically corresponds to the current generated sentences $Y_{1,\dots,t}$ and possibly empty object X in sequence generation.

3. Proposed Model

The model is illustrated in Figure 3, with the first building block an autoencoder (AE) structure (the encoder-decoder in Figure 3) for sentence feature extraction and generation. The encoder is shared for sentences from both training data and generated data, as explained in detail below. Overall, sequence generation can be formulated as a sequential decision-making problem. At each timestep t , the agent, also called a generator (which corresponds to the LSTM decoder in Figure 3), takes the current LSTM state as input, denoted as s_t . The policy $\pi^\phi(\cdot|s_t)$ parameterized by ϕ is a conditional generator, to generate the next token (action) given s_t , the observation representing the current generated sequence. At each time step, an immediate reward r_t is also revealed, which is calculated based on the output of the guider network and used to update the sequence generator as described below. The objective of sequence generation is to maximize the total reward as in (3). We detail the components for our proposed model in the following subsections.

3.1. Guider Network as Environment Dynamics

The guider network, implemented as an RNN with LSTM units, is adopted to model environment dynamics to better assist sequence generation. The idea is to train a guider network such that its predicted sequence features at each time step are used to construct intermediate rewards in an RL setting, which in turn are used to optimize the sentence generator. Denote the guider network as $G^\psi(s_{t-1}^G, \mathbf{f}_t)$, with parameters ψ and input arguments $(s_{t-1}^G, \mathbf{f}_t)$ at time t to explicitly write out the dependency on the *guider network* latent state s_{t-1}^G from the previous time step. Here \mathbf{f}_t is the input to the LSTM guider, which represents the feature of the current generated sentence after an encoder network. Specifically, let the current generated sentence be $Y_{1..t}$ (forced to be the same as parts of a training sequence in training), with \mathbf{f}_t calculated as: $\mathbf{f}_t = \text{Enc}(Y_{1..t})$, where $\text{Enc}(\cdot)$ denotes the encoder transformation, implemented with a convolutional neural network (CNN) (Zhang et al., 2017); see Figure 3. The initial state of the guider network is the encoded feature of a true input sequence by the same CNN, i.e., $s_0^G = \text{Enc}(\mathbf{X})$. Importantly, the input to the guider network, at each time point, is defined by features from the entire sequence generated to that point. This provides an important "guide" to the traditional LSTM decoder, accounting for the global properties of the generated text.

Sequence-to-Sequence Generation with Planning We first explain how one uses the guider network to guide next-word generation for the generator (the LSTM decoder in Figure 3). Our framework is inspired by the MPC method (Nagabandi et al., 2017), and can be regarded as a type of plan-ahead attention mechanism. Given the feature \mathbf{f}_t at time t from the current input sequence, the guider network produces a prediction $G^\psi(s_{t-1}^G, \mathbf{f}_t)$ as a future feature representation, by feeding \mathbf{f}_t into the LSTM guider. Since the training of the guider network is based on real data (detailed in the next paragraph), the predicted feature contains global-structure information of the training sequences. To utilize such information to predict the next word, we combine the predicted feature with the output of the decoder by constructing an attention-like mechanism. Specifically, we first apply a linear transformation φ on the predicted feature $G^\psi(s_{t-1}^G, \mathbf{f}_t)$, forming a weight vector $\mathbf{w}_t \triangleq \varphi(G^\psi(s_{t-1}^G, \mathbf{f}_t))$. The weight \mathbf{w}_t is applied to the output \mathbf{O}_t of the LSTM decoder by an element-wise multiplication operation. The result is then fed into a softmax layer to generate the next token y_t . Formally, the generative process is written as

$$\mathbf{O}_t = g(s_{t-1}), \quad (4)$$

$$\mathbf{w}_t = \varphi(G^\psi(s_{t-1}^G, \mathbf{f}_t)), \quad (5)$$

$$y_t \sim \text{Multi}(1, \text{softmax}(\mathbf{O}_t \cdot \mathbf{w}_t)), \quad (6)$$

$$s_t^G = h^G(s_{t-1}^G, \mathbf{f}_t), \quad s_t = h(s_{t-1}, e(y_t)). \quad (7)$$

Guider Network Optimization In training, given a sequence of feature representations $(\mathbf{f}_1, \mathbf{f}_2, \dots, \mathbf{f}_T)$ for a training sentence, we seek to update the guider network such that it is able to predict \mathbf{f}_{t+c} given \mathbf{f}_t , where $c > 0$ is the number of steps looked ahead. We implement this by forcing the predicted feature, $G^\psi(s_t^G, \mathbf{f}_t)$, to match both the sentence feature \mathbf{f}_{t+c} (first term in (8)) and the corresponding feature-changing direction (second term in (8)). This is formalized by maximizing an objective function of the following form at time t :

$$J_G^\psi = \mathcal{D}_{\cos}(\mathbf{f}_{t+c}, G^\psi(s_{t-1}^G, \mathbf{f}_t)) + \mathcal{D}_{\cos}(\mathbf{f}_{t+c} - \mathbf{f}_t, G^\psi(s_{t-1}^G, \mathbf{f}_t) - \mathbf{f}_t), \quad (8)$$

where $\mathcal{D}_{\cos}(\cdot, \cdot)$ denotes the cosine similarity¹. By minimizing (8), an ideal guider network should be able to predict the true next words conditioned on the current word in a sequence. As a result, the prediction is used to construct an intermediate reward, which is then used to update the generator (the LSTM decoder), as described further below.

3.2. Feature-Matching Rewards and Generator Optimization

As in many RL-based sequence-generation methods, such as SeqGAN (Yu et al., 2017) and LeakGAN (Guo et al., 2017), the generator is updated based on policy-gradient methods. As a result, collecting rewards in the generation process is critical. Though (Yu et al., 2017) has proposed to use rollout to get rewards for each generated word, the variance of the rewards is typically too high to be useful practically. In addition, the computational cost may be too expensive for practical use. We describe how to use the proposed guider network to define intermediate rewards below, leading to a definition of feature-matching reward.

Feature-matching rewards We first define an intermediate reward to generate a particular word. The idea is to match the ground-truth features from the CNN encoder in Figure 3 with those generated from the guider network. Equation (8) indicates that the further the generated feature is from the true feature, the smaller the reward should be. To this end, for each time t , we define the intermediate reward for generating the current word as:

$$r_t^g = \frac{1}{2c} \sum_{i=1}^c (\mathcal{D}_{\cos}(\mathbf{f}_t, \hat{\mathbf{f}}_t) + \mathcal{D}_{\cos}(\mathbf{f}_t - \mathbf{f}_{t-i}, \hat{\mathbf{f}}_t - \mathbf{f}_{t-i})), \quad (9)$$

where $\hat{\mathbf{f}}_t = G^\psi(s_{t-c-1}^G, \mathbf{f}_{t-c})$ is the predicted feature. Intuitively, $\mathbf{f}_t - \mathbf{f}_{t-i}$ measures the difference between the generated sequences in feature space; the reward will be high if it matches the predicted feature transition $\hat{\mathbf{f}}_t - \mathbf{f}_{t-i}$ from the guider network. At the last step of sequence generation, i.e., $t = T$, the corresponding reward measures the

¹In our case, cosine similarity works better than l^2 -norm.

quality of the whole generated sequence, thus it is called a final reward. The final reward is defined differently from the intermediate reward, which will be discussed below for both the unconditional- and conditional-generation cases.

