Enhancing Semantic Code Search with Multimodal Contrastive Learning and Soft Data Augmentation

Ensheng Shi\textsuperscript{a,†} Wenchao Gu\textsuperscript{b,‡} Yanlin Wang\textsuperscript{c,§} Lun Du\textsuperscript{c}  
Hongyu Zhang\textsuperscript{d} Shi Han\textsuperscript{c} Dongmei Zhang\textsuperscript{e} Hongbin Sun\textsuperscript{a,§}  
\textsuperscript{a}Xi’an Jiaotong University \hspace{1cm} \textsuperscript{b}The Chinese University of Hong Kong \hspace{1cm} \textsuperscript{c}Microsoft Research \hspace{1cm} \textsuperscript{d}The University of Newcastle \hspace{1cm} \textsuperscript{e}The Chinese University of Hong Kong  
s1530129650@stu.xjtu.edu.cn, wcg@cs.cuhk.edu.hk  
\{yanlwang, lun.du, shihan, dongmeiz\}@microsoft.com  
hongyu.zhang@newcastle.edu.au, hsun@mail.xjtu.edu.cn

ABSTRACT

Code search aims to retrieve the most semantically relevant code snippet for a given natural language query. Recently, large-scale code pre-trained models such as CodeBERT and GraphCodeBERT learn generic representations of source code and have achieved substantial improvement on code search task. However, the high-quality sequence-level representations of code snippets have not been sufficiently explored. In this paper, we propose a new approach with multimodal contrastive learning and soft data augmentation for code search. Multimodal contrastive learning is used to pull together the representations of code-query pairs and push apart the unpaired code snippets and queries. Moreover, data augmentation is critical in contrastive learning for learning high-quality representations. However, only semantic-preserving augmentations for source code are considered in existing work. In this work, we propose to do soft data augmentation by dynamically masking and replacing some tokens in code sequences to generate code snippets that are similar but not necessarily semantic-preserving as positive samples for paired queries. We conduct extensive experiments to evaluate the effectiveness of our approach on a large-scale dataset with six programming languages. The experimental results show that our approach significantly outperforms the state-of-the-art methods. We also adapt our techniques to several pre-trained models such as RoBERTa and CodeBERT, and significantly boost their performance on the code search task.

1 INTRODUCTION

Code search plays an important role in software development and maintenance [55, 64]. To implement a certain functionality, developers often search and reuse previously-written relevant code from open source repositories such as Github or from a large local codebase [53, 60, 79]. With the advent of a large corpus of open source code, finding the relevant code snippets among massive code snippets based on natural language queries has become one of the key challenges in this field [1].

Early studies [46, 51–53] on code search mainly focus on the lexical information of the code snippets and leverage information retrieval methods to find the relevant code snippets for a given search query. Deep end-to-end code search approaches [6, 14, 23, 26, 29, 42, 44, 45, 63, 70, 77, 82] leverage the neural networks to embed code snippets and queries into a shared high-dimensional vector space and measure their semantic similarity through vector distances. Recently, pre-trained source code models [17, 27], which are pre-trained on a large multi-programming-language dataset, improve the understanding of code semantics and achieve better code search performance. For example, GraphCodeBERT [27] improves the performance of end-to-end supervised approaches [33] by about 70% on MRR and achieves an average MRR of 0.713 on the CodeSearchNet [33] dataset. Nevertheless, code search performance still needs to be further improved.

Contrastive learning approaches [28], which pull together the representations of relevant (positive) samples and push apart the irrelevant (negative) samples, are well suitable for solving the code search problem, which expects the paired code snippet and query to have close representations and unpaired code snippet and query to have different representations. Recently, some studies [5, 37] apply contrastive learning approaches for unsupervised code representation learning and achieve good performance on downstream tasks. Their core idea is to generate the semantic-preserving code transformations as a form of program augmentation and pre-train a model to identify semantically similar and dissimilar code snippets through the approaches related to contrastive learning. At the same time, some studies [78, 80] also apply contrastive learning to learn the cross-modal representations for videos/images and texts.

For contrastive learning, data augmentation is critical for learning high-quality representations [7, 8, 38, 67]. Previous studies [5, 37] on unsupervised code representation learning mainly use the semantic-preserving program transformations (§ 5.2.2) to augment source code. However, some program transformations require certain constraints and such constraints may not be always satisfied in one code snippet. For example, the Statements Permutation [5] requires that the swapped two statements have no data dependency on each other in a basic block and Loop Exchange [5] requires that...
there is at least one loop statement in code snippets. Some simple semantic-preserving program transformations such as Variable Renaming have a limited benefit for code search task [5, 37].

In this work, we adapt the multimodal contrastive learning framework [78] and soft data augmentation (SoDa) for code search. Our proposed model CoCoSoDa (stand for semantic Code search with multimodal Contrastive learning and Soft Data augmentation) aims to minimize the distance between the representations of code-query pair and maximize the distance between representations of the query (code snippet) and other many unpaired code snippets (queries). To learn the better sequence-level representation of the code snippet and query instead of focusing on token-level semantic modeling\(^1\), we propose SoDa adopting dynamic masking technique [48] to generate similar code snippets or queries (shown in the left of Figure 1) as a data augmentation approach. SoDa is easy to build and scale to any programming language. The overall framework of CoCoSoDa are shown in Figure 1. Specifically, first, we sample randomly \( r \) (\( r \) is stable in 5% - 25%) of the tokens from the code snippet or query. These tokens are pre-processed as follows: 80% are replaced with a [MASK] token, 10% with a random token from the vocabulary built from the training dataset in advance, and 10% left unchanged. Second, we use a large-scale code pre-trained model GraphCodeBERT to initialize the code/query encoder and momentum code/query encoder, and feed the original and augmented samples (codes or queries) to the encoders and momentum encoders, respectively, to obtain the representation of samples. Third, a multimodal contrastive learning loss is used to pull close semantically similar representations and push apart dissimilar representations.

