

TreeGAN: Syntax-Aware Sequence Generation with Generative Adversarial Networks

Xinyue Liu

Worcester Polytechnic Institute
xliu4@wpi.edu

Xiangnan Kong

Worcester Polytechnic Institute
xkong@wpi.edu

Lei Liu

Apple
magieliulei@gmail.com

Kuorong Chiang

Huawei
kuorong.chiang@gmail.com

Abstract—Generative Adversarial Networks (GANs) have shown great capacity on image generation, in which a discriminative model guides the training of a generative model to construct images that resemble real images. Recently, GANs have been extended from generating images to generating sequences (e.g., poems, music and codes). Existing GANs on sequence generation mainly focus on general sequences, which are grammar-free. In many real-world applications, however, we need to generate sequences in a formal language with the constraint of its corresponding grammar. For example, to test the performance of a database, one may want to generate a collection of SQL queries, which are not only similar to the queries of real users, but also follow the SQL syntax of the target database. Generating such sequences is highly challenging because both the generator and discriminator of GANs need to consider the structure of the sequences and the given grammar in the formal language. To address these issues, we study the problem of syntax-aware sequence generation with GANs, in which a collection of real sequences and a set of pre-defined grammatical rules are given to both discriminator and generator. We propose a novel GAN framework, namely TreeGAN, to incorporate a given Context-Free Grammar (CFG) into the sequence generation process. In TreeGAN, the generator employs a recurrent neural network (RNN) to construct a parse tree. Each generated parse tree can then be translated to a valid sequence of the given grammar. The discriminator uses a tree-structured RNN to distinguish the generated trees from real trees. We show that TreeGAN can generate sequences for any CFG and its generation fully conforms with the given syntax. Experiments on synthetic and real data sets demonstrated that TreeGAN significantly improves the quality of the sequence generation in context-free languages.

Index Terms—Generative Adversarial Networks, GANs, Tree Generation, Sequence Generation, Context-Free Language

I. INTRODUCTION

Generative Adversarial Network (GAN) is an unsupervised learning framework that consists of a generative network and a discriminative network. We called them the generator (G) and the discriminator (D) respectively. D learns to distinguish whether a data instance is from real world or synthetic. G attempts to confuse D by producing high-quality synthetic instances. D and G in a GAN framework are trained against each other iteratively until they reach the Nash equilibrium. A well-trained GAN yields a generator that is capable of producing high quality data instances that look like real ones.

Inspired by the enormous success in image generation and related fields, GANs [1] have recently been extended to sequence generation tasks [2, 3]. GANs for sequence generation

have many important applications in real world. For instance, in order to build a good query optimizer for a database, researcher may want to generate a large amount of high quality synthetic SQL queries to benchmark the optimizer. Unlike image generation tasks, most languages have their inherent grammar or syntax. Existing GAN models [2, 3, 7] for sequence generation mainly focus on grammar-free settings as illustrated in Figure 1a. These methods attempt to learn the complex underlying syntax and grammatical pattern from the data, which is usually highly challenging and requires a large amount of real data samples to achieve a reasonable performance. In many formal languages, the grammatical rules or syntax (e.g., SQL syntax, Python PL syntax) are pre-defined. Incorporating such syntax in GAN training should yield a better sequence generator with syntax-awareness and significantly reduce the searching space during the training phase. Existing syntax aware sequence generation models [4] are mainly trained via maximum likelihood estimation (MLE), which highly relies on the quality and quantity of the real data samples. Some studies [2, 5] show that the adversarial training could further improve the generation performance based on MLE. Even though the existing syntax-aware generation methods incorporate the grammatical information, the generation could be suboptimal.

To tackle above issues, we study the problem of sequence generation under a pre-defined grammar using GANs. We illustrate this problem setting in Figure. 1b, in which a corpus of real sequences (top left box) and a set of grammatical rules (top right box) are given as the input. The goal is to learn a generative net G that can construct high-quality sequences following the given grammar while resembling the real sequences via adversarial training. We focus on Context-Free Grammars (CFGs) according to the well-known Chomsky hierarchy [8], which can apply to many existing formal languages. A formal definition of CFGs are provided in Section II-C. To the best of our knowledge, we make the very first effort to build a syntax aware GAN for sequence generation.

Although GANs have been successfully applied on many tasks, learning such a syntax-aware generative network is not an easy task, which has several challenges:

- **Guarantee the syntax correctness:** The difficulty of ensuring the syntax validity lies in the nature of sequence generator: it generates tokens one by one in a sequential

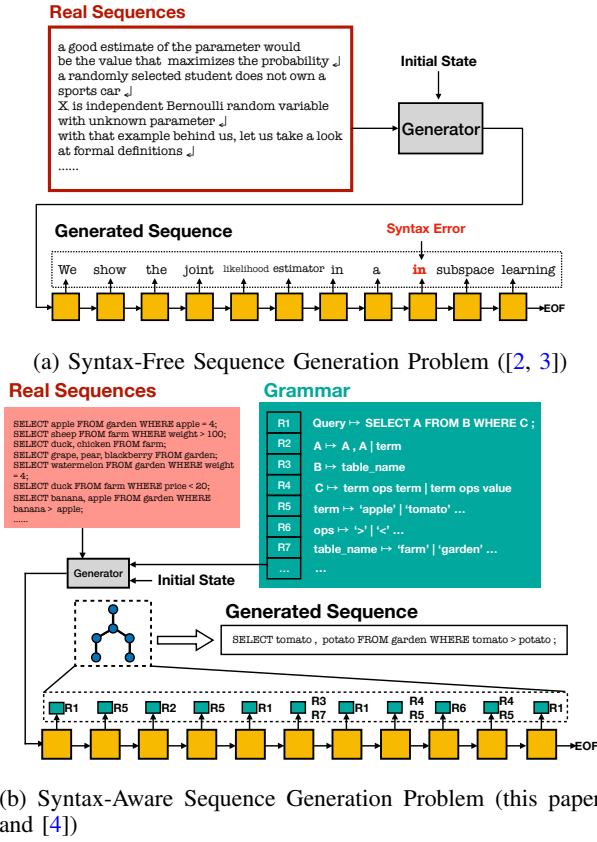


Fig. 1: Comparison of two problem settings. (a) Syntax-free sequence generation problem. Only a set of real sequences are used for training the generator, and the generated sequence may exhibit syntax error. (b) Syntax-aware sequence generation problem. Besides a set of real sequences (top left box), a set of syntax rules are given as the prior knowledge (top right boxes, e.g., “ $A \mapsto A, A$ ” and “ $B \mapsto \text{table_name}$ ”). At each step, the generator follows one or multiple pre-defined rules (the small boxes in the middle of output arrows, “R1”, “R2”, etc.) to construct a sequence (dashed box) that resembles the real sequences and follow the grammar.

order. Most syntax models employ a top-down structure like trees to abstract the grammatical information. To fully achieve the syntax awareness, the sequence generator have to follow a certain grammatical tree structure. However, the structure of grammatical trees can vary a lot, it is impossible for a sequence generator to cover all the possibilities.

- **Tracking the syntax state of incomplete phrase:** RNN is usually used as the generator in sequence generation, which stores a summary of the generated tokens in its hidden state at each step. However, such summary does not keep track of the syntax information in the partially generated sequence, which leads to possible syntax errors in the entire sequence. To build a syntax-aware generator, we need a mechanism that enables RNN to store full syntax information and track the state while generating sequences.