Note a token generated at time t will influence not only the rewards received at that time but also the rewards at subsequent time steps. Thus we propose to define the cumulative reward, $\sum_{i=t}^T \gamma^i r_i^g$ with γ a discount factor, as a *feature-matching reward*. Intuitively, this encourages the generator to focus on achieving higher long-term rewards. Finally, in order to apply policy gradient to update the generator, we combine the feature-matching reward with the problem-specific final reward, to form a Q -value reward specified below. We consider unconditional- and conditional-sequence generation in the following.

Unconditional generation This case corresponds to generating sequences from scratch. Similar to SeqGAN, the final reward is defined as the output of a discriminator, evaluating the quality of the whole generated sequence, *i.e.*, the smaller the output, the less likely the generation is a true sequence. As a result, we combine the adversarial reward², denoted as r^f with the feature-matching rewards as follows, to define a Q -value reward:

$$Q_t = \left(\sum \gamma^i r_i^g \right) \times r^f. \quad (10)$$

Conditional generation ^{$i=t$} This case corresponds to generating sequences conditioned on some input features, such as image features in image captioning. Following ideas from self-critical sequence training (SCST) (Rennie et al., 2016), the final reward r_s^f (also called the self-critical reward) is constructed by constituting a baseline reward, denoted as $\hat{r}^f(Y')$, for variance reduction:

$$r_s^f = r^f(Y) - \hat{r}^f(Y'), \quad (11)$$

where $r^f(Y)$ is the reward of a sampled sentence Y by the current generator, and $\hat{r}^f(Y')$ is the reward of a sentence obtained by choosing the words with the highest probabilities at each step t , *i.e.*, a greedy decoding. The Q -value reward is defined as:

$$Q_t = \begin{cases} r_s^f \sum_{i=t}^T \gamma^i r_i^g & \text{if } r_s^f > 0 \\ r_s^f \sum_{i=t}^T \gamma^i (1 - r_i^g) & \text{otherwise} \end{cases} \quad (12)$$

Generator optimization The sequence generator is initialized by pre-training on training sequences with an autoencoder structure, based on MLE training. After that, the final Q -value reward Q_t is used as a reward for each time t , with standard policy gradient optimization methods to update the generator. Specifically, the policy gradient is

$$\begin{aligned} \nabla_{\phi} J &= \mathbb{E}_{(s_{t-1}, y_t) \sim \rho_{\pi}} [Q_t \nabla_{\phi} \log p(y_t | s_{t-1}; \phi, \varphi)] \\ \nabla_{\varphi} J &= \mathbb{E}_{(s_{t-1}, y_t) \sim \rho_{\pi}} [Q_t \nabla_{\varphi} \log p(y_t | s_{t-1}; \phi, \varphi)], \end{aligned}$$

²This reward is given by the discriminator in GAN framework, and $r^f \in [0, 1]$

where $p(y_t | s_{t-1}; \phi, \varphi)$ is the probability of generating y_t given s_{t-1} in the generator.

Discussions For unconditional generation, the feature-matching reward is typically good enough, since the task focuses more on sentence structure, which is well-reflected by the feature-matching reward. For conditional generation, however, a final reward in terms of an evaluation score (*e.g.*, the BLEU score) is more important because the reference information of the conditioned variable is encoded into the score. This final reward thus guides the generator to generate semantically consistent sentence w.r.t. the conditioned variable.

Model-based vs. Model-free Sequence generation seeks to generate the next word (action) given the current (sub-)sequence (state). The generator is considered as an agent and learns a policy to predict the next word given its current state. The state of an RNN is a function of the agent’s true state (which is unknown), but in general does not encode all agent-state information. In previous work (Ranzato et al., 2016), a metric reward is given and the generator is trained to only maximize the metric reward by trial, and falls into the model-free category. In this work, the guider network models the environment dynamics, and is trained by minimizing the cosine similarity between the prediction and the ground truth on real texts. In generator training, it maximizes the reward which is determined by the metric and guider network, and thus falls into the model-based category. To be more specific, our model learns a policy with model-based boosting. The model predictive control scheme is also included in our method, where the guider network is used to help next-word selection at each time-step.

3.3. Other Training Details

Encoder as a feature extractor For unconditional generation, the feature extractor that generates inputs for the guider network shares the CNN part of the encoder. We stop gradients from the guider network to the encoder CNN in the training process. For conditional generation, we use a pre-trained feature extractor, trained similarly to the unconditional generation and fixed later on.

Training procedure As with many RL-based models (Bahdanau et al., 2017; Rennie et al., 2016; Sutskever et al., 2014), warm starting with a pre-trained model is important. Thus we first pre-train the encoder-decoder part based on the training data with an MLE loss. After pre-training, we use RL training to fine-tune the pre-trained generator. We adaptively transfer the training from MLE loss to RL loss, similar to (Paulus et al., 2017; Ranzato et al., 2016).

Initial states We use the same initial state for both the sequence generator and the guider network. For conditional generation, the initial state is the encoded latent code of the conditional information for both training and testing. For unconditional generation, it is the encoded latent code of a

target sequence in training and random noise in testing.

4. Related Work

We first review related works that combine RL and GAN for text generation.

As one of the most representative models in this direction, SeqGAN (Yu et al., 2017) adopts Monte-Carlo search to calculate rewards. However, such a method introduces high variance in policy optimization. There were a number of works proposed subsequently to improve the reward-generation process. For example, RankGAN (Lin et al., 2017) proposes to replace the reward from the GAN discriminator with a ranking-based reward, MaliGAN (Che et al., 2017) modifies the GAN objective and proposes techniques to reduce gradient variance, MaskGAN (Fedus et al., 2018) uses a filling technique to define a Q -value reward for sentence completion, and LeakGAN (Guo et al., 2017) tries to address the sparse-reward issue for long-text generation with hierarchical RL by utilizing the leaked information from a GAN discriminator. One problem of LeakGAN is that it tends to overfit on training data, yielding generated sentences that are often not diverse. By contrast, by relying on a model-based RL approach, our method learns global-structure information, which generates more-diverse sentences, and can be extended to conditional sequence generation.

