RETRIEVAL-AUGMENTED GENERATION FOR CODE SUMMARIZATION VIA HYBRID GNN

Anonymous authors
Paper under double-blind review

ABSTRACT

Source code summarization aims to generate natural language summaries from structured code snippets for better understanding code functionalities. However, automatic code summarization is challenging due to the complexity of the source code and the language gap between the source code and natural language summaries. Most previous approaches either rely on retrieval-based (which can take advantage of similar examples seen from the retrieval database, but have low generalization performance) or generation-based methods (which have better generalization performance, but cannot take advantage of similar examples). This paper proposes a novel retrieval-augmented mechanism to combine the benefits of the both worlds. Furthermore, to mitigate the limitation of Graph Neural Networks (GNNs) on capturing global graph structure information of source code, we propose a novel attention-based dynamic graph to complement the static graph representation of the source code, and design a hybrid message passing GNN for capturing both the local and global structural information. To evaluate the proposed approach, we release a new challenging benchmark, crawled from diversified large-scale open-source *C* projects (total **95k+** unique functions in the dataset). Our method achieves the state-of-the-art performance, improving existing methods by **1.65**, **1.76** and **1.81** in terms of BLEU-4, ROUGE-L and METEOR.

1 INTRODUCTION

With software growing in size and complexity, developers tend to spend nearly 90% (Wan et al., 2018) effort on software maintenance (*e.g.*, version iteration and bug fix) in the completed life cycle of software development. Source code summary, in the form of natural language, plays a critical role in comprehension and maintenance process and greatly reduces the effort of reading and comprehending programs. However, manually writing code summaries is tedious and time-consuming, and with the acceleration of software iteration, it has become a heavy burden for software developers. Hence, source code summarization which automates concise descriptions of programs is meaningful.

Automatic source code summarization is a crucial yet far from settled problem. The key challenges include: 1) the source code and the natural language summary are heterogeneous, which means they may not share common lexical tokens, synonyms, or language structures and 2) the source code is complex with complicated logic and variable grammatical structure, making it hard to learn the semantics. Conventionally, information retrieval (IR) techniques have been widely used in code summarization (Eddy et al., 2013; Haiduc et al., 2010; Wong et al., 2015; 2013). Since code duplication (Kamiya et al., 2002; Li et al., 2006) is common in "big code" (Allamanis et al., 2018), early works summarize the new programs by retrieving the similar code snippet in the existing code database and use its summary directly. Essentially, the retrieval-based approaches transform the code summarization to the code similarity calculation task, which may achieve promising performance on similar programs, but are limited in generalization, *i.e.* they have poorer performance on programs that are very different from the code database.

To improve the generalization performance, recent works focus on generation-based approaches. Some works explore Seq2Seq architectures (Bahdanau et al., 2014; Luong et al., 2015) for generate summaries from the given source code. The Seq2Seq-based approaches (Iyer et al., 2016; Hu et al., 2018a; Alon et al., 2018) usually treat the source code or abstract syntax tree parsed from the source code as a sequence and follow a paradigm of encoder-decoder with attention mechanism for generating a summary. However, these works only rely on sequential models, which are struggling to capture the rich semantics of source code *e.g.*, control dependencies and data dependencies. In addition, generation-based approaches typically cannot take advantage of similar examples from the retrieval database, as retrieval-based approaches do.

To better learn the semantics of the source code, Allamanis et al. (Allamanis et al., 2017) lighted up this field by representing programs as graphs. Some follow-up works (Fernandes et al., 2018) attempted to encode more code structures (*e.g.*, control flow, program dependencies) into code graphs with graph neural networks (GNNs), and achieved the promising performance than the sequence-based approaches. Existing works (Allamanis et al., 2017; Fernandes et al., 2018) usually convert code into graph-structured input during preprocessing, and directly consume it via modern neural networks (*e.g.*, GNNs) for computing node and graph embeddings. However, most GNN-based encoders only allow message passing among nodes within a k-hop neighborhood (where k is usually a small number such as 4) to avoid over-smoothing (Zhao & Akoglu, 2019; Chen et al., 2020), thus capture only local neighborhood information and ignore global interactions among nodes. Even there are some works (Li et al., 2019) that try to address this challenging with deep GCNs (i.e., 56 layers) (Kipf & Welling, 2016) by the residual connection (He et al., 2016), however, the computation cost cannot endure in the program especially for a large and complex program. For example, on our benchmark, the average/max node size of functions are 70/200 and the average node degree is 1.77.

To address these challenges, we propose a framework for automatic code summarization, namely Hybrid-GNN *(HGNN)*. Specifically, from the source code, we first construct a code property graph (CPG) based on abstract syntax tree (AST) with different types of edges (*i.e*., Flow To, Reach). In order to combine the benefits of both retrieval-based and generation-based methods, we propose a novel *retrieval-based augmentation mechanism* to retrieve the source code that is most similar to the current program from the retrieval database (excluding the current program itself), and add the retrieved code as well as the corresponding summary as auxiliary information for training the model. In order to go beyond local graph neighborhood information, and capture global interactions in the program, we further propose an attention-based dynamic graph by learning global attention scores (*i.e*., edge weights) in the augmented static CPG. Then, a hybrid message passing (HMP) is performed on both static and dynamic graphs. We also release a new code summarization benchmark by crawling data from popular and diversified projects containing **95k+** functions in *C* programming language and make it public [1]. We highlight our main contributions as follows:

- We propose a general-purpose framework for automatic code summarization, which combines the benefits of both retrieval-based and generation-based methods via a novel retrieval-based augmentation mechanism.
- We innovate a Hybrid-GNN by fusing the static graph (based on code property graph) and dynamic graph (via structure-aware global attention mechanism) to mitigate the limitation of the GNN on capturing global graph information.
- We release a new challenging *C* benchmark for the task of source code summarization.
- We conduct an extensive experiment to evaluate our framework. The proposed approach achieves the state-of-the-art performance and improves existing approaches by **1.65**, **1.76** and **1.81** in terms of BLEU-4, ROUGE-L and METEOR metrics.

2 Hybrid-GNN Framework

In this section, we introduce the proposed framework Hybrid-GNN *(HGNN)*, as shown in Figure 1, which mainly includes four components: 1) Retrieval-augmented Static Graph Construction *(c.f.,* Section 2.2), which incorporates retrieved code-summary pairs to augment original code for learning. 2) Attention-based Dynamic Graph Construction (*c.f.,* Section 2.3), which allows message passing among any pair of nodes via a global attention mechanism. 3) *HGNN*, (*c.f.,* Section 2.4), which incorporates information from both static graphs and dynamic graphs with Hybrid Message Passing. 4) Decoder (*c.f.,* Section 2.5), which utilizes an attention-based LSTM (Hochreiter & Schmidhuber, 1997) model to generate a summary.

