RetCL: A Selection-based Approach for Retrosynthesis via Contrastive Learning

Hankook Lee Sungsoo Ahn Seung-Woo Seo You Young Song
Eunho Yang Sung-Ju Hwang Jinwoo Shin

August 6, 2021

- Introduction
- Search procedure
- Training scheme with contrastive learning
- Experiments

Notation

- P, R: product and reactant molecules.
- $\mathcal{C}$: the candidate set.
- $\mathcal{R}$: the reactant set.
- $\prod$: the space of permutations.

Introduction

The RetCL framework:

- selection-based
- template-free

The Search Procedure of RetCL

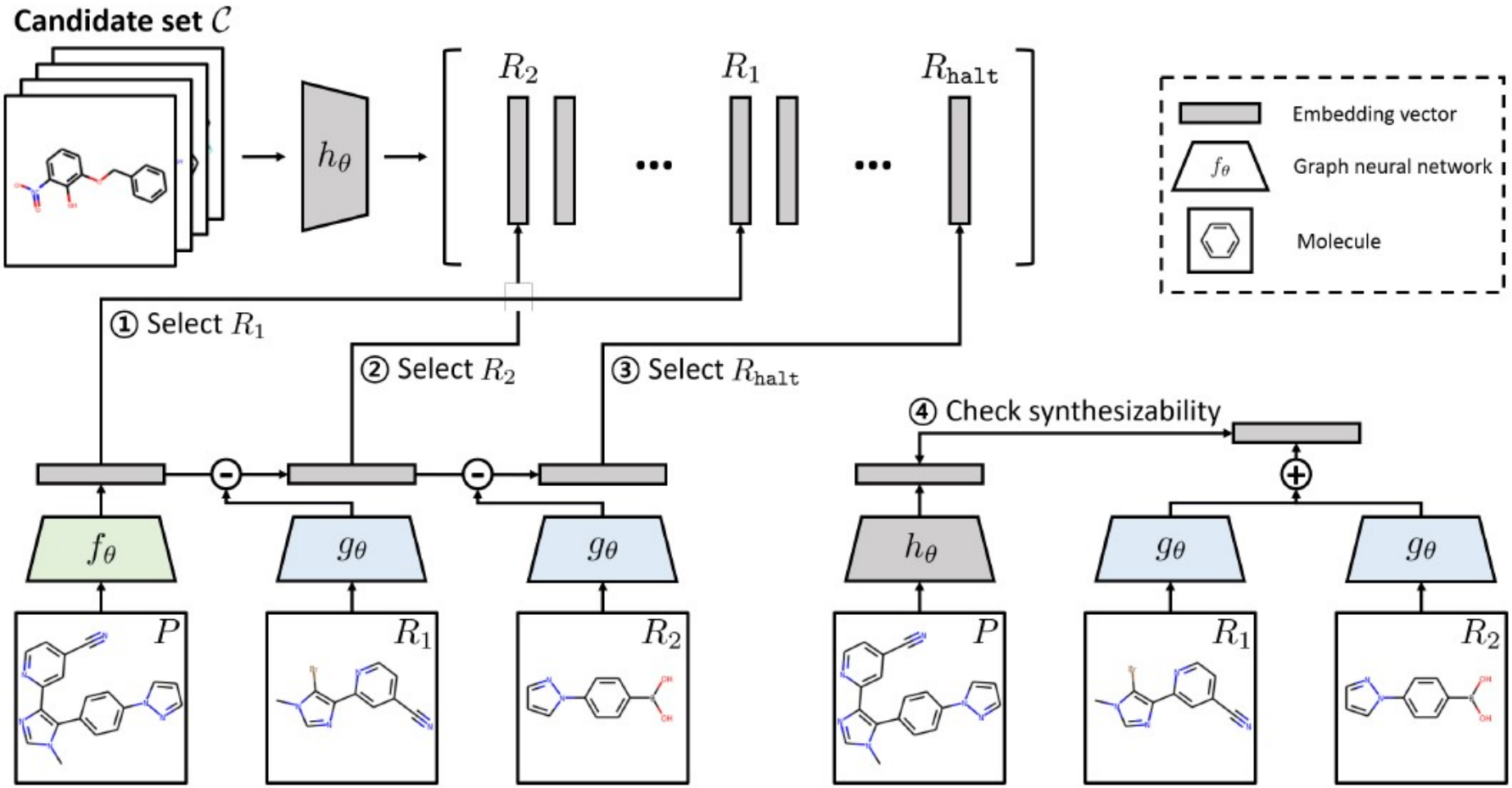

Figure 1: Illustration of the search procedure in RetCL.

