A Noise-tolerant Differentiable Learning Approach for Single Occurrence Regular Expression with Interleaving

Rongzhen Ye[1], Tianqu Zhuang[1], Hai Wan[1,*], Jianfeng Du[2,*], Weilin Luo[1], Pingjia Liang[1]
[1]Sun Yat-sen University, Guangzhou, China [2]Guangdong University of Foreign Studies, China

Motivation

Problem. Learning *single occurrence regular expression with interleaving* (SOIRE) in *full expressive power* from a set of text strings with *noise*.

Challenge.

1. *heavy computation in searching* to get the full expressive power
2. *wrong search bias* resulting from noisy data

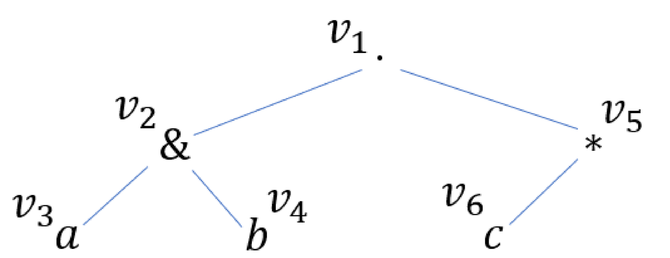
Contributions

1. We have proposed a noise-tolerant differentiable learning approach `SOIREDL`. We first train the neural network that simulates SOIRE matching to classify the given strings, and then interpret the target SOIRE from the parameters.
2. Theoretically, the faithful encodings learnt by `SOIREDL` one-to-one correspond to SOIREs for a bounded size.
3. Experimental results have demonstrated higher performance compared with the SOTA approaches.

Filter Matching

Filter matching for a SOIRE r and a string s is to check if r matches $filter(s, \alpha(r))$, where $\alpha(r)$ denotes the set of symbol in r, and function $filter(s, V)$ returns a string that only retains symbols in V, where $V \subseteq \Sigma$. Let $g^t_{i,j} \in \{0, 1\}$ denote whether r^t matches $filter(s_{i,j}, \alpha(r^t))$, where $s_{i,j}$ denotes the substring of s from i to j, and where $s_{1,0} = \epsilon$ specially.

Example 1. *For $(a\&b)c^*$ and $s = dbac$,*



- *filter matching is to check if $(a\&b)c^*$ matches $filter(dbac, \{a, b, c\}) = bac$, as $\alpha((a\&b)c^*) = \{a, b, c\}$.*
- *$g^2_{1,2}$ denotes if $a\&b$ (r^2) matches ba and $g^2_{1,2} = 1$.*

SOIRETM

Theorem 1. *Given a SOIRE r and a string s, $r \models s$ iff $filter(s, \alpha(r)) = s$ and $r \models filter(s, \alpha(r))$.*

The steps of `SOIRETM`:

1. build the syntax tree of r
2. check if $filter(s, \alpha(r)) = s$
3. calculate $g^t_{i,j}$ from shorter substrings to longer ones and from bottom to top of the syntax tree (dynamic programming)
4. return $g^1_{1,|s|}$ (filter matching)

Theorem 2. *Given a SOIRE r and a string s, $r \models s$ iff $\mathtt{SOIRETM}(r, s) = 1$.*

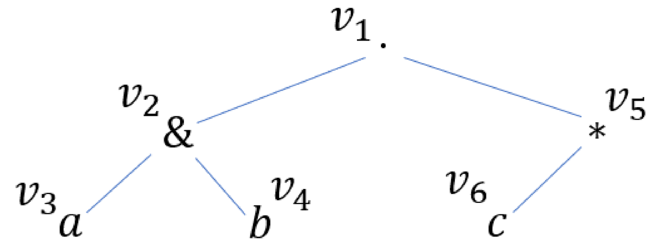
From `SOIRETM` to Neural Network

The trainable parameters $\theta = (w, u)$:

- $w \in [0, 1]^{T \times |\mathbb{B}|}$, w^t_a denotes the probability of vertex t representing a symbol in Σ or an ordinary operator or the none operator.
- $u \in [0, 1]^{T \times T}$, $u^t_{t'}$ denotes the probability of vertex t choosing vertex t' as its right child.

where $\mathbb{B} = \Sigma \cup \{?, *, +, \cdot, \&, |, \text{none}\}$, T the bounded size of the target SOIRE.

Example 2.



