Exploring Representation-Level Augmentation for Code Search

Haochen Li[1] **Chunyan Miao**[1,2*] **Cyril Leung**[1,2] **Yanxian Huang**[3]
Yuan Huang[3] **Hongyu Zhang**[4] **Yanlin Wang**[3]

[1]School of Computer Science and Engineering, Nanyang Technological University, Singapore
[2]China-Singapore International Joint Research Institute (CSIJRI), China
[3]School of Software Engineering, Sun Yat-sen University, China
[4]The University of Newcastle, Australia
{haochen003,ascymiao,cleung}@ntu.edu.sg, huangyx353@mail2.sysu.edu.cn
{huangyuan5,wangylin36}@mail.sysu.edu.cn, hongyu.zhang@newcastle.edu.au

Abstract

Code search, which aims at retrieving the most relevant code fragment for a given natural language query, is a common activity in software development practice. Recently, contrastive learning is widely used in code search research, where many data augmentation approaches for source code (e.g., semantic-preserving program transformation) are proposed to learn better representations. However, these augmentations are at the raw-data level, which requires additional code analysis in the preprocessing stage and additional training costs in the training stage. In this paper, we explore augmentation methods that augment data (both code and query) at representation level which does not require additional data processing and training, and based on this we propose a general format of representation-level augmentation that unifies existing methods. Then, we propose three new augmentation methods (linear extrapolation, binary interpolation, and Gaussian scaling) based on the general format. Furthermore, we theoretically analyze the advantages of the proposed augmentation methods over traditional contrastive learning methods on code search. We experimentally evaluate the proposed representation-level augmentation methods with state-of-the-art code search models on a large-scale public dataset consisting of six programming languages. The experimental results show that our approach can consistently boost the performance of the studied code search models. Our source code is available at https://github.com/Alex-HaochenLi/RACS.

1 Introduction

In software development, developers often search and reuse commonly used functionalities to improve their productivity (Nie et al., 2016; Shuai et al., 2020). With the growing size of large-scale codebases such as GitHub, retrieving semantically relevant code fragments accurately becomes increasingly important in this field (Allamanis et al., 2018; Liu et al., 2021).

Traditional approaches (Nie et al., 2016; Yang and Huang, 2017; Rosario, 2000; Hill et al., 2011; Satter and Sakib, 2016; Lv et al., 2015; Van Nguyen et al., 2017) leverage information retrieval techniques to treat code snippets as natural language text and match certain terms in code with queries, hence suffering from the vocabulary mismatch problem (McMillan et al., 2011; Robertson et al., 1995). Deep siamese neural networks first embed queries and code fragments into a joint embedding space, then measure similarity by calculating dot product or cosine distance (Lv et al., 2015; Cambronero et al., 2019; Gu et al., 2021). Recently, with the popularity of large scale pre-training techniques, some big models for source code (Guo et al., 2021; Feng et al., 2020; Guo et al., 2022; Wang et al., 2021; Jain et al., 2021; Li et al., 2022) with various pre-training tasks are proposed and significantly outperform previous models.

Contrastive learning is widely adopted by the above-mentioned models. It is suitable for code search because the learning objective aims to push apart negative query-code pairs and pull together positive pairs at the same time. In contrastive learning, negative pairs are usually generated by In-Batch Augmentation (Huang et al., 2021). For positive pairs, besides labeled ones, some researchers proposed augmentation approaches to generate more positive pairs (Bui et al., 2021; Jain et al., 2021; Fang et al., 2020; Gao et al., 2021; He et al., 2020). The main hypothesis behind these approaches is that augmentations do not change the original semantics. However, these approaches are resource-consuming (Yin et al., 2021; Jeong et al., 2022). Models have to embed the data again for the augmented data.

To solve this problem, some researchers proposed representation-level augmentation, which

*Corresponding author.

Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, pages 4924 - 4936
December 7-11, 2022 ©2022 Association for Computational Linguistics

augments the representations of the original data. For example, linear interpolation, a representation-level augmentation method, is adopted by many classification tasks in NLP (Guo et al., 2019; Sun et al., 2020; Du et al., 2021). The augmented representation captures the structure of the data manifold and hence could force model to learn better features, as argued by Verma et al. (2021). These augmentation approaches are also considered to be semantic-preserving.

The representation-level augmentation methods are not investigated on the code search task before. To the best of our knowledge, Jeong et al. (2022) is the only work to bring representation-level augmentation approaches to a retrieval task. Besides linear interpolation, it also proposes another approach called stochastic perturbation for document retrieval. Although these augmentation methods bring improvements to model performance, they are not yet fully investigated. The relationships between the existing methods and how they affect model performance remain to be explored.

In this work, we first unify linear interpolation and stochastic perturbation into a general format of representation-level augmentation. We further propose three augmentation methods (linear extrapolation, binary interpolation, and Gaussian scaling) based on the general format. Then we theoretically analyze the advantages of the proposed augmentation methods based on the most commonly used InfoNCE loss (Van den Oord et al., 2018). As optimizing InfoNCE loss equals to maximizing the mutual information between positive pairs, applying representation-level augmentation leads to tighter lower bounds of mutual information. We evaluate representation-level augmentation on several Siamese networks across several large-scale datasets. Experimental results show the effectiveness of the representation-level augmentation methods in boosting the performance of these code search models. To verify the generalization ability of our method to other tasks, we also conduct experiments on the paragraph retrieval task, and the results show that our method can also improve the performance of several paragraph retrieval models.

In summary, our contributions of this work are as follows:

- We unify previous representation-level augmentation methods to propose a general format. Based on this general format, we propose three novel augmentation methods.
- We conduct theoretical analysis to show that representation-level augmentation has tighter lower bounds of mutual information between positive pairs.
- We apply representation-level augmentation on several code search models and evaluate them on the public CodeSearchNet dataset with six programming languages. Improvement of MRR (Mean Reciprocal Rank) demonstrates the effectiveness of the representation-level augmentation methods.

The rest of the paper is organized as follows. We introduce related work of code search and data augmentation in Section 2. Section 3 introduces the main part, including the general format of representation-level augmentation, new augmentation methods, and their application on code search. In Section 4, we analyze the theoretical lower bounds of mutual information and study why our approach works. In Section 5 and Section 6, we conduct extensive experiments to show the effectiveness of our approach. Then we discuss the generality of our approach in Section 7 and Section 8 concludes this paper.

2 Related Work

2.1 Code search

As code search can significantly improve the productivity of software developers by reusing functionalities in large codebases, finding the semantic-relevant code fragments precisely is one of the key challenges in code search.

Traditional approaches leverage information retrieval techniques that try to match some keywords between queries and codes (McMillan et al., 2011; Robertson et al., 1995). These approaches suffer from vocabulary mismatch problem where models fail to retrieve the relevant codes due to the difference in semantics.