- Introduction
- Search Procedure
- Training Scheme with Contrastive Learning
- Experiments

Search Procedure

- **Object**: To find a reactant-set $\mathcal{R} = \{R_1, \ldots, R_n\}$
- **Input**: The product P and the candidate set $\mathcal{C}$
- First, select each reactant R_i sequentially from the candidate set $\mathcal{C}$ based on the backward selection score $\psi(R|P, \mathcal{R}_{\text{given}})$.
- Then, repeat the first step to get many reactant-sets.
- Finally, rank the chosen reactant-sets $\mathcal{R}_1, \ldots, \mathcal{R}_T$ based on the backward selection score $\psi(R|P, \mathcal{R}_{\text{given}})$ and the forward score $\phi(P|\mathcal{R})$.

Score Design

$$\psi(R|P, \mathcal{R}_{\text{given}}) = \text{CosSim}\left(f_\theta(P) - \sum_{\mathcal{S}\in\mathcal{R}_{\text{given}}} g_\theta(S),\ h_\theta(R)\right)$$

$$\phi(P|\mathcal{R}) = \text{CosSim}\left(\sum_{R\in\mathcal{R}} g_\theta(R),\ h_\theta(P)\right)$$

where CosSim is the cosine similarity and $f_\theta, g_\theta, h_\theta$ are embedding functions from a molecule to a fixed-sized vector with parameters θ.

Score Design

The overall score on a chemical reaction $\mathcal{R} \to P$ is defined as

$$\text{score}(P, \mathcal{R}) = \frac{1}{n+2} \left(\max_{\pi \in \prod} \sum_{i=1}^{n+1} \psi(R_{\pi(i)} | P, \{R_{\pi(1)}, \ldots, R_{\pi(i-1)}\}) + \phi(P|\mathcal{R}) \right)$$

where $R_{n+1} = R_{\text{halt}}$ and $\prod$ is the space of permutations defined on the integers $1, \ldots, n+1$ satisfying $\pi(n+1) = n+1$.

- Introduction
- Search Procedure
- Training Scheme with Contrastive Learning
- Experiments

Two Classification Tasks

$$p(R|P, \mathcal{R}_{\text{given}}, \mathcal{C}) = \frac{\exp(\psi(R|P, \mathcal{R}_{\text{given}})/\tau)}{\sum_{R' \in \mathcal{C} \setminus \{P\}} \exp(\psi(R'|P, \mathcal{R}_{\text{given}})/\tau)}$$

$$q(P|\mathcal{R}, \mathcal{C}) = \frac{\exp(\phi(P|\mathcal{R})/\tau)}{\sum_{P' \in \mathcal{C} \setminus \mathcal{R}} \exp(\phi(P'|\mathcal{R})/\tau)}$$

where τ is a hyperparameter for temperature scaling and $\mathcal{C}$ is the given candidate set of molecules.

Loss Function

The Losses defined on a reaction of the product P and the reactant-set $\mathcal{R} = \{R_1, \ldots, R_n\}$:

$$\mathcal{L}_{\text{backward}}(P, \mathcal{R}|\theta, \mathcal{C}) = -\max_{\pi \in \prod} \sum_{i=1}^{n+1} \log\, p(R_{\pi(i)}|P, \{R_{\pi(1)}, \ldots, R_{\pi(i-1)}\}, \mathcal{C})$$

$$\mathcal{L}_{\text{forward}}(P, \mathcal{R}|\theta, \mathcal{C}) = -\log\, q(P|\mathcal{R}, \mathcal{C})$$

where $R_{n+1} = R_{\text{halt}}$ and $\prod$ is the space of permutations defined on the integers $1, \ldots, n+1$ satisfying $\pi(n+1) = n+1$.