We evaluate the effectiveness of CoCoSoDa on a large-scale dataset CodeSearchNet [33] with six programming language (Ruby, JavaScript, Go, Python, Java, PHP) and compare CoCoSoDa with eight state-of-the-art approaches. We also apply CoCoSoDa to three large-scale pre-trained models, including natural language pre-trained model RoBERTa [48], code pre-trained models RoBERTa (code) [17] and CodeBERT [17]. In addition, we conduct the ablation study to study the effectiveness of each component of CoCoSoDa. We also assign different hyperparameters to check their impact on code search. Experimental results show that: (1) CoCoSoDa significantly outperforms existing SOTA approaches on the code search task (§ 5.1). (2) SoDa is a simple yet effective data augmentation approach to learn the sequence-level representation compared with the semantic-preserving transformations (§ 5.2.2). (3) CoCoSoDa can be easily adapted to other pre-trained models and significantly boost their performance (§ 5.3). (4) CoCoSoDa and GraphCodeBERT can benefit from larger queue size and mini-batch size, respectively (§ 5.4). (5) CoCoSoDa performs stably over a range of hyperparameters: learning rate is from \( 5e^{-6} \) to \( 5e^{-3} \), momentum coefficient \( m \) is between 0.930 and 0.999, masked ratio \( r \) is from 5% to 25%, and temperature hyperparameter \( \tau \) varies from 0.01 to 0.1 (§ 5.4).

We summarize the contributions of this paper as follows:

- We propose a new approach incorporating multimodal contrastive learning for code search task. It can pull together the representations of matched code-query pairs and push apart the representations of unmatched code-query pairs.
- We propose a simple yet effective soft data augmentation method SoDa that utilizes sequence dynamic masking for data augmentation. Unlike existing approaches that use hard data augmentation with semantic-preserving code transformations, we demonstrate that soft data augmentation can boost code search performance by learning richer sequence-level representations of codes and queries.
- We conduct extensive experiments to evaluate the effectiveness of our approach on a large-scale multi-programming-language dataset. The results show that our approach significantly outperforms baselines.

2 RELATED WORK

2.1 Code Search

Learning the representation of code is an emerging topic and has been found to be useful in many software engineering tasks, such as code summarization [35, 40, 61, 62, 81], code search [15, 25, 26, 29, 82], code completion [4, 57, 59, 65, 73], commit message generation [10, 49, 66, 68, 76]. Among them, code search plays an important role in software development and maintenance [55, 64]. Code search aims to find the most semantically matching code snippet for the given query. Traditional approaches [46, 51–53] based on retrieval information mainly focus on the lexical information of the source code and apply keywords matching methods to search the relevant code snippets for the given query. In recent years, deep learning-based approaches leverage the neural network to learn the semantic representations of the source code and natural language to improve the understanding of code snippets and queries. Gu et al. [26] is the first to use the deep neural network to encode the code and query into a shared vector space and measure the similarity of them using vector distance. Subsequently, various types of model structures are applied to code search, including sequential models [23, 29, 63, 70, 77], convolutional neural network [42, 44, 82], tree neural network [70], graph models [45, 70], and transformers [14, 82].

Recently, large-scale code pre-trained models [17, 27], which are pre-trained on a massive source code dataset, improve the understanding of code semantics and achieve significant improvements in code search task. RoBERTa (code) [17], CodeBERT [17] and GraphCodeBERT [27] are all based on RoBERTa architecture but pre-trained with different pre-trained tasks. RoBERTa (code) is pre-trained with masked language modeling (MLM), which is to predict the original tokens which are masked. CodeBERT is pre-trained with MLM and replaced token detection (RTD), which uses a discriminator to identify the replaced token. GraphCodeBERT takes source code paired with summarization and the corresponding data flow as the input and is pre-trained with MLM, data flow edge prediction, and node alignment tasks. Our approach can be easily applied to these pre-trained models and boost their performance.

2.2 Contrastive Learning for Code Representation Learning

Contrastive learning approaches [28], which pull close the similar representations and push apart different representations, have been
Figure 1: The framework of CoCoSoDa.

3 PROPOSED APPROACH

In this section, we illustrate our model CoCoSoDa for code search. The overall architecture is shown in Figure 1. In general, our approach adopts a pre-trained model as the base code/query encoder and finetunes it based on multimodal contrastive learning and soft data augmentation. To avoid using a large mini-batch size that requires expensive computation resources and could bring generalization issues, we adopt the momentum contrastive learning algorithm as our main framework. CoCoSoDa is comprised of the following four components:

- **Pre-trained code/query encoder** captures the semantic information of a code snippet or natural language query and maps it into a high-dimensional embedding. We use the GraphCodeBERT, a bimodal pre-trained model for both programming and natural languages, as the code/query encoder.
- **Momentum code/query encoder** encodes the samples (code snippets or queries) of current and previous mini-batches to enrich the negative samples. To keep the consistency of representation, the momentum encoder is updated by linear interpolation of the encoder and momentum encoder.
• **Soft data augmentation** is to dynamically mask or replace some tokens in a sample (code/query) to generate a similar sample as a form of data augmentation. This is a simple yet effective augmentation approach compared to static masking [11, 48] or other semantic-preserving program transformations [5, 37].

• **Multimodal contrastive learning loss function** is defined for contrastive learning. The optimized objective is to minimize the distance of the representations of the paired code snippet and query and maximize the distance between the query (code snippet) and other unpaired code snippets (queries).

### 3.1 Pre-trained Encoder and Momentum Encoder

In this section, we introduce the base model, input samples, output representation and update mechanism of encoder and momentum encoder. As the pre-trained models such as GraphCodeBERT [17] have achieved substantial improvement in code search, we take GraphCodeBERT as the base code/query encoder. Following previous study [27], to obtain the whole sequence representations of the query/code, we insert a special token [CLS] at the beginning of the input code/query sequence and using the embedding of [CLS] at the last layer as the whole sequence-level representation. Then we use a 2-layer Multi-Layer Perception (MLP) projector to map the sequence-level representation of code/query to shared latent space.