Later, deep neural models for code search are proposed. They could be divided into two categories, early fusion and late fusion. Late fusion approaches (Gu et al., 2018; Husain et al., 2019) use a siamese network to embed queries and codes into a shared vector space separately, then calculate dot product or cosine distance to measure the semantic similarity. Recently, following the idea of late fusion, some transformer-based models with

specifically designed pre-training tasks are proposed (Feng et al., 2020; Guo et al., 2021, 2022). They significantly outperform previous models by improving the understanding of code semantics. Instead of calculating representations of queries and codes independently, early fusion approaches model the correlations between queries and codes during the embedding process (Li et al., 2020). Li et al. (2020) argues that early fusion makes it easier to capture implicit similarities. For applications of an online code search system, late fusion approach facilitates the use of neural models because the code representations can be calculated and stored in advance. During run time, only query representations need to be computed. Thus, in this work, we focus on late fusion approaches.

2.2 Data augmentation

Data augmentation has long been considered crucial for learning better representations in contrastive learning. The augmented data are considered to have the same semantics with the original data. For the augmentation of queries, synonym replacement, random insertion, random swap, random deletion, back-translation, spans technique and word perturbation can be potentially used to generate individual augmentations (Wei and Zou, 2019; Giorgi et al., 2021; Fang et al., 2020). For the augmentation of code fragments, Bui et al. (2021) proposed six semantic-preserving transformations: *Variable Renaming*, *Permute Statement*, *Unused Statement*, *Loop Exchange*, *Switch to If* and *Boolean Exchange*. These query and code augmentation approaches have one thing in common, that is, the transformation is applied to the original input data. Another category is augmenting during the embedding process. Models can generate different representations of the same data by leveraging time-varying mechanisms. MoCo (He et al., 2020) encodes data twice by the same model with different parameters. SimCSE (Gao et al., 2021) leverages the property of dropout layers by randomly deactivating different neurons for the same input. Methods described in this paragraph are resource-consuming because models embed twice to get representations of original data and augmented one.

For representation-level augmentation on NLP tasks, linear interpolation is widely used on classification tasks in previous work (Guo et al., 2019; Sun et al., 2020; Du et al., 2021). They take the interpolation result as noised data and want models to classify the noised one into the original class. Verma et al. (2021) theoretically analyzed how interpolation noise benefits classification tasks and why it is better than Gaussian noise. Jeong et al. (2022) is the first to introduce linear interpolation and perturbation to the document retrieval task. However, the effect and intrinsic relationship of these two methods are not fully investigated.

3 Approach

In this section, we unify the linear interpolation and stochastic perturbation into a general format. Based on it, we propose three other augmentation methods for the code retrieval task, linear extrapolation, binary interpolation and Gaussian scaling. Then, we explain how to apply representation-level augmentation with InfoNCE loss in code retrieval.

3.1 General format of representation-level augmentation

For simplicity, we take code augmentation as an example to elaborate the details. The calculation process is similar when applying to query augmentations. Given a data distribution $\mathbf{D} = \{x_i\}_{i=1}^K$ where x_i is a code snippet, K is the size of the dataset. We use an encoder function $h : \mathbf{D} \rightarrow \mathbf{H}$ to map codes to representations $\mathbf{H}$.

Linear interpolation Linear interpolation randomly interpolate h_i with another chosen sample h_j from $\mathbf{H}$:

$$h_i^+ = \lambda h_i + (1 - \lambda) h_j \quad (1)$$

where λ is a coefficient sampled from a random distribution. For example, λ can be sampled from a uniform distribution $\lambda \sim U(\alpha, 1.0)$ with high values of α to make sure that the augmented data has similar semantics with the original code x_i.

Stochastic perturbation Stochastic perturbation aims at randomly deactivating some features of representation vectors. In order to do so, masks are sampled from a Bernoulli distribution $B(e, p)$, where e is the embedding dimension. p is a low probability value since we only deactivate a small proportion of features. For implementation, we could use Dropout layers.

General format of representation-level augmentation We revisit the above two augmentation approaches and unify them into a general format, which could be described as:

$$h^+ = \alpha \odot h + \beta \odot h' \quad (2)$$

where $h, h' \in \mathbf{H}$, α and β are coefficient vectors. For linear interpolation, $\alpha = \lambda, \beta = 1 - \lambda$, $h, h' \in \mathbf{H}$, and $h \neq h'$. For stochastic perturbation, $\alpha \in \{0, \frac{1}{1-p}\}^e$ where elements of α are sampled from a Bernoulli distribution $B(p)$, $\beta = \frac{1}{1-p} - \alpha$, $h \in \mathbf{H}$, and $h' = 0$.

3.2 New augmentation methods

Based on the general format, we provide three new augmentation methods for the code retrieval task.

Binary interpolation Binary interpolation randomly swaps some features with another chosen sample. The difference between binary interpolation and stochastic perturbation is that the former swaps with other samples while the latter swaps with zero vector. Specifically, for binary interpolation, $\alpha \in \{0, 1\}^e$ where elements of α are sampled from a Bernoulli distribution $B(p)$, $\beta = 1 - \alpha$, $h, h' \in \mathbf{H}$, and $h \neq h'$.

Linear extrapolation Wang and Isola (2020) concludes that optimizing contrastive loss makes feature vectors roughly uniformly distributed on the hypersphere. As linear interpolation generates augmented data inside the hypersphere, linear extrapolation oppositely generates outside ones. λ is sampled from a uniform distribution $\lambda \sim U(1.0, \alpha)$ with small values of α. Other settings are the same as linear interpolation.

Gaussian scaling Gaussian scaling generates scaling coefficients for each feature in the representation, which can be considered as a type of perturbation noise. Compared with directly adding Gaussian noise, the proposed scaling noise captures the structure of the data manifold, which may force networks to learn better representations. If we describe Gaussian scaling in general format, then $\alpha = 1$, $\beta \sim N(0, \sigma)$ with small values of σ, $h = h' \in \mathbf{H}$.

3.3 Contrastive learning with representation-level augmentation for code search

Contrastive learning seeks to satisfy that similarities between positive pairs are greater than that between negative pairs, which is suitable for code search. In previous works, in order to optimize the objective, several loss functions are proposed, including triplet loss (Mikolov et al., 2013), max-margin loss (Hadsell et al., 2006), and logistic loss (Weinberger and Saul, 2009). In this work, we consider InfoNCE loss (Van den Oord et al., 2018) because of its better performance hence dominant adaption in current contrastive learning models. For the effect of representation-level augmentation on other loss functions, we empirically analyze it in Appendix B.