- **Syntax-aware discriminator:** Discriminator is a crucial component of a GAN framework, and should be designed specifically based on the nature of studied task. DCGANs [5] employs Convolutional Neural Networks (CNN) [9] as the discriminator to achieve better performance on image representation and generation, while MaskGAN [6] uses LSTM [10] as the discriminator to train a sequence generator that fills in missing text. In our problem, simply using LSTM or CNN as the discriminative model could miss critical grammatical pattern, which makes the GAN framework yields weak generator. Hence, a tailored discriminative model should be designed carefully for syntax-aware sequence generation task to encode the rich grammar information of the sequences properly, and to guide the generator to better capture the underlying syntax pattern.
- **Pre-training:** A proper pre-training is usually required for both generator and discriminator. However, it is unclear that how to design a suitable pre-training strategy for the syntax-aware sequence generation task since we are the first to investigate this problem using GANs.

To tackle above challenges, we propose a novel GAN model called TreeGAN. Instead of generating sequences directly, TreeGAN absorbs a set of grammatical rules and learns to generate parse trees. Each generated tree corresponds to a sequence that is valid according to the given grammar. This approach imposes hard restrictions on the generator, and the syntax correctness of generated sequences is guaranteed. We show how these restrictions can be applied in Section III-B. Consequently, the vanilla RNN/LSTM is no longer the optimal choice for the discriminator since the generator of TreeGAN is generating trees instead of plain sequences. To better distinguish the fake parse trees from real parse trees, we use TreeLSTM [11] to guide the tree generator during the adversarial training, the details are presented in Sec. III-C. The corresponding pre-training strategies are discussed in Sec. III-D. The contributions of this work are summarized as follows,

- We transform the sequence generation problem into the parse tree generation task to effectively incorporate the structural information. We show that each sequence under a CFG could be translated to a corresponding parse tree, which is used in the proposed TreeGAN to guide the generator producing real look parse trees.
- We propose a tree generator that employs LSTM to generating parse trees that follow a pre-defined context-free grammar.
- We propose an adversarial training framework called TreeGAN, in which a Tree-structured LSTM model [11] is used as the discriminator to guide the tree generator constructing parse trees.
- Extensive experiments performed on synthetic data sets and real data sets demonstrate that the proposed TreeGAN framework can produce high-quality texts/sequences follow the pre-defined context-free grammar.

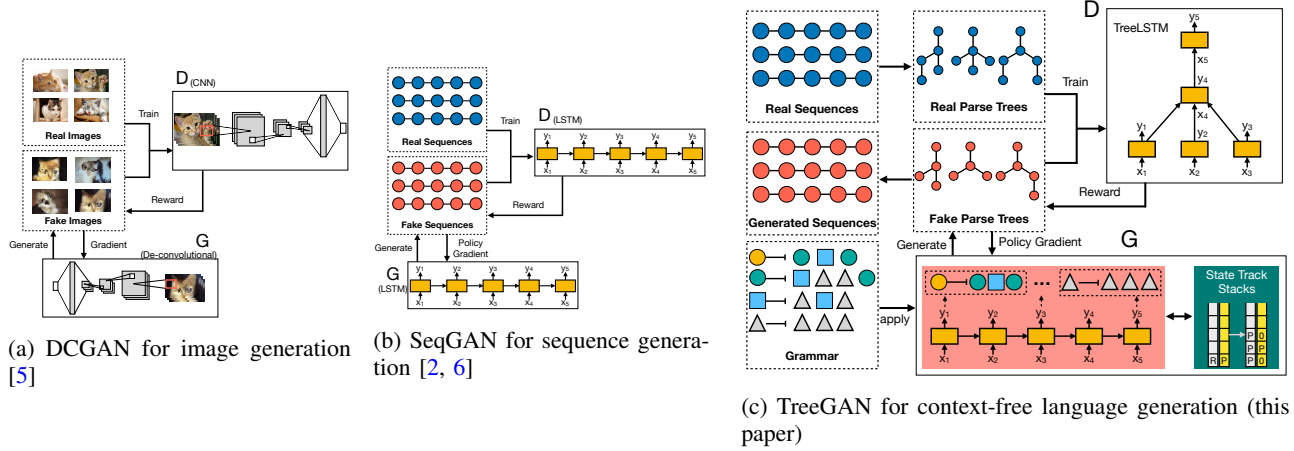


Fig. 2: The comparison of related GAN models. “D” represents the discriminator of the GAN model and “G” denotes the generator. (a) DCGAN [5]; (b) SeqGAN [2] and MaskGAN [6]; (c) TreeGAN (this paper).

TABLE I: Summary of Notations.

Symbol	Definition
\mathbb{G}	a context free grammar (CFG)
\mathcal{V}	the set of non-terminal variables in a CFG
\mathcal{T}	the set of terminal tokens in a CFG
\mathcal{P}	the set of production rule in a CFG
$P \in \mathcal{P}$	the production rule
S	the start token in a CFG
G_θ	a generator that parametrized by θ
D_ϕ	a discriminator that parametrized by ϕ
\mathbf{x}	the input feature vector
\mathbf{h}	the hidden state vector
\mathbf{W}	the weight matrix for the input feature
\mathbf{U}	the weight matrix for the hidden state
\mathbf{b}	the bias vector
i	the input gate of LSTM
f	the forget gate of LSTM
o	the output gate of LSTM
\mathbf{u}	the memory cell of LSTM before input gate
\mathbf{c}	the memory cell of LSTM after input gate
Ψ	the probability output after fully connected layer
\mathbf{M}	the mask matrix of TreeGAN
Ω	the stack of TreeGAN

The rest of this paper is organized as follows. We compare our work with the related work in Section VI. We set the problem formulations in Section II. We show how to solve the proposed problems in Section III. The experimental results for both synthetic data and real data are shown in Section IV. Then we conclude the paper in Section VII.

II. PROBLEM FORMULATION

A. Notation

Throughout this paper, we use capital alphabet in boldface, e.g. \mathbf{X} , to denote a matrix, and x_{ij} refers to the entry of \mathbf{X} at i -th row and j -th column. We use lowercase alphabet in boldface, e.g. \mathbf{x} , to denote a column-based vector, and x_i refers to the i -th entry of \mathbf{x} . We use calligraphic letters to denote sets, e.g. $\mathcal{A}, \mathcal{B}, \mathcal{C}$. The important notations used in this paper are summarized in Table I.

B. Syntax Aware Sequence Generation

The syntax-aware sequence generation problem is defined as follows.

Definition II.1. Given a dataset of real-world structured sequences $\mathcal{X} = \{X_1, \dots, X_N\}$, where all $X_n \in \mathcal{X}$ follows a grammar \mathbb{G} , train a θ -parameterized generative net G_θ to construct a sequence $Y_{1:T} = (y_1, \dots, y_T)$ with $y_t \in \mathcal{V}$, where \mathcal{V} is the set of vocabulary of tokens.