RL techniques can also be used in other ways for sequence generation. For example, (Ranzato et al., 2016) trains a Seq2Seq model by directly optimizing the BLEU/ROUGE scores with the REINFORCE algorithm. To reduce variance of the vanilla REINFORCE, (Bahdanau et al., 2017) adopts the actor-critic framework for sequence prediction. Furthermore, (Rennie et al., 2016) trains a baseline with a greedy decoding scheme for the REINFORCE method. Note all these methods can only obtain rewards after a whole sentence is generated. Finally, planning techniques in RL have also been explored to improve sequence generation (Gulcehre et al., 2017; Serdyuk et al., 2018). Compared to these related works, the proposed guider network can provide a planning-ahead mechanism and intermediate rewards for RL training. Also, we consider using Q -value as the reward to encourage the generator to focus on long-term rewards.

5. Experiments

We test the proposed framework on unconditional and conditional sequence generation tasks, and analyze the results to understand the performance gained by the guider network. We also perform an ablation investigation on the improvements brought by each part of our proposed method and style transfer in the Supplementary Material (SM). All experiments are conducted on a single Tesla P100 GPU and

implemented with TensorFlow or Theano. Details of the datasets, the experimental setup and model architectures are provided in the SM.

5.1. Unconditional Text Generation

We focus on adversarial text generation, and compare our approach with a number of related works (Guo et al., 2017; Lin et al., 2017; Yu et al., 2017; Zhang et al., 2017; Zhu et al., 2018). In this setting, a discriminator in the GAN framework is added to the model in Figure 3 to guide the generator to generate high-quality sequences. This is implemented by defining the final reward to be the output of the discriminator. All baseline experiments are implemented on the texygen platform (Zhu et al., 2018). We adopt the BLEU score, referenced by test set (test-BLEU, higher value implies better quality) and themselves (self-BLEU, lower value implies better diversity) (Zhu et al., 2018) to evaluate quality of generated samples, where test-BLEU evaluates the reality of generated samples, and self-BLEU measures the diversity. A good generator should achieve both a high test-BLEU score and a low self-BLEU score. In practice, we use $\Delta t = c = 4$. We call the proposed method guider-matching GAN (GMGAN) for unconditional text generation. A detailed description of GMGAN is provided in the SM.

Short Text Generation: COCO Image Captions For this task, we use the COCO Image Captions Dataset (Lin et al., 2014), in which most sentences are of length about 10. Since we consider unconditional text generation, only image captions are used as the training data. After preprocessing, we use 120,000 random sample sentences as the training set, and 10,000 as the test set. The BLEU scores with different methods are listed in Tables 1. We observe that GMGAN performs significantly better than the baseline models. Specifically, besides achieving higher test-BLEU scores, the proposed method can also generate samples with very good diversity in terms of self-BLEU scores. LeakGAN represents the state-of-the-art in adversarial text generation, however, its diversity measurement is relatively poor (Zhu et al., 2018). We suspect the high BLEU score achieved by LeakGAN is due to its mode collapse on some good samples, resulting in high self-BLEU scores. Other baselines achieve lower self-BLEU scores since they cannot generate reasonable sentences.

Long Text Generation: EMNLP2017 WMT Following (Zhu et al., 2018), we use the News section in the EMNLP2017 WMT4 Dataset as our training data. The dataset consists of 646,459 words and 397,726 sentences. After preprocessing, the training dataset contains 5,728 words and 278,686 sentences. The BLEU scores with different methods are provided in Tables 2. Compared with other methods, LeakGAN and GMGAN achieves comparable test-BLEU scores, demonstrating high quality of the

Method	Test-BLEU-2	3	4	5	Self-BLEU-2	3	4
SeqGAN (Yu et al., 2017)	0.820	0.604	0.361	0.211	0.807	0.577	0.278
RankGAN (Lin et al., 2017)	0.852	0.637	0.389	0.248	0.822	0.592	0.230
GSGAN (Kusner et al., 2016)	0.810	0.566	0.335	0.197	0.785	0.522	0.230
TextGAN (Zhang et al., 2017)	0.910	0.728	0.484	0.306	0.806	0.548	0.217
LeakGAN (Guo et al., 2017)	0.922	0.797	0.602	0.416	0.912	0.825	0.689
GMGAN (ours)	0.949	0.823	0.635	0.421	0.746	0.511	0.319

Table 1. Test-BLEU and Self-BLEU scores on Image COCO. (Higher Test-BLEU and lower Self-BLEU is better).

Method	Test-BLEU-2	3	4	5	Self-BLEU-2	3	4
SeqGAN (Yu et al., 2017)	0.630	0.354	0.164	0.087	0.728	0.411	0.139
RankGAN (Lin et al., 2017)	0.723	0.440	0.210	0.107	0.672	0.346	0.119
GSGAN (Kusner et al., 2016)	0.723	0.440	0.210	0.107	0.807	0.680	0.450
TextGAN (Zhang et al., 2017)	0.777	0.529	0.305	0.161	0.806	0.662	0.448
LeakGAN (Guo et al., 2017)	0.923	0.757	0.546	0.335	0.837	0.683	0.513
GMGAN (ours)	0.923	0.727	0.491	0.303	0.814	0.576	0.328

Table 2. Test-BLEU and Self-BLEU scores on EMNLP WMT. (Higher Test-BLEU and lower Self-BLEU is better).

Method	COCO Image Captions	EMNLP2017 WMT News
LeakGAN	(1) A bathroom with a black sink and a white toilet next to a tub. (2) A man throws a Frisbee across the grass covered yard.	(1) "I'm a fan of all the game, I think if that's something that I've not," she said, adding that he would not be decided. (2) The UK is Google's largest non-US market, he has added "20, before the best team is amount of fewer than one or the closest home or two years ago.
GMGAN	(1) Bicycles are parked near a row of large trees near a sidewalk. (2) A married couple posing in front of a piece of birthday cake.	(1) "Sometimes decisions are big, but they're easy to make," he told The Sunday Times in the New Year. (2) A BBC star has been questioned by police on suspicion of sexual assault against a 23-year-old man , it was reported last night.

Table 3. Examples of generated samples with different methods on COCO and EMNLP datasets.

generated sentences. Again, LeakGAN tends to over-fit on training data, leading to much higher (worse) self-BLEU scores. Our proposed GMGAN shows good diversity of long text generation with lower self-BLEU scores. Other baselines obtain both low self-BLEU and test-BLEU scores, leading to more random generations.

Human Evaluation Besides quantitatively evaluating the results using BLEU scores, we also conduct a human evaluation on the WMT News dataset (Caccia et al., 2018). We randomly sample 100 sentences generated by each model. Ten native English speakers on Amazon Mechanical Turk are asked to rate each sentence in a scale from 0 to 5 in terms of their readability. Detailed settings of human evaluation are provided in the SM. The averaged human rating scores are shown in Table 4, indicating GMGAN achieves higher human scores compared to other methods.