We start from the vanilla InfoNCE loss. Suppose we have a batch of B samples consisting of queries and codes. We encode queries and codes to get query representations $Q = \{q_i\}_{i=1}^B$ and code representations $C = \{c_j\}_{j=1}^B$ using the encoder function h. (q_i, c_j) are positive pairs when $i = j$ and negative pairs otherwise. Therefore, for each query, we could generate 1 positive pair and $B - 1$ negative pairs. The InfoNCE loss tries to maximize the similarity between one positive pair and minimize the similarity between other negative pairs. Here we use dot product as the measurement of similarity and in Section 6.4 we will discuss the effect of using dot product compared to cosine distance. The loss can be described as:

$$L = -\mathbb{E}\left[\log \frac{\exp(q_i \cdot c_i)}{\exp(q_i \cdot c_i) + \sum_{j \neq i}^{B} \exp(q_i \cdot c_j)}\right] \quad (3)$$

Then, we conduct representation-level augmentation. We randomly choose one out of the five augmentation methods and augment the original queries Q and original codes C for N times. In each augmentation, the augmentation approach is fixed, but the coefficients α and β are randomized hence different. After augmentation, we get augmented query and code sets, $Q^+ = \{q_{ni}\}_{n=1,i=1}^{n=N,i=B}$ and $C^+ = \{c_{nj}\}_{n=1,j=1}^{n=N,j=B}$. We follow the hypothesis of other representation-level augmentation methods that the augmented representation still preserves or is similar to the original semantics. Therefore, for a certain query q_i, we consider q_i and q_{ni},$\forall n \in [1, N]$ share the same semantic meaning. And this similarly applies to c_j and c_{nj}, $\forall n \in [1, N]$. Since (q_i, c_j) is labeled as a positive pair when $i = j$, $\forall n \in [1, N]$ (q_i, c_{nj}) and (q_{ni}, c_j) are also naturally labeled as positive pairs. Similarly, we get $(q_{ni}, c_{nj, j \neq i})$ as negative pairs. Thus, compared with vanilla InfoNCE loss, we now have $(N+1)^2 B$ positive pairs in total and for each query $q \in Q \cup Q^+$ we can generate $(B-1)(N+1)$ negative pairs.

4 Theoretical Analysis

In this section, we mathematically analyze the effect of InfoNCE loss with or without representation-level augmentation and prove that optimizing InfoNCE loss with representation-level augmentation leads to mutual information with tighter lower bounds between positive pairs. Here we take linear interpolation as an example to demonstrate the benefits. Other forms of representation-level augmentation are left to future work. The mutual information between a query q and a code fragment c is:

$$I(q,c) = \mathbb{E}_{q,c}\left[\log \frac{p(c|q)}{p(c)}\right] \tag{4}$$

Theorem 1 *Optimizing InfoNCE loss $\mathcal{L}_{\mathcal{N}}$ improves lower bounds of mutual information $I(q,c)$ for a positive pair:*

$$I(q,c) \geq \log(B) - \mathcal{L}_{\mathcal{N}} \tag{5}$$

where $q \in Q$, $c \in C$, and B is the size of sets.

Proof of Theorem 1 is presented in Appendix A.1. This is proved in the original paper of InfoNCE loss (Van den Oord et al., 2018). Here, to better extend the proof of Theorem 2, we prove it in another way.

Theorem 2 *Optimizing InfoNCE loss $\mathcal{L}_{\mathcal{N}}$ with representation-level augmentation improves a tighter lower bounds of mutual information $I(q,c)$ for a positive pair:*

$$\begin{aligned} I(q,c) \geq & \frac{1}{\alpha^2}(\log(NB) - \mathcal{L}_{\mathcal{N}} \\ & - \alpha\beta \cdot I(q,c^-) - \alpha\beta \cdot I(q^-,c) \\ & - \beta^2 \cdot I(q^-,c^-)) \end{aligned} \tag{6}$$

where $q,q^- \in Q$,$c,c^- \in C$, (q,c^-), (q^-,c) and (q^-,c^-) are all negative pairs, α and β are coefficients in Equation 2, B is the size of sets, and N is the augmentation time.

Proof of Theorem 2 is presented in Appendix A.2. Since we interpolate q with other samples q^- in the batch (c with c^-), the mutual information between (q,c^-), (q^-,c) and (q^-,c^-) are also incorporated into the optimizing process. As defined in Eq.1, β is a small value that close to 0. According to Eq.4, for negative pairs $p(c|q)$ can be expressed as $\frac{p(c)p(q)}{p(q)}$ due to the independence of sampling q and c. Considering this, we can see that the optimal mutual information between negative pairs is also 0. Note that we interpolate q and c with different samples to make sure that (q^-,c^-) is a negative pair. Thus, the last three terms in Eq.6 can be ignored. Comparing $\frac{1}{\alpha^2}(\log(NB) - \mathcal{L}_{\mathcal{N}})$ with $\log(B) - \mathcal{L}_{\mathcal{N}}$, we can see that representation-level augmentation improves the lower bounds of mutual information.

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

Table 1: CodeSearchNet dataset statistics.

5 Experimental Setup

In this section, we describe datasets, baselines, and implementation details.

5.1 Datasets

To evaluate the effectiveness of representation-level augmentation, we use a large-scale benchmark dataset **CodeSearchNet** (CSN) (Husain et al., 2019) that is widely used in previous studies (Guo et al., 2021; Feng et al., 2020; Guo et al., 2022). It contains six programming languages including Ruby, JavaScript, Go, Python, Java, and PHP. The statistics of the dataset are shown in Table 1. For the training set, it contains positive-only query-code pairs. For validation and test sets, they only have queries and the model retrieves the correct code fragments from a fixed codebase. Here we follow (Guo et al., 2021) to filter out low-quality examples (such as code that cannot be successfully parsed into Abstract Syntax Trees).

We measure the performance using Mean Reciprocal Rank (MRR) which is widely adopted in previous studies. MRR is the average of reciprocal ranks of a true code fragment for a given query Q. It is calculated as:

$$MRR = \frac{1}{|Q|}\sum_{i=1}^{|Q|}\frac{1}{Rank_i} \tag{7}$$

where $Rank_i$ is the rank of the correct code fragment that is related to the i-th query.

5.2 Baselines

Since representation-level augmentation is orthogonal to siamese networks, we apply it to several models:

- **RoBERTa (code)** is pre-trained with mask language modeling (MLM) task on code corpus (Husain et al., 2019).
- **CodeBERT** is a bi-modal pre-trained model pre-trained on two tasks: MLM and replaced token detection (Feng et al., 2020). Note that in this work we refer CodeBERT to the siamese network architecture described in the appendix of the original paper.
- **GraphCodeBERT** takes the structure information of codes into account. (Guo et al., 2021) develops two structure-based pre-training tasks: data flow edge prediction and node alignment.
- **UniXCoder** leverages cross-model contents like AST and comments to enhance code representation (Guo et al., 2022).