[1] C-Code-Summarization Benchmark

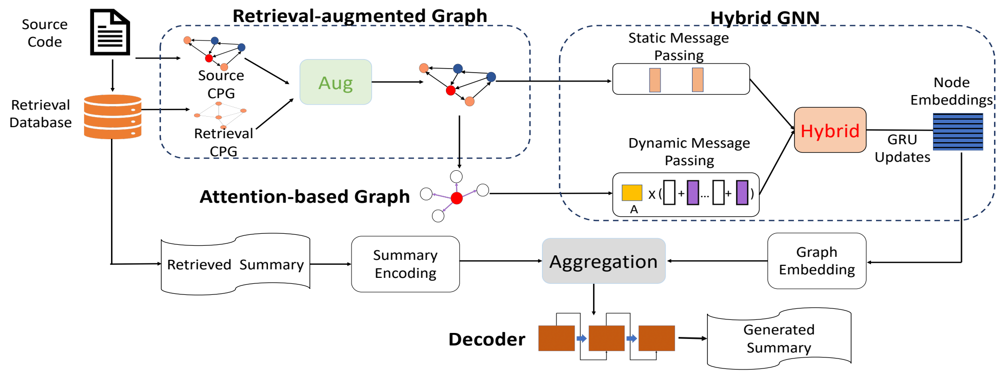

Figure 1: The framework of our Hybrid-GNN. Best viewed in color.

2.1 Problem Formulation

In this work, we focus on generating summaries for the given functions (Wan et al., 2018; Zhang et al., 2020). We define a dataset as $D = \{(c, s)|c \in C, s \in S\}$, where c is the source code of a function in the function set C and s represents its targeted summary in S. The task of code summarization is, given a source code c, to generate the best summary consisting of a sequence of tokens $\hat{s} = (t_1, t_2, ..., t_T)$ which maximizes the conditional likelihood $\hat{s} = \text{argmax}_s P(s|c)$. In this paper, we follow the problem setting and propose the technique to learn the mapping from the source code to the natural language summary.

2.2 Retrieval-augmented Static Graph

2.2.1 Graph Initialization

The source code of a function can be represented as Code Property Graph (CPG) (Yamaguchi et al., 2014), which is built on the abstract syntax tree (AST) with different type of edges (*i.e.*, Flow To, Control, Define/Use, Reach). Formally, one raw function c could be represented by a multi-edged graph $g(\mathcal{V}, \mathcal{E})$, where $\mathcal{V}$ is the set of AST nodes, $(v, u) \in \mathcal{E}$ denotes the edge between the node v and the node u. A node v consists of two parts: the *node sequence* and the *node type*. An illustrative example is shown in Figure 2. For example, in the red node, $a\%2 == 0$ is the node sequence and $Condition$ is the node type. An edge (v, u) has a type, named *edge type*, *e.g.*, AST type and Flow To type. For more details about the CPG, please refer to Appendix A.

Initialization Representation. Given a CPG, we utilize a BiLSTM to encode its nodes. We represent each token of the node sequence, each node type and each edge type using the learned embedding matrix $\boldsymbol{E}^{seqtoken}$, $\boldsymbol{E}^{nodetype}$ and $\boldsymbol{E}^{edgetype}$, respectively. Then nodes and edges of the CPG can be encoded as:

$$\begin{aligned} \boldsymbol{h}_1, ..., \boldsymbol{h}_l &= \text{BiLSTM}(\boldsymbol{E}^{seqtoken}_{v,1}, ..., \boldsymbol{E}^{seqtoken}_{v,l}) \\ encode\_node(v) &= \text{linear}(\text{concat}[\boldsymbol{E}^{nodetype}_v; \boldsymbol{h}^{\rightarrow}_1; \boldsymbol{h}^{\leftarrow}_l]) \\ encode\_edge(v,u) &= \boldsymbol{E}^{edgetype}_{v,u} \;\; if \;\; (v,u) \in \mathcal{E} \;\; else \;\; \boldsymbol{0} \end{aligned} \tag{1}$$

where l is the number of tokens in the node sequence of v. For the sake of simplicity, in the following section, we use $\boldsymbol{h}_v$ and $\boldsymbol{e}_{v,u}$ to represent the embedding of the node v and the edge (v, u), respectively, *i.e.*, $encode\_node(v)$ and $encode\_edge(v, u)$. Given the source code c of a function as well as the CPG $g(\mathcal{V}, \mathcal{E})$, $\boldsymbol{H}_c \in \mathbb{R}^{m \times d}$ denotes the initial node matrix of the CPG, where m is the total number of nodes in the CPG and d is the dimension of the node embedding.

2.2.2 Retrieval-based Augmentation

While retrieval-based methods can perform reasonably well on examples that are similar to those examples from a retrieval database, they typically have low generalization performance and might perform poorly on dissimilar examples. On the contrary, generation-based methods usually have better generalization performance, but cannot take advantage of similar examples from the retrieval database.

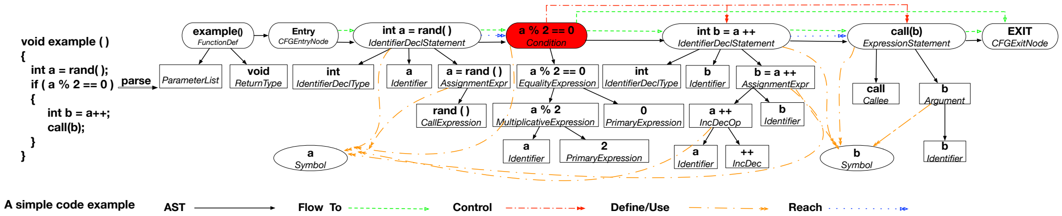

Figure 2: An example of Code Property Graph (CPG).

In this work, we propose to combine the benefits of the two worlds, and design a retrieval-augmented generation framework for the task of code summarization.

In principle, the goal of code summarization is to learn a mapping from source code c to the natural language summary $s = f(c)$. In other words, for any source code c', a code summarization system can produce its summary $s' = f(c')$. Inspired by this observation, conceptually, we can derive the following formulation $s = f(c) - f(c') + s'$. This tells us that we can actually compute the semantic difference between c and c', and further obtain the desired summary s for c by considering both the above semantic difference and s' which is the summary for c'. Mathmatically, our goal becomes to learn a function which takes as input c, c' and s' and outputs the summary s for c, that is, $s = g(c, c', s')$. This motivates us to design our Retrieval-based Augmentation mechanism, as detailed below.

Step 1: Retrieving. For each sample $(c, s) \in D$, we retrieve the most similar sample: $(c', s') = \text{argmax}_{(c',s') \in D'} sim(c, c')$, where $c \neq c'$, D' is a given retrieval database and $sim(c, c')$ is the text similarity. Following Zhang et al. (2020), we utilize Lucene for retrieval and calculate the similarity score z between the source code c and the retrieved code c' by dynamic programming (Bellman, 1966) $z = 1 - \frac{dis(c,c')}{max(|c|,|c'|)}$, where $dis(c, c')$ is the text edit distance.

