# Pyramid Attention For Source Code Summarization

**Lei Chai** and **Ming Li***
National Key Laboratory for Novel Software Technology,
Nanjing University, Nanjing 210023, China
{chail, lim}@lamda.nju.edu.cn

## Abstract

This paper presents a multi-granularity method for the task of source code summarization, which generates a concise functional description for the given code snippet. We notice that skilled programmers write and read source codes hierarchically and pay close attention to conceptual entities like statements, tokens, sub-tokens, and the mapping relations between them. The entities have specific emphasis according to their granularities, e.g., statements in coarse-granularity reveal the global logical semantics of code, and the sub-tokens in fine-granularity are more related to the textual semantics. Driven by this observation, we argue that a multi-granularity formulation incorporating these conceptual entities may benefit the code summarization task. Concretely, the source code is transformed into a pyramidal representation, and then a pyramid attention mechanism is applied for efficient feature aggregation among different hierarchies in it. We instantiate our multi-granularity method using the proposed pyramid attention and name it PA-former (Pyramid Attention transformer), which is evaluated on two source code summarization benchmarks where it surpasses the prior works and achieves new state-of-the-art results. Our code and data are available at https://github.com/leichainju/pa-former.

## 1 Introduction

Automatic source code summarization is drawing increasing attention as its promising application prospects in software development and maintenance. This task is challenging due to the complex syntax structure in programs and the arbitrariness of variable naming. Summarization methods need to exploit textual and grammatical information in programs to learn a meaningful representation for concise description generation. Following the early stage works [12, 30] that treated the code as a sequence of text and directly adopted the well-developed seq2seq [28] models for this task, recent leading methods [27, 25] further exploit the explicit structure (i.e., abstract syntax tree or pair-wise relations extracted from it) in programs to boost the model performance.

Current structure-aware code summarization methods are either in a hybrid way [27, 1, 13, 3] or in a structure-guided way [25, 35, 9, 19]. Works in hybrid fashion encode the AST-based structure using structure-aware models like GGNN [16] or TreeLSTM [29], and then combine the structural representation with the textual representation learned by sequential models for the decoding process. The structure-guided approaches use the pair-wise relations between tokens (or sub-tokens) extracted from code structure as inductive biases to guide the learning process. While these methods achieve excellent results, they only model the source code from a single-granularity perspective, which can be further improved in a multi-granularity manner.

To understand a given code snippet, a skilled programmer hierarchically breaks it down into fine-grained entities (i.e., sub-tokens) and builds up conceptual semantics gradually from the fine-grained

*Ming Li is the corresponding author.

36th Conference on Neural Information Processing Systems (NeurIPS 2022).

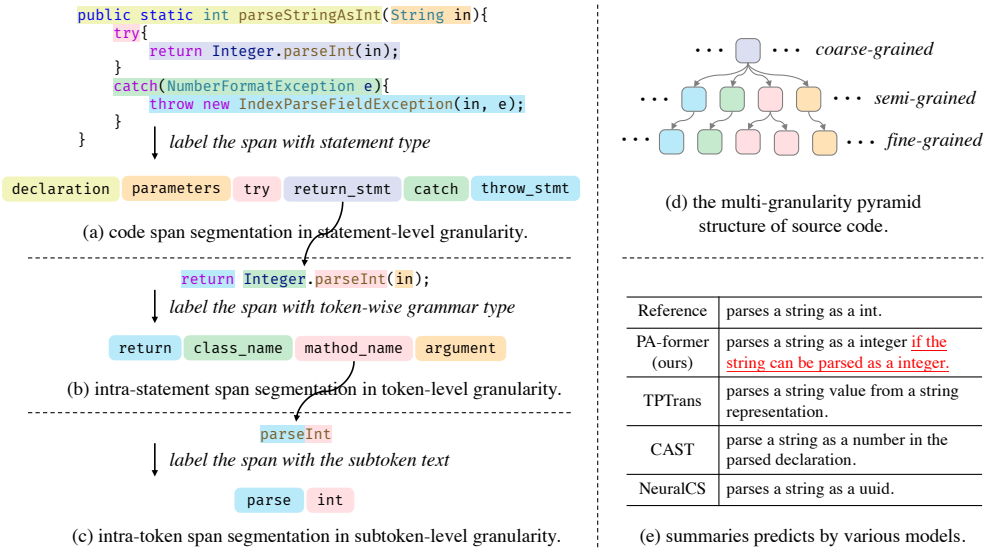


Figure 1: Example of human-like code comprehension. Skilled programmers hierarchically break the source code down to fine-grained entities (i.e., sub-tokens) and gradually build up conceptual semantics understanding for the code from fine-grained entities to coarse-grained ones.