Candidate Set

Because the denominators of $p(\cdot)$ and $q(\cdot)$ require summation over the large set of candidate set $\mathcal{C}$.
For each mini-batch of reactions $\mathcal{B}$ sampled from the training dataset:

$$\mathcal{C}_{\mathcal{B}} = \{M | \exists (\mathcal{R}, P) \in \mathcal{B} \text{ such that } M = P \text{ or } M \in \mathcal{R}\}$$

$$\mathcal{L}(\mathcal{B}|\theta) = \frac{1}{|\mathcal{B}|} \sum_{(\mathcal{R},P)\in\mathcal{B}} (\mathcal{L}_{\text{backward}}(P, \mathcal{R}|\theta, \mathcal{C}_{\mathcal{B}}) + \mathcal{L}_{\text{forward}}(P, \mathcal{R}|\theta, \mathcal{C}_{\mathcal{B}}))$$

Hard Negative Mining

$$\tilde{\mathcal{C}_{\mathcal{B}}} = \mathcal{C}_{\mathcal{B}} \cup \bigcup_{M \in \mathcal{C}_{\mathcal{B}}} \{\text{Top-K nearest neighbors of M from } \mathcal{C}\}$$

where K is a hyperparameter controlling hardness of the contrastive task. The nearest neighbors are defined with respect to the cosine similarity on $\{h_\theta(M)\}_{M \in \mathcal{C}}$.

- Introduction
- Search Procedure
- Training Scheme with Contrastive Learning
- Experiments

Datasets

- Datasets: USPTO-50k.
- The candidate set: choose the candidate set of commercially available molecules $\mathcal{C}$ as the all reactants in the entire USPTO database
- Evaluation metric: top-k exact match accuracy.

The top-k exact match accuracy

Category	Method	Top-1	Top-3	Top-5	Top-10	Top-20	Top-50
Reaction type is unknown							
Template-free	Transformer (Karpov et al., 2019)	37.9	57.3	62.7	-	-	-
	SCROP (Zheng et al., 2019)	43.7	60.0	65.2	68.7	-	-
	Transformer (Chen et al., 2019)	44.8	62.6	67.7	71.1	-	-
	G2Gs (Shi et al., 2020)	**48.9**	**67.6**	**72.5**	**75.5**	-	-
Template-based	retrosim (Coley et al., 2017b)	37.3	54.7	63.3	74.1	82.0	85.3
	neuralsym (Segler & Waller, 2017)	44.4	65.3	72.4	78.9	82.2	83.1
	GLN (Dai et al., 2019)	**52.5**	**69.0**	**75.6**	**83.7**	**89.0**	**92.4**
Selection-based	Bayesian-Retro (Guo et al., 2020)	47.5	67.2	77.0	80.3	-	-
	RetCL (Ours)	**71.3**	**86.4**	**92.0**	**94.1**	**95.0**	**96.4**
Reaction type is given as prior							
Template-free	seq2seq (Liu et al., 2017)	37.4	52.4	57.0	61.7	65.9	70.7
	Transformer† (Chen et al., 2019)	54.1	70.0	74.2	77.8	80.4	83.3
	SCROP (Zheng et al., 2019)	59.0	74.8	78.1	81.1	-	-
	G2Gs (Shi et al., 2020)	**61.0**	**81.3**	**86.0**	**88.7**	-	-
Template-based	retrosim (Coley et al., 2017b)	52.9	73.8	81.2	88.1	91.8	92.9
	neuralsym (Segler & Waller, 2017)	55.3	76.0	81.4	85.1	86.5	86.9
	GLN (Dai et al., 2019)	**64.2**	**79.1**	**85.2**	**90.0**	**92.3**	**93.2**
Selection-based	Bayesian-Retro (Guo et al., 2020)	55.2	74.1	81.4	83.5	-	-
	RetCL (Ours)	**78.9**	**90.4**	**93.9**	**95.2**	**95.8**	**96.7**

Figure 2: The top-k exact match accuracy (%).