5.3 Implementation details

For all the settings of these models except the training epoch, we follow the original paper. For representation-level augmentation, since linear extrapolation and linear interpolation is similar, we implement it as one approach. During training, we randomly choose one augmentation approach out of four with equal probability for a batch and augment data 5 times. We re-sample data and coefficients to augment the original data in each augmentation. Specifically, for linear interpolation and extrapolation, we sample α from a uniform distribution $U \sim (0.9, 1.1)$. For perturbation, we set the probability p of the Dropout layer as 0.1. For binary interpolation, we sample α from a Bernoulli distribution $B(p = 0.25)$. For Gaussian scaling, we sample β from a normal distribution $N(0, 0.1)$. We set the training epoch as 30. All experiments are conducted on a GeForce RTX A6000 GPU.

6 Results

In this section, we first show the overall performance on code search when applying representation-level augmentation, and then individually analyze each augmentation method. Then, we demonstrate the relationship between loss and MRR and the effect of augmentation on vector distribution. Finally, we take an ablation study to analyze the impact of augmentation times.

6.1 Overall results

The overall performance evaluation results are shown in Table 2. In each iteration, we randomly choose one augmentation method. We can see that representation-level augmentation is a universal approach that can consistently improve the code search performance. Optimizing a tighter lower bound of mutual information brings about 2% gain of MRR on average. The robust improvements have no relationships with certain models or certain programming languages.

6.2 Effectiveness of individual augmentation approach

To evaluate the effectiveness of the five augmentation approaches, we apply them alone and test them on the CSN-Python dataset, as shown in Table 3. Note that we follow the same experimental settings described in Section 5. As the results show, every augmentation approach can improve baselines and the improvements brought by these augmentation approaches are stable. The combination of these approaches does not boost the performance compared with individually applying one of these augmentations. Since all these approaches can be derived from the general format, they have no distinct difference and hence share the similar effect on improving the mutual information between positive pairs.

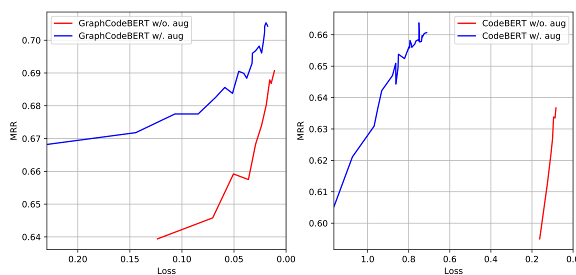

Figure 1: The relationship between loss and MRR with or without representation-level augmentation. *aug* is short for representation-level augmentation.

6.3 Relationship between Loss and MRR

In theoretical analysis, we get the conclusion that the relationship between the mutual information of positive pairs and loss is $I(q, c) \geq \log(B) - \mathcal{L}_N$ while for representation-level augmentation

Model	Ruby		JavaScript		Go		Python		Java		PHP	
	Original	w/ RA	Original	w/ RA	Original	w/ RA	Original	w/ RA	Original	w/ RA	Original	w/ RA
RoBERTa (code)	0.641	**0.665**	0.583	**0.612**	0.867	**0.892**	0.610	**0.663**	0.634	**0.674**	0.584	**0.617**
CodeBERT	0.648	**0.664**	0.594	**0.608**	0.878	**0.890**	0.636	**0.654**	0.663	**0.674**	0.615	**0.619**
GraphCodeBERT	0.705	**0.721**	0.647	**0.671**	0.896	**0.903**	0.690	**0.708**	0.691	**0.708**	0.648	**0.656**

Table 2: Performance of different approaches under MRR. "w/ RA" stands for "with representation-level augmentation".

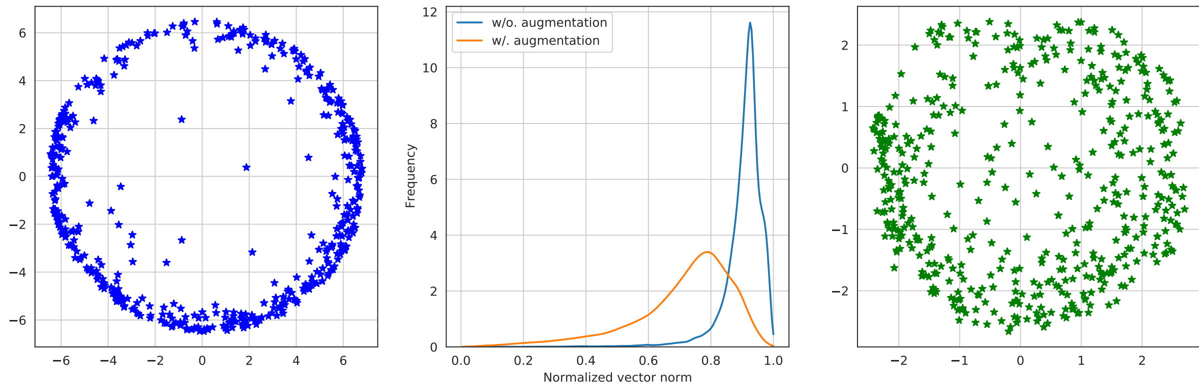

Figure 2: Visualization of code vector distribution with and without representation-level augmentation. **Left**: without augmentation. **Right**: with augmentation. **Middle**: distribution of vector norms.

Augmentations	RoBERTa	CodeBERT	GraphCodeBERT
no augmentations	0.629	0.636	0.690
linear interpolation	0.644	0.648	0.702
linear extrapolation	0.640	0.646	0.704
stochastic perturbation	0.658	0.648	0.698
binary interpolation	0.655	**0.655**	0.705
Gaussian scaling	0.657	0.649	0.696
all augmentations	**0.663**	0.654	**0.708**

Table 3: Results of individual augmentation on CSN-Python dataset.

Augmentations	MRR
UniXCoder	0.721
UniXCoder + all augmentations	0.699
UniXCoder - normalization	0.708
UniXCoder - normalization + all augmentations	**0.728**

Table 4: MRR of UniXCoder on CSN-Python under different settings.

$I(q, c) \geq \frac{1}{\alpha^2}(\log(NB) - \mathcal{L}_\mathcal{N})$. Thus, we argue that the effect of representation-level augmentation is improving the lower bounds of mutual information. However, this only comes true when loss can decrease to similar values under such two conditions. To prove that, after each epoch, we save the model parameters, record the loss of the training set, test models on the CSN-Python test set, and plot the relationship between loss and MRR with or without representation-level augmentation, as shown in Figure 1. We can see that when the loss is the same, representation-level augmentation always leads to better performance (GraphCodeBERT). Even when loss cannot decrease to the same value without augmentation, it outperforms the vanilla contrastive learning (CodeBERT). We believe that other than improving the lower bounds, capturing the explicit relations between augmented data and the original one can also lead to higher mutual information between positive pairs.