Step 2: Retrieved Code-based Augmentation. Given the retrieved source code c' for the current sample c, we adopt a fusion strategy to inject retrieved semantics into the current sample. The fusion strategy is based on their initial graph representations ($\boldsymbol{H}_c$ and $\boldsymbol{H}_{c'}$) with an attention mechanism:

- To capture the relevance between c and c', we design an attention function, which computes the attention score matrix $\boldsymbol{A}^{aug}$ based on the embedding of each pair of nodes in CPGs of c and c':

$$\boldsymbol{A}^{aug} \propto \exp(\text{ReLU}(\boldsymbol{W}\boldsymbol{H}_c)\text{ReLU}(\boldsymbol{W}\boldsymbol{H}_{c'})^T) \tag{2}$$

  where $\boldsymbol{W} \in \mathbb{R}^{d \times d}$ is the matrix with d-dim embedding size and ReLU is the rectified linear unit.

- Multiply the attention matrix $\boldsymbol{A}^{aug}$ with the retrieved representation $\boldsymbol{H}_{c'}$ to inject retrieved features into $\boldsymbol{H}_c$:

$$\boldsymbol{H}'_c = z\boldsymbol{A}^{aug}\boldsymbol{H}_{c'} \tag{3}$$

  where z is the similarity score, which is introduced to weaken the negative impact of c' on the original training data c, *i.e.*, when the similarity of c and c' is low.

- Finally, we merge $\boldsymbol{H}'_c$ and the original $\boldsymbol{H}_c$ to get the final representation of c.

$$\boldsymbol{comp} = \boldsymbol{W}_c\boldsymbol{H}_c + \boldsymbol{W}'_c\boldsymbol{H}'_c \tag{4}$$

  where $\boldsymbol{W}_c, \boldsymbol{W}'_c \in \mathbb{R}^{d \times d}$ are weighted matrices and $\boldsymbol{comp}$ is the retrieval augmented node representation.

Step 3: Retrieved Summary-based Augmentation. We further encode the retrieved summary s' with another BiLSTM model. We represent each token t'_i of s' using the learned embedding matrix $\boldsymbol{E}^{seqtoken}$. Then s' can be encoded as:

$$\boldsymbol{h}_{t'_1}, ..., \boldsymbol{h}_{t'_T} = \text{BiLSTM}(\boldsymbol{E}^{seqtoken}_{t'_1}, ..., \boldsymbol{E}^{seqtoken}_{t'_T}) \tag{5}$$

where $h_{t'_i}$ is the state of the BiLSTM model for the token t'_i in s' and T is the length of s'. We also multiply the similarity score z to $[\boldsymbol{h}_{t'_1}, ..., \boldsymbol{h}_{t'_T}]$ and concatenate with the graph encoding results (i.e., the outputs of the GNN encoder) as the input $[\text{GNN}_{\text{output}}, z\boldsymbol{h}_{t'_1}, ..., z\boldsymbol{h}_{t'_T}]$ to the decoder.

2.3 ATTENTION-BASED DYNAMIC GRAPH

Due to that GNN-based encoders usually consider the k-hop neighborhood, the global relation among nodes in the static graph (see Section 2.2.1) may be ignored. In order to better capture the global semantics of source code, based on the static graph, we propose to dynamically construct a graph via structure-aware global attention mechanism, which allows message passing among any pair of nodes. The attention-based dynamic graph can better capture the global dependency among nodes, and thus supplement the static graph.

Structure-aware Global Attention. The construction of the dynamic graph is motivated by the structure-aware self-attention mechanism proposed in Zhu et al. (2019). Given the static graph, we compute a corresponding dense adjacency matrix $\boldsymbol{A}^{dyn}$ based on a structure-aware global attention mechanism and the constructed graph namely *attention-based dynamic graph*. Unlike the self-attention mechanisms in Zhu et al. (2019), we consider not only the node semantics but also the edges in the static graph (*i.e.*, the CPG of the training data) when computing attention scores between any pair of nodes.

$$\boldsymbol{A}^{dyn}_{v,u} = \frac{\mathrm{ReLU}(\boldsymbol{h}_v \boldsymbol{W}^Q)(\mathrm{ReLU}(\boldsymbol{h}_u \boldsymbol{W}^K) + \mathrm{ReLU}(\boldsymbol{e}_{v,u} \boldsymbol{W}^R))}{\sqrt{d}} \tag{6}$$

where $\boldsymbol{h}_v, \boldsymbol{h}_u \in \boldsymbol{comp}$ are the augmented node embedding for any node pair (v, u) in the CPG. Note that the global attention considers each pair of nodes of the CPG, regardless of whether there is an edge between them. $\boldsymbol{e}_{v,u} \in \mathbb{R}^{d_e}$ is the edge embedding and $\boldsymbol{W}^Q, \boldsymbol{W}^K \in \mathbb{R}^{d \times d}$, $\boldsymbol{W}^R \in \mathbb{R}^{d_e \times d}$ are parameter matrices, d_e and d are the dimensions of edge embedding and node embedding, respectively. The adjacency matrix $\boldsymbol{A}^{dyn}$ normalizes with softmax function, which will be used to compute dynamic message passing (see Section 2.4).

$$\boldsymbol{A}^{dyn} = \mathrm{softmax}(\boldsymbol{A}^{dyn}) \tag{7}$$

2.4 HYBRID GNN

To better incorporate the information of the static graph and the dynamic graph, we propose the Hybrid Message Passing (HMP), which are performed on both retrieval-augmented static graph and attention-based dynamic graph.

Static Message Passing. We incorporate the edge type embedding to encode the static graph. For every node v at each computation hop k, we apply an aggregation function to calculate the aggregated vector $\boldsymbol{h}^k_v$ by consdering a set of neighboring node embeddings computed from the previous hop.

$$\boldsymbol{h}^k_v = \mathrm{SUM}(\{\boldsymbol{h}^{k-1}_u + \boldsymbol{e}_{v,u} \boldsymbol{W}^P | \forall u \in \mathcal{N}_{(v)}\}) \tag{8}$$

where $\boldsymbol{W}^P \in \mathbb{R}^{d_e \times d}$ is the weighted matrix and $\mathcal{N}_{(v)}$ is a set of the neighboring nodes which are directly connected with v. For each node v, $\boldsymbol{h}^0_v$ is the initial augmented node embedding of v, *i.e.*, $\boldsymbol{h}_v \in \boldsymbol{comp}$.

Dynamic Message Passing. The node information and edge information are propagated on the attention-based dynamic graph with the adjacency matrices $\boldsymbol{A}^{dyn}$, defined as

$$\boldsymbol{h}^{'k}_v = \sum_{j=1}^{m} \boldsymbol{A}^{dyn}_{v,v_j} (\boldsymbol{h}^{'k-1}_{v_j} \boldsymbol{W}^V + \boldsymbol{e}_{v,v_j} \boldsymbol{W}^F) \tag{9}$$

where m is the total number of nodes, v_j is the j^{th} node, $\boldsymbol{W}^V \in \mathbb{R}^{d \times d}$, $\boldsymbol{W}^F \in \mathbb{R}^{d_e \times d}$ are learned matrices, and $\boldsymbol{e}_{v,v_j}$ is the edge embedding between the edge v, v_j. Similarly, $\boldsymbol{h}^{'0}_{v_j}$ is the initial augmented node embedding of v in $\boldsymbol{comp}$.