Methods	MLE	SeqGAN	RankGAN	GSGAN
Human scores	2.45±0.14	2.57±0.15	2.91±0.17	2.48±0.14
Methods	textGAN	LeakGAN	GMGAN	Real
Human scores	3.11±0.16	3.47±0.15	3.89±0.15	4.21±0.14

Table 4. Results of human evaluation on EMNLP2017 WMT.

As examples, Table 3 illustrates some generated samples by GMGAN and its baselines. The performance on the two datasets indicates that the generated sentences of GM-

GAN are of higher global consistency and better readability than SeqGAN and LeakGAN. More generated examples are provided in the SM.

5.2. Conditional Generation

We conduct experiments on image captioning (Karpathy & Fei-Fei, 2015). We investigate the benefits brought by the proposed method in (12). In image captioning, instead of using a discriminator to define final rewards for generated sentence, we adopt evaluation metrics computed based on human references. The final rewards appear more important as they contain reference (ground-truth) information. Feature-matching rewards work as a regularizer of the final rewards. We call our model in this setting a guider-matching sequence training (GMST) model. An overview of GMST is provided in the SM.

Image Captioning We test our proposed model on the MS COCO dataset (Lin et al., 2014), which contains 123,287 images in total. Each image is annotated with at least 5 captions. Following Karpathy's split (Karpathy & Fei-Fei, 2015), 5,000 images are used for both validation and testing. We report BLEU- k (k from 1 to 4) (Papineni et al., 2002), CIDEr (Vedantam et al., 2015), and METEOR (Banerjee & Lavie, 2005) scores. We consider two settings: (i) using a pre-trained 152-layer ResNet (He

Method	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	CIDEr
Soft Attention (Xu et al., 2015)	70.7	49.2	34.4	24.3	23.9	-
Show & Tell (Vinyals et al., 2015)	-	-	-	27.7	23.7	85.5
SCN-LSTM (Gan et al., 2017)	72.8	56.6	43.3	33.0	25.7	101.2
Top-Down (Anderson et al., 2018)	77.2	—	—	36.2	27.0	113.5
<i>No attention, Greedy, Resnet-152</i>						
AE	69.5	51.7	37.2	26.5	23.1	83.9
SCST (BLEU-4)	71.1	54.8	41.6	31.6	23.1	87.5
GMST (BLEU-4)	71.2	54.8	41.8	32.1	23.4	87.9
SCST (CIDEr)	73.9	56.1	41.2	30.0	24.3	98.6
GMST (CIDEr)	73.8	56.3	41.3	30.3	24.4	100.1
<i>No attention, Greedy, Tag</i>						
AE	70.9	53.6	39.4	28.8	24.4	91.3
AE-g	71.0	53.9	39.6	28.9	24.3	92.8
SCST (BLEU-4)	73.2	57.1	43.9	33.6	24.5	95.9
GMST (BLEU-4)	73.4	57.5	44.3	33.9	24.5	97.1
SCST (CIDEr)	75.8	58.6	43.6	32.1	25.4	105.5
GMST (CIDEr)	76.1	59.0	44.1	32.6	25.5	107.4

Table 5. Results for image captioning on the MS COCO dataset; the higher the better for all metrics.

et al., 2016) for feature extraction, where we take the output of the 2048-way *pool5* layer from ResNet-152, pretrained on the ImageNet dataset; (ii) using semantic tags detected from the image as features (Gan et al., 2017). We use an LSTM with 512 hidden units with mini-batches of size 64. Adam (Kingma & Ba, 2014) is used for optimization, with learning rate 2×10^{-4} . We pretrain the captioning model for the maximum 20 epochs, then use the reinforcement learning to train it for 20 epochs and test on the best model on the validation set. The results are summarized in Table 5. When comparing an AutoEncoder (AE) with a variant implemented by adding a guider network (AE-g), reasonable improvements are observed. We compare the proposed GMST with SCST, one of the state-of-the-art methods. Note the main difference between GMST and SCST is that the former employs our proposed feature-matching reward, while the latter only considers the final reward provided by evaluation metrics. GMST achieves higher scores compared with SCST on its optimized metrics. The gain of GMST compared with SCST comes from the immediate rewards, which can maintain the semantic consistency and sentence structure, preventing language-fluency damage caused by only focusing on evaluation metrics. Comparison of generated examples is provided in the SM.

Style Transfer Our framework naturally provides a way for style transfer, where the guider network plays the role of sentiment selection, and the generator only focus on generating meaningful sentence without considering the sentiments. Details of the experiments settings are provided in SM. We test the proposed framework on the non-parallel text-style-transfer task, where the goal is to transfer one sentence in one style (e.g., positive) to a similar sentence but with a different style (e.g., negative). For a fair comparison, we use the same data and its split method as in (Shen et al., 2017). Specifically, there are 444,000,

63,500, and 127,000 sentences with either positive or negative sentiments in the training, validation and test sets, respectively. To measure whether the original sentences (in the test set) have been transferred to the desired sentiment, we follow the settings of (Shen et al., 2017) and employ a pretrained CNN classifier, which achieves an accuracy of 97.4% on the validation set, to evaluate the transferred sentences. Results are shown in Table 6. It can be observed that our proposed model exhibits higher transfer accuracy, indicating the guider network provides good sentiment guidance for the generator. Other techniques (Hu et al., 2017; Yang et al., 2018; Fu et al., 2018; Prabhumoye et al., 2018) can be applied for further improvement.

Method	Accuracy
VAE (Shen et al., 2017)	23.2%
Cross-align (Shen et al., 2017)	78.4%
CVAE (Hu et al., 2017)	86.5%
Ours	92.7%

Table 6. Sentiment accuracy of transferred sentences.

6. Conclusions

We propose an RL-based method for learning a sequence model, by introducing a guider network to model the generation environment. The guider network provides a plan-ahead mechanism for next-word selection. Furthermore, feature rewards are calculated based on the guider network, to overcome the sparse-reward problem in previous methods; they are used to optimize the generator via policy-gradient method. Our proposed models are validated on both unconditional and conditional sequence generation, including adversarial text generation, image captioning and style transfer. We obtain state-of-the-art results in terms of generation quality and diversity for unconditional generation, and achieve improved performance on several conditional-generation tasks.