6.4 Impact on vector distribution

In code search, we measure the similarity by calculating the dot product between query representations and code representations. Here we analyze the influence of representation-level augmentation by visualizing the code vector distribution. We add a linear layer to embed the representations to a two-dimensional vector, as shown in Figure 2. Besides, we plot the two-norms of vectors with Gaussian kernel density estimation.

As Wang and Isola (2020) concluded, the optimization of InfoNCE loss makes vectors evenly distributed, which corresponds to the left image of Figure 2. Vectors roughly have the same two-norm. After applying augmentation, vectors still follow a circular pattern but their two-norms are

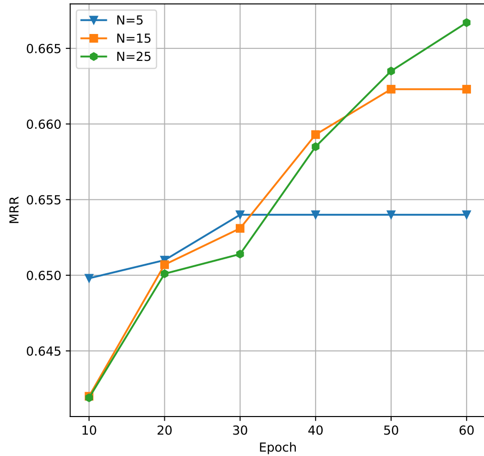

Figure 3: Performance of CodeBERT on each epoch interval with different augmentation times N.

changed. We argue that the InfoNCE loss degrades dot product to cosine distance since all vectors have similar norms while representation-level augmentation leverages the norms of vectors to distinguish some hard examples. We can see that in the middle image of Figure 2, vector norms are much more uniformly distributed than that without augmentation.

However, this impact also reveals a limitation of representation-level augmentation. Some previous studies argue the necessity of using vector normalization, otherwise, the *Softmax* distribution will be made arbitrarily sharp. For example, the last step of UniXCoder is normalization. We evaluate UniXCoder on CSN-Python, as shown in Table 4. When we apply augmentations to UniXCoder with normalization, the performance is worse. However, if we remove the normalization step, augmentations can boost performance just like for other models.

6.5 Impact of the number of augmentation times

To analyze the impact of augmentation times N, we conduct experiments on CodeBERT with $N = 5, 15, 25$, respectively. We take 10 epochs as an interval, save the best model in each interval and test them on the test set, as shown in Figure 3. We can see that augmenting more times leads to a relatively better performance that is even greater than the result reported in Table 2. However, better results require more training time. According to $I(q, c) \geq \frac{1}{\alpha^2}(\log(NB) - \mathcal{L}_\mathcal{N})$, bigger N is better, but time cost of minimizing $\mathcal{L}_\mathcal{N}$ also increases significantly. As we could see in the figure, a 5-times augmentation takes 30 epochs to converge while a 15-times augmentation takes 50 epochs. We train CodeBERT with $N = 25$ for 60 epochs and MRR is still increasing. Therefore, we argue that for the application of representation-level augmentation, we should find a balance between performance and time cost.

Model	FiQA-2018		NFCorpus	
	Original	w/ RA	Original	w/ RA
DistilBERT	0.352	**0.400**	0.481	**0.505**
RoBERTa	0.343	**0.356**	0.367	**0.389**

Table 5: The performance of different approaches on MRR@1000. Here we follow the widely used metrics on passage retrieval tasks. MRR@1000 only considers the top 1000 returned passages.

7 Discussion

According to the proof of Theorem 2, the advantage of applying representation-level augmentation should be task-agnostic. To show its generalization ability, we evaluate our proposed approaches on two passage retrieval benchmark datasets, NFCorpus (Boteva et al., 2016) and FiQA-2018[1]. The difference between code search and passage retrieval is that the retrieved items are changed from code snippets written in programming language to passages written in English. We take DistilBERT (Sanh et al., 2019) and RoBERTa (Liu et al., 2019) for experiments. We implement our approach based on an open-sourced framework BEIR (Thakur et al., 2021). The two models are fine-tuned on two datasets for 20 epochs, respectively. The settings of augmentation are the same as those in the code search task. For other hyper-parameters, we follow the settings that are provided by the framework. Results are shown in Table 5, which confirms that representation-level augmentation also improves the performance of passage retrieval models.

8 Conclusion

In this work, we unify existing approaches to propose a general format of representation-level augmentation in code search. Based on the general format, we propose three other augmentation methods. We further theoretically analyze the effect of representation-level augmentation by proving that it helps optimize a tighter lower bound of mutual information between positive pairs. We evaluate our

[1] https://sites.google.com/view/fiqa/

approach on several models and datasets and the results demonstrate the effectiveness of the proposed approach.

Limitations

As discussed in Section 6, representation-level augmentation mainly has two limitations. First, it cannot boost the performance for models with vector normalization. We find that representation-level augmentation improves performance by leveraging the norms of vectors. With normalization, the norm of vectors is fixed and hence augmentation cannot bring performance gains. Second, as discussed in Section 6.5, although augmenting more times can lead to better performance, the time spent on training also increases. Balancing performance and time-cost will be our future work.

9 Acknowledgement

We thank the anonymous reviewers for their helpful comments and suggestions. This research is supported in part by the National Research Foundation, Prime Minister's Office, Singapore under its NRF Investigatorship Programme (NRFI Award No. NRF-NRFI05-2019-0002). Any opinions, findings and conclusions or recommendations expressed in this material are those of the authors and do not reflect the views of National Research Foundation, Singapore. This research is also supported in part by the China-Singapore International Joint Research Institute (CSIJRI), Guangzhou, China (Award No. 206-A021002) and the Joint NTU-WeBank Research Centre on Fintech (Award No. NWJ-2020-007), Nanyang Technological University, Singapore.

References

Miltiadis Allamanis, Earl T Barr, Premkumar Devanbu, and Charles Sutton. 2018. A survey of machine learning for big code and naturalness. *ACM Computing Surveys (CSUR)*, 51(4):1–37.

Vera Boteva, Demian Gholipour Ghalandari, Artem Sokolov, and Stefan Riezler. 2016. A full-text learning to rank dataset for medical information retrieval. In *Advances in Information Retrieval - 38th European Conference on IR Research*, volume 9626, pages 716–722.

Nghi DQ Bui, Yijun Yu, and Lingxiao Jiang. 2021. Self-supervised contrastive learning for code retrieval and summarization via semantic-preserving transformations. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 511–521.

Jose Cambronero, Hongyu Li, Seohyun Kim, Koushik Sen, and Satish Chandra. 2019. When deep learning met code search. In *Proceedings of the 2019 27th ACM Joint Meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering*, pages 964–974.