Hybrid Message Passing. Given the static/dynamic aggregated vectors $\boldsymbol{h}^k_v / \boldsymbol{h}^{'k}_v$ for static and dynamic graphs, we fuse both vectors and feed the resulting vector to a Gated Recurrent Unit (GRU) to update node representations.

$$\boldsymbol{f}^k_v = \mathrm{GRU}(\boldsymbol{f}^{k-1}_v, \mathrm{Fuse}(\boldsymbol{h}^k_v, \boldsymbol{h}^{'k}_v)) \tag{10}$$

where $\boldsymbol{f}^0_v$ is the augmented node initialization in $\boldsymbol{comp}$. The fusion function Fuse is designed as a gated sum of two inputs.

$$\mathrm{Fuse}(\boldsymbol{a}, \boldsymbol{b}) = \boldsymbol{z} \odot \boldsymbol{a} + (1 - \boldsymbol{z}) \odot \boldsymbol{b} \quad \boldsymbol{z} = \sigma(\boldsymbol{W}_z[\boldsymbol{a}; \boldsymbol{b}; \boldsymbol{a} \odot \boldsymbol{b}; \boldsymbol{a} - \boldsymbol{b}] + \boldsymbol{B}_z) \tag{11}$$

where $\odot$ is the component-wise multiplication, σ is a sigmoid function and z is a gating vector. After n hops of GNN computation, we obtain the final node representation $\boldsymbol{f}_v^n$ and then apply max-pooling over all nodes $\{\boldsymbol{f}_v^n | \forall v \in \mathcal{V}\}$ to get the graph representation.

2.5 DECODER

The decoder is similar with other state-of-the-art Seq2seq models (Bahdanau et al., 2014; Luong et al., 2015) where an attention-based LSTM decoder is used. The decoder takes the input of the concatenation of the node representation and the representation of the retrieved summary (*i.e.*, $s' = (t'_1, ..., t'_T)$): $[\boldsymbol{f}_{v_1}^n; ...; \boldsymbol{f}_{v_m}^n; z\boldsymbol{h}_{t'_1}; ...; z\boldsymbol{h}_{t'_T}]$, where n is the number of hops and m is the number of nodes in the CPG. The initial hidden state of decoder is the fusion of graph representation and the weighted (i.e., multiply similarity score z) final state of retrieved summary.

We train the model with regular cross-entropy loss, defined as $\mathcal{L} = \sum_1^t -\log P(s_t^* | c, s_{<t}^*)$, where s_t^* is the word at the t-th position of the ground-truth output and c is the source code of the function. To alleviate the exposure bias, we utilize schedule teacher forcing (Bengio et al., 2015). During the inference, we use beam search to generate final results.

3 EXPERIMENTS

3.1 SETUP

We evaluate our proposed framework against a number of state-of-the-art methods. Specifically, we classify the selected baseline methods into three groups: 1) Retrieval-based approaches: TF-IDF (Haiduc et al., 2010) and NNGen (Liu et al., 2018), 2) Sequence-based approaches: CODE-NN (Iyer et al., 2016; Barone & Sennrich, 2017), Transformer (Ahmad et al., 2020), Hybrid-DRL (Wan et al., 2018), Rencos (Zhang et al., 2020) and Dual model (Wei et al., 2019), 3) Graph-based approaches: SeqGNN (Fernandes et al., 2018). In addition, we implemented two another graph-based baselines: GCN2Seq and GAT2Seq, which respectively adopt the Graph Convolution (Kipf & Welling, 2016) and Graph Attention (Velickovic et al., 2018) as the encoder and a LSTM as the decoder for generating summaries. Note that Rencos (Zhang et al., 2020) combines the retrieval information into Seq2Seq model, we classify it into Sequence-based approaches. More detailed description about baselines and the configuration of *HGNN* can be found in the Appendix B and C.

Existing benchmarks (Barone & Sennrich, 2017; Hu et al., 2018b) are all based on high-level programming language *i.e.*, Java, Python. Furthermore, they have been confirmed to have extensive duplication, making model overfit to the training data that overlapped with the testset (Fernandes et al., 2018; Allamanis, 2019). We are the first to explore neural summarization on *C* programming language and make our benchmark public to benefit the academia and industry. We crawled from popular *C* repositories on GitHub and extract function-summary pairs based on the documents of functions. After a strict deduplication process, we kept **95k+** unique function-summary pairs and name it *C* Code Summarization Dataset (CCSD). To further test the model generalization ability, we construct in-domain functions and out-of-domain functions by dividing the projects into two sets, denoted as a and b. For each project in a, we randomly select some of functions in this project as the training data and the unselected functions are the in-domain validation/test data. All functions in projects b are regarded as out-of-domain test data. Finally, we obtain 82,656 training functions, 4,340 in-domain validation functions, 4,124 in-domain test functions and 2,264 out-of-domain test functions. For the retrieval augmentation, we also use the training set as the retrieval database, *i.e.*, $D' = D$ (see Step 1 in Section 2.2). More details about data processing, please refer to Appendix D.

Similar to previous works (Zhang et al., 2020; Wan et al., 2018; Fernandes et al., 2018; Iyer et al., 2016), BLEU (Papineni et al., 2002), METEOR (Banerjee & Lavie, 2005) and ROUGE-L (Lin, 2004) are used as our automatic evaluation metrics. These metrics are popular in machine translation, text summarization. Except for these similarity-based metrics, we also conduct a human evaluation study to evaluate semantic similarity. We invite 5 Ph.D students and 10 master students as the volunteers, who have rich C programming experiences. The volunteers are asked to rank summaries generated from the anonymized approaches from 1 to 5 (*i.e.*, 1: Poor, 2: Marginal, 3: Acceptable, 4: Good, 5: Excellent) based on the relevance of the generated summary to the source code and the degree of similarity between the generated summary and the actual summary. Specifically, we randomly

Table 1: Automatic evaluation results (in %) on the CCSD test set.

Methods	In-domain			Out-of-domain			Overall		
	BLEU-4	ROUGE-L	METEOR	BLEU-4	ROUGE-L	METEOR	BLEU-4	ROUGE-L	METEOR
TF-IDF	15.20	27.98	25.91	5.50	15.37	13.12	12.19	23.49	21.34
NNGen	15.97	28.14	26.11	5.74	16.33	14.27	12.76	23.93	21.96
CODE-NN	9.02	26.94	22.54	4.77	20.91	18.52	7.77	25.15	21.11
Hybrid-DRL	9.29	30.00	24.59	6.30	24.19	**21.85**	8.42	28.64	23.62
Transformer	12.91	28.04	18.47	5.75	18.62	15.00	10.69	24.65	17.27
Dual Model	11.49	29.20	25.25	5.25	21.31	18.34	9.61	26.40	22.80
Rencos	14.47	31.61	28.55	6.50	22.81	18.74	11.74	28.47	24.41
GCN2Seq	9.79	26.59	22.58	4.06	18.96	16.03	7.91	23.67	20.08
GAT2Seq	10.52	26.17	22.89	3.80	16.94	13.96	8.29	22.63	19.47
SeqGNN	10.51	29.84	25.04	4.94	20.80	18.17	8.87	26.34	22.97
HGNN w/o augment & dynamic	12.00	29.06	25.23	4.65	21.06	18.05	9.64	26.00	22.48
HGNN w/o augment & static	11.87	29.36	25.27	5.31	21.90	18.65	9.75	26.88	23.12
HGNN w/o augment	12.43	30.05	25.75	5.56	22.64	18.27	9.87	27.04	23.16
HGNN w/o summary augment	13.37	30.36	26.13	5.81	22.97	19.05	10.34	27.43	23.82
HGNN w/o code augment	15.10	32.19	27.83	6.94	23.80	20.44	12.01	28.79	24.93
HGNN w/o static	15.65	32.72	28.78	6.98	24.03	21.16	12.78	29.20	25.48
HGNN w/o dynamic	15.34	32.13	28.01	6.91	23.95	20.53	12.21	29.07	25.14
HGNN	**16.24**	**33.62**	**29.60**	**7.62**	**24.77**	20.78	**13.39**	**30.23**	**26.22**

choose 50 functions for per approach with the corresponding generated summaries and ground-truths. After the summarizes are ranked, we calculate the average score for each function. Higher scores mean better quality.