In the MoCo [30] framework, there is a momentum encoder encoding the samples of the current and previous mini-batches. Specifically, the momentum encoder maintains a queue by enqueuing the samples in the current mini-batch and dequeuing the samples in the oldest mini-batch. Here, we also take GraphCodeBERT as the momentum code/query encoder. The difference of the update mechanism between the encoder and momentum encoder is that the encoder is updated by the back-propagation algorithm while the momentum encoder is updated by linear interpolation of the encoder and the momentum encoder. Thus, compared with the memory bank approach [75], which fixes and saves the representations of all samples of the training dataset in advance, the momentum encoder can generate consistent representations and has been demonstrated to be effective [30]. For the end-to-end approach [23, 27], it has one encoder and takes other samples in the current mini-batch as negative samples. Thus, it requires a large mini-batch size in order to expand the number of negative samples [30]. Therefore, end-to-end approaches require larger memory computational resources than our approach.

We denote the parameters of the code encoder as \( \theta_{ce} \) and the momentum code encoder as \( \theta_{mce} \), with parameters being the weights of GraphCodeBERT and 2-layer MLP projector. Therefore, \( \theta_{mce} \) is updated by:

\[
\theta_{mce} = m \theta_{mce} + (1 - m) \theta_{ce}
\]

where \( m \in [0, 1) \) is a momentum coefficient. Similarly, we denote the parameters of the query encoder and moment query encoder as \( \phi_{qe} \) and \( \phi_{mqe} \). Then \( \phi_{mqe} \) is updated by:

\[
\phi_{mqe} = m \phi_{mqe} + (1 - m) \phi_{qe}
\]

Both \( \theta_{ce} \) and \( \phi_{qe} \) are learnable parameters and updated by the back-propagation algorithm.

### 3.2 Soft Data Augmentation

In this section, we introduce soft data augmentation (SoDa), which is a simple data augmentation approach without external constraints for source code or queries. We first introduce how to obtain soft data augmentation and then introduce how to use the augmented data.

As shown in the left of Figure 1, we adopt the dynamic masking technique [48] to implement SoDa. Specifically, we randomly select \( r \) (default value is 15%) of the tokens from code snippets or queries. Then the tokens are pre-processed as follows: 80% of them are replaced with a [MASK] token, 10% with a random token, and 10% left unchanged. Here, dynamic means that in data processing, the masking and replacement operations are performed at each iteration rather than only performed once as in static making [48]. We will further study the impact of these two strategies in § 5.2.2.

We denote the SoDa module as \( G_{so\text{da}} \) which performs the dynamic masking operation for the given input sequence. Specifically, we first perform the dynamic masking operation for the code snippets \( C = (c_1, ..., c_n) \) and queries \( Q = (q_1, ..., q_b) \) in a mini-batch with batch size \( bs \) by:

\[
c_i^* = G_{so\text{da}}(c_i), \quad q_i^* = G_{so\text{da}}(q_i) \quad (i = 1, ..., bs)
\]

where \( c_i^* \) and \( q_i^* \) are the augmented samples of the code snippet \( c_i \) and query \( q_i \), respectively. Then code snippet \( c_i \) and query \( q_i \) are fed into the code/query encoder and augmented samples \( c_k^* \) and \( q_k^* \) (\( k = 1, ..., K \) and \( K \) is the queue size) in the current and previous mini-batches are fed to the momentum code/query encoder by:

\[
\begin{align*}
\nu_c &= \tilde{f}_{theta_c}(c_i), \\
\nu_q &= \tilde{f}_{theta_q}(q_i), \\
\nu_{c_k} &= \tilde{f}_{theta_{mce}}(c_k^*) \\
\nu_{q_k} &= \tilde{f}_{theta_{mqe}}(q_k^*)
\end{align*}
\]

where, \( \nu_c, \nu_q, \nu_{c_k} \), and \( \nu_{q_k} \) are final sequence-level representations of the code snippet \( c_i \), query \( q_i \), augmented code snippet \( c_k^* \), and augmented query \( q_k^* \), respectively.

### 3.3 Multimodal Contrastive Learning Loss Function

Multimodal contrastive learning loss function is used to optimize the parameters of the model. Specifically, given a query \( q_i \), we denote the paired \( c_i \) or \( c_i^* \) as \( c_i^* \) and unpaired \( c_k^* \) as \( c_k^* \) (\( i = 1, ..., bs \) and \( k = 1, ..., K \)). For the query \( q_i \), with similarity measured by dot product [27], we define the multimodal contrastive learning loss [56, 70] as:

\[
L_{q_i} = -\log \frac{\exp(\nu_{q_i} \cdot \nu_{q_i}^*/r)}{\exp(\nu_{q_i} \cdot \nu_{q_i}^*/r) + \sum_{k=1}^{K} \exp(\nu_{q_i} \cdot \nu_{c_k}^*/r)}
\]

where \( r \) is the temperature hyperparameter [30, 75] and is set to 0.07 following previous works [30, 37]. Intuitively, this optimization objective is to maximize the semantic similarity of the query and its paired code snippet and minimize the semantic similarity of the query and its unpaired code snippets. In the same way, for a code snippet \( c_i \), we define multimodal contrastive learning loss as:

\[
L_{c_i} = -\log \frac{\exp(\nu_{c_i} \cdot \nu_{c_i}^*/r)}{\exp(\nu_{c_i} \cdot \nu_{c_i}^*/r) + \sum_{k=1}^{K} \exp(\nu_{c_i} \cdot \nu_{q_k}^*/r)}
\]

where \( q_i^* \) is the paired query of input code snippet \( c_i \), and \( q_k^- \) denotes the unpaired query. To this end, the overall multimodal
We conduct experiments on a large-scale benchmark dataset CodeSearchNet [33] as used in Guo et al. [27]. It contains six programming languages - Ruby, JavaScript, Go, Python, Java, and PHP. This dataset is widely used in previous studies [14, 17, 23, 26, 27, 29, 33, 42, 44, 45, 63, 70, 70, 77, 82]. The statistics of the dataset are shown in Table 1. The training data contains code-query pairs, while valid and test set only have natural language queries. Following previous studies [26, 27, 32], the model is to retrieve the correct code snippets from the Candidate Codes for the given queries when performing the evaluation.

### 4 EXPERIMENTAL DESIGN

#### 4.1 Datasets

We conduct experiments on a large-scale benchmark dataset CodeSearchNet [33] as used in Guo et al. [27]. It contains six programming languages - Ruby, JavaScript, Go, Python, Java, and PHP. This dataset is widely used in previous studies [14, 17, 23, 26, 27, 29, 33, 42, 44, 45, 63, 70, 70, 77, 82]. The statistics of the dataset are shown in Table 1. The training data contains code-query pairs, while valid and test set only have natural language queries. Following previous studies [26, 27, 32], the model is to retrieve the correct code snippets from the Candidate Codes for the given queries when performing the evaluation.