References

- Anderson, P., He, X., Buehler, C., Teney, D., Johnson, M., Gould, S., and Zhang, L. Bottom-up and top-down attention for image captioning and visual question answering. In *CVPR*, 2018.
- Bahdanau, D., Brakel, P., Xu, K., and Bengio, Y. An actor-critic algorithm for sequence prediction. In *ICLR*, 2017.
- Banerjee, S. and Lavie, A. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *ACL Workshop*, 2005.
- Caccia, M., Caccia, L., Fedus, W., Larochelle, H., Pineau, J., and Charlin, L. Language gans falling short. *arXiv:1811.02549*, 2018.
- Che, T., Li, Y., Zhang, R., Hjelm, R. D., Li, W., Song, Y., and Bengio, Y. Maximum-likelihood augmented discrete generative adversarial networks. In *arXiv:1702.07983*, 2017.
- Cho, K., Van Merriënboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., and Bengio, Y. Learning phrase representations using rnn encoder-decoder for statistical machine translation. In *EMNLP*, 2014.
- Fedus, W., Goodfellow, I., and Dai, A. M. Maskgan: Better text generation via filling in the _ . *ICLR*, 2018.
- Fu, Z., Tan, X., Peng, N., Zhao, D., and Yan, R. Style transfer in text: Exploration and evaluation. In *AAAI*, 2018.
- Gan, Z., Gan, C., He, X., Pu, Y., Tran, K., Gao, J., Carin, L., and Deng, L. Semantic compositional networks for visual captioning. In *CVPR*, 2017.
- Gu, S., Lillicrap, T., Sutskever, I., and Levine, S. Continuous deep q-learning with model-based acceleration. In *ICML*, 2016.
- Gulcehre, C., Dutil, F., Trischler, A., and Bengio, Y. Plan, attend, generate: Character-level neural machine translation with planning. In *NIPS*, 2017.
- Guo, J., Lu, S., Cai, H., Zhang, W., Yu, Y., and Wang, J. Long text generation via adversarial training with leaked information. In *AAAI*, 2017.
- He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition. In *CVPR*, 2016.
- Hochreiter, S. and Schmidhuber, J. Long short-term memory. *Neural computation*, 1997.
- Hu, Z., Yang, Z., Liang, X., Salakhutdinov, R., and Xing, E. P. Controllable text generation. In *ICML*, 2017.
- Karpathy, A. and Fei-Fei, L. Deep visual-semantic alignments for generating image descriptions. In *CVPR*, 2015.
- Kingma, D. P. and Ba, J. Adam: A method for stochastic optimization. In *ICLR*, 2014.
- Kurutach, T., Clavera, I., Duan, Y., Tamar, A., and Abbeel, P. Model-ensemble trust-region policy optimization. In *ICLR Workshop*, 2018.
- Kusner, M. J., Hernández-Lobato, and Miguel, J. Gans for sequences of discrete elements with the gumbel-softmax distribution. *arXiv preprint arXiv:1611.04051*, 2016.
- Li, P., Bing, L., and Lam, W. Actor-critic based training framework for abstractive summarization. In *arXiv:1803.11070*, 2018.
- Lin, K., Li, D., He, X., Zhang, Z., and Sun, M.-T. Adversarial ranking for language generation. In *NIPS*, 2017.
- Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., and Zitnick, C. L. Microsoft coco: Common objects in context. In *ECCV*, 2014.
- Nagabandi, A., Kahn, G., Fearing, R. S., and Levine, S. Neural network dynamics for model-based deep reinforcement learning with model-free fine-tuning. In *ICRA*, 2017.
- Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. Bleu: a method for automatic evaluation of machine translation. In *ACL*, 2002.
- Pathak, D., Agrawal, P., Efros, A. A., and Darrell, T. Curiosity-driven exploration by self-supervised prediction. In *ICML*, 2017.
- Paulus, R., Xiong, C., and Socher, R. A deep reinforced model for abstractive summarization. In *ICLR*, 2017.
- Prabhumoye, S., Tsvetkov, Y., Salakhutdinov, R., and Black, A. W. Style transfer through back-translation. In *ACL*, 2018.
- Ranzato, M., Chopra, S., Auli, M., and Zaremba, W. Sequence level training with recurrent neural networks. In *ICLR*, 2016.
- Ren, Z., Wang, X., Zhang, N., Lv, X., and Li, L.-J. Deep reinforcement learning-based image captioning with embedding reward. In *CVPR*, 2017.
- Rennie, S. J., Marcheret, E., Mroueh, Y., Ross, J., and Goel, V. Self-critical sequence training for image captioning. In *CVPR*, 2016.
- Rush, A. M., Chopra, S., and Weston, J. A neural attention model for abstractive sentence summarization. *arXiv:1509.00685*, 2015.

- Serdyuk, D., Ke, N. R., Sordoni, A., Trischler, A., Pal, C., and Bengio, Y. Twin networks: Matching the future for sequence generation. In *ICLR*, 2018.
- Shen, T., Lei, T., Barzilay, R., and Jaakkola, T. Style transfer from non-parallel text by cross-alignment. In *NIPS*, 2017.
- Sutskever, I., Vinyals, O., and Le, Q. V. Sequence to sequence learning with neural networks. In *NIPS*, 2014.
- Sutton, R. S. and Barto, A. G. *Reinforcement learning: An introduction*. MIT Press, 1998.
- Vedantam, R., Lawrence Zitnick, C., and Parikh, D. Cider: Consensus-based image description evaluation. In *CVPR*, 2015.
- Vinyals, O., Toshev, A., Bengio, S., and Erhan, D. Show and tell: A neural image caption generator. In *CVPR*, 2015.
- Xu, K., Ba, J., Kiros, R., Cho, K., Courville, A., Salakhudinov, R., Zemel, R., and Bengio, Y. Show, attend and tell: Neural image caption generation with visual attention. In *ICML*, 2015.
- Yang, Z., Hu, Z., Dyer, C., Xing, E. P., and Berg-Kirkpatrick, T. Unsupervised text style transfer using language models as discriminators. In *NeurIPS*, 2018.
- Yu, L., Zhang, W., Wang, J., and Yu, Y. Seqgan: Sequence generative adversarial nets with policy gradient. In *AAAI*, 2017.
- Zhang, Y., Gan, Z., Fan, K., Chen, Z., Henao, R., Shen, D., and Carin, L. Adversarial feature matching for text generation. In *ICML*, 2017.
- Zhu, Y., Lu, S., Zheng, L., Guo, J., Zhang, W., Wang, J., and Yu, Y. Texygen: A benchmarking platform for text generation models. In *SIGIR*, 2018.

A. Extensive Experiments

Ablation Study We conduct ablation studies on long text generation to investigate the improvements brought by each part of our proposed method. First, we test the benefits of using the guider-network information for word selection. Among the methods compared, VAE-g is the standard VAE with the guider network as a pre-trained baseline. We compare RL training with *i)* only final rewards³, *ii)* only feature-matching rewards, and *iii)* combining both rewards, namely GMGAN. The results are shown in Table 7 and Table 8. We observe that guider network plays an important role on improving the baselines. RL training with final rewards given by a discriminator typically damages the generation quality; whereas feature-matching reward produces sentences with much better diversity due to the ability of exploration.