Yangkai Du, Tengfei Ma, Lingfei Wu, Fangli Xu, Xuhong Zhang, Bo Long, and Shouling Ji. 2021. Constructing contrastive samples via summarization for text classification with limited annotations. In *Findings of the Association for Computational Linguistics*, pages 1365–1376.

Hongchao Fang, Sicheng Wang, Meng Zhou, Jiayuan Ding, and Pengtao Xie. 2020. Cert: Contrastive self-supervised learning for language understanding. *arXiv preprint arXiv:2005.12766*.

Zhangyin Feng, Daya Guo, Duyu Tang, Nan Duan, Xiaocheng Feng, Ming Gong, Linjun Shou, Bing Qin, Ting Liu, Daxin Jiang, and Ming Zhou. 2020. Codebert: A pre-trained model for programming and natural languages. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1536–1547.

Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2021. Simcse: Simple contrastive learning of sentence embeddings. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6894–6910.

John M. Giorgi, Osvald Nitski, Bo Wang, and Gary D. Bader. 2021. Declutr: Deep contrastive learning for unsupervised textual representations. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021*, pages 879–895.

Jian Gu, Zimin Chen, and Martin Monperrus. 2021. Multimodal representation for neural code search. In *2021 IEEE International Conference on Software Maintenance and Evolution (ICSME)*, pages 483–494. IEEE.

Xiaodong Gu, Hongyu Zhang, and Sunghun Kim. 2018. Deep code search. In *2018 IEEE/ACM 40th International Conference on Software Engineering (ICSE)*, pages 933–944. IEEE.

Daya Guo, Shuai Lu, Nan Duan, Yanlin Wang, Ming Zhou, and Jian Yin. 2022. Unixcoder: Unified cross-modal pre-training for code representation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*, pages 7212–7225.

Daya Guo, Shuo Ren, Shuai Lu, Zhangyin Feng, Duyu Tang, Shujie Liu, Long Zhou, Nan Duan, Alexey Svyatkovskiy, Shengyu Fu, Michele Tufano, Shao Kun

Deng, Colin B. Clement, Dawn Drain, Neel Sundaresan, Jian Yin, Daxin Jiang, and Ming Zhou. 2021. Graphcodebert: Pre-training code representations with data flow. In *9th International Conference on Learning Representations, ICLR 2021*.

Hongyu Guo, Yongyi Mao, and Richong Zhang. 2019. Augmenting data with mixup for sentence classification: An empirical study. *arXiv preprint arXiv:1905.08941*.

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'06)*, volume 2, pages 1735–1742. IEEE.

Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross B. Girshick. 2020. Momentum contrast for unsupervised visual representation learning. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9726–9735.

Emily Hill, Lori Pollock, and K Vijay-Shanker. 2011. Improving source code search with natural language phrasal representations of method signatures. In *ASE*. IEEE.

Junjie Huang, Duyu Tang, Linjun Shou, Ming Gong, Ke Xu, Daxin Jiang, Ming Zhou, and Nan Duan. 2021. Cosqa: 20, 000+ web queries for code search and question answering. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021,*, pages 5690–5700.

Hamel Husain, Ho-Hsiang Wu, Tiferet Gazit, Miltiadis Allamanis, and Marc Brockschmidt. 2019. Codesearchnet challenge: Evaluating the state of semantic code search. *arXiv preprint arXiv:1909.09436*.

Paras Jain, Ajay Jain, Tianjun Zhang, Pieter Abbeel, Joseph Gonzalez, and Ion Stoica. 2021. Contrastive code representation learning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021*, pages 5954–5971.

Soyeong Jeong, Jinheon Baek, Sukmin Cho, Sung Ju Hwang, and Jong C. Park. 2022. Augmenting document representations for dense retrieval with interpolation and perturbation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics, ACL 2022*, pages 442–452.

Wei Li, Haozhe Qin, Shuhan Yan, Beijun Shen, and Yuting Chen. 2020. Learning code-query interaction for enhancing code searches. In *2020 IEEE International Conference on Software Maintenance and Evolution (ICSME)*, pages 115–126. IEEE.

Xiaonan Li, Yeyun Gong, Yelong Shen, Xipeng Qiu, Hang Zhang, Bolun Yao, Weizhen Qi, Daxin Jiang, Weizhu Chen, and Nan Duan. 2022. Coderetriever: Unimodal and bimodal contrastive learning. *arXiv preprint arXiv:2201.10866*.

Chao Liu, Xin Xia, David Lo, Cuiyun Gao, Xiaohu Yang, and John Grundy. 2021. Opportunities and challenges in code search tools. *ACM Computing Surveys (CSUR)*, 54(9):1–40.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Fei Lv, Hongyu Zhang, Jian-guang Lou, Shaowei Wang, Dongmei Zhang, and Jianjun Zhao. 2015. Codehow: Effective code search based on api understanding and extended boolean model (e). In *2015 30th IEEE/ACM International Conference on Automated Software Engineering (ASE)*, pages 260–270. IEEE.

Collin McMillan, Mark Grechanik, Denys Poshyvanyk, Qing Xie, and Chen Fu. 2011. Portfolio: finding relevant functions and their usage. In *Proceedings of the 33rd International Conference on Software Engineering*, pages 111–120.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Liming Nie, He Jiang, Zhilei Ren, Zeyi Sun, and Xiaochen Li. 2016. Query expansion based on crowd knowledge for code search. *IEEE Transactions on Services Computing*, 9(5):771–783.

Stephen E Robertson, Steve Walker, Susan Jones, Micheline M Hancock-Beaulieu, Mike Gatford, et al. 1995. Okapi at trec-3. *Nist Special Publication Sp*, 109:109.

Barbara Rosario. 2000. Latent semantic indexing: An overview. *Techn. rep. INFOSYS*, pages 1–16.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*.

Abdus Satter and Kazi Sakib. 2016. A search log mining based query expansion technique to improve effectiveness in code search. In *ICCIT*, pages 586–591. IEEE.

Jianhang Shuai, Ling Xu, Chao Liu, Meng Yan, Xin Xia, and Yan Lei. 2020. Improving code search with co-attentive representation learning. In *Proceedings of the 28th International Conference on Program Comprehension*, pages 196–207.

Lichao Sun, Congying Xia, Wenpeng Yin, Tingting Liang, Philip S. Yu, and Lifang He. 2020. Mixup-transformer: Dynamic data augmentation for NLP tasks. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 3436–3440.

Nandan Thakur, Nils Reimers, Andreas Rücklé, Abhishek Srivastava, and Iryna Gurevych. 2021. BEIR: A heterogeneous benchmark for zero-shot evaluation of information retrieval models. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks 1*.