entities to coarse-grained ones. We now use the example shown in Figure 1 to illustrate the multi-granularity process for code understanding. Concretely, we first segment the whole code into several coarse-grained spans according to punctuations and grammars, where a span corresponds to a statement conceptually (Figure 1(a)). Given the coarse-grained sequence, we can roughly summarize the code as *return something if something is done*. After that, we dive into each statement and do intra-statement segmentation for it (Figure 1(b)). Now we can detail the first *something* as *return something produced by a method invocation*. To further clarify *which method is invoked* and *what is produced*, we go deep into each token and further segment token into sub-tokens. Here, we can figure out that *a int parsing method is invoked* and *a int is produced* (Figure 1(c)). We finally roll back the details from fine-grained entities to coarse-grained ones and obtain the overall understanding of the given code.

Driven by the above analysis, we propose our multi-granularity formulation for source code summarization by mimicking human behavior. We first construct a pyramidal input, which is divided into fine-grained layer, semi-grained layer and coarse-grained layer from bottom to top (Figure 1(d)). And a novel *pyramid attention* mechanism is proposed for efficient feature aggregation among these layers and produces a pyramidal representation with rich semantics. Because both coarse-grained logical patterns and fine-grained textual features are considered simultaneously, our model can predict a more comprehensive summary. Taking the code illustrated in Figure 1(e) as an example, compared with current leading methods, our model notices the *try ... catch ...* structure in this code snippet and uses the additional *if the string can be parsed as an int.* to tell the captured exception handling pattern.

Without bells and whistles, vanilla Transformer [31] equipped with the proposed **P**yramid **A**ttention (PA-former) achieves about 5% performance improvement (evaluated by BLEU [24]) on a Java dataset. PA-former is evaluated on two source code summarization benchmarks where it surpasses the prior works and reaches new state-of-the-art results. And ablation studies are conducted to show the efficiency of the proposed method.

## 2 Related Work

**Source code representation.** Learning-based source code comprehension is drawing increasing attention as its promising application prospects in the field of software development and software maintenance. A meaningful representation can be used for various downstream tasks, such as code clone detection [34, 33], code classification [23, 37, 3], code summarization [10, 30, 3], code search

[7], etc. Early learning-based works treated source code as a sequence of text ignoring the structure features. Mainstream base models like CNN [2], Attention-based LSTM [12] and Transformer [30] were all used directly to capture the textual patterns in programs.

Many follow-ups leveraged both textual and structure information to learn a more meaningful code representation. Tree-based models like TBCNN [23], Code-GRU [17], ASTNN [37] were proposed to represent the source code by considering the tree structure. Some works [1] extended ASTs into graphs by introducing control flows and data flows, then a GGNN was applied to represent the code graphs. The set of paths [3, 4] extracted from AST was also used as a representation of the corresponding code. Some works [11, 8, 10] flatten the ASTs into a sequence, and then the off-the-shelf sequential models are applied to learn the textual and syntax features. Currently, most attention is focused on guiding the attention computation in Transformer by using the structure bias in programs [25, 9, 38].

**Source code summarization.** To generate a functional description for the given code snippet, code summarization models should capture both textual and structural information in programs. Some early works [12, 30, 32] just treated the source code as sequence or flatten the AST using a structure-based traversal (SBT) method [10] and then performed a conventional seq2seq learning. The hybrid works represented both textual and structural features separately, then a hybrid mechanism is used to combine the learned representations for summary generation. HDeepcom [11] applied the SBT method for structural representation and CAST [27] learning the structure via hierarchical splitting and reconstruction of abstract syntax trees. Many Transformer-based methods [27, 1, 13, 3] used structure information to guide the attention computation. TPTrans [25] encoded the pairwise path between code tokens and incorporated the path embeddings like relative position encoding. Unlike the above-mentioned methods which view the source code as a single-granularity data, we propose a multi-granularity source code summarization method.