Method	MLE	VAE-g	Final	Feature	GMGAN
BLEU-2	0.761	0.920	0.843	0.914	0.923
BLEU-3	0.468	0.723	0.623	0.704	0.727
BLEU-4	0.230	0.489	0.390	0.457	0.491
BLEU-5	0.116	0.289	0.221	0.276	0.303

Table 7. BLEU scores on EMNLP2017 WMT.

Method	MLE	VAE-g	Final	Feature	GMGAN
BLEU-2	0.664	0.812	0.778	0.798	0.814
BLEU-3	0.338	0.589	0.525	0.563	0.576
BLEU-4	0.113	0.360	0.273	0.331	0.328

Table 8. Self-BLEU scores on EMNLP2017 WMT.

Illustrations of Feature Matching Rewards Figure 2(a) illustrates the feature-matching rewards during the generation, where it shows an example of failure generation at the initial RL-training stage, when two sentences are combined by the word ‘was’. It is grammatically wrong to select ‘was’ for the generator, thus the guider network generates a negative rewards. We can see that the rewards becomes lower with more time steps, which is consistent with the exposure bias. Figure 2(b) shows a successful generation, where the rewards given by the guider are relatively high (usually larger than 0.5). These observations validate that: (i) exposure bias exists in MLE training. (ii) RL training with exploration can help reducing the effects of exposure bias. (iii) Our proposed feature-matching rewards can provide meaningful guidance to maintain sentence structure and fluency.

More Generated Samples of Text Generation Table 12 lists more generated samples on the proposed GMGAN and

³We only use RL training for 200 batches, as the performance keeps dropping with more training time.

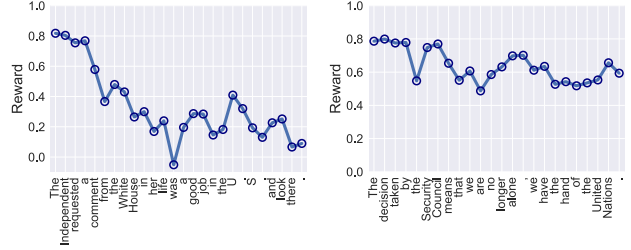


Figure 2. Feature Matching Rewards Illustration.

its baselines. From the experiments, we can see, (i) SeqGAN tends to generate shorter sentences, and the readability and fluency is very poor. (ii) LeakGAN tends to generate very long sentences, and usually longer than the original sentences. However, even with good locality fluency, its sentences usually are not semantically consistent. By contrast, our proposed GMGAN can generate sentences with similar length to the original sentences, and has good readability and fluency. This is also validated in the Human evaluation experiment.

B. Extension to Style Transfer

Our framework naturally provides a way for style transfer, where the guider network plays the role of sentiment selection, and the generator only focus on generating meaningful sentence without considering the sentiments. To make the guider network focus on the guidance of sentiments, we use the label vector l as the initial state s_0^G of the guider network. Especially, at each step t , we feed the current sentence representation f_t and label l into the guider network:

$$\hat{f}_{t+c} = G^{\psi}(s_{t-1}^G, [f_t, l]). \quad (13)$$

For the generator, we put an adversarial regularizer on the encoded latent $s_0(X)$ and penalize it if it contains the sentiment information. Intuitively, the generator gives candidate words represented by O_t , while the guider based on the sentiment information to make choice implicitly by w_t . So the sentiment information are contained in w_t , while the contents of the original sentences are represented by O_t .

Generated Samples of Style Transfer We show more examples of style transfer of our proposed methods in Table 13 and Table 14, which contains 30 sentences of sentiment transfer. As can be seen, our proposed method can maintain most contents of the original sentences after the style transfer.

Table 9. Human evaluation rating criterion.

Scores	Criterion
5 (Best)	It is consistent, informative, grammatically correct.
4	It is grammatically correct and makes sense.
3	It is mostly meaningful and with small grammatical error.
2	It needs some time to understand and has grammatical errors.
1 (Worst)	Meaningless, not readable.

C. Comments on the Guider network

C.1. Final Rewards of Unconditional Sequence Generation

One can adopt many GAN variants to define final rewards. However, it is typically computational expensive and difficult to generalize. For example, it is extremely difficult to train a single discriminator to discriminate the partial generated and real sentences in the generative process. The CNN extracts features of the current sentence, which is an abstraction in the feature space. The LSTM guider network then takes this sequence of features to produce predictions. In unconditional generation, the evaluation metric such as BLEU is not designed to provide reasonable guidance for the generator, since the references are all real sentences. Besides, even with a positive or negative feedback, every word generation is not consistently positive or negative.

C.2. Guider Network and Model-based RL

Guider network can be regarded as a model of the sequence-generation environments, namely the model of dynamics. It takes current s_t and a_t as input, and outputting an estimate of the next state $s_{t+\Delta t}$ at time $t + \Delta t$. In the sequence generation setting, when $\Delta t = 1$, we can exactly get the feature representation of the current generated sentence if the guider does not help the word selection. If not, we cannot exactly get this feature extraction since the guider’s prediction partly determine next token. In practice, we use $\Delta t = c = 4$, to give the guider planning ability, to help for word selection and guide sentence generation.

C.3. Settings of human evaluation

We perform human evaluation using Amazon Mechanical Turk, evaluating the text quality based on readability and meaningfulness (whether sentences make sense). We ask the worker to rate the input sentence with scores scaling from 1 to 5, with criterion listed in Table 9. We require all the workers to be native English speakers, with approval rate higher than 90% and at least 100 assignments completed.

D. Experimental Setup

D.1. Adversarial Text Generation

For Image COCO, the learning rate of the generator is 0.0002, the learning rate of the guider 0.0002, the maximum length of sequence is 25. For WMT, the learning rate of the guider 0.0002, the learning rate of the generator 0.0002, the maximum length of sequence is 50. We use $c=4$ chosen from [2,3,4,5,8] and $\gamma = 0.25$ chosen from [0.1, 0.25, 0.5, 0.75, 0.99]. We use Adam (Kingma & Ba, 2014) optimization algorithm to train the guider, generator and discriminator.

For both tasks, the LSTM state of dimension for the generator is 300, and the LSTM state of dimension for the guider is 300. The dimension of word-embedding is 300. The output dimension of the linear transformation connecting guider and generator is 600×10 . The learning rate of Discriminator is 0.001.

D.2. Conditional Generation

For Image Captioning, the learning rate of the guider 0.0002, the learning rate of the generator 0.0002, the maximum length of sequence is 25. For Style transfer, the learning rate of the guider 0.0001, the learning rate of the generator 0.0001, the maximum length of sequence is 15.

D.3. Network Structure of Models

Table 10. Architecture of Encoder.