Aaron Van den Oord, Yazhe Li, and Oriol Vinyals. 2018. Representation learning with contrastive predictive coding. *arXiv e-prints*, pages arXiv–1807.

Thanh Van Nguyen, Anh Tuan Nguyen, Hung Dang Phan, Trong Duc Nguyen, and Tien N Nguyen. 2017. Combining word2vec with revised vector space model for better code retrieval. In *ICSE-C*, pages 183–185. IEEE.

Vikas Verma, Thang Luong, Kenji Kawaguchi, Hieu Pham, and Quoc V. Le. 2021. Towards domain-agnostic contrastive learning. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 10530–10541.

Tongzhou Wang and Phillip Isola. 2020. Understanding contrastive representation learning through alignment and uniformity on the hypersphere. In *International Conference on Machine Learning*, pages 9929–9939. PMLR.

Xin Wang, Yasheng Wang, Fei Mi, Pingyi Zhou, Yao Wan, Xiao Liu, Li Li, Hao Wu, Jin Liu, and Xin Jiang. 2021. Syncobert: Syntax-guided multi-modal contrastive pre-training for code representation. *arXiv preprint arXiv:2108.04556*.

Jason W. Wei and Kai Zou. 2019. EDA: easy data augmentation techniques for boosting performance on text classification tasks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019*, pages 6381–6387.

Kilian Q Weinberger and Lawrence K Saul. 2009. Distance metric learning for large margin nearest neighbor classification. *Journal of machine learning research*, 10(2).

Yangrui Yang and Qing Huang. 2017. Iecs: Intent-enforced code search via extended boolean model. *Journal of Intelligent & Fuzzy Systems*, pages 2565–2576.

Wenpeng Yin, Huan Wang, Jin Qu, and Caiming Xiong. 2021. Batchmixup: Improving training by interpolating hidden states of the entire mini-batch. In *Findings of the Association for Computational Linguistics*, volume ACL/IJCNLP 2021, pages 4908–4912.

A Proof

A.1 Proof of Theorem 1

$$
\begin{aligned}
\mathcal{L}_{\mathcal{N}} &= -\mathbb{E}\left[\log \frac{\exp(q_i \cdot c_i)}{\exp(q_i \cdot c_i) + \sum_{j\neq i}^{B} \exp(q_i \cdot c_j)}\right] \\
&= \mathbb{E}\left[\log \left(1 + \frac{\sum_{j\neq i}^{B} \exp(q_i \cdot c_j)}{\exp(q_i \cdot c_i)}\right)\right] \\
&\geq \mathbb{E}\left[\log \frac{\sum_{j\neq i}^{B} \exp(q_i \cdot c_j)}{\exp(q_i \cdot c_i)}\right] \\
&= \mathbb{E}\left[\log \left(\sum_{j\neq i}^{B} \exp(q_i \cdot c_j)\right) - \log \exp(q_i \cdot c_i)\right] \\
&\approx \mathbb{E}\Big[\log\left[B \cdot \mathbb{E}[\exp(q_i \cdot c_j)]\right]\Big] \\
&\quad - \mathbb{E}\Big[\log \exp(q_i \cdot c_i)\Big]
\end{aligned} \tag{8}
$$

According to the original paper of InfoNCE loss (Van den Oord et al., 2018), the optimal value for $\exp(q \cdot c)$ is given by $\frac{p(c|q)}{p(c)}$, by substituting it, we can get:

$$
\begin{aligned}
\mathcal{L}_{\mathcal{N}} \geq \mathbb{E}\Big[\log\Big[B \cdot \mathbb{E}[\frac{p(c_j|q_i)}{p(c_j)})]\Big]\Big] \\
- \mathbb{E}\Big[\log \frac{p(c_i|q_i)}{p(c_i)}\Big]
\end{aligned} \tag{9}
$$

Since q_i and c_j are negative pairs and sampled independently, $\mathbb{E}[\frac{p(c_j|q_i)}{p(c_j)}] = \mathbb{E}[\frac{p(c_j,q_i)}{p(c_j)p(q_i)}] = \mathbb{E}[\frac{p(c_j)p(q_i)}{p(c_j)p(q_i)}] = 1$. According to the definition of mutual information described in Eq.4, the mutual information $I(q_i, c_i)$ is the second term. Thus, we get:

$$
\mathcal{L}_{\mathcal{N}} \geq \log B - I(q_i, c_i) \tag{10}
$$

Therefore, $I(q, c) \geq \log(B) - \mathcal{L}_{\mathcal{N}}$.

A.2 Proof of Theorem 2

By applying augmentation, the expectation of loss can be divided into four parts that correspond to four types of pairs: original query and original code, augmented query and original code, original query and augmented code, and augmented query and augmented code. It can be expressed as:

$$\mathcal{L}_{\mathcal{N}} = \mathbb{E}\Bigg[-\mathbb{E}\left[\log \frac{\exp(q_i \cdot c_i)}{\exp(q_i \cdot c_i) + \sum_{j\neq i}^{BN}\exp(q_i \cdot c_j)}\right] - \mathbb{E}\left[\log \frac{\exp(q_i \cdot c_i^*)}{\exp(q_i \cdot c_i^*) + \sum_{j\neq i}^{BN}\exp(q_i \cdot c_j)}\right] - \mathbb{E}\left[\log \frac{\exp(q_i^* \cdot c_i)}{\exp(q_i^* \cdot c_i) + \sum_{j\neq i}^{BN}\exp(q_i^* \cdot c_j)}\right] - \mathbb{E}\left[\log \frac{\exp(q_i^* \cdot c_i^*)}{\exp(q_i^* \cdot c_i^*) + \sum_{j\neq i}^{BN}\exp(q_i^* \cdot c_j)}\right]\Bigg] \tag{11}$$

where $q_i \in Q$, $c_i \in C$, $q_i^* \in Q^+$, $c_i^* \in C^+$, and $c_j \in C \cup C^+$. Next, we will analyze the four terms individually. For the first term, original query and original code, it is the same as proved in Appendix A.1. For the second term, original query and augmented code, the derivation can be formulated as:

$$\begin{aligned}
&-\mathbb{E}\left[\log \frac{\exp(q_i \cdot c_i^*)}{\exp(q_i \cdot c_i^*) + \sum_{j\neq i}^{BN}\exp(q_i \cdot c_j)}\right] \\
&= \mathbb{E}\left[\log\left(1 + \frac{\sum_{j\neq i}^{BN}\exp(q_i \cdot c_j)}{\exp(q_i \cdot c_i^*)}\right)\right] \\
&\geq \mathbb{E}\left[\log\left(\frac{\sum_{j\neq i}^{BN}\exp(q_i \cdot c_j)}{\exp(q_i \cdot (\alpha \cdot c_i + \beta \cdot c_j))}\right)\right] \\
&= \mathbb{E}\left[\log\left(\frac{\sum_{j\neq i}^{BN}\exp(q_i \cdot c_j)}{\exp(q_i \cdot c_i \cdot \alpha) \cdot \exp(q_i \cdot c_j \cdot \beta))}\right)\right] \\
&= \mathbb{E}\Big[\log \sum_{j\neq i}^{BN}\exp(q_i \cdot c_j) - \alpha \cdot \log\exp(q_i \cdot c_i) \\
&\quad - \beta \cdot \log\exp(q_i \cdot c_j)\Big]
\end{aligned} \tag{12}$$