C. Grammar

In this paper, we study the sequence generation problem in context-free grammars (CFGs), which is formulated in the well-known Chomsky hierarchy [8]. CFGs can apply to many existing formal languages, such as palindrome and SQL. A CFG is formally defined as $\mathbb{G} = (\mathcal{V}, \mathcal{T}, \mathcal{P}, S)$, where \mathcal{V} is a set of non-terminal variables, $\mathcal{T} = \mathcal{V} \cup \{\epsilon\}$ the set of terminal variables¹, \mathcal{P} the set of production rules, and $S \in \mathcal{V}$ the start symbol. Each production rule $P \in \mathcal{P}$ follows the form:

$$\mathcal{V} \mapsto (\mathcal{T} \cup \mathcal{V})^+ \quad (1)$$

For example, the context free grammar defines palindromes of 0s and 1s are $\mathbb{G}_{pal} = (\{P\}, \{0, 1, \epsilon\}, \mathcal{A}, P)$, where \mathcal{A} consists of production rules: $\{P \mapsto \epsilon, P \mapsto 0, P \mapsto 1, P \mapsto 0P0, P \mapsto 1P1\}$. Accordingly, the palindrome “010010” could be derived by applying following procedures sequentially:

Step 1	$P \mapsto 0P0$	$[P \mapsto 0P0]$
Step 2	$0P0 \mapsto 01P10$	$[P \mapsto 1P1]$
Step 3	$01P10 \mapsto 010P010$	$[P \mapsto 0P0]$
Step 4	$010P010 \mapsto 010010$	$[P \mapsto \epsilon]$

¹ ϵ denotes the empty token, alternatively it can be considered as a special symbol that not included in the set of terminal variables.

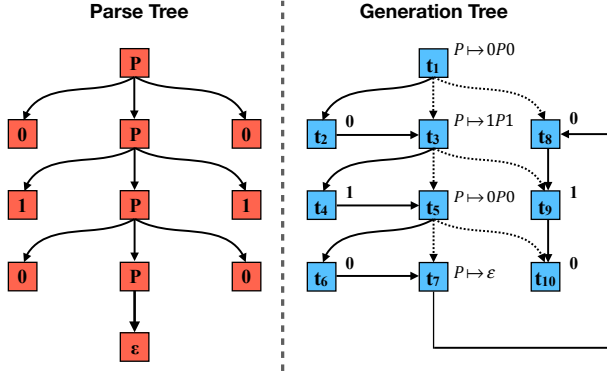


Fig. 3: **Left:** the parse tree for sequence “010010”. **Right:** the action sequence used to generate the parse tree shown on the left. The solid arrow denotes the chronological order of the action flow, and the dashed arrow denotes the input of parent embedding (see Sec. III-B).

D. Parse Tree

For each derivation of a CFG sequence, there is a corresponding tree representation called parse tree. The parse tree for any sequence follow context free grammar $\mathbb{G} = (\mathcal{V}, \mathcal{T}, \mathcal{P}, S)$ are trees with following properties:

- 1) The root node is labeled by S .
- 2) The interior node is labeled by a variable in \mathcal{V} .
- 3) Every leaf is labeled by a terminal in \mathcal{T} .
- 4) If a node labeled A , and its children are labeled N_1, \dots, N_k from left to right. Then $A \mapsto N_1, \dots, N_k$ is a production rule in \mathcal{P} .

If we concatenate leaves of a parse tree from left to right and top to bottom, we obtain a *yield* of the tree, which is equivalent to the string derived from the root variable. The parse tree of palindrome sequence 010010 is illustrated as on L.H.S. in Figure 3. If we concatenate the leaves of the parse tree shown in Figure 3, we can obtain the sequence “010ε010”, which is equivalent to “010010” since ϵ refers to the empty token.

Theorem 1. Let $\mathbb{G} = (\mathcal{V}, \mathcal{T}, \mathcal{P}, S)$ be a CFG. If a sequence Y can be derived using the production rules from \mathcal{P} and the derivation starts with S , then there is a parse tree with root S that yields Y .

Proof. It is equivalent to the proof of Theorem 5.12 in [12]. \square

Lemma 1. If sequence X follows a context free grammar $\mathbb{G} = (\mathcal{V}, \mathcal{T}, \mathcal{P}, S)$, there is a sequence of productions $Z = (P_1, \dots, P_k)$ that derives Y , where $P_1, \dots, P_k \in \mathcal{P}$. Such mapping can be denoted as $Z \Leftrightarrow X$.

Proof. From Theorem 1 we know there exists a parse tree Q yields Y , traversing Q via a depth-first search order yields a sequence of productions Z that derives Y . \square

Given Lemma 1, we can find a set of production sequences $\mathcal{D} = \{D_1, \dots, D_N\}$ for $\mathcal{X} = \{X_1, \dots, X_N\}$, where $D_1 \Leftrightarrow$

$X_1, \dots, D_N \Leftrightarrow X_N$. How to parse each X_n into D_n is out of the scope of this paper and will not be discussed here.

Now we can transform the original syntax-aware sequence generation problem defined in Section II-B into a parse tree generation problem.

Definition II.2 (Parse Tree Generation Problem). Given a CFG defined as $\mathbb{G} = (\mathcal{V}, \mathcal{T}, \mathcal{P}, S)$, and $\mathcal{D} = \{D_1, \dots, D_N\}$ where all the production rules in $\{Z_1, \dots, Z_N\}$ are from \mathcal{P} , the goal is to train a θ -parameterized generative net G_θ to construct a sequence $Z_{1:T} = (P_1, \dots, P_T)$ with $P_t \in \mathcal{P}$.

Additionally, we also train a ϕ -parameterized discriminative net D_ϕ to guide G_θ to improve the generating quality. Specifically, $D_\phi(Z)$ is a probability indicating how likely Z is a real data sample.

III. METHODOLOGY

In this section, we introduce the technical details of our proposed method TreeGAN. In section III-A, We first briefly review the key components of conventional GANs, including its objective function and the optimization approach. Then we present the detailed design of the tree generator of TreeGAN in section III-B, we show how could the generator keep a lossless track information of the syntax state while generating the sequence. Section III-C presents the tailored discriminator we used for TreeGAN. Finally, in section III-D we introduce our pre-training strategy, which gives the adversarial training phase a better start point.

A. Generative Adversarial Network

GAN[1] aims to obtain the equilibrium of the following optimization objective

$$\mathcal{L}(\theta, \phi) = -\mathbb{E}_{X \sim p_x} \log D_\phi(X) - \mathbb{E}_{Y \sim G_\theta} \log (1 - D_\phi(Y)) \quad (2)$$

where \mathcal{L} is minimized w.r.t. D_ϕ and is maximized w.r.t. G_θ . X are sampled from the real-data distribution p_x . Since the first term of Eq. (2) does not depend on G_θ , we only need to consider the second term when training the generator. However, applying GAN on sequence data has a problem: the gradient of loss from D_ϕ w.r.t the output of G_θ is not meaningful for discrete tokens [1, 2]. Thus, we follow the approach proposed in SeqGAN[2] to use the policy gradient[13] to guide the learning of G_θ . The reward of G_θ when given a start state s_0 is :

$$\mathcal{J}(\theta) = \mathbb{E}_{Y \sim G_\theta} \log \left(G_\theta(y_1 | s_0) \prod_{t=2}^T G_\theta(y_t | Y_{1:t-1}) \right) R(Y_{1:T}), \quad (3)$$

where $R(\cdot)$ is the reward function for a generated sequence, here we consider the estimated probability of being real by the discriminative net D_ϕ as the reward. Formally it is defined as

$$R(Y_{1:T}) = D_\phi(Y_{1:T}) \quad (4)$$

Hence, for sequence generation task, the objective of training the discriminative net is $\arg \min_\phi \mathcal{L}(\theta, \phi)$, where θ is

fixed. And the objective of training the generative net is $\arg \min_{\theta} \mathcal{J}(\theta)$.