**Multi-scale feature learning.** The multi-scale feature processing is widely used in the computer vision community. FPN [21] leverages the pyramid-shaped features of ConvNet and creates a feature pyramid that has rich semantics at all scales. Recently, the multi-scale feature hierarchies were also introduced into Vision Transformer (ViT) by a variety of works like MViT [6], Swin [22] and HRFormer [36], which show strong evidence that Transformer could work well with the multi-scale features.

## 3 Approach

As shown in Figure 2, given a source code, we first construct sequences with three granularities and form them into a pyramid-shaped structure based on the mapping constraints between different granularities. After that, the proposed pyramid attention mechanism (in Figure 3) is applied layer by layer in the encoding stage to produce a feature pyramid which is used by a vanilla Transformer decoder to generate the summary. The details of the proposed pyramidal input constructor and pyramid attention mechanism are introduced in this section.

### 3.1 Pyramidal Input Constructor

The pyramidal input is constructed following the human code comprehension process described in Section 1. Entities in different granularities have their own emphases, telling specific attributes of source code from different viewpoints. For instance, the coarse-grained statement sequence reveals the global logical semantics of code, the semi-grained token sequence shows the code grammar details, while the fine-grained sub-token split from each token is attached with textual information. We introduce the details of the proposed pyramid input constructor in the following.

**Fine-grained sequence.** Each leaf node in the AST corresponds to a token (like `boolean`, `value`) in the source code. Since method names and variable names in programs are usually composed by multiple natural language words with *CamelCase* or *snake_case*, we can further split these combined tokens into sub-tokens, which reveal the textual information of source code. To construct the fine-grained layer (i.e., sub-token sequence), we first obtain the leaf nodes sequence of AST by depth-first pre-order traversal and then split each token into sub-tokens. Such a sub-token technique is widely used by source code summarization methods [30, 25, 27] and proved to be effective. Formally, let

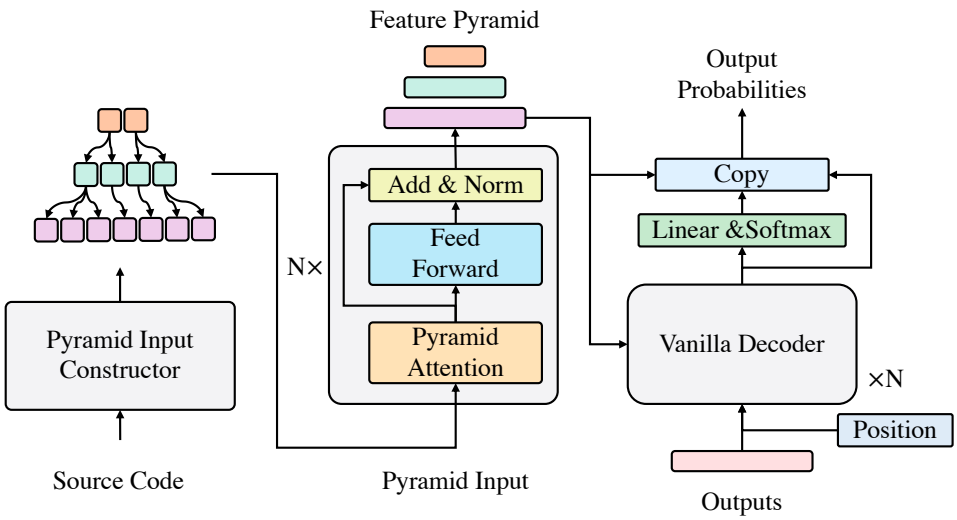


Figure 2: The proposed multi-granularity source code summarization pipeline. We first construct a pyramidal input with three granularities, and then a novel pyramid attention based encoder is applied to produce a feature pyramid which is used for decoding.

$\mathcal{T} = [t_1, t_2, \ldots, t_N]$ ($\mathcal{T}$ for text) denote the ordered sub-token sequence, that is, the fine-grained input representation situated in bottom of the pyramidal input. For the example in Figure 1, $\mathcal{T}$ is `[public, static, int, parse, String, As, Int, (, ...]`.