(Sub-)sequence to latent features
Input $300 \times \text{Seq. Length Sequences}$
5×300 conv. 300 ReLU, stride 2
5×1 conv. 600 ReLU, stride 2
MLP output 600, ReLU

The LSTM state of dimension for the generator is 300, and the LSTM state of dimension for the guider is 300. The dimension of word-embedding is 300.

Table 11. Architecture of Discriminator.

Sequence to a scalar value
Input $300 \times \text{Seq. Length Sequences}$
5×300 conv. 300 ReLU, stride 2
5×1 conv. 600 ReLU, stride 2
MLP output 1, ReLU

E. Algorithm Overview

Algorithm 1 Guider Matching Generative Adversarial Network (GMGAN)

Require: generator policy π^ϕ ; discriminator D_θ ; guider network G^ψ ; a sequence dataset $\mathcal{S} = \{X_{1...T}\}$.

- 1: Initialize G^ψ , π^ϕ , D^θ with random weights.
 - 2: Pretrain generator π^ϕ , guider G^ψ and discriminator D^θ with MLE loss.
 - 3: **repeat**
 - 4: **for** g-steps **do**
 - 5: Generate a sequence $Y_{1...T} \sim \pi^\phi$.
 - 6: Compute Q_t via (5), and update π^ϕ with policy gradient via (8).
 - 7: **end for**
 - 8: **for** d-steps **do**
 - 9: Generate a sequences from π^ϕ .
 - 10: Train discriminator D_θ .
 - 11: **end for**
 - 12: **until** GMGAN converges
-

Algorithm 2 Guider Matching Sequence Training (GMST)

Require: generator policy π^ϕ ; discriminator D_θ ; guider network G^ψ ; a sequence dataset $\mathcal{S} = \{Y_{1...T}\}$ and its condition information $\mathcal{I} = \{X\}$

- 1: Initialize G^ψ , π^ϕ , D^θ with random weights.
 - 2: Pretrain generator π^ϕ , guider G^ψ and discriminator D^θ with MLE loss.
 - 3: **repeat**
 - 4: Generate a sequence $Y_{1...T} \sim \pi^\phi$.
 - 5: Compute evaluation scores based on references.
 - 6: Compute Q_t^s via (6), and update π^ϕ with policy gradient via (8).
 - 7: **until** GMST converges
-





	<p>Res152-SCST: a group of zebras standing in a field .</p> <p>Res152-GMST: a herd of zebras standing in a field of grass .</p> <p>Tag-SCST: a zebra and a zebra drinking water from a field of grass .</p> <p>Tag-GMST: a group of zebras drinking water in the field of grass .</p>		<p>Res152-SCST: a group of people walking down a skateboard .</p> <p>Res152-GMST: a group of people standing on a street with a skateboard .</p> <p>Tag-SCST: a woman walking down a street with a skateboard .</p> <p>Tag-GMST: a black and white photo of a man riding a skateboard .</p>
	<p>Res152-SCST: a baby sitting next to a baby giraffe .</p> <p>Res152-GMST: a little baby sitting next to a baby holding a teddy bear .</p> <p>Tag-SCST: a black and white photo of a woman holding a teddy bear .</p> <p>Tag-GMST: a black and white photo of a man and a woman holding a teddy bear .</p>		<p>Res152-SCST: a traffic light on a street with a in the .</p> <p>Res152-GMST: a traffic light on the side of a street .</p> <p>Tag-SCST: a traffic light on a street with a green .</p> <p>Tag-GMST: a red traffic light sitting on the side of a road .</p>

Figure 3. Examples of image captioning on MS COCO.