B. Tree Generator

Inspired by the model proposed in [4], we consider the tree generation problem as generating a sequence of actions. The actions can be categorized into two types, which are (1) the production rules as defined in Eq. 1 and (2) the terminal tokens in \mathcal{V} . The R.H.S. of Figure 3 illustrates the generation process of the parse tree on L.H.S of Figure 3. Each node in R.H.S. of Figure 3 refers to an action and actions are connected by solid arrows that indicate the chronological order of them. The generation proceeds in depth-first, left-to-right order. Thus, in order to generate the parse tree shown in Figure 3, the tree generator G_{θ} produces the following actions sequentially,

$$P \mapsto 0P0, 0, P \mapsto 1P1, 1, P \mapsto 0P0, 0, P \mapsto \epsilon, 0, 1, 0$$

G_{θ} starts from the root node at step t_1 and proceeds by choosing different production rules to expand the tree, and at leaves, the model generates terminal tokens to close the tree branches.

We employ a vanilla LSTM to implement our tree generator:

$$\begin{aligned} i_t &= \sigma(\mathbf{W}^{(i)}x_t + \mathbf{U}^{(i)}h_{t-1} + \mathbf{b}^{(i)}), \\ f_t &= \sigma(\mathbf{W}^{(f)}x_t + \mathbf{U}^{(f)}h_{t-1} + \mathbf{b}^{(f)}), \\ o_t &= \sigma(\mathbf{W}^{(o)}x_t + \mathbf{U}^{(o)}h_{t-1} + \mathbf{b}^{(o)}), \\ u_t &= \tanh(\mathbf{W}^{(u)}x_t + \mathbf{U}^{(u)}h_{t-1} + \mathbf{b}^{(u)}), \\ c_t &= i_t \odot u_t + f_t \odot c_{t-1}, \\ h_t &= o_t \odot \tanh(c_t), \end{aligned} \quad (5)$$

where, i_t, f_t, o_t, c_t, h_t are the input gate, the forget gate, the output gate, the memory cell and the hidden state at time step t respectively. u_t is the memory cell before input gate at step t , and \odot denotes the element-wise multiplication.

For a data sample $D = (d_1, \dots, d_T)$, the input vector at time step t is $x_t = (a_{t-1}, p_t)$, where a_{t-1} is the action embedding vector for d_{t-1} and p_t is the parent embedding vector for d_t .

Action Embedding: Two action embedding matrices $\mathbf{W}^{(P)}$ and $\mathbf{W}^{(V)}$ are initialized before train the generator G_{θ} . Each row in $\mathbf{W}^{(P)}$ ($\mathbf{W}^{(V)}$) corresponds to an embedding vector for an action of production rules (terminal tokens).

Parent Embedding: The tree generator uses the parent feeding illustrated on R.H.S. in Figure 3 to inherit the information encoded in the parent action along the generation tree. As shown in Figure 3, when generating action at t_5 , the embedding of its parent action at t_3 will be used. The parent action step $p(t)$ is formally defined as the time step at which the action node at time step t is initiated. Specifically, in Figure 3, the action node at time step t_2, t_3, t_9 are all initiated at t_1 when G_{θ} generates the production $P \mapsto 0P0$. In this case, $p(t_2) = p(t_3) = p(t_9) = t_1$.

Generation State Tracking: As we discussed in Section I, the conventional RNN stores lossy summarization in its hidden state h , which only contains incomplete syntax information

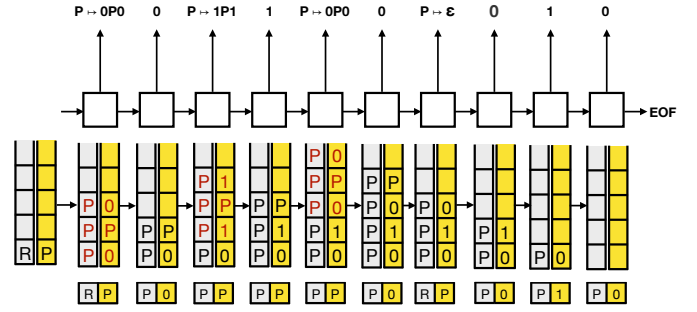


Fig. 4: The generation process of the parse tree shown in Figure 3. The generator maintains a parent stack (grey columns) and a children stack (yellow columns), the current node (yellow boxes below the stacks) and its parent (grey boxes) are popped from the two stacks respectively at each generation step. The red elements in the stack refer to the ones are pushed at each step. The generation terminates when both stacks are empty.

of the generated part of a sequence. For example, at time step t_2 in Figure 3, a conventional RNN may generate action $P \mapsto 1P1$ or $P \mapsto 0P0$, which violates the pre-defined grammar. Thus, we need an extra control on the RNN to track the generation state accurately. The output of the LSTM at time step t is denoted as $o_t \in \mathbb{R}^L$, where L is the size of the set of actions. At the output layer of each time step, the generator samples an action from the multinomial distribution denoted by $\text{softmax}(o_t) = (\hat{o}_t^{(1)}, \dots, \hat{o}_t^{(L)})$, where $\hat{o}_t^{(k)}$ corresponds to the probability of sampling action a_k at time step t_t . A mask matrix $\mathbf{M}^{(G)} \in \{0, 1\}^{(|\mathcal{V}| \times |\mathcal{P} \cup \mathcal{T}|)}$ can be derived for grammar \mathcal{G} . The k -th row in $\mathbf{M}^{(G)}$, which is denoted as $\mathbf{M}^{(G)}(k)$, marks the valid actions for $v_k \in \mathcal{V}$ as 1s and the invalid ones as 0s. Thus, when the generator G_{θ} reaches the time step t_t where the corresponding node is non-terminal node $v_k \in \mathcal{V}$, then the following masking is performed before it generates the token for step t_t :

$$\tilde{o}_t = \text{softmax}(o_t) \odot \mathbf{M}^{(G)}(k) \quad (6)$$

Hence, the probability of invalid actions for v_k is reset to 0 in \tilde{o}_t . In the other cases, when G_{θ} reaches a time step where the corresponding node is a terminal node $y_k \in \mathcal{T}$, then y_k is directly generated. By applying such masking process, our tree generator can no longer sample actions that violate the syntax.

Tracking Algorithm: The remaining problem is how the tree generator identifies the node type and retrieves the parent action at time step t_t . As shown in Figure 4, we maintain two stacks $\Omega^{(P)}$ and $\Omega^{(C)}$ for *parent tracking* and *children tracking* respectively, which is analogous to the well-known pushdown automata (PDA). At the beginning of generation, the stacks are initialized as $\Omega^{(P)} = [\Gamma, R]$ and $\Omega^{(C)} = [\Gamma, S]$, where Γ is the empty stack symbol that cannot be popped and R is the pseudo-root symbol. At each step t_t of generation, the following stack operations are performed sequentially: $P \xleftarrow{\text{pop}} \Omega^{(P)}$, $C \xleftarrow{\text{pop}} \Omega^{(C)}$, where P is the corresponding parent action and C is the head variable for time step t_t . If

$C \in \mathcal{T}$, then C is generated directly and no further stack operations are required before next time step.