#### 4.2 Baselines

To evaluate the effectiveness of our approach, we compare CoCoSoDa with four deep end-to-end approaches: NBow, CNN, BiRNN, and SelfAtnn proposed by Husains et al. [33] and four pre-training-based approaches: RoBERTa [48], RoBERTa (code) [17], CodeBERT [17], and GraphCodeBERT [27].

- **NBow, CNN, BiRNN, and SelfAtnn** use various encoding models such as neural bag-of-words [36], 1D convolutional neural network [39], bi-directional GRU [9], and multi-head attention [69] to obtain the representation of code snippets and queries. Then they use inner product of the representations of code snippets and queries to measure their similarity.

- **RoBERTa and RoBERTa (code)** are built on a multi-layer bidirectional Transformer [69] encoder and pre-trained with masked language modeling (MLM) task, which is to predict the original tokens of the masked positions. The formal is pre-trained on natural language corpus [48], while the latter is pre-trained on source code corpus [33].

- **CodeBERT** is a bimodal pre-trained model pre-trained with large-scale code-text pairs on two tasks: MLM and replaced token detection (RTD), which uses a discriminator to identify the replaced token.

- **GraphCodeBERT** considers the code structure information for pre-training and achieves the state-of-art results among code search baselines. The pre-training tasks include MLM, data flow edge prediction, and node alignment.

In our experiments, we train the four deep end-to-end approaches from scratch, and for the four pre-trained approaches, we initialize them with the pre-trained models and fine-tune them according to their original paper descriptions and the available source code [17, 27, 33].

#### 4.3 Experimental Settings

Following GraphCodeBERT [17], we use Transformer with 12 layers, 768 dimensional hidden states, and 12 attention heads. The vocabulary sizes of code and queries are set to 50,265. Max sequence lengths of code snippets and queries are 128 and 256, respectively. For optimizer, we use AdamW with the learning rate 2e-5. Following previous studies [27, 32], the code encoder and query encoder share parameters to reduce the number of total parameters. Following MoCo [30], the temperature hyperparameter $\tau$ is set as 0.07 and momentum coefficient $m$ is 0.999. The queue size is set to 4096 for relatively larger datasets (Go, Python, Java, and PHP) and 1024 for relatively smaller datasets (Ruby and JavaScript). The batch size is 64 and the maximum number of epochs is 20. In addition, we run the experiments 3 times with random seeds 0,1,2 and display the mean value in the paper. All experiments are conducted on a machine with 252 GB main memory and 4 Tesla V100 32GB GPUs.

#### 4.4 Evaluation Metrics

We measure the performance of our approach using four metrics: MRR (Mean Reciprocal Rank) and top-k recall ($R@k$, $k=1,5,10$), which are widely used in previous studies [14, 17, 23, 26, 27, 29, 33, 42, 44, 45, 63, 70, 70, 77, 82].

MRR is the average of reciprocal ranks of the correct code snippets for given queries $Q$. $R@k$ measures the percentage of queries that the paired code snippets exist in the top-k returned ranked lists. They are calculated as follows:

$$MRR = \frac{1}{|Q|} \sum_{i=1}^{|Q|} \frac{1}{\text{Rank}_i}$$

$$R@k = \frac{1}{|Q|} \sum_{i=1}^{|Q|} \delta(\text{Rank}_i \leq k)$$

where $\text{Rank}_i$ is the rank of the paired code snippet related to the $i$-th query. $\delta$ is an indicator function that returns 1 if $\text{Rank}_i \leq k$ otherwise returns 0.

#### 5 EXPERIMENTAL RESULTS

##### 5.1 RQ1: What is the Effectiveness of CoCoSoDa?

We evaluate the effectiveness of our model CoCoSoDa by comparing it to four recent deep end-to-end code search models (NBow, CNN, BiRNN, and SelfAtnn) and four pre-trained models (RoBERTa, RoBERTa (code), CodeBERT, and GraphCodeBERT) introduced in § 4.2 on the CodeSearchNet dataset with six programming languages. The experimental results are shown in Table 2. We display the result with MRR metric due to space limitation. Results under other metrics are put on the anonymous replication package. The following conclusion hold for other metrics.

We can see that the four pre-trained models perform better than the four deep end-to-end models trained from scratch, which shows the effectiveness of the pre-training technique. Since GraphCodeBERT considers the data flow information of source code, it

| Language | Training | Validation | Test | Candidate Codes |
|----------|----------|------------|------|-----------------|
| Ruby     | 24,927   | 1,400      | 1,261| 4,360           |
| JavaScript | 58,025   | 3,885      | 3,291| 13,981          |
| Java     | 164,923  | 5,183      | 10,955| 40,347         |
| Go       | 167,288  | 7,325      | 8,122| 28,120          |
| PHP      | 241,241  | 12,982     | 14,014| 52,660         |
| Python   | 251,820  | 13,914     | 14,918| 43,827         |

contrastive learning loss function for a mini-batch is:

$$L = \sum_{i=1}^{b_t} (L_q + L_c)$$

We apply AdamW [50] algorithm to optimize the loss functions.
Table 2: The performance of different approaches on MRR. JS is short for JavaScript. CoCoSoDa outperforms significantly (statistical significance $p < 0.01$).

| Model      | Ruby | JS | Go | Python | Java | PHP | Avg. |
|------------|------|----|----|--------|------|-----|------|
| NBow       | 0.162| 0.157| 0.330| 0.161| 0.171| 0.152| 0.189|
| CNN        | 0.276| 0.224| 0.680| 0.242| 0.263| 0.260| 0.324|
| BiRNN      | 0.213| 0.193| 0.688| 0.290| 0.304| 0.338| 0.338|
| SelfAtt    | 0.275| 0.287| 0.723| 0.398| 0.404| 0.426| 0.419|
| RoBERTa    | 0.587| 0.523| 0.855| 0.590| 0.605| 0.561| 0.620|
| RoBERTa (code) | 0.631| 0.57 | 0.864| 0.621| 0.636| 0.581| 0.650|
| CodeBERT   | 0.679| 0.621| 0.885| 0.672| 0.677| 0.626| 0.693|
| GraphCodeBERT | 0.703| 0.644| 0.897| 0.692| 0.691| 0.649| 0.713|
| CoCoSoDa   | **0.712**| **0.666**| **0.913**| **0.717**| **0.726**| **0.669**| **0.734**|

performs the best among four pre-trained models. CoCoSoDa takes GraphCodeBERT as the base code/query encoder and outperforms GraphCodeBERT\(^2\) because we apply multimodal contrastive learning and soft data augmentation. Both of them contribute to learning a good sequence-level representation.