**Semi-grained sequence.** Attribute of the parent node of each token (i.e. leaf) in parsed AST represents its grammatical property. We construct the semi-grained sequence by replacing the element in ordered token sequence with its grammatical property directly. Formally, let $\mathcal{G} = [g_1, g_2, \ldots, g_M]$ ($\mathcal{G}$ for grammar) denote the ordered token grammatical sequence, that is, the semi-grained input representation situated in middle of the pyramidal input. For the example in Figure 1, $\mathcal{G}$ is `[modifier, modifier, type, identifier, (, ...]`. In addition, a one-to-many mapping relationship $\mathcal{M}_{\mathcal{G}\to\mathcal{T}}$ between token and sub-token is introduced. $\mathcal{M}_{\mathcal{G}\to\mathcal{T}}(g_i, t_j) = 1$ if the sub-token $t_j$ is split from $g_i$, which will be used to guide the information aggregation process from fine-grained hierarchy to semi-grained hierarchy.

**Coarse-grained sequence.** We use the nodes corresponding to statements to form our coarse-grained sequence, which reveals the logic information of source code. As the statement type is limited, we directly identify the AST node as statement-level node if the tag belongs to the pre-defined statement type set (see Appendix C for more details). Formally, let $\mathcal{L} = [l_1, l_2, \ldots, l_K]$ ($\mathcal{L}$ for logic) denote the ordered statements sequence, that is, the coarse-grained input representation situated in top of the pyramidal input. For the code in Figure 1, $\mathcal{L}$ is `[declaration, parameters, try, return_statement, ...]`. Similar to the semi-grained sequence, we also construct a one-to-many mapping function $\mathcal{M}_{\mathcal{L}\to\mathcal{G}}$ between statement and token, which bridge the coarse-grained hierarchy with the semi-grained hierarchy. $\mathcal{M}_{\mathcal{L}\to\mathcal{G}(l_i, g_j)} = 1$ if the statement $l_i$ is the most recent ancestor of token $g_j$ among $\mathcal{L}$ in AST.

In summary, we construct a sequence set $\{\mathcal{T}, \mathcal{G}, \mathcal{L}\}$ in three granularities which helps the model to capture patterns from different scales and extract mapping relationships $\{\mathcal{M}_{\mathcal{L}\to\mathcal{G}}, \mathcal{M}_{\mathcal{G}\to\mathcal{T}}\}$ between adjacent granularities which introduce inductive bias to aid the model to learn accurate semantics. Thus, the sequences and the mapping relationships all together form the pyramidal input representation of the given code.

### 3.2 Pyramid Attention

The pyramidal layers $\{\mathcal{T}, \mathcal{G}, \mathcal{L}\}$ complement each other following the mapping relationships $\{\mathcal{M}_{\mathcal{L}\to\mathcal{G}}, \mathcal{M}_{\mathcal{G}\to\mathcal{T}}\}$. We propose a novel pyramid attention mechanism for efficient feature aggrega-

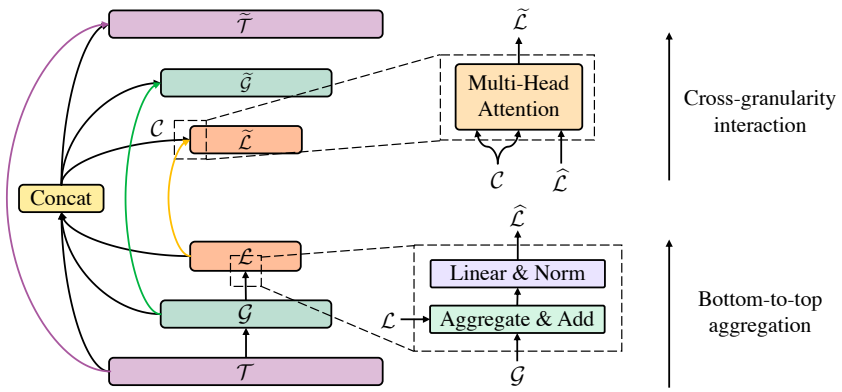


Figure 3: Pyramid attention mechanism. The pyramid attention takes a pyramid-shaped data described in Section 3.1 as input, and outputs a multi-scale representation without changing the original pyramid shape. The whole process consists of a bottom-to-top aggregation pathway and a cross-granularity information interaction process.

tion among these layers. As illustrated in Figure 3, pyramid attention module includes a bottom-to-top aggregation pathway and a cross-granularity information interaction process. We describe the details in the following.