When $T = 6$, $w^1_{\cdot}$, $w^2_{\&}$, w^3_a, w^4_b, w^5_*, w^6_c, u^1_5, u^2_4 are 1s, whereas other parameters are 0s. When $T = 8$, w^7_{none}, w^8_{none} are 1s in addition to the above parameters.

Four main parts in conversion.

- ρ^t_a, the differentiable version of $\alpha(r^t)$. ($\sigma_{01}(x) = \min(\max(x, 0), 1)$)

$$\rho^t_a = \sigma_{01}\Big(w^t_a + \sum_{o \in \{?,*,+,\cdot,\&,|\}} w^t_o \rho^{t+1}_a + \sum_{o \in \{\cdot,\&,|\}} w^t_o \sum_{t'=t+2}^{T} u^t_{t'} \rho^{t'}_a\Big)$$

- $flag^{t,t'}_{i,j}$, the probability that there does not exist a symbol occurring in both $s_{i,j}$ and $\alpha(r^t)$ but not occurring in $\alpha(r^{t'})$.

$$flag^{t,t'}_{i,j} = 1 - \sigma_{01}\Big(\sum_{a \in \Sigma} \sigma_{01}(1[a \in s_{i,j}] + (\rho^t_a - \rho^{t'}_a) - 1)\Big)$$

- $g^t_{i,j}$, whether r^t matches $filter(s_{i,j}, \alpha(r^t))$

$$g^t_{i,j} = \sum_{a \in \Sigma} w^t_a \cdot 1[filter(s_{i,j}, a) = a]$$
$$+ \sum_{o \in \{?,*,+\}} w^t_o p^t_{i,j}(o) + \sum_{o \in \{\cdot,\&,|\}} w^t_o \sum_{t'=t+2}^{T} u^t_{t'} p^t_{i,j}(o, t')$$

- return value of `SOIRETM`, combining $filter(s, \alpha(r)) = s$ and $g^1_{1,|s|}$.

$$\hat{y} = g^1_{1,|s|} - \max_{a \in \Sigma} \sigma_{01}(1[a \in s] - \rho^1_a)$$

Objective function: $\frac{1}{2}(\hat{y} - y)^2$, where y is ground-truth label for r matching s.

Faithful Encoding

Definition 1 (Faithful encoding). *An encoding $\theta = (w, u)$ of SOIREs with length T is said to be faithful if it satisfies all the following conditions:*

1. *$\forall 1 \le t \le T, w^t$ is a one-hot vector.*
2. *$\forall 1 \le t \le T, u^t$ is either a one-hot vector or an all-zero vector.*
3. *$\forall 1 \le t \le T, \sum_{t'=t+2}^{T} u^t_{t'} + \sum_{a \in \Sigma \cup \{?,*,+,none\}} w^t_a = 1$.*
4. *$\forall 1 \le t \le T - 1, w^{t+1}_{none} - w^t_{none} \ge 0$.*
5. *$\forall 2 \le t \le T, \sum_{a \in \{?,+,*,\cdot,\&,|\}} w^{t-1}_a + \sum_{t'=1}^{t-2} u^{t'}_t + w^t_{none} = 1$.*
6. *$\forall 3 \le t \le T, \forall 1 \le p \le t-2, (t-1-p)u^p_t + \sum_{p'=p+1}^{t-1} \sum_{t'=t+1}^{T} u^{p'}_{t'} \le t - 1 - p$.*
7. *$\forall a \in \Sigma, \sum_{t=1}^{T} w^t_a \le 1$.*

Interpretation

Theorem 3. *Given a bounded size $T \in \mathbb{Z}^+$, prefix notations of SOIREs r with $|r| \le T$ and faithful encodings θ with length T have a one-to-one correspondence, i.e., $\mathtt{Enc2Pre}(\theta) = \mathtt{PreForm}(r)$.*

Based on this correspondence, we apply beam search to find a faithful encoding nearby the learnt encoding and then interpret it to the target SOIRE.