Sequence Generation with a Guider Network

Method	Generated Examples
Real Data	<p>What this group does is to take down various different websites it believes to be criminal and leading to terrorist acts .</p> <p>Over 1 , 600 a day have reached Greece this month , a higher rate than last July when the crisis was already in full swing .</p> <p>" We ' re working through a legacy period , with legacy products that are 10 or 20 years old , " he says .</p> <p>' The first time anyone says you need help , I ' m on the defensive , but that ' s all that I know .</p> <p>Out of those who came last year , 69 per cent were men , 18 per cent were children and just 13 per cent were women .</p> <p>He has not played for Tottenham ' s first team since and it is now nearly two years since he completed a full Premier League match for the club .</p> <p>So you have this man who seems to represent this way to live and how to be a good citizen of the world .</p> <p>CNN : You made that promise , but it wasn ' t until 45 years later that you acted on it .</p> <p>This is a part of the population that is notorious for its lack of interest in actually showing up when the political process takes place .</p> <p>They picked him off three times and kept him out of the end zone in a 22 - 6 victory at Arizona in 2013 .</p> <p>The treatment was going to cost £ 12 , 000 , but it was worth it for the chance to be a mum .</p> <p>But if black political power is so important , why hasn ' t it made more of a difference in the lives of poor black people in Baltimore such as Gray ?</p> <p>Local media reported the group were not looking to hurt anybody , but they would not rule out violence if police tried to remove them .</p> <p>The idea was that couples got six months ' leave per child with each parent entitled to half the days each .</p> <p>The 55 to 43 vote was largely split down party lines and fell short of the 60 votes needed for the bill to advance .</p> <p>Taiwan ' s Defence Ministry said it was " aware of the information , " and declined further immediate comment , Reuters reported .</p> <p>I ' m racing against a guy who I lost a medal to - but am I ever going to get that medal back ?</p> <p>Others pushed back their trips , meaning flights early this week are likely to be even more packed than usual .</p> <p>" In theory there ' s a lot to like , " Clinton said , " but ' in theory ' isn ' t enough .</p> <p>If he makes it to the next election he ' ll lose , but the other three would have lost just as much .</p>
SeqGAN	<p>Following the few other research and asked for " based on the store to protect older , nor this .</p> <p>But there , nor believe that it has reached a the person to know what never - he needed .</p> <p>The trump administration later felt the alarm was a their doctors are given .</p> <p>We have been the time of single things what people do not need to get careful with too hurt after wells then .</p> <p>If he was waited same out the group of fewer friends a more injured work under it .</p> <p>It will access like the going on an " go back there and believe .</p> <p>Premier as well as color looking to put back on a his is .</p> <p>So , even though : " don ' t want to understand it at an opportunity for our work .</p> <p>I was shocked , nor don ' t know if mate , don ' t have survived ,</p> <p>So one point like ten years old , but a sure , nor with myself more people substantial .</p> <p>And if an way of shoes of crimes the processes need to run the billionaire .</p> <p>Now that their people had trained and people the children live an actor , nor what trump had .</p> <p>However , heavily she been told at about four during an innocent person .</p>
LeakGAN	<p>The country has a reputation for cheap medical costs and high - attack on a oil for more than to higher its - wage increase to increase access to the UK the UK women from the UK ' s third nuclear in the last couple of weeks .</p> <p>I ' ve been watching it through , and when the most important time it is going to be so important .</p> <p>I ' m hopeful that as that process moves along , that the U . S . Attorney will share as much as far as possible .</p> <p>The main thing for should go in with the new contract , so the rest of the Premier League is there to grow up and be there , " she said .</p> <p>I think the main reason for their sudden is however , I didn ' t get any big thing , " he says , who is the whole problem on the U . S . Supreme Court and rule had any broken .</p> <p>The average age of Saudi citizens is still very potential for the next year in the past year , over the last year he realised he has had his massive and family and home .</p> <p>" I think Ted is under a lot of people really want a " and then the opportunity to put on life for security for them to try and keep up .</p> <p>The new website , set to launch March 1 , but the U . S is to give up the time the case can lead to a more than three months of three months to be new home .</p> <p>It ' s a pub ; though it was going to be that , but , not , but I am not the right thing to live , " she said .</p> <p>" I ' m not saying method writing is the only way to get in the bedroom to get through the season and we ' ll be over again , " he says .</p> <p>I ' m not suggesting that our jobs or our love our years because I have a couple of games where I want it to be .</p> <p>The German government said 31 suspects were briefly detained for questioning after the New Year ' s Eve trouble , among them not allowed to stay in the long - term .</p> <p>It was a punishment carried out by experts in violence , and it was hard to me he loved the man and he ' s got off to support me in the future .</p> <p>" I ' ve known him , all that just over the last two weeks and for the last 10 years , I ' ll have one day of my life , " she said .</p> <p>The main idea behind my health and I think we saw in work of our country was in big fourth - up come up with a little you ' ve ever .</p> <p>he Kings had needed scoring from the left side , too , and King has provided that since his return are the of the first three quarters of the game .</p> <p>The average number of monthly passengers arriving at the University of January 1 . 1 million people and another average visit men were on the year .</p> <p>It ' s going to be a good test for us and we are on the right way to be able to get through it on every day on the year .</p>
GMGAN	<p>" I ' m actually going to take my baby shopping and get him go back and never let them go into the property .</p> <p>The ban will continue in several European countries that were occupied by Nazi Germany , including Austria and the Netherlands .</p> <p>But 2015 was also a year when people again took to the streets to protest corruption - people across the globe sent a strong signal to those in power : it is .</p> <p>But the best advice , especially if you ' re starting out , you can get a feel that works for them sure .</p> <p>She said : " To those for those who were in music , time you never say to live here .</p> <p>We ' re certainly going to have to prepare and coach the team a lot better than we did that night .</p> <p>But the benefits from the UK - are being prepared to have used in Australian middle for alcohol consumption and six weeks or less than the year in easier times .</p> <p>It ' s a very well - I ' ve really had and it makes me feel like a lot of beach for what I ' m doing , " she says .</p> <p>We ' re creating the space for them to think about what their choices are , because at the end of the day , the players will be OK .</p> <p>He said E . On customers could wait for the small reduction in their bill or shop around and save more than £ 300 a year .</p> <p>" I ' m actually going to take my baby shopping and get him go back and never let them go into the property .</p> <p>' I was actually surprised that he got some other as a big that I would be as a big storm , " she said .</p> <p>He returned to work on his father ' s maintenance crews - while his head is about to the Democratic nomination , a 15 - year - old who did not learn as a way to everything play behind responsible for the guns , they sell with us later .</p> <p>But the eye of the storm was China , where the main index in Shanghai lost 19 % of its value in the same period .</p> <p>The capital could get 15 - 20 inches , Philadelphia could see 12 to 18 and New York City and Long Island could get 8 to 10 .</p> <p>A Virginia couple was surprised after receiving a letter their son sent almost 11 years ago while serving in Iraq .</p> <p>I ' m looking forward to going to battle with those guys all year long and for the rest of our careers . A new ISIS video experience just - aged after the three - year period , and two of three inmates - are thought strongly , the U . N ? and campaign .</p> <p>" I ' m well aware that little ones can get into trouble and well , " Ms Turner ' s business .</p> <p>It ' s a wonder the producers of this year ' s show did not sign up someone from the world of sport .</p>

Table 12. Generated Examples on EMNLP2017 WMT.

Original:	i 'm so lucky to have found this place !
Transferred:	i 'm so embarrassed that i picked this place .
Original:	awesome place , very friendly staff and the food is great !
Transferred:	disgusting place , horrible staff and extremely rude customer service .
Original:	this was my first time trying thai food and the waitress was amazing !
Transferred:	this was my first experience with the restaurant and we were absolutely disappointed .
Original:	thanks to this place !
Transferred:	sorry but this place is horrible .
Original:	the staff was warm and friendly .
Transferred:	the staff was slow and rude .
Original:	great place and huge store .
Transferred:	horrible place like ass screw .
Original:	the service is friendly and quick especially if you sit in the bar .
Transferred:	the customer service is like ok - definitely a reason for never go back ..
Original:	everything is always delicious and the staff is wonderful .
Transferred:	everything is always awful and their service is amazing .
Original:	best place to have lunch and or dinner .
Transferred:	worst place i have ever eaten .
Original:	best restaurant in the world !
Transferred:	worst dining experience ever !
Original:	you 'll be back !
Transferred:	you 're very disappointed !
Original:	you will be well cared for here !
Transferred:	you will not be back to spend your money .
Original:	they were delicious !
Transferred:	they were overcooked .
Original:	seriously the best service i 've ever had .
Transferred:	seriously the worst service i 've ever experienced .
Original:	it 's delicious !
Transferred:	it 's awful .

Table 13. Sentiment transfer samples on Yelp dataset (positive → negative).

Original:	gross !
Transferred:	amazing !
Original:	the place is worn out .
Transferred:	the place is wonderful .
Original:	very bland taste .
Transferred:	very fresh .
Original:	terrible service !
Transferred:	great customer service !
Original:	this place totally sucks .
Transferred:	this place is phenomenal .
Original:	this was bad experience from the start .
Transferred:	the food here was amazing good .
Original:	very rude lady for testing my integrity .
Transferred:	very nice atmosphere for an amazing lunch !
Original:	they recently renovated rooms but should have renovated management and staff .
Transferred:	great management and the staff is friendly and helpful .
Original:	this store is not a good example of sprint customer service though .
Transferred:	this store is always good , consistent and they 're friendly .
Original:	one of my least favorite ross locations .
Transferred:	one of my favorite spots .
Original:	horrible in attentive staff .
Transferred:	great front desk staff !
Original:	the dining area looked like a hotel meeting room .
Transferred:	the dining area is nice and cool .
Original:	never ever try to sell your car at co part !
Transferred:	highly recommend to everyone and recommend this spot for me !
Original:	i ordered the filet mignon and it was not impressive at all .
Transferred:	i had the lamb and it was so good .

Table 14. Sentiment transfer samples on Yelp dataset (negative \rightarrow positive).