**Bottom-to-top aggregation.** The bottom-to-top pathway does a fine-to-coarse information aggregation, which ensures to enrich the higher level representation using the lower level semantics. Given the pyramid input data: $\{\mathcal{T}, \mathcal{G}, \mathcal{L}\}$ and the mapping function $\{\mathcal{M}_{\mathcal{L}\to\mathcal{G}}, \mathcal{M}_{\mathcal{G}\to\mathcal{T}}\}$, we aggregate the representation of $\mathcal{T}$ into $\mathcal{G}$ w.r.t. $\mathcal{M}_{\mathcal{G}\to\mathcal{T}}$. The formulation of the aggregation process is given as:

$$\widehat{\boldsymbol{g}}_i = \text{LayerNorm}\left(\boldsymbol{g}_i + \boldsymbol{W}_{\text{agg}} \frac{\sum_{j=1}^{n} \mathcal{M}_{\mathcal{G}\to\mathcal{T}}(\boldsymbol{g}_i, \boldsymbol{t}_j)\boldsymbol{t}_j}{\sum_{j=1}^{n} \mathcal{M}_{\mathcal{G}\to\mathcal{T}}(\boldsymbol{g}_i, \boldsymbol{t}_j)} + \boldsymbol{b}_{\text{agg}}\right) \tag{1}$$

Where $\boldsymbol{W}_{\text{agg}} \in \mathbb{R}^{D\times D}, b_{\text{agg}} \in \mathbb{R}^D$ are learnable parameters, $\boldsymbol{g}_i, \boldsymbol{t}_i \in \mathbb{R}^D$ are the $i$-th representation vectors in $\mathcal{G}$ and $\mathcal{T}$ respectively, $D$ represents the dimension of the representation, and $\widehat{\boldsymbol{g}}_i$ is the aggregated representation of $\boldsymbol{g}_i$. The aggregation process from $\mathcal{G}$ into $\mathcal{L}$ w.r.t. $\mathcal{M}_{\mathcal{L}\to\mathcal{G}}$ is the same as above. We further conduct a aggregation directly from $\mathcal{T}$ into $\mathcal{L}$ w.r.t. $\mathcal{M}_{\mathcal{L}\to\mathcal{T}}$, which is obtained by:

$$\mathcal{M}_{\mathcal{L}\to\mathcal{T}}(\boldsymbol{l}_i, \boldsymbol{t}_j) = \vee_{k\in\{1,\ldots,m\}} \left(\mathcal{M}_{\mathcal{L}\to\mathcal{G}}(\boldsymbol{l}_i, \boldsymbol{g}_k) \wedge \mathcal{M}_{\mathcal{G}\to\mathcal{T}}(\boldsymbol{g}_k, \boldsymbol{t}_j)\right) \tag{2}$$

In summary, we update $\{\mathcal{G}, \mathcal{L}\}$ into $\{\widehat{\mathcal{G}}, \widehat{\mathcal{L}}\}$ using $\mathcal{T}$ and $\{\mathcal{T}, \mathcal{G}\}$ respectively. The mapping-based aggregation is reasonable and necessary as the semantics of the element in high levels come from the combination of its descendants' semantics. The semi-grained sequence and coarse-grained sequence are initialized as the grammatical type of the tokens and statements. Figuratively speaking, they are like torsos with empty skeletons, and we need to fill them with the textual information to make them complete. There is a ablation experiment demonstrating its effectiveness in Section 4.3.

**Cross-granularity interaction.** The aggregated semantics of the elements in semi/coarse-grained sequences tend to be local. To alleviate this dilemma, we further make the elements in a specific granularity to query all elements at three granularities, which guarantees that the model can capture both fine-grained and coarse-grained features.

Formally, we now have $\{\mathcal{T}, \widehat{\mathcal{G}}, \widehat{\mathcal{L}}\}$ after aggregation. Taking sequence $\widehat{\mathcal{G}} \in \mathbb{R}^{M\times D}$ of length $M$ in certain granularity as an example, $\mathcal{G}$ will query all of $\{\mathcal{T}, \widehat{\mathcal{G}}, \widehat{\mathcal{L}}\}$ in the cross-granularity interaction process. We first concatenate the pyramid representation as

$$\mathcal{C} = \text{Concat}([\mathcal{T}, \widehat{\mathcal{G}}, \widehat{\mathcal{L}}]) \in \mathbb{R}^{(K+M+N)\times D} \tag{3}$$

here $N, M, K$ are the sequence lengths of $\mathcal{T}, \widehat{\mathcal{G}}, \widehat{\mathcal{L}}$ respectively. A relative position encoded multi-head attention [26] is applied here and we use the relative position map $\mathcal{R}^{\widehat{\mathcal{G}}} \in \mathbb{Z}^{M\times|\mathcal{C}|}$ to make the