When $C \in \mathcal{V}$, the embedding of the action at previous time step t_{t-1} and the embedding of P are fetched respectively to build the input vector $\mathbf{x}_t = (\mathbf{a}_{t-1}, \mathbf{p}_t)$. After applying Eq. (5), an action that takes the form $(C \mapsto H) \in \mathcal{P}$ is generated based upon the masked probability vector $\tilde{\mathbf{o}}_t$, where $H \in (\mathcal{V} \cup \mathcal{T})^+$ is a sequence of variables. Before moving forward to next time step, the following stack operations are performed, $C \xrightarrow{\text{push}} \Omega^{(P)}$, $\text{reversed}(H) \xrightarrow{\text{push}} \Omega^{(C)}$, where we push the variable C into the parent stack, and push the variables in H into the children stack in a reversed order.

Close A Generation: If $\Omega^{(P)} = \Omega^{(C)} = [\Gamma]$ at the beginning of a time step, it indicates that all interior nodes have been expanded and all leaves are labeled with a terminal token in the tree, then the generator closes the generation by producing an end symbol.

C. Tree Discriminator

Since we require the discriminator encode the rich grammar information of a sequence, it should capture the structure and the semantics of the corresponding parse tree. Thus, we use the Child-Sum Tree-LSTM [11] as the discriminator of TreeGAN. The formulation is as follows,

$$\begin{aligned} \tilde{\mathbf{h}}_j &= \sum_{k \in Ch(j)} \mathbf{h}_k \\ \mathbf{i}_j &= \sigma(\mathbf{W}^{(i)} \mathbf{x}_j + \mathbf{U}^{(i)} \tilde{\mathbf{h}}_j + \mathbf{b}^{(i)}), \\ \mathbf{f}_{jk} &= \sigma(\mathbf{W}^{(f)} \mathbf{x}_j + \mathbf{U}^{(f)} \mathbf{h}_k + \mathbf{b}^{(f)}), \\ \mathbf{o}_j &= \sigma(\mathbf{W}^{(o)} \mathbf{x}_j + \mathbf{U}^{(o)} \tilde{\mathbf{h}}_j + \mathbf{b}^{(o)}), \\ \mathbf{u}_j &= \tanh(\mathbf{W}^{(u)} \mathbf{x}_j + \mathbf{U}^{(u)} \tilde{\mathbf{h}}_j + \mathbf{b}^{(u)}), \\ \mathbf{c}_j &= \mathbf{i}_j \odot \mathbf{u}_j + \sum_{k \in Ch(j)} \mathbf{f}_{jk} \odot \mathbf{c}_k, \\ \mathbf{h}_j &= \mathbf{o}_j \odot \tanh(\mathbf{c}_j), \end{aligned} \quad (7)$$

where $Ch(j)$ refers to the set of children of node j . This model is also called Child-Sum Tree-LSTM, in which a tree proceeds from leaves to the root. Moreover, \mathbf{h}_r denotes the final hidden state for a given tree where r is the root node of the tree, and it encodes the entire tree and can be used for classification. A fully connected linear layer is appended after the output of Tree-LSTM to obtain the confidence:

$$\Psi = \text{sigmoid}(\mathbf{W}^{(c)} \mathbf{h}_r + \mathbf{b}^{(c)}), \quad (8)$$

where $\Psi \in (0, 1)$ refers to the probability of the encoded tree being a real instance.

D. Pre-Training

Before starting the adversarial training, pre-training of D_ϕ and G_θ are usually required to reach a good initialization, which can facilitate the convergence later in adversarial training. We initialize the tree generator parameters using conventional maximum likelihood estimation (MLE). As to the tree discriminator initialization, we let the discriminator distinguish the twisted trees from the real trees. We randomly swap two

subtrees of different head types for each real parse tree in the corpus to construct the twisted tree counterparts. The swapping operation breaks the syntax of the real parse tree, which guides the discriminator to learn correct syntax patterns.

IV. SYNTHETIC STUDY

Due to the lack of well documented syntax and schema (for SQL) in real datasets, we first test the effectiveness of the proposed model on three synthetic datasets with pre-defined syntax and schema as the ground-truth. In this section, we will first introduce the detailed experimental settings, compared methods and the evaluation metrics used on synthetic study. We attempt to answer the following research questions within this section.

- **RQ1:** Does TreeGAN correctly capture the syntax information?
- **RQ2:** How good does TreeGAN capture the underlying semantical pattern (e.g. schema)?
- **RQ3:** Could TreeGAN generates sequences of better quality when compared to other baselines?

A. Dataset

We prepare three different synthetic datasets with controlled syntax and schema (for SQL datasets only).

- **PLD:** A dataset of palindrome in english alphabet (26 capital letters and 26 lowercase letters).
- **SQL-A:** A dataset of SQL queries (SELECT queries) with a small set of grammatical rules.
- **SQL-B:** A dataset of SQL queries with larger set of grammatical rules.

Note that the proposed TreeGAN uses only the grammar (syntax) but not the schema to train the sequence generator. The synthetic datasets and the corresponding grammatical rules and schema will be public available after this paper is accepted.

TABLE II: Summary of Synthetic Datasets

Dataset	# Training	# Test	# Vocab.	# Prod. Rules
PLD	10,000	1,000	160	106
SQL-A	50,000	5,000	1000	231
SQL-B	100,000	5,000	5000	422

B. Compared Methods

We test the following methods to demonstrate the effectiveness of the proposed method.

- **TreeGAN (Our):** it uses the tree generator described in Sec. III and the Child-Sum Tree-LSTM as the discriminative model.
- **TreeGAN- (Our):** A variation of TreeGAN that uses LSTM as the discriminative model instead of Tree-LSTM.
- **TreeGen [4]:** Tree generator without adversarial training, using MLE training.
- **SeqGAN [2]:** The original Sequence GAN that proposed for general purpose sequence generation task.

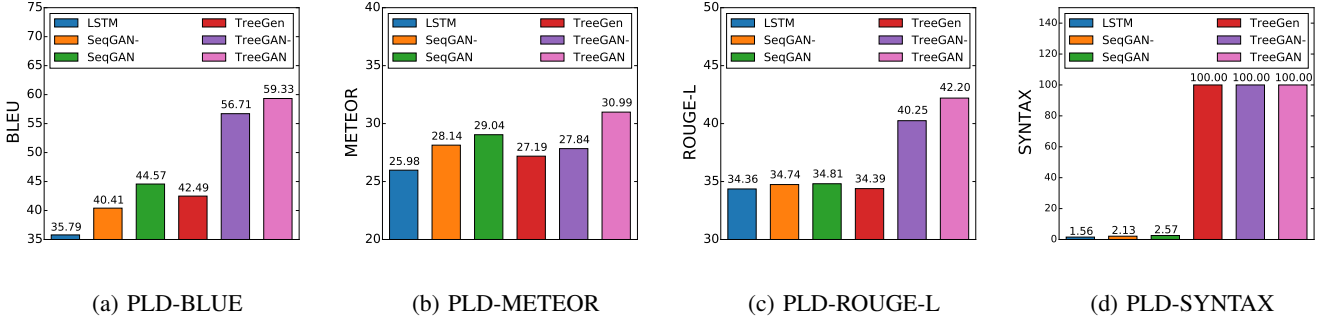


Fig. 5: Quantitative Evaluation on PLD Dataset.