3.2 Comparison with the baselines

Table 1 shows the evaluation results including two parts: the comparision with baselines and the ablation study. Consider the comparison with state-of-the-art baselines, in general, we find that our proposed model outperforms existing methods by a significant margin on both in-domain and out-of-domain datasets, and shows good generalization performance. Compared with others, on in-domain dataset, the simple retrieval-based approaches could achieve competitive performance on BLEU-4, however ROUGE-L and METEOR are fare less than ours. They also do not perform well on out-of-domain dataset. Even without augmentation (*HGNN w/o augment*), our approach still outperforms the graph-based approaches (*i.e.*, GCN2Seq, GAT2Seq and SeqGNN), which further demonstrates the effectiveness of Hybrid-GNN for additionally capturing global graph information. Compared with Sequence-based approaches, *HGNN w/o augment* outperforms the majority models, except for Hybrid-DRL and Rencos. For Hybrid-DRL with a better performance on the out-of-domain functions, we ascribe to the advantages of deep reinforcement learning to optimize evaluation metrics. Compared with Rencos that also considers the retrieved information in the Seq2Seq model, we could find that its performance is still lower than *HGNN*. On the overall dataset including both of in-domain and out-of-domain data, our model achieves **13.39**, **30.23** and **26.22**, outperforming existing methods by **1.65**, **1.76** and **1.81** in terms of BLEU-4, ROUGE-L and METEOR metrics.

3.3 Ablation Study

We also conduct an ablation study to evaluate the impact of different components of our framework, *e.g.*, retrieval-based augmentation, static graph and dynamic graph in the last row of Table 1. Overall, we found that: retrieval-augmented mechanism significantly contributed to the overall model performance (*HGNN* vs. *HGNN w/o augment*). More specifically, we noticed that summary-based augmentation has the most impact (*HGNN* vs. *HGNN w/o summary augment*). Besides, considering both summary and code augmentation further significantly improved the performance compared to considering only summary augmentation (*HGNN* vs. *HGNN w/o code augment*). The summary-based augmentation is more useful, we conjecture that it depends on the specific task: 1) this task is to generate summary and 2) the code and summary are heterogeneous data. Thus, summary-based augmentation could provide a more direct signal for generating better summaries. However, the code-based augmentation could further improve the performance by enhancing the semantic learning of the program. Combining them together, our method achieves the best result. Similarly, consider results in *HGNN w/o static* and *HGNN w/o dynamic*, we see that: 1) their performance decreases, which demonstrates the effectiveness of the Hybrid-GNN and 2) the performance without dynamic graph is worse than the performance without static graph, which demonstrates the usefulness of

Table 2: Human evaluation results on the CCSD test set.

Metrics	NNGen	Transformer	Rencos	SeqGNN	***HGNN***
Relevance	3.16	3.17	3.31	3.46	**3.64**
Similarity	3.08	3.02	3.16	3.14	**3.47**

Table 3: Examples of generated summaries on the CCSD test set.

Example	Example 1	Example 2
Source Code	`static void strInit(Str *p){ p->z = 0; p->nAlloc = 0; p->nUsed = 0; }`	`void ReleaseCedar(CEDAR *c){ if (c == NULL) return; if (Release(c->ref) == 0) CleanupCedar(c); }`
Ground-Truth	initialize a str object	release reference of the cedar
NNGen	free the string	release the virtual host
Transformer	initialize the string	release cedar communication module
Rencos	initialize a floating poing string	release of the cancel object
SeqGNN	initialize the string	release cedar communication mode
HGNN	**initialize a str object**	**release reference of cedar**

dynamic graph that captures the global structural information. Similarly, we also evaluate the performance without augmentation and static graph/dynamic graph (see *HGNN w/o augment& static* and *HGNN w/o augment& dynamic*). Compared with *HGNN w/o augment*, the results further confirm the effectiveness of the Hybrid-GNN.

3.4 Human Evaluation

As shown in Table 2, we perform a human evaluation on the overall dataset to assess the quality of the generated summaries by our approach, NNGen, Transformer, Rencos and SeqGNN in terms of relevance and similarity. As depicted in Table 1, NNGen, Rencos and SeqGNN are the best retrieval-based, sequence-based, and graph-based approaches, respectively. We also compare with Transformer as it has been widely used in natural language processing. Inspection on the results, our method can generate better summaries which are more relevant with the source code and more similar with the ground-truth summaries.

3.5 Case Study

To perform qualitative analysis, we present two examples with generated summaries by different methods from the overall data set, shown in Table 3. We can see that, in the first example, our approach learns more code semantics, *i.e.*, *p* is a self-defined struct variable. Thus, we could generate a token *object* for the variable *p*. However, other models can only produce *string*. Example 2 is a more difficult function with the functionality to "release reference", as compared to other baselines, our approach effectively captures the functionality and generates more precise summary.

3.6 Extension on the public dataset

We conducted additional experiments on a public dataset, i.e., the Python Code Summarization Dataset (PCSD), which was also used in Rencos (the most competitive baseline in our paper). The total number of code samples in PCSD is 109,726 (Barone & Sennrich, 2017). This number is comparable to the size (i.e., $\sim$ 95k) of our own CCSD benchmark. We follow the setting of Rencos and split PSCD into the training set, validation set and testing set with fractions of 60%, 20% and 20%. We construct the static graph and compare our methods on PCSD against various competitive baselines, i.e., NNGen, CodeNN, Rencos and Transformer, which are either retrieval-based, generation-based or hybrid methods. The results are shown in Table 4. We can see that our method outperforms NNGen, CODENN, Rencos and Transformer by 0.95, 3.27 and 1.12 in terms of BLEU-4, ROUGE-L and METEOR. We also perform the ablation study on PSCD to demonstrate the usefulness of the static graph (i.e., HGNN w/o dynamic) and dynamic graph (i.e., HGNN w/o static). The results also demonstrate that both static graph and dynamic graph can contribute to our framework. In summary, the results on both our released benchmark (C benchmark) and existing benchmark (PCSD) demonstrate the effectiveness of our method.

Table 4: Evaluation on PCSD.