## 5.2 RQ2: How Much do Different Components Contribute?

In this section, we study the contributions of the two main components: multimodal contrastive learning framework and soft data augmentation. We take the Java dataset as an example since Java is the most studied language in the recent 81 studies [47].

### 5.2.1 Multimodal contrastive learning ablation study

To study the effectiveness of the multimodal contrastive learning framework (short for MultiMoCo), we remove the soft data augmentation component from Figure 1 and feed the original samples (code snippets and queries) to the code/query encoder and momentum code/query encoder. We evaluate the pre-trained GraphCodeBERT, fine-tuned GraphCodeBERT and MultiMoCo in terms of the four metrics: MRR, R@1, R@5, R@10.

The experimental results are shown in Table 3, we can see that pre-trained GraphCodeBERT without fine-tuning performs poorly due to the representation degeneration problem [18, 41]. That is, the high-frequency tokens dominate the sequence representation [41], resulting in the poor sequence-level semantic representation of the code snippet and query.

For GraphCodeBERT (fine-tuned) and MultiMoCo, they are all optimized by minimizing the distance of the representation of paired code snippet and query. However, MultiMoCo aims to maximize the distance between representations of the queries and other unpaired code snippets in the queue (including the current and previous mini-batches), while GraphCodeBERT only aims to maximize the distance between representations of the queries and other unpaired code snippets in the current mini-batch. We discuss the impact of the queue size and mini-batch size on the model performance in § 5.4. Since the queue is decoupled with mini-batch,

### Table 3: The gain of multimodal contrastive learning.

| Model                        | MRR  | R@1  | R@5  | R@10 |
|------------------------------|------|------|------|------|
| GraphCodeBERT (Pre-trained)  | 0.004| 0.003| 0.004| 0.006|
| GraphCodeBERT (Fine-tuned)   | 0.691| 0.593| 0.814| 0.865|
| MultiMoCo                    | **0.710**| **0.616**| **0.829**| **0.874**|

### Table 4: The gain of data augmentation technique for GraphCodeBERT and MultiMoCo.

| Base Model       | Data Augmentation | MRR  | R@1  | R@5  | R@10 |
|------------------|-------------------|------|------|------|------|
| GraphCodeBERT    | 6-SP Transformations | 0.707| 0.597| 0.817| 0.869|
|                  | SoDa (static)     | 0.704| 0.608| 0.824| 0.871|
|                  | SoDa (dynamic)    | 0.715| 0.622| 0.833| 0.879|
| MultiMoCo        | 6-SP Transformations | 0.710| 0.615| 0.832| 0.879|
|                  | SoDa (static)     | 0.707| 0.612| 0.828| 0.875|
|                  | SoDa (dynamic)    | **0.726**| **0.630**| **0.845**| **0.887**|
|                  | Variable Renaming | 0.709| 0.614| 0.614| 0.876|
|                  | Unused Statement  | 0.708| 0.611| 0.830| 0.878|
|                  | Permute Statement | 0.711| 0.616| 0.834| 0.877|
|                  | Loop Exchange     | 0.713| 0.618| 0.832| 0.877|
|                  | Switch to If      | 0.711| 0.615| 0.834| 0.879|
|                  | Boolean Exchange  | 0.712| 0.617| 0.835| 0.879|

MultiMoCo can use the large queue size with more negative samples, learn the good representations, and perform well in the code search task.

### 5.2.2 Soft data augmentation ablation study

In this section, we study the impact of different data augmentations. We first briefly describe these augmentations including dynamic masking (denoted as SoDa (dynamic)), static masking (denoted as SoDa (static)), and six semantic-preserving program transformations (denoted as 6-SP Transformations) and then analyze them based on the experimental results.

SoDa (dynamic) and SoDa (static) introduced in § 3.2 are to obtain the similar but not semantically equivalent code snippets, while previous studies [5, 37] try to use the semantic-preserving transformations. The following are the popular 6-SP Transformations:

- **Variable Renaming** is to rename a variable with a random token selected from variable vocabulary built from training set.
- **Permute Statement** is to swap two statements, which have no data dependency on each other in a basic block.
- **Unused Statement** is to insert an dead code snippet such as `log.warn(...)`.
- **Loop Exchange** is to replace a loop with an equivalent loop.
- **Switch to If** is to replace a switch statement with an equivalent if statement.
- **Boolean Exchange** is to switch the value of a boolean variable and propagates this change in the method. For example, changing "boolean removed = false; if (!removed) log.warn(...)" to "boolean removed = true; if (removed) log.warn(...)"

\( ^{2}\)We conducted t-test between our model and four pre-trained models, and the results show the improvements are statistically significant with $p < 0.01$. 


We use the *mpa* tool [58] to obtain the above semantic-preserving program transformations.

Next, we introduce how to use augmented samples. One way is to directly treat augmentation code snippets (queries) paired their original queries (code snippets) as augmented training samples and feed them to GraphCodeBERT. This way ignores the relationship between the augmented samples and original samples. Another way is to integrate different data augmentation approaches into MultiMoCo. Specifically, as shown in Figure 1, we replace the soft data augmentation module with different data augmentation approaches to generate the positive samples $q_k^*$ and $q_k^r$ and feed them to momentum code encoder and momentum query encoder, respectively.