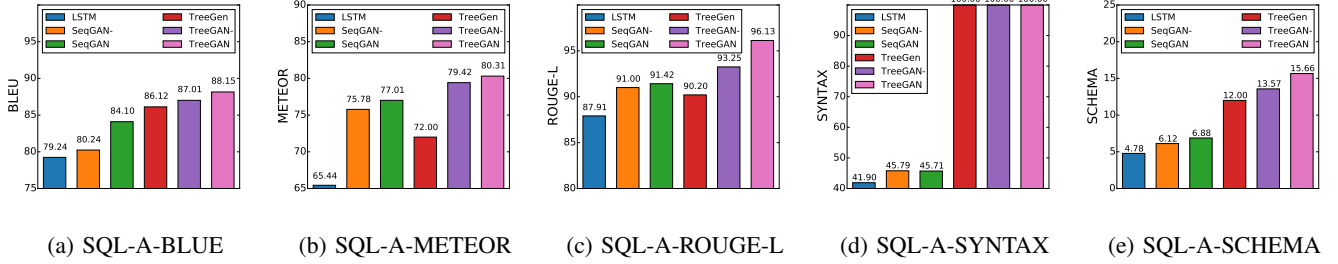


Fig. 6: Quantitative Evaluation on SQL-A.

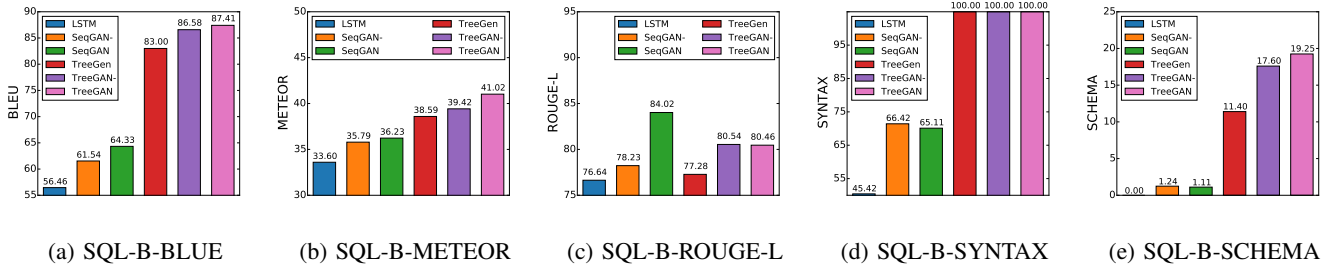


Fig. 7: Quantitative Evaluation on SQL-B.

- **SeqGAN-** [2]: A variation of Sequence GAN that uses LSTM as the discriminative model instead of CNN.
- **LSTM** [10]: LSTM generator employs Maximum Likelihood Estimation as the training strategy.

All compared methods are implemented using PyTorch² in Python. The batch size is set to 64 for all models.

C. Experimental Settings

For each dataset we used in this section, we first transform each sequence into a sequence of actions that pre-defined in the given syntax. Note that each sequence of actions represents the yield of syntax parse tree for the corresponding sequence. Then we randomly select 10% of data samples to form the test (reference) set, and use the remaining 90% as the training set. For all GAN models include the proposed TreeGAN, we perform 50 epochs of pre-training before starting the adversarial training. And the adversarial training last up to 50 epochs or until the policy gradient loss converges. We

use grid search to find the best hyper-parameter of TreeGAN. Default hyper-parameter are used for the compared methods unless otherwise stated. All the generated trees of TreeGAN are translated into sequence for evaluation purpose. We report the evaluation scores based on the generations of trained generative net against the samples in the test (reference) set. We used the number of generations produced by the trained generator to be the same as the size of test set in each dataset.

D. Evaluation Metrics

We include commonly used metrics such as BLEU-3 [14], METEOR score [15] and ROUGE-L score [16]. Since neither of these metrics is designed to measure how well the generated sequences fit the target grammar, we propose two additional metrics to evaluate them. The first one measures the percentage of the generated sequences that are grammatically correct (labeled as SYNTAX). For SQL generation tasks, we additionally report the percentage of generated sequences which obey the schema (labeled as SCHEMA, evaluate the correctness of entity and relation for the generated SQL).

²<http://pytorch.org>

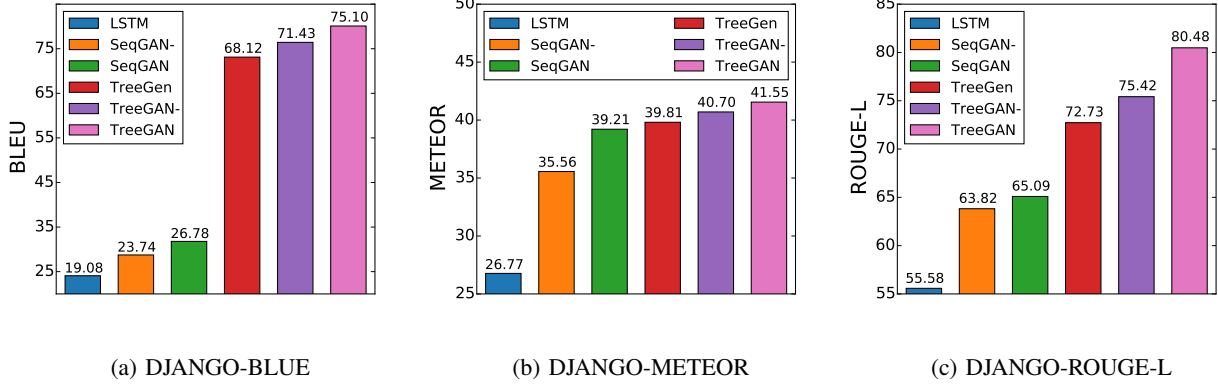


Fig. 8: Quantitative Evaluation on Django

E. Quantitative Results

Figure 5, Figure 6 and Figure 7 show the quantitative results on synthetic dataset PLD, SQL-A, SQL-B respectively.

To answer the **RQ1**, we demonstrate the SYNTAX scores in Figure 5(d), Figure 6(d) and Figure 7(d), in which we observe that the tree-based frameworks, including the proposed TreeGAN, achieve 100% syntax correctness regards the pre-defined grammar while other baselines perform badly in terms of this syntax correctness. This results show that the proposed TreeGAN could fully capture the given syntax information and generate grammatically correct sequence.

As to the **RQ2**, we could read Figure 6(e) and Figure 7(e). We discover that even though without explicit input of schema, the proposed TreeGAN has higher chance for capturing the underlying semantic pattern, given at least 3.66% and 7.85% improvement on SCHEMA in SQL-A and in SQL-B respectively.

More generally, we use three popular NLP metrics to evaluate the quality of the generated sequences, which are presented in Figure 5(a)-(c), Figure 6(a)-(c) and Figure 7(a)-(c). From these figures we can clearly see the superiority of TreeGAN, who consistently outperforms the compared methods in terms of the BLEU, METEOR and ROUGE-L (except the case of ROUGE-L in SQL-B, where TreeGAN still obtain competitive results). These results should clearly answer the **RQ3** we raised earlier, the quality of sequences generated by TreeGAN is better than the generations of compared methods.