Methods	BLEU-4	ROUGE-L	METEOR
NNGen	21.60	31.61	15.96
CODE-NN	16.39	28.99	13.68
Transformer	17.06	31.16	14.37
Rencos	22.24	36.00	18.26
HGNN w/o static	21.82	38.61	18.36
HGNN w/o dynamic	21.75	38.37	18.42
HGNN	**23.19**	**39.27**	**19.38**

4 Related Work

Source Code Summarization Early works (Eddy et al., 2013; Haiduc et al., 2010; Wong et al., 2015; 2013) for code summarization focused on using information retrieval to retrieve summaries. Later works attempted to employ attentional Seq2Seq model on the source code (Iyer et al., 2016) or some variants from code text, i.e., AST (Hu et al., 2018a; Alon et al., 2018) to generate summaries. However, these works are based on sequential models, ignoring rich code semantics. Some latest attempts (LeClair et al., 2020; Fernandes et al., 2018) embedded program semantics into graph neural networks. However, these works mainly use simple representations, which are limited to learn full program semantics.

Graph Neural Networks Over the past few years, GNNs (Li et al., 2015; Hamilton et al., 2017; Kipf & Welling, 2016) have attracted increasing attention with many successful applications in computer vision (Norcliffe-Brown et al., 2018), natural language processing (Xu et al., 2018a). Because by design GNNs can model graph-structured data, recently, some works have extended the widely used Seq2Seq architectures to Graph2Seq architectures for various tasks including machine translation (Beck et al., 2018), and graph (e.g., AMR, SQL)-to-text generation (Zhu et al., 2019; Xu et al., 2018b). Some works have also attempted to encode programs with graphs for diverse tasks e.g., VARNAMING/VARMISUSE (Allamanis et al., 2017), Source Code Vulnerability Detection (Zhou et al., 2019). As compared to these works, we innovate a hybrid message passing GNN performed on both static graph and dynamic graph for better message fusion.

5 Conclusion and Future Work

In this paper, we proposed a framework for automatic code summarization. A novel retrieval-augmented mechanism is proposed for combining the benefits of both retrieval-based and generation-based approaches. Moreover, to capture global semantics among nodes, we developed a hybrid message passing GNN based on both static and dynamic graphs. The evaluation shows that our approach improves state-of-the-art techniques substantially. Future directions include exploring more effective ways to learn graph structures, combining other information, *e.g.*, API knowledge for code summarization, and graph robustness analysis for GNN.

References

Wasi Uddin Ahmad, Saikat Chakraborty, Baishakhi Ray, and Kai-Wei Chang. A transformer-based approach for source code summarization. *arXiv preprint arXiv:2005.00653*, 2020.

Miltiadis Allamanis. The adverse effects of code duplication in machine learning models of code. In *Proceedings of the 2019 ACM SIGPLAN International Symposium on New Ideas, New Paradigms, and Reflections on Programming and Software*, pp. 143–153, 2019.

Miltiadis Allamanis, Marc Brockschmidt, and Mahmoud Khademi. Learning to represent programs with graphs. *arXiv preprint arXiv:1711.00740*, 2017.

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

Uri Alon, Shaked Brody, Omer Levy, and Eran Yahav. code2seq: Generating sequences from structured representations of code. *arXiv preprint arXiv:1808.01400*, 2018.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*, 2014.

Satanjeev Banerjee and Alon Lavie. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, pp. 65–72, 2005.

Antonio Valerio Miceli Barone and Rico Sennrich. A parallel corpus of python functions and documentation strings for automated code documentation and code generation. *arXiv preprint arXiv:1707.02275*, 2017.

Daniel Beck, Gholamreza Haffari, and Trevor Cohn. Graph-to-sequence learning using gated graph neural networks. *arXiv preprint arXiv:1806.09835*, 2018.

Richard Bellman. Dynamic programming. *Science*, 153(3731):34–37, 1966.

Samy Bengio, Oriol Vinyals, Navdeep Jaitly, and Noam Shazeer. Scheduled sampling for sequence prediction with recurrent neural networks. In *Advances in Neural Information Processing Systems*, pp. 1171–1179, 2015.

Deli Chen, Yankai Lin, Wei Li, Peng Li, Jie Zhou, and Xu Sun. Measuring and relieving the over-smoothing problem for graph neural networks from the topological view. In *AAAI*, pp. 3438–3445, 2020.

Brian P Eddy, Jeffrey A Robinson, Nicholas A Kraft, and Jeffrey C Carver. Evaluating source code summarization techniques: Replication and expansion. In *2013 21st International Conference on Program Comprehension (ICPC)*, pp. 13–22. IEEE, 2013.

Patrick Fernandes, Miltiadis Allamanis, and Marc Brockschmidt. Structured neural summarization. *arXiv preprint arXiv:1811.01824*, 2018.

Luca Franceschi, Michele Donini, Paolo Frasconi, and Massimiliano Pontil. Forward and reverse gradient-based hyperparameter optimization. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pp. 1165–1173. JMLR. org, 2017.

Sonia Haiduc, Jairo Aponte, Laura Moreno, and Andrian Marcus. On the use of automated text summarization techniques for summarizing source code. In *2010 17th Working Conference on Reverse Engineering*, pp. 35–44. IEEE, 2010.

Will Hamilton, Zhitao Ying, and Jure Leskovec. Inductive representation learning on large graphs. In *Advances in neural information processing systems*, pp. 1024–1034, 2017.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8): 1735–1780, 1997.

Xing Hu, Ge Li, Xin Xia, David Lo, and Zhi Jin. Deep code comment generation. In *Proceedings of the 26th Conference on Program Comprehension*, pp. 200–210, 2018a.

Xing Hu, Ge Li, Xin Xia, David Lo, Shuai Lu, and Zhi Jin. Summarizing source code with transferred api knowledge. 2018b.

Srinivasan Iyer, Ioannis Konstas, Alvin Cheung, and Luke Zettlemoyer. Summarizing source code using a neural attention model. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 2073–2083, 2016.

Siyuan Jiang, Ameer Armaly, and Collin McMillan. Automatically generating commit messages from diffs using neural machine translation. In *2017 32nd IEEE/ACM International Conference on Automated Software Engineering (ASE)*, pp. 135–146. IEEE, 2017.

Toshihiro Kamiya, Shinji Kusumoto, and Katsuro Inoue. Ccfinder: a multilinguistic token-based code clone detection system for large scale source code. *IEEE Transactions on Software Engineering*, 28(7):654–670, 2002.

Thomas N Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907*, 2016.

Alexander LeClair, Sakib Haque, Linfgei Wu, and Collin McMillan. Improved code summarization via a graph neural network. *arXiv preprint arXiv:2004.02843*, 2020.

Guohao Li, Matthias Müller, Ali K. Thabet, and Bernard Ghanem. Deepgcns: Can gcns go as deep as cnns? In *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*, pp. 9266–9275. IEEE, 2019. doi: 10.1109/ICCV.2019.00936. URL https://doi.org/10.1109/ICCV.2019.00936.

Yujia Li, Daniel Tarlow, Marc Brockschmidt, and Richard Zemel. Gated graph sequence neural networks. *arXiv preprint arXiv:1511.05493*, 2015.