Similar to Eq.9, we substitute exponential function with probability, and we could get that the second term is greater than $\log NB - \alpha I(q_i, c_i) - \beta I(q_i, c_j)$. The only difference between the second term and the third term is that in the derivation of second term we decompose c_i^* into $\alpha \cdot c_i + \beta \cdot c_j$ while for the third one we decompose q_i^* into $\alpha \cdot q_i + \beta \cdot q_j$. Therefore, the third term is greater than $\log NB - \alpha I(q_i, c_i) - \beta I(q_j, c_i)$.

For the fourth term, it can be described as:

$$\begin{aligned}
&-\mathbb{E}\left[\log \frac{\exp(q_i^* \cdot c_i^*)}{\exp(q_i^* \cdot c_i^*) + \sum_{j\neq i}^{BN}\exp(q_i^* \cdot c_j)}\right] \\
&\geq \mathbb{E}\left[\log\left(\frac{\sum_{j\neq i}^{BN}\exp(q_i^* \cdot c_j)}{\exp((\alpha \cdot q_i + \beta \cdot q_j)) \cdot (\alpha \cdot c_i + \beta \cdot c_j'))}\right)\right] \\
&= \mathbb{E}\Big[\log \sum_{j\neq i}^{BN}\exp(q_i^* \cdot c_j) - \alpha^2 \cdot \log\exp(q_i \cdot c_i) \\
&\quad - \alpha\beta \cdot \log\exp(q_i \cdot c_j') - \alpha\beta \cdot \log\exp(q_j \cdot c_i) \\
&\quad - \beta^2 \cdot \log\exp(q_j \cdot c_j')\Big] \\
&\geq \log NB - \alpha^2 I(q_i, c_i) - \alpha\beta I(q_i, c_j') \\
&\quad - \alpha\beta I(q_j, c_i) - \beta^2 I(q_j, c_j')
\end{aligned} \tag{13}$$

Then we remove the expectation of Eq.A.2 by multiplying the corresponding probabilities of four terms, we get:

$$\begin{aligned}
\mathcal{L}_{\mathcal{N}} \geq\ & \frac{B^2}{(N+1)^2B^2}\Big(\log NB - I(q_i, c_i)\Big) \\
&+ \frac{NB^2}{(N+1)^2B^2}\Big(\log NB - \alpha I(q_i, c_i) - \beta I(q_i, c_j)\Big) \\
&+ \frac{NB^2}{(N+1)^2B^2}\Big(\log NB - \alpha I(q_i, c_i) - \beta I(q_j, c_i)\Big) \\
&+ \frac{N^2B^2}{(N+1)^2B^2}\Big(\log NB - \alpha^2 I(q_i, c_i) - \alpha\beta I(q_i, c_j') \\
&- \alpha\beta I(q_j, c_i) - \beta^2 I(q_j, c_j')\Big) \\
=\ & \log NB - \Big(\frac{(\alpha N+1)^2}{(N+1)^2} I(q_i, c_i) \\
&+ \frac{N\beta + N^2\alpha\beta}{(N+1)^2} I(q_i, c_j') + \frac{N\beta + N^2\alpha\beta}{(N+1)^2} I(q_j, c_i) \\
&+ \frac{N^2\beta^2}{(N+1)^2} I(q_j, c_j')\Big) \\
\geq\ & \log NB - \Big(\alpha^2 I(q_i, c_i) + \alpha\beta \cdot I(q_i, c_j') \\
&+ \alpha\beta \cdot I(q_j, c_i) + \beta^2 \cdot I(q_j, c_j')\Big)
\end{aligned} \tag{14}$$

Therefore, we get:

$$\begin{aligned}
I(q_i, c_i) \geq & \frac{1}{\alpha^2}\Big(\log(NB) - \mathcal{L}_{\mathcal{N}} \\
& - \alpha\beta \cdot I(q_i, c_j') - \alpha\beta \cdot I(q_j, c_i) \\
& - \beta^2 \cdot I(q_j, c_j')\Big)
\end{aligned} \tag{15}$$

B Representation-level augmentation on other loss functions

Here, we investigate whether representation-level augmentation also benefits other loss functions. We apply representation-level augmentation on triplet loss (Mikolov et al., 2013) and logistic loss (Weinberger and Saul, 2009), because these two loss functions are also used prior to InfoNCE loss. Triplet

loss learns to maximize the distances between negative pairs while minimizing the distances between positive pairs, which can be written as:

$$
\begin{aligned}
L_{triplet} = - \frac{1}{N} \sum_{i=1}^{N} max(0, \\
||q_i - c_i||_2^2 - ||q_i - c_{j,j \neq i}||_2^2 + \epsilon)
\end{aligned} \quad (16)
$$

And logistic loss is also called NCE loss, which can be described as:

$$
\begin{aligned}
L_{logistic} = - \frac{1}{N} \sum_{i=1}^{N} \Big[\log \sigma(q_i \cdot c_i) \\
- \frac{1}{N-1} \sum_{j \neq i} \log \sigma(q_i \cdot c_j) \Big]
\end{aligned} \quad (17)
$$

where definitions of variables follow Equation 3, σ represents sigmoid function, and we set $\epsilon = 5$ in our experiments.

We finetune CodeBERT and GraphCodeBERT with the two loss functions for 20 epochs on Ruby and Javascript dataset. Results of triplet loss and logistic loss are shown in Table 6 and Table 7, respectively.

Model	Ruby		JavaScript	
	Original	w/ RA	Original	w/ RA
CodeBERT	0.594	**0.610**	0.525	**0.533**
GraphCodeBERT	0.683	**0.698**	0.620	**0.637**

Table 6: Performance of triplet loss under MRR. “w/ RA” stands for “with representation-level augmentation”.

Model	Ruby		JavaScript	
	Original	w/ RA	Original	w/ RA
CodeBERT	0.513	**0.522**	0.503	**0.513**
GraphCodeBERT	0.696	**0.705**	0.612	**0.632**

Table 7: Performance of logistic loss under MRR. “w/ RA” stands for “with representation-level augmentation”.

As we can see, representation-level augmentation can also improve performance when other loss functions are used. We believe this is because the augmentation makes the model more robust.