F. Qualitative Results

Table III samples several generations on SQL-B dataset for qualitative evaluation. We mainly compare the generations of TextGAN with SeqGAN to demonstrate the advantage of employing tree structure generator and discriminator in GANs on sequence. Consistent with the results we have seen in quantitative evaluation, SeqGAN's generations could not perfectly follow the underlying grammar and exhibit syntax errors (highlighted in red). As shown in Table III, the generation ④ of SeqGAN mistakenly applies 'count' aggregation on a numerical value and does not close the 'from' clause correctly. Similar syntax errors can also be observed in the generation

TABLE III: SQL query generation in SQL-B. Syntax errors are highlighted in red color.

Real SQL Queries

- ① select count(authenticated) from America where alight>3;
- ② select driftpin, min(deject) from Danmark where driftpin=16;
- ③ select hedy from Hungary;

Queries generated by SeqGAN[2]

- ④ select count(17), min(acoustically) from;
- ⑤ select max(cookstove), gainfully, min ()), min(buttonhole) from America;
- ⑥ select aalesund from Brazil where hanuman acoustically Hungary;

Queries generated by TreeGAN (our method)

- ⑦ select min(jakarta) from Jamaica;
- ⑧ select min(endogenous) from Brazil where epigraphical=1;
- ⑨ select hedy from Hungary where deject!=2;

⑤ and ⑥ by SeqGAN. Meanwhile, TreeGAN incorporates the pre-defined grammar, and all its generations are valid. We randomly select some examples in Table III. The generation ⑦ and ⑧ mimic the ground truth ① and ② well and capture the underlying schema correctly. Generation ⑨ resembles ground truth ③ and extend it with an extra 'where' clause. These observations contribute to **RQ1** and **RQ3**, and re-confirm the answers we obtained in the analysis of quantitative results.

V. EXPERIMENTS ON REAL DATA

To better demonstrate the superiority of the proposed TreeGAN, we also perform experiments on real dataset. We try to additionally answer the following research questions through this section.

- **RQ4:** Does TreeGAN achieve similar performance on real dataset with more complex syntax as in synthetic dataset?
- **RQ5:** What is the limitation of TreeGAN on sequence generation?

A. Dataset

We test our proposed model on the python code dataset [17] from django³ project. It is a collection of lines of python code, and each performs a functional task. We use Python AST package and Astor package⁴ to construct and parse the AST corresponds to each line of code in the dataset. The code in Django dataset is diverse and spanning a wide variety of real-world use cases such as I/O operations, exception handling, and mathematical computation. We follow the same setting and same evaluation metrics (SYNTAX is not reported due to the freeness of Python grammar) as in the previous section.

B. Quantitative Results

Figure 8 shows the results quantitative evaluation on Django dataset, from which we discover TreeGAN achieves 6.82% improvement against TreeGen and 18.14% improvement against SeqGAN in terms of BLEU score. As to the METEOR score, TreeGAN improves the performance 1.75% against TreeGen and 2.34% against SeqGAN. We also discover obvious improvement has been made by TreeGAN against TreeGen and SeqGAN in terms of ROUGE-L. Hence, TreeGAN exhibits similar advantages on the real data as in the synthetic study, which gives a positive answer towards **RQ4**.

C. Qualitative Results

Table IV shows generations from TreeGAN and SeqGAN on Django dataset. Similar to the results obtained in the synthetic study, we found although SeqGAN could mimic the real Python code, it exhibits several types of syntax errors (highlighted in **red**). Generation ④ indicates SeqGAN sometimes could not correctly fill the function arguments, generation ⑤ exhibits a misunderstanding of import statement, while generation ⑥ demonstrates SeqGAN has difficulty in pairing the parentheses. Meanwhile, generation ⑦ and ⑧ show the capability of TreeGAN on learning the usage of assignment statement, function call, conditional statement, *etc.* These observations indicate that on code generation tasks, TreeGAN could effectively plug in complex grammatical rules and generate valid code snippets, which re-confirm the answer we obtained for **RQ4**.

We also discuss the limitation of TreeGAN (**RQ5**), which could shed a light on future extension. From generation ⑨, we can see TreeGAN has difficulty in understanding the concept of inheritance and the member function, where the parent class of 'META' does not have the member function called 'new_file()', the call is invalid and causes a running-time error. There are about 3.7% of generations by TreeGAN exhibit the similar semantic error in our experiments. It is not difficult to identify that this semantic error in Python is the counterpart of the schema error in SQL. It is possible for our model to learn these semantic pattern from the data, but it may need a better way to guide the learning process for fully capturing the semantic, which could be a future direction for this work.

TABLE IV: Python code generation in Django. Syntax errors are highlighted in **red** color.

Real Python Code

- ① f.write(pickle.dumps(expiry,-1))
- ② db = router.db_for_read(self.cache_model_class)
- ③ if connections[db].features.needs_datetime_string_cast and not isinstance(expires, datetime)

Code generated by SeqGAN [2]

- ④ name=self._save(, name, content, self)
- ⑤ from django.ImproperlyConfigured **import 0**
- ⑥ return urljoin(self.base_url, filepath_to_uri()))

Code generated by TreeGAN (our method)

- ⑦ table = connections[db].ops.quote_name(self._table)
- ⑧ if exp is None or exp>time.time()
- ⑨ super(META, self).new_file(file_name, *args)

D. Final Remarks

Through the analysis of synthetic study and experiments on real data, we can finally answer the following two research questions to summarize our experiments:

- **RQ6:** Is TreeGAN a better *GAN model* on sequence generation?
- **RQ7:** Is TreeGAN a better *syntax-aware model* on sequence generation?

The quantitative and qualitative comparisons between TreeGAN and SeqGAN (and its variation) should support a positive answer towards the **RQ6**. As to the **RQ7**, we can conclude that employing GAN improves the generation quality by comparing TreeGAN with TreeGen (as shown in synthetic study and real data experiments). By convincing TreeGAN is a better GAN model and a better syntax-aware model on sequence generation, we justify that both *syntax-aware* and *GAN* are indispensable components toward a more useable sequence generation.

VI. RELATED WORK

Our work is related to both syntax aware sequence generation and generative adversarial networks (GANs), we briefly discuss them respectively in this section.

A. Syntax Aware Sequence Generation

Most works on this line require descriptive input such as text specification [4, 18–20], and their overall goal is to generate code that performs the corresponding task(s) described in the input text. Our proposed model is different from these existing models in several aspects: (1) Our model does not require any descriptive text as the input. (2) Our model employs a GAN training framework to improve the generation quality. (3) Our model targets at generating arbitrary sequences that follow the pre-defined syntax and resemble the real sequences. Some other methods on code generation focus on specific languages [21, 22], but our model is generalized to fit any context-free grammar. Besides, there are several probabilistic generation models [23, 24], which are mainly based on Bayesian estimation while our work is based on neural networks.

³<https://www.djangoproject.com/>

⁴<http://astor.readthedocs.io/en/latest/>

B. Generative Adversarial Networks (GANs)

Figure 2 illustrates the comparison between TreeGAN and the related GAN generation methods. GAN was first proposed in [1], and it exhibits superb performance on image generation [25, 26] and image synthesis [27, 28]. Later on, [2] studied GAN on sequence generation using policy gradient and Monte Carlo search. [3] alternatively employs feature matching to perform the similar task. [7] additionally consider adding control on the sentiment and tenses of the generated text. However, neither of these works consider the existing grammar or syntax of the target language, and the generated text may exhibit syntax errors. Our work makes the first effort to incorporate the grammatical knowledge into GAN model on sequence and text generation.