Zhenmin Li, Shan Lu, Suvda Myagmar, and Yuanyuan Zhou. Cp-miner: Finding copy-paste and related bugs in large-scale software code. *IEEE Transactions on software Engineering*, 32(3): 176–192, 2006.

Chin-Yew Lin. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pp. 74–81. Association for Computational Linguistics, July 2004.

Zhongxin Liu, Xin Xia, Ahmed E Hassan, David Lo, Zhenchang Xing, and Xinyu Wang. Neural-machine-translation-based commit message generation: how far are we? In *Proceedings of the 33rd ACM/IEEE International Conference on Automated Software Engineering*, pp. 373–384, 2018.

Minh-Thang Luong, Hieu Pham, and Christopher D Manning. Effective approaches to attention-based neural machine translation. *arXiv preprint arXiv:1508.04025*, 2015.

Will Norcliffe-Brown, Stathis Vafeias, and Sarah Parisot. Learning conditioned graph structures for interpretable visual question answering. In *Advances in Neural Information Processing Systems*, pp. 8344–8353, 2018.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting on association for computational linguistics*, pp. 311–318. Association for Computational Linguistics, 2002.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pp. 5998–6008, 2017.

Petar Velickovic, Guillem Cucurull, A. Casanova, A. Romero, P. Liò, and Yoshua Bengio. Graph attention networks. *ArXiv*, abs/1710.10903, 2018.

Yao Wan, Zhou Zhao, Min Yang, Guandong Xu, Haochao Ying, Jian Wu, and Philip S Yu. Improving automatic source code summarization via deep reinforcement learning. In *Proceedings of the 33rd ACM/IEEE International Conference on Automated Software Engineering*, pp. 397–407, 2018.

Bolin Wei, Ge Li, Xin Xia, Zhiyi Fu, and Zhi Jin. Code generation as a dual task of code summarization. In *Advances in Neural Information Processing Systems*, pp. 6563–6573, 2019.

Edmund Wong, Jinqiu Yang, and Lin Tan. Autocomment: Mining question and answer sites for automatic comment generation. In *2013 28th IEEE/ACM International Conference on Automated Software Engineering (ASE)*, pp. 562–567. IEEE, 2013.

Edmund Wong, Taiyue Liu, and Lin Tan. Clocom: Mining existing source code for automatic comment generation. In *2015 IEEE 22nd International Conference on Software Analysis, Evolution, and Reengineering (SANER)*, pp. 380–389. IEEE, 2015.

Kun Xu, Lingfei Wu, Zhiguo Wang, and Vadim Sheinin. Graph2seq: Graph to sequence learning with attention-based neural networks. *arXiv preprint arXiv:1804.00823*, 2018a.

Kun Xu, Lingfei Wu, Zhiguo Wang, Mo Yu, Liwei Chen, and Vadim Sheinin. Sql-to-text generation with graph-to-sequence model. *arXiv preprint arXiv:1809.05255*, 2018b.

Fabian Yamaguchi, Nico Golde, Daniel Arp, and Konrad Rieck. Modeling and discovering vulnerabilities with code property graphs. In *2014 IEEE Symposium on Security and Privacy*, pp. 590–604. IEEE, 2014.

Jian Zhang, Xu Wang, Hongyu Zhang, Hailong Sun, and Xudong Liu. Retrieval-based neural source code summarization. In *Proceedings of the 42nd International Conference on Software Engineering. IEEE*, 2020.

Lingxiao Zhao and Leman Akoglu. Pairnorm: Tackling oversmoothing in gnns. *arXiv preprint arXiv:1909.12223*, 2019.

Yaqin Zhou, Shangqing Liu, Jingkai Siow, Xiaoning Du, and Yang Liu. Devign: Effective vulnerability identification by learning comprehensive program semantics via graph neural networks. In *Advances in Neural Information Processing Systems*, pp. 10197–10207, 2019.

Jie Zhu, Junhui Li, Muhua Zhu, Longhua Qian, Min Zhang, and Guodong Zhou. Modeling graph structure in transformer for better amr-to-text generation. *arXiv preprint arXiv:1909.00136*, 2019.

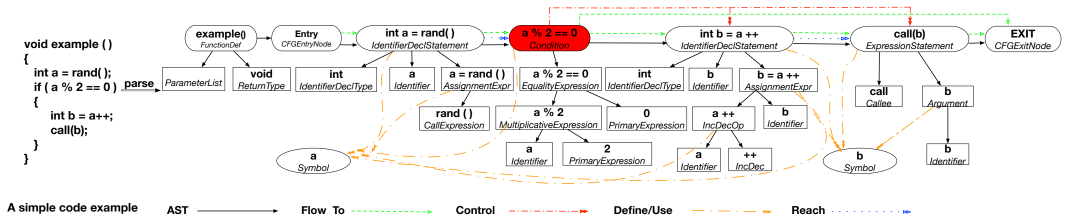

Figure 3: A example of code property graph (CPG).

Appendices

A Details on Code Property Graph

Code Property Graph (CPG) (Yamaguchi et al., 2014), which is constructed on abstract syntax tree (AST), combines different edges (i.e., "Flow to", "Control") to represent the semantics of the program. We describe each representation combining with Figure 3 as follows:

- **Abstract Syntax Tree (AST).** AST contains syntactic information for a program and omits irrelevant details that have no effect on the semantics. Figure 3 shows the completed AST nodes on the left simple program and each node has a code sequence in the first line and type attribute in the second line. The black arrows represent the child-parent relations among ASTs.
- **Control Flow Graph (CFG).** Compared with AST highlighting the syntactic structure, CFG displays statement execution order, i.e., the possible order in which statements may be executed and the conditions that must be met for this to happen. Each statement in the program is treated as an independent node as well as a designated entry and exit node. Based on the keywords *if*, *for*, *goto*, *break* and *continue*, control flow graphs can be easily built and "Flow to" with green dashed arrows in Figure 3 represents this flow order.
- **Program Dependency Graph (PDG).** PDG includes **data dependencies** and **control dependencies**: 1) data dependencies are described as the definition of a variable in a statement reaches the usage of the same variable at another statement. In Figure 3, the variable "*b*" is defined in the statement "*int b = a++*" and used in "*call (b)*". Hence, there is a "Reach" edge with blue arrows point from "*int b = a++*" to "*call (b)*". Furthermore, Define/Use edge with orange double arrows denotes the definition and usage of the variable. 2) different from CFG displaying the execution process of the complete program, control dependencies define the execution of a statement may be dependent on the value of a predicate, which more focus on the statement. For instance, the statements "*int b = a++*" and "*call(b)*" are only performed if a is even. Therefore, a red double arrow "Control" points from "*if (a % 2) == 0*" to "*int b = a++*" and "*call(b)*".

B Details on Baseline Methods

We compare our approach with existing baselines. They can be divided into three groups: Retrieval-based approaches, Sequence-based approaches and Graph-based approaches. For papers that provide the source code, we directly reproduce their methods on CCSD dataset. Otherwise, we reimplement their approaches with reference to the papers.