model granular-aware. The multi-head cross-attention (MHCA) process is calculated as:

$$\widetilde{\mathcal{G}} = \widehat{\mathcal{G}} + \Big[\text{MultiHead}(\widehat{\mathcal{G}}, \mathcal{C})\Big] \mathbf{W}_o \in \mathbb{R}^{M \times D} \tag{4}$$

$$\text{MultiHead}(\widehat{\mathcal{G}}, \mathcal{C}) = \text{Concat}([\text{Head}_1(\widehat{\mathcal{G}}, \mathcal{C}), \ldots, \text{Head}_H(\widehat{\mathcal{G}}, \mathcal{C})]) \in \mathbb{R}^{M \times D} \tag{5}$$

$$\text{Head}_h(\widehat{\mathcal{G}}, \mathcal{C}) = [\boldsymbol{z}_1, \boldsymbol{z}_2, \ldots, \boldsymbol{z}_M] \in \mathbb{R}^{M \times \frac{D}{H}} \tag{6}$$

$$\boldsymbol{z}_i = \sum_{j=1}^{|\mathcal{C}|} \frac{\exp \boldsymbol{\alpha}_{ij}}{\sum_{j'=1}^{|\mathcal{C}|} \exp \boldsymbol{\alpha}_{ij'}} \left(\mathcal{C}_j \mathbf{W}_v^h + \widehat{\mathcal{R}}_{ij} \mathbf{W}_{rel_v}^{\widehat{\mathcal{G}}}\right) \in \mathbb{R}^{\frac{D}{H}} \tag{7}$$

$$\boldsymbol{\alpha}_{ij} = \frac{\widehat{\mathcal{G}}_i \mathbf{W}_q^h}{\sqrt{D/H}} \left(\mathcal{C}_j \mathbf{W}_k^h + \widehat{\mathcal{R}}_{ij} \mathbf{W}_{rel_k}^{\mathcal{G}}\right)^\top \in \mathbb{R} \tag{8}$$

$$\widehat{\mathcal{R}} = \text{Embedding}_{rel}^{\widehat{\mathcal{G}}}(\mathcal{R}^{\widehat{\mathcal{G}}}) \in \mathbb{R}^{M \times |\mathcal{C}| \times \frac{D}{H}} \tag{9}$$

where $\mathbf{W}_q^h, \mathbf{W}_k^h, \mathbf{W}_v^h \in \mathbb{R}^{\frac{D}{H} \times \frac{D}{H}}, h \in [1, \ldots, H]$ are projection matrices of Query, Key and Value for $h$-th head. $\mathbf{W}_o \in \mathbb{R}^{D \times D}$ is projection matrix for attention output. $H$ represents the number of heads and $\widehat{\mathcal{X}}$ represents the output of multi-head self-attention process. $\mathbf{W}_{rel_k}^{\widehat{\mathcal{G}}}, \mathbf{W}_{rel_k}^{\widehat{\mathcal{G}}} \in \mathbb{R}^{\frac{D}{H} \times \frac{D}{H}}$ are projection matrices of Key and Value of relative position embedddings, the $\widehat{\mathcal{G}}$ means that thay are specific to $\widehat{\mathcal{G}}$.

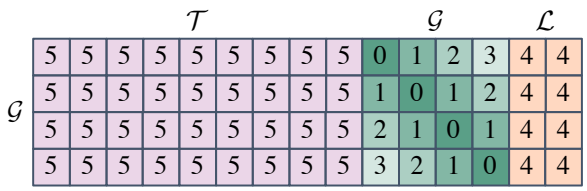


Figure 4: a toy example of granular-aware relative position map for $\mathcal{G}$ cross-granularity interaction.