VII. CONCLUSION AND FUTURE WORK

We proposed a syntax-aware GAN model called TreeGAN for sequence generation. We transform the problem into parse tree generation to incorporate the rich grammar information, and both the generator and discriminator are well-tailored to encode the syntax properly. The experiments on both synthetic datasets and real-world datasets demonstrate that TreeGAN is a promising adversarial learning framework for syntax-aware sequence generation.

We plan to extend this work in two directions in the future. The first extension will focus on incorporating pre-defined schema information (e.g. SQL schema) into the GAN model, which allows the generator to fully compatible with the finer level semantics of the target formal language and extend the incorporated grammar to the scope of context sensitive grammar (CSG). The second extension will consider a topic-wise TreeGAN, which could not only generate valid sequences under the given grammar, but ensure all the generated sequences are describing the given target topic. With the proposed TreeGAN and these potential extensions, we could make GANs a more practicable and versatile tool for automatically composing sequence.

REFERENCES

- [1] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Proc. 28th Advances in Neural Information Processing Systems (NIPS'14)*, 2014, pp. 2672–2680.
- [2] L. Yu, W. Zhang, J. Wang, and Y. Yu, "Seqgan: sequence generative adversarial nets with policy gradient," in *Proc. 31st AAAI Conf. on Artificial Intelligence (AAAI '17)*, 2017, pp. 2852–2858.
- [3] Y. Zhang, Z. Gan, K. Fan, Z. Chen, R. Henao, D. Shen, and L. Carin, "Adversarial feature matching for text generation," in *Proc. 34th Int. Conf. Machine Learning (ICML'17)*, 2017.
- [4] P. Yin and G. Neubig, "A syntactic neural model for general-purpose code generation," in *Proc. 57th Annual Meeting of the Association for Computational Linguistics (ACL'17)*, 2017.
- [5] A. Radford, L. Metz, and S. Chintala, "Unsupervised representation learning with deep convolutional generative adversarial networks," in *Proc. 4th International Conf. on Learning Representations (ICLR'16)*, 2016.
- [6] W. Fedus, I. Goodfellow, and A. Dai, "Maskgan: Better text generation via filling in the _," *arXiv preprint arXiv:1801.07736*, 2018.
- [7] Z. Hu, Z. Yang, X. Liang, R. Salakhutdinov, and E. P. Xing, "Toward controlled generation of text," in *Proc. 34th Int. Conf. Machine Learning (ICML'17)*, 2017, pp. 1587–1596.
- [8] N. Chomsky, "Three models for the description of language," *IRE Transactions on information theory*, vol. 2, no. 3, pp. 113–124, 1956.
- [9] A. Krizhevsky, I. Sutskever, and G. Hinton, "Imagenet classification with deep convolutional neural networks," in *Proc. 26th Advances in Neural Information Processing Systems (NIPS'12)*, 2012, pp. 1097–1105.
- [10] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.
- [11] T. Kai, S. Richard, and D. Christopher, "Improved semantic representations from tree-structured long short-term memory networks," in *Proc. 55th Annual Meeting of the Association for Computational Linguistics (ACL'15)*, 2015.
- [12] J. E. Hopcroft, R. Motwani, and J. Ullman, *Introduction to Automata Theory, Languages, and Computation*. Addison Wesley, 2006.
- [13] R. Sutton, D. McAllester, S. Singh, and Y. Mansour, "Policy gradient methods for reinforcement learning with function approximation," in *Proc. 14th Advances in Neural Information Processing Systems (NIPS'00)*, 2000, pp. 1057–1063.
- [14] K. Papineni, S. Roukos, T. Ward, and W. Zhu, "Bleu: a method for automatic evaluation of machine translation," in *Proc. 42nd Annual Meeting of the Association for Computational Linguistics (ACL'02)*, 2002, pp. 311–318.
- [15] M. Denkowski and A. Lavie, "Meteor universal: Language specific translation evaluation for any target language," in *Proc. of the 9th Workshop on Statistical Machine Translation (WMT'14)*, 2014, pp. 376–380.
- [16] C. Lin, "Rouge: A package for automatic evaluation of summaries," in *Proc. of Workshop on Text Summarization Branches Out, Post Conference Workshop of ACL'04*, 2004.
- [17] Y. Oda, H. Fudaba, G. Neubig, H. Hata, S. Sakti, T. Toda, and S. Nakamura, "Learning to generate pseudo-code from source code using statistical machine translation (t)," in *Proc. 30th IEEE/ACM International Conf. Automated Software Engineering (ASE'15)*, 2015, pp. 574–584.
- [18] T. Lei, F. Long, R. Barzilay, and M. Rinard, "From natural language specifications to program input parsers," in *Proc. 53rd Annual Meeting of the Association for Computational Linguistics (ACL'13)*, 2013, pp. 1294–1303.
- [19] W. Ling, P. Blunsom, E. Grefenstette, K. Hermann, T. Kočiský, F. Wang, and A. Senior, "Latent predictor networks for code generation," in *Proc. 56th Annual Meeting of the Association for Computational Linguistics (ACL'16)*, 2016, pp. 599–609.
- [20] M. Balog, A. Gaunt, M. Brockschmidt, S. Nowozin, and D. Tarlow, "Deepcoder: Learning to write programs," in *Proc. 4th International Conf. on Learning Representations (ICLR'16)*, 2016.
- [21] M. Raza, S. Gulwani, and N. Milic-Frayling, "Compositional program synthesis from natural language and examples," in *Proc. 24th Int. Joint Conf. on Artificial Intelligence (IJCAI'15)*, 2015, pp. 792–800.
- [22] E. Parisotto, A. Mohamed, R. Singh, L. Li, D. Zhou, and P. Kohli, "Neuro-symbolic program synthesis," *arXiv preprint arXiv:1611.01855*, 2016.
- [23] C. Maddison and D. Tarlow, "Structured generative models of natural source code," in *Proc. 31st Int. Conf. Machine Learning (ICML'14)*, 2014, pp. 649–657.
- [24] T. Nguyen, A. Nguyen, H. Nguyen, and T. Nguyen, "A statistical semantic language model for source code," in *Proc. 9th Joint Meeting on Foundations of Software Engineering (SIGSOFT'13)*, 2013, pp. 532–542.
- [25] P. Isola, J. Zhu, T. Zhou, and A. Efros, "Image-to-image translation with conditional adversarial networks," in *Proc. Conf. on Computer Vision and Pattern Recognition (CVPR'17)*, 2017.

- [26] E. Denton, S. Chintala, R. Fergus *et al.*, “Deep generative image models using a laplacian pyramid of adversarial networks,” in *Proc. 29th Advances in Neural Information Processing Systems (NIPS’15)*, 2015, pp. 1486–1494.
- [27] S. Reed, Z. Akata, X. Yan, L. Logeswaran, B. Schiele, and H. Lee, “Generative adversarial text to image synthesis,” in *Proc. 33rd Int. Conf. Machine Learning (ICML’16)*, 2016, pp. 1060–1069.
- [28] H. Zhang, T. Xu, H. Li, S. Zhang, X. Huang, X. Wang, and D. Metaxas, “Stackgan: Text to photo-realistic image synthesis with stacked generative adversarial networks,” in *IEEE Int. Conf. Comput. Vision (ICCV’17)*, 2017, pp. 5907–5915.