All experimental results are shown in Table 4. For GraphCodeBERT (first row of Table 4), all data augmentation can improve the performance of the model in terms of all the four metrics: MRR, R@1, R@5, and R@10 because they can enrich the data diversity. However, 6-SP Transformations and SoDa (static) have a limited benefit for the model MultiMoCo (second row of Table 4). For example, compared with MultiMoCo, MultiMoCo based on 6-SP Transformations improves the performance on R@5 and R@10, and MultiMoCo based on SoDa (static) improves the performance on R@10. However, MultiMoCo based on SoDa (dynamic) boosts the performance of the MultiMoCo on four metrics. These results demonstrate the effectiveness of the SoDa (dynamic).

Furthermore, we conduct the ablation study on the effectiveness of each semantic-preserving program transformation individually when integrated into MultiMoCo. The six semantic-preserving program transformations can be generally divided into three groups according to the changed information: Variable Renaming and Unused Statement change the lexical information of the code snippets; Statement Permutation, Loop Exchange and Switch to If change the structure information; and Boolean Exchange changes the logic of source code. The experimental results are shown in the third row of Table 4, we can see that each semantic-preserving program transformation has a slight impact for MultiMoCo. Specifically, the simple program transformations—Variable Renaming and Unused Statement only change the lexical information of source code and degrade the performance of the MultiMoCo. Other transformations (such as Switch to If and Boolean Exchange) are to change the structure and logic information of source code and improve the performance of the MultiMoCo. Thus, increasing the diversity of source code data structures and logic is beneficial for learning a good representation of source code.

In summary, we can conclude that SoDa (dynamic) is a simple yet effective data augmentation and is better than SoDa (static) and 6-SP Transformations for code search. For each semantic-preserving program transformation, increasing the diversity of code structures and logic information has more gains on code search than changing the lexical information of source code.

### 5.3 RQ3: What is the Performance of Our Approach on other Pre-trained Models?

We further study the performance of our approach on the other three pre-trained models introduced in § 4.2, including a natural language pre-trained model RoBERTa and two source code pre-trained models RoBERTa (code) and CodeBERT. Specifically, we use these pre-trained models as the code/query encoders and momentum code/query encoders in Figure 1. For the input code snippet and query sequence, we insert a special token [CLS] and take the embedding of [CLS] as the sequence-level representations of the code snippet or query. Experimental settings are same as in § 4.3.

The results are shown in Table 5. RoBERTacoeCosSoDa, RoBERTa (code)CoCosSoDa and CodeBERTCoCosSoDa significantly outperform RoBERTa, RoBERTa (code) and CodeBERT respectively on all the six programming languages in terms of four metrics. These results demonstrate that our approach can be generalized to other pre-trained models and boost their performance. Besides, RoBERTa, which is pre-trained on the natural language corpus and fine-tuned with our method, achieves comparable performance with GraphCodeBERT on Go dataset. RoBERTa (code)CoCosSoDa and CodeBERTCoCosSoDa also slightly outperform GraphCodeBERT on Go, Python and Java datasets. These indicate that our approach is orthogonal to the pre-trained technique and can lead to independent improvements.

![Figure 2: The performance of CoCosSoDa under different queue sizes and GraphCodeBERT under mini-batch sizes.](image-url)

### 5.4 RQ4 :What is the Impact of Different Hyperparameters?

In this section, we study the impact of different hyperparameters: queue size $K$, learning rate, momentum coefficient $m$, masking ratio $r$, and temperature hyperparameter $\tau$.

We show the performance of CoCosSoDa on different queue sizes and GraphCodeBERT (fine-tuned) on different mini-batch sizes in Figure 2. We can see that CoCosSoDa and GraphCodeBERT benefit from larger queue size and mini-batch size respectively, which is consistent with previous studies [7, 30]. However, GraphCodeBERT is limited by the larger mini-batch size because large mini-batch optimization could bring some new challenges [22, 43], especially on the generalization issue. Besides, a larger mini-batch requires more memory resources. Here, the max mini-batch size of 32GB GPU can afford is 64 for GraphCodeBERT.

We also evaluate our model CoCosSoDa on different learning rates, momentum coefficient $m$, masking ratio $r$, and temperature hyperparameter $\tau$. The experimental results are shown in Figure 3. From the results of varying learning rate (the top left of Figure 3), we can
Table 5: Results on other pre-trained models. RoBERTa-C is short for RoBERTa (code). CoCoSoDa improves the performance of the pre-trained models significantly (statistical significance $p < 0.01$).

| PL    | Metric | RoBERTa | RoBERTaCoCoSoDa | RoBERTa-C | RoBERTa-C CoCoSoDa | CodeBERT | CodeBERT CoCoSoDa |
|-------|--------|---------|-----------------|-----------|--------------------|----------|------------------|
| Ruby  | MRR    | 0.587   | **0.619**       | 0.631     | **0.695**          | 0.679    | **0.688**        |
|       | R@1    | 0.469   | **0.511**       | 0.524     | **0.598**          | 0.583    | **0.587**        |
|       | R@5    | 0.717   | **0.756**       | 0.761     | **0.819**          | 0.800    | **0.808**        |
|       | R@10   | 0.785   | **0.819**       | 0.821     | **0.869**          | 0.853    | **0.862**        |
| JavaScript | MRR | 0.523   | **0.562**       | 0.570     | **0.635**          | 0.621    | **0.630**        |
|       | R@1    | 0.413   | **0.448**       | 0.452     | **0.528**          | 0.514    | **0.524**        |
|       | R@5    | 0.652   | **0.696**       | 0.716     | **0.770**          | 0.752    | **0.761**        |
|       | R@10   | 0.730   | **0.774**       | 0.794     | **0.831**          | 0.814    | **0.821**        |
| Go    | MRR    | 0.855   | **0.893**       | 0.864     | **0.906**          | 0.885    | **0.903**        |
|       | R@1    | 0.800   | **0.850**       | 0.811     | **0.863**          | 0.837    | **0.859**        |
|       | R@5    | 0.926   | **0.946**       | 0.930     | **0.958**          | 0.944    | **0.955**        |
|       | R@10   | 0.949   | **0.962**       | 0.952     | **0.972**          | 0.962    | **0.971**        |
| Python | MRR | 0.590   | **0.646**       | 0.621     | **0.697**          | 0.672    | **0.695**        |
|       | R@1    | 0.480   | **0.539**       | 0.511     | **0.601**          | 0.574    | **0.595**        |
|       | R@5    | 0.727   | **0.775**       | 0.756     | **0.818**          | 0.792    | **0.814**        |
|       | R@10   | 0.793   | **0.835**       | 0.819     | **0.869**          | 0.850    | **0.867**        |
| Java  | MRR    | 0.605   | **0.657**       | 0.636     | **0.705**          | 0.677    | **0.702**        |
|       | R@1    | 0.499   | **0.553**       | 0.528     | **0.608**          | 0.580    | **0.607**        |
|       | R@5    | 0.737   | **0.785**       | 0.770     | **0.825**          | 0.796    | **0.821**        |
|       | R@10   | 0.796   | **0.842**       | 0.831     | **0.873**          | 0.852    | **0.868**        |
| PHP   | MRR    | 0.561   | **0.614**       | 0.581     | **0.636**          | 0.626    | **0.638**        |
|       | R@1    | 0.450   | **0.507**       | 0.467     | **0.528**          | 0.521    | **0.532**        |
|       | R@5    | 0.694   | **0.750**       | 0.715     | **0.770**          | 0.753    | **0.768**        |
|       | R@10   | 0.764   | **0.813**       | 0.783     | **0.831**          | 0.814    | **0.823**        |