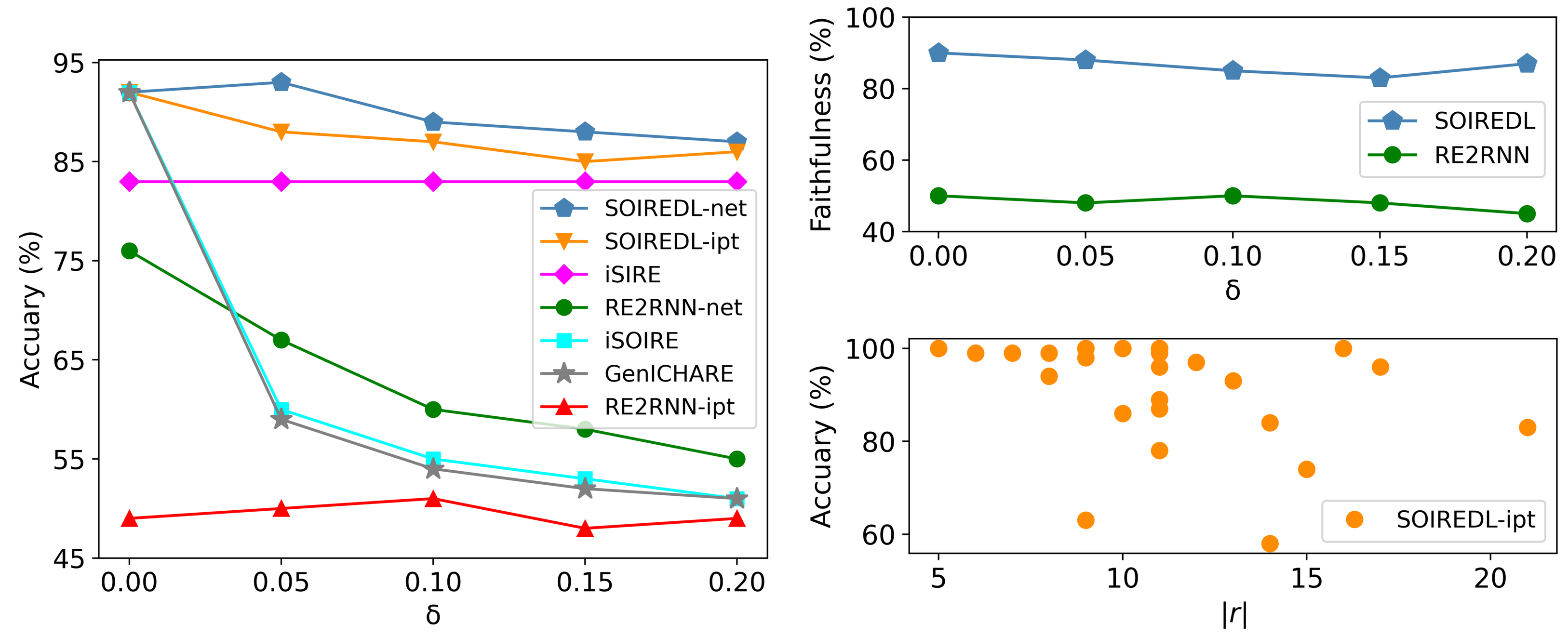
- The interpretation steps are conducted from bottom to top of the syntax tree. The score of each SOIRE is defined as the geometric mean of the probabilities of all operators and symbols.
- calculate the accuracy of each SOIRE in the last step on the training set and pick out the SOIRE with the highest accuracy.

Result Analysis

Competitor.

- Positive strings only: `iSOIRE` (Li et al. 2019) for RSOIREs, `GenICHARE` (Zhang et al. 2018) for ICHAREs
- Both positive and negative strings: `iSIRE` (Li et al. 2020) for SIREs, `RE2RNN` (Jiang et al. 2020) for automatons, `SOIREDL` (Ours) for SOIREs

Comparisons on Noisy Data, Faithfulness and Scalability.



- `SOIREDL` gets the SOTA results and is the most robust on noisy data.
- The neural network of `SOIREDL` and its interpreted SOIRE are more consistent in performing SOIRE matching.
- The accuracy of `SOIREDL` decreases when the size of the ground-truth SOIRE increases.
  - This may be due to the difficulty for a neural network to capture the long-distance dependency in SOIRE matching.

Acknowledgments

We thank Kunxun Qi for discussion on the paper and anonymous referees for helpful comments. This paper was supported by the National Natural Science Foundation of China (No. 62276284, 61976232, 61876204), the National Key Research and Development Program of China (No. 2021YFA1000504), Guangdong Basic and Applied Basic Research Foundation (No. 2022A1515011355), Guangzhou Science and Technology Project (No. 202201011699), Guizhou Science Support Project (No. 2022-259), as well as Humanities and Social Science Research Project of Ministry of Education (No. 18YJCZH006).