B.1 Retrieval-based approaches

TF-IDF (Haiduc et al., 2010) is the abbreviation of Term Frequency-Inverse Document Frequency, which is adopted in the early code summarization (Haiduc et al., 2010). It transforms programs into weight vectors by calculating term frequency and inverse document frequency. We retrieve the summary of the most similar programs by calculating the cosine similarity on the weight vectors.

NNGen (Liu et al., 2018) is a retrieved-based approach to produce commit messages for code changes. We reproduce such an algorithm on code summarization. Specifically, we retrieve the most similar top-k code snippets on a bag-of-words model and prioritizes the summary in terms of BLEU-4 scores in top-k code snippets.

B.2 Sequence-based approaches

CODE-NN (Iyer et al., 2016; Barone & Sennrich, 2017) adopts an attention-based Seq2Seq model to generate summaries on the source code.

Transformer (Ahmad et al., 2020) adopts the transformer architecture (Vaswani et al., 2017) with self-attention to capture long dependency in the code for source code summrization.

Hybrid-DRL (Wan et al., 2018) is a reinforcement learning-based approach, which incorporates AST and sequential code snippets into a deep reinforcement learning framework and employ evaluation metrics e.g., BLEU as the reward.

Dual Model (Wei et al., 2019) propose a dual training framework by training code summarization and code generation tasks simultaneously to boost each task performance.

Rencos (Zhang et al., 2020) is the retrieval-based Seq2Seq model for code summarization. it utilized a pretrained Seq2Seq model during the testing phase by computing a joint probability conditioned on both the original source code and retrieved source code for the summary generation.

B.3 Graph-based approaches

We also compared with some latest GNN-based works, employing graph neural network for source code summarization.

GCN2Seq, GAT2Seq modify Graph Convolution Network (Kipf & Welling, 2016) and Graph Attention Network (Velickovic et al., 2018) to perform convolution operation and attention operation on the code property graph for learning and followed by a LSTM to generate summaries.

SeqGNN (Fernandes et al., 2018) combines GGNNs and standard sequence encoders for summarization. They take the code and relationships between elements of the code as input. Specially, a BiLSTM is employed on the code sequence to learn representations and each source code token is modelled as a node in the graph, and employed GGNN for graph-level learning. Since our node sequences are sub-sequence of source code rather than individual token, we adjust to slice the output of BiLSTM and concatenate each token representation in node sequences as node initial representation for summarization.

C Model Settings

We embed the most frequent 40,000 words in the training set with 512-dims and set the hidden size of BiLSTM to 256 and the concatenated state size for both directions is 512. The dropout is set to 0.3 after word embedding layer and BiLSTM. We set GNN hops to 3 for the best performance. The optimizer is selected with Adam with an initial learning rate 0.001. We also use teacher forcing strategy with forcing probability equals to 0.8 and forcing decay is set to 0.99. The batch size is set to 64 and early stop for 10. The beam search width is set to 5 as usual. All experiments are conducted on the dgx server with four Nvidia Graphics Tesla V100 and each epoch takes averagely 20min. All hyperparameters are tuned with grid search (Franceschi et al., 2017) on the validation set.

D Details on Data Preparation

It is non-trivial to obtain high-quality datasets for code summarization. We noticed that despite some previous works (Barone & Sennrich, 2017; Hu et al., 2018b) released their datasets, however, they are all based on high-level programming languages i.e. Java, Python. Furthermore, they have been confirmed to have extensive duplication to make model overfit to the training data that overlapped with the test set (Fernandes et al., 2018; Allamanis, 2019). We are the first to explore summarization on *C* programming language and make our benchmark public to benefit the community research.

Table 5: More Examples of generated summaries on the CCSD test set.

Example	Example 1	Example 2
Source Code	`void hv_ringbuffer_cleanup (struct hv_ring_buffer_info *ring_info){ mutex_lock(&ring_info ->ring_buffer_mutex); vunmap(ring_info->ring_buffer); ring_info->ring_buffer=NULL; mutex_unlock(&ring_info ->ring_buffer_mutex); }`	`void BSP_LCD_DrawRect(uint16_x Xpos, uint16_t Ypos, uint16_t Width, uint16_t Height){ BSP_LCD_DrawHLine(Xpos, Ypos, Width); BSP_LCD_DrawHLine(Xpos, (Ypos+Height), Width); BSCP_LCD_DrawVLine(Xpos,Ypos,Height); BSP_LCD_DrawVLine((Xpos+Width),Ypos,Height); }`
Ground-Truth	cleanup the ring buffer	draws a rectangle
NNGen	fini ring also free the buffer for the ring	generate a 16 bit luma map from an 8 bit image
Transformer	drop a ring mapping of ring buffer	x y relative to shape origin
Rencos	release dma buffers from ring buffer	draws a range of display panel
SeqGNN	release resources related to a ring buffer	screen to draw the screen
HGNN	**clean up the ring buffer**	**draw a rectangle**
Example	Example 3	Example 4
Source Code	`static void udc_dd_free( struct lpc32xx_udc *udc, struct lpc32xx_usbd_dd_gad *dd){ dma_pool_free(udc->dd_cache, dd, dd->this_dma); }`	`static bool build_cookie(private_ike_mobike_t *this, message_t *message){ rng_t *rng; chunk_free(&this->cookie2); rng=lib->crypto->create_rng(lib->crypo, RNG_STRONG); if(!rng||rng->allocate_bytes(rng,COOKIE2_SIZE, &this->cookie2)){ DESTROY_IF(rng); return FALSE; } message->add_notify(message, FALSE, COOKIE2, this->cookie2); rng->destroy(rng); return True }`
Ground-Truth	free a dma descriptor	build a cookie and add it to the message
NNGen	allocate a dma descriptor	initialize seeds for spi generation
Transformer	free a dma descriptor	build as10x command header
Rencos	allocate a dma descriptor	initialize seeds for spi generation
SeqGNN	free the device	build a new task
HGNN	**free a dma descriptor**	**build a message cookie**

We crawled from popular *C* repositories (e.g., Linux and Redis) on GitHub, and then extracted separate function-summary pairs from these projects. Specifically, we extracted functions and associated comments marked by special characters "/**" and "*/" over the function declaration. These comments can be considered as explanations of the functions. We filtered out functions with line exceeding 1000 and any other comments inside the function, and the first sentence was selected as summary. A similar practice can be found in (Jiang et al., 2017). We totally collected **360k** raw function-summary pairs. Furthermore, functions with token size greater than 150 were removed for computational efficiency and there were **130k** functions left. Since duplication is very common in existing datasets (Fernandes et al., 2018), we performed a strict de-duplication process followed by Allamanis (2019) and removed functions with text similarity over 80% and finally kept **95k+** unique functions. We name this dataset *C* Code Summarization Dataset (CCSD). To testify model generalization ability, we randomly selected some projects as the out-of-domain test set with 2,264 examples and the remaining were randomly split into train/development/test with 82,656/4,340/4,124 examples. The open-source code analysis platform for *C* Joern (Yamaguchi et al., 2014) was applied to construct code property graphs.

E MORE EXAMPLES

We show more examples in Table 5 and find that *HGNN* can generate more high-quality summries based on our approch.