see that performance is generally stable for small learning rate [54] (from $5e^{-6}$ to $5e^{-5}$). The learning rates that are larger than $7e^{-5}$ have obvious impacts on the model performance. The results of different momentum coefficient $m$ are shown in the top right of Figure 3. We can see that performance increases when the momentum coefficient $m$ becomes larger. This is because a large momentum coefficient is beneficial to obtain the consistent representation for the queue [30]. The momentum coefficient that is smaller than 0.9 has a significant impact on performance. These findings are consistent with the previous work [30]. From the results of varying masked ratio $r$ (the bottom left of Figure 3), we can see that the performance is insensitive to the masked ratio $r$ when the masked ratio $r$ is between 5% and 25%. A larger masked ratio such as 50% brings considerable performance degradation. It is reasonable because the larger masked ratio causes the code snippet to lose too much information. The results of different temperature hyperparameter $\tau$ are shown in the bottom right of Figure 3. We can see that performance is stable when the temperature hyperparameter $\tau$ varies from 0.01 to 0.1.

In general, our model is stably over a range of these hyperparameters (learning rate is from $5e^{-6}$ to $5e^{-5}$, momentum coefficient is between 0.930 and 0.999, masked ratio $r$ is from 5% to 25%, and temperature hyperparameter $\tau$ varies from 0.01 to 0.1).

5.5 Case Study

In this section, we show some cases to demonstrate the effectiveness of our model CoCoSoDa.

Figure 4 shows the results returned by CoCoSoDa and Graph-CodeBERT for the query "Transform a hexadecimal String to a byte array." from the Java dataset. The query includes two operation objects: hexadecimal String and byte array and one action Transform. To implement the functionality of the query, we usually take the hexadecimal String as an input parameter and use "toXXX("") to perform the "Transform" action. Our model Co-CoSoDa can successfully understand the semantics of the whole
The advantages of CoCoSoDa mainly come from soft data augmentation and multimodal contrastive learning. The soft data augmentation destroys a sequence within a certain range and generates a similar sample as the "positive" example. It can help the model learn a representation of the code snippet and query from a global (sequence-level) view rather than simply aggregate the token-level semantics. Thus, our model tends to return code snippets related to the sequence-level functionality rather than the token-level similarities. For multimodal contrastive learning, it can pull together the representations of the code-query pair and push apart the representations of queries and many unpaired code snippets. Therefore, CoCoSoDa can learn the better representations of code and query and perform well on code search tasks.

6.2 Limitations & Threats to Validity

Although CoCoSoDa has an overall advantage, our model could still return inaccurate results, especially for the code snippets that use the third-library API or self-defined methods. This is because CoCoSoDa only considers the information of the code snippet itself rather than other contexts such as other methods in the enclosing class or project [2, 72]. In our future work, more contextual information (such as enclosing class/project and called API/methods) could be considered in our model to further improve the performance of CoCoSoDa.

We also have identified the following threats to our approach: Programming Languages. Due to the heavy effort to evaluate the model on all programming languages, we conduct our experiment with as many programming languages as possible on the existing build datasets. Our model on different programming languages would have different results. In the future, we will evaluate the effectiveness of our approach with more other programming languages.

Pre-trained models. To demonstrate that our approach is orthogonal to the pre-trained technique and can lead to independent improvements for code search tasks, we have adopted and evaluated our approach on four pre-trained models including a natural
In this paper, we present CoCoSoDa, which leverages multimodal constructive learning and soft data augmentation for code search. Soft data augmentation helps CoCoSoDa learn the representation of the code snippets from the sequence-level point of view. Multimodal constructive learning can pull together the representations of code-query pair and push apart the unpaired code snippets and queries. Thus, CoCoSoDa can learn sequence-level representations of code snippets and queries. We conduct extensive experiments on large-scale benchmark dataset with six programming languages (Ruby, JavaScript, Go, Python, Java, PHP) and the results confirm its effectiveness. We also apply CoCoSoDa on other pre-trained models RoBERTa and RoBERTa (code), and CodeBERT and boost their performance on code search. We also assign different hyperparameters to check their impact on code search and find that CoCoSoDa performs stably over a range of hyperparameters. In our future work, more contextual information (such as enclosing class/proj-
ject and called API/methods) could be considered in our model to further improve the performance of CoCoSoDa.

7 CONCLUSION

In this paper, we present CoCoSoDa, which leverages multimodal constructive learning and soft data augmentation for code search. Soft data augmentation helps CoCoSoDa learn the representation of the code snippets from the sequence-level point of view. Multimodal constructive learning can pull together the representations of code-query pair and push apart the unpaired code snippets and queries. Thus, CoCoSoDa can learn sequence-level representations of code snippets and queries. We conduct extensive experiments on large-scale benchmark dataset with six programming languages (Ruby, JavaScript, Go, Python, Java, PHP) and the results confirm its effectiveness. We also apply CoCoSoDa on other pre-trained models RoBERTa and RoBERTa (code), and CodeBERT and boost their performance on code search. We also assign different hyperparameters to check their impact on code search and find that CoCoSoDa performs stably over a range of hyperparameters. In our future work, more contextual information (such as enclosing class/proj-
ject and called API/methods) could be considered in our model to further improve the performance of CoCoSoDa.