For the construction of relative position map $\mathcal{R}$, we calculate the clipped relative distance with max distance as 32 for the pair within granularity and assign the relative distance as a default value (e.g. max distance + 1 and max distance + 2) for all the pair across the specific granularity. Figure 4 shows a toy example of granular-aware relative position map with max distance as 3 for $\mathcal{G}$ cross-granularity interaction. The cross-granularity interaction for $\mathcal{T}$ and $\mathcal{L}$ is similar to $\widehat{\mathcal{G}}$. We share attention parameters (i.e. $\mathbf{W}_q^h, \mathbf{W}_k^h, \mathbf{W}_v^h$ and $\mathbf{W}_o$ ) across three granularities,but assign specific relative path parameters for each granularity. In summary, in this cross-granularity interaction process, semantics of the feature pyramid are updated as:

$$\widetilde{\mathcal{L}}, \widetilde{\mathcal{G}}, \widetilde{\mathcal{T}} = \text{MHCA}(\widehat{\mathcal{L}}, \mathcal{C}), \text{MHCA}(\widehat{\mathcal{G}}, \mathcal{C}), \text{MHCA}(\mathcal{T}, \mathcal{C}) \tag{10}$$

### 3.3 Pyramid Attention Transformer

As illustrated in Figure 2, the overall model is similar to the vanilla Transformer. The self-attention module in encoder is replaced with the proposed pyramid attention module for an efficient process of the multi-scale input. Notably, we only use the fine-grained sequence in the pyramidal representation for decoding, which already contains enough semantics for summary generation and reduces the complexity of the model design. Following prior works [30], we also apply a copy mechanism in our model which reduces the risk caused by out-of-vocabulary and speeds up training convergence.

Although there is overlap between the vocabularies of different granularity, the meanings of these common words are quite different when differentiated at different granularities. For instance, `else` maybe just a natural word in the textual sequence, but a keyword for control in a grammatical sequence. Thus, we assign individual embedding layer for each granularity to avoid such confusion.

For efficient parameter usage, a parameter sharing strategy is also applied in the cross-granularity interaction process as stated in Section 3.2, We find the self-attention mechanism is strong enough to handle this multi-role mission. Thus, our model doesn't add too many parameters compared to

Transformer. What's more, such a parameter sharing strategy also brings efficient model implementation. We only need to concatenate these three sequences and the relative position maps separately, then directly adopt a relative position encoded multi-head self-attention layer to implement the cross-granularity interaction module.

# 4 Experiments

## 4.1 Experiments Setup

**Dataset and Metrics.** To demonstrate the effectiveness of the proposed method, we conduct experiments on two widely-used and well-developed java datasets: EMSE-DeepCom[2] [11] which is collected from GitHubs Java repositories and FunCom[3] [14] which has ∼2 million java method-comment pairs. We filter out the examples that cannot be parsed properly or are of too large size for both datasets. And deduplication is done on EMSE-DeepCom to avoid data leakage. Table 1 shows the statistics of the datasets.

Table 1: Dataset statistics. #train and #test mean the number of examples for training and testing, #sub-token means the max sequence length when the code is tokenized as sub-token, and the rest are similar to this.

| Name | #train | #test | #sub-token | #token | #statement | #summary |
|---|---|---|---|---|---|---|
| EMSE-Deepcom | 295,967 | 12,226 | 196 | 160 | 16 | 24 |
| FunCom | 1,017,964 | 53,936 | 256 | 196 | 32 | 32 |

We use 6 metrics to evaluate the model performance on the datasets. BLEU [24], Rouge-L [20] and Meteor [5] are widely used to evaluate the quality of text similarity in the domain of natural language generation. Another 3 metrics are used to qualify the token-wise prediction performance. We use beam search with a beam size of 3 for all models in inference phase.

**Data preprocessing.** All data preprocessing of code is based on the parsed syntax tree. We obtain the AST using the open-sourced tool Tree-sitter[4]. Please refer to 3.1 for details of the proposed pyramid data construction. We re-implement the data preprocessing pipeline for all compared baselines based on the AST parsed by Tree-sitter, please refer to Appendix C for more details.

**Baselines.** Early works viewed source code as a sequence of text and used the sequential model like LSTM or Transformer, we mainly compare with CODE-NN [12] and NeuralCodeSum [30] for this type of methods. A few works only learned the structure features for this generation task, we compare with TreeLSTM [29] here. Most works both exploited the textual and structural information for this task. One group of works learned to represent these two types of information separately, then a hybrid mechanism is used to combine them in the decoding phase. We do comparison with HDeepCom [11], ASTAttnGRU [13] and CAST [27] for the hybrid fashion. Another line of methods used structure relations as an inductive bias to guide the information interactions in the model. Here, we compare with SiT [35], GREAT [9] and TPTrans [25]. For clarity, the backbone and taxonomy of each method are listed in Table 2.