REFERENCES

[1] Miltiadis Allamanis, Earl T. Barr, Premkumar T. Devanbu, and Charles Sutton. 2018. A Survey of Machine Learning for Big Code and Naturalness. ACM Comput. Surv. 51, 4 (2018), 81:1–87.

[2] Aalaksh Bansal, Sakib Haque, and Collin Mcmillan. 2021. Project-Level Encoding Baselines with Momentum Contrastive Learning. In IJCAI: Proceedings of the 29th International Joint Conference on Artificial Intelligence, IJCAI’20.

[3] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Pathaksouza, Aapo Kyrola, Andrew Talleck, Yann Dauphinais, Ilya Sutskever, and Dario Amodei. 2020. Language Models Are Few-Shot Learners. In NeurIPS.

[4] Marcel Bruch, Martin Monperrus, and Mira Mezini. 2009. Learning from examples using RNN Encoder-Decoder for Statistical Machine Translation. In EMNLP.

[5] Luis Fernando Cortes-Coy, Mario Linares-Vásquez, Jairo Aponte, and Denys Poshyvanyk. 2014. On Automatically Generating Commit Messages via Summarization of Source Code Changes. In SCAM. IEEE Computer Society, 275–284.

[6] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In NAACL-HLT (1). Association for Computational Linguistics, 4171–4186.

[7] Simeon Devries and Graham W. Taylor. 2017. Improved Regularization of Convolutional Neural Networks with Cutout. CoRR abs/1708.04552 (2017).

[8] Zhangyin Deng, Luca Buratti, Saurabh Pujar, Alessandro Morari, Baishalki Ray, and Saikat Chakraborty. 2021. Contrastive Learning for Source Code with Structural and Functional Properties. CoRR abs/2110.03868 (2021).

[9] Jun Du, Xiaohou Shi, Yanlin Wang, Ensheng Shi, Shi Han, and Dongmei Zhang. 2021. Is a Single Model Enough? MuCoS: A Multi-Model Ensemble Learning Approach for Semantic Code Search. In CIKM. ACM, 2994–2998.

[10] Jun Du, Xiaohou Shi, Yanlin Wang, Ensheng Shi, Shi Han, and Dongmei Zhang. 2021. Is a Single Model Enough? MuCoS: A Multi-Model Ensemble Learning Approach for Semantic Code Search. In Proceedings of the 30th ACM International Conference on Information & Knowledge Management. ACM, 2994–2998.

[11] Jun Du, Xiaohou Shi, Yanlin Wang, Ensheng Shi, Shi Han, and Dongmei Zhang. 2021. Is a Single Model Enough? MuCoS: A Multi-Model Ensemble Learning Approach for Semantic Code Search. In ICLR (Poster). OpenReview.net.

[12] Tianyu Guo, Xingcheng Yao, and Danqi Chen. 2021. SimCSE: Simple Contrastive Learning of Sentence Embeddings. In EMNLP (1). Association for Computational Linguistics, 6894–6910.

[13] Spyros Gidaris, Praveer Singh, and Nikos Komodakis. 2018. Unsupervised Representation Learning by Predicting Image Rotations. In ICLR (Poster). OpenReview.net.

[14] John M. Giorgi, Oswald Nitski, Bo Wang, and Gary D. Bader. 2021. DeCLUTER: Deep Contrastive Learning for Unsupervised Textual Representations. In ACL-JCNLP (1). Association for Computational Linguistics, 879–895.

[15] Priya Goyal, Piotr Dollár, Ross B. Girshick, Pieter Noordhuis, Lukasz Wesolowski, Aapo Kyrola, Andrew Talleck, Yangqing Jia, and Kaiming He. 2017. Accurate, Large minibatch SGD: Training ImageNet in 1 Hour. CoRR abs/1708.02673.

[16] Jun Gu, Zimin Chen, and Martin Monperrus. 2021. Multimodal Representation for Neural Code Search. In IEEE International Conference on Software Maintenance and Evolution, ICSME 2021, Luxembourg, September 27 - October 1, 2021. IEEE, 483–494. https://doi.org/10.1109/ICSME52107.2021.00084

[17] Wenchao Gu, Zongjie Li, Cuiyun Gao, Chaozhong Wang, Hongyu Zhang, Zenglin Xu, and Michael R. Lyu. 2021. CBRaDe: Deep code retrieval based on semantic Dependency Learning. Neural Networks 141 (2021), 385–394.

[18] Xiaodong Gu, Hongyu Zhang, and Sunghun Kim. 2018. Deep code search. In ICSE. ACM, 933–944.

[19] Daya Guo, Shuo Ren, Shuai Lu, Zhiyang Feng, Duyu Tang, Shujie Liu, Long Zhou, Nan Duan, Alexey Syvatskovskiy, Shengyu Fu, Michele Tufano, Shao Kun Deng, Colin B. Clement, Dawn Drain, Niel Sundaresan, Jin Yan, Daxin Jiang, and Ming Zhou. 2021. GraphCodeBERT: Pre-training Code Representations with Data Flow. In 9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021. OpenReview.net. https://openreview.net/forum?id=lJoLC4ez43fZ

[20] Raia Hadsell, Sumit Chopra, and Yann LeCun. 2006. Dimensionality Reduction by Learning an Invariant Mapping. In 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2006), 17-22 June 2006, New York, NY, USA. IEEE Computer Society, 1735–1742. https://doi.org/10.1109/CVPR.2006.100

[21] Rajarshi Haldar, Lingfei Wu, Jinjun Xiong, and Julia Hockenmaier. 2020. A Multi-Perspective Architecture for Semantic Code Search. In ACL.

[22] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross B. Girshick. 2020. Moment contrast for unsupervised visual representation learning. In CVPR (Poster). OpenReview.net.

[23] Junjie Huang, Duyu Tang, Nan Duan, Daxin Jiang, Ming Zhou, and Nan Duan. 2021. CoQA: 20, 000+ Web Queries for Code Search and Question Answering. In ACL.

Conference acronym 'XX', June 03–05, 2018, Woodstock, NY Shi, et al.