And a copy mechanism is added for all transformer-based models. We directly adopt the model implementation script if the code is publicly available, and re-implement the method according to the description in the corresponding paper if the code is not provided or not available. What's more, all baselines are integrated under a unified PyTorch-based code summarization framework developed by ourselves for fair comparisons. We provide implement details for baselines in Appendix A.3.

**Hyperparameters and training setup.** For fair comparisons, all the Transformer-based models use the default Transformer configurations with embedding dimension as 512, feedforward dimension as 2048, head number as 8, and layer number for encoder/decoder as 6 and all RNN-based models use the hidden dimension with 512. For embeddings, we use the embedding dimension with

[2] `https://github.com/xing-hu/EMSE-DeepCom`

[3] `http://leclair.tech/data/funcom/`

[4] `https://tree-sitter.github.io/tree-sitter/`

Table 2: Comparisons with other code summarization methods on a middle-size dataset and a large-size dataset.

| Methods | Backbone | Hybrid | Structure | Epochs | BLEU | Rouge-L | Meteor | Precision | Recall | F1 |
|---|---|---|---|---|---|---|---|---|---|---|
| RMSE-Deepcom (middle-size) | | | | | | | | | | |
| CODE-NN [12] | LSTM | | | 24 | 28.448 | 43.506 | 17.886 | 47.266 | 45.165 | 44.774 |
| TreeLSTM [29] | LSTM | | ✓ | 24 | 28.992 | 43.985 | 18.179 | 48.486 | 45.115 | 45.287 |
| HDeepCom [11] | GRU | ✓ | ✓ | 24 | 32.179 | 49.029 | 21.528 | 54.273 | 50.584 | 50.754 |
| ASTAttnGRU [13] | GRU | ✓ | ✓ | 24 | 33.041 | 49.761 | 22.205 | 54.539 | 51.754 | 51.467 |
| SiT [35] | Transformer | | ✓ | 30 | 35.689 | 53.750 | 24.196 | 60.784 | 54.243 | 55.717 |
| GREAT [9] | Transformer | | ✓ | 30 | 36.382 | 53.606 | 24.181 | 60.008 | 54.230 | 55.460 |
| NeuralCodeSum [30] | Transformer | | | 30 | 37.133 | 54.800 | 25.051 | 61.356 | 55.413 | 56.682 |
| CAST [27] | Transformer | ✓ | ✓ | 30 | 37.195 | 54.868 | 25.069 | 61.601 | 55.368 | 56.747 |
| TPTrans [25] | Transformer | | ✓ | 30 | 37.248 | 54.996 | 25.022 | 62.022 | 55.354 | 56.884 |
| PA-former (ours) | Transformer | | | 30 | **38.848** | **56.095** | **25.895** | **62.498** | **56.642** | **57.903** |
| Funcom (large-size) | | | | | | | | | | |
| CODE-NN [12] | LSTM | | | 24 | 31.862 | 48.896 | 19.108 | 54.018 | 49.241 | 49.924 |
| TreeLSTM [29] | LSTM | | ✓ | 24 | 31.462 | 48.293 | 18.869 | 53.237 | 48.902 | 49.332 |
| HDeepCom [11] | GRU | ✓ | ✓ | 24 | 35.063 | 53.350 | 22.645 | 59.460 | 53.693 | 54.808 |
| ASTAttnGRU [13] | GRU | ✓ | ✓ | 24 | 37.001 | 55.034 | 23.753 | 61.244 | 55.315 | 56.523 |
| SiT [35] | Transformer | | ✓ | 36 | 42.121 | 59.330 | 26.819 | 65.562 | 59.427 | 60.839 |
| GREAT [9] | Transformer | | ✓ | 36 | 43.286 | 60.364 | 27.439 | 66.186 | 60.580 | 61.831 |
| NeuralCodeSum [30] | Transformer | | | 36 | 43.355 | 60.405 | 27.540 | 66.488 | 60.459 | 61.860 |
| TPTrans [25] | Transformer | | ✓ | 36 | 43.450 | 60.566 | 27.607 | 66.765 | 60.568 | 62.030 |
| CAST [27] | Transformer | ✓ | ✓ | 36 | 43.580 | 60.524 | 27.665 | 66.458 | 60.657 | 61.976 |
| PA-former (ours) | Transformer | | | 36 | **44.649** | **61.450** | **28.274** | **67.210** | **61.593** | **62.863** |

512 for all approaches. The size of vocabularies is limited by 50000 over code sequences and 30000 over summary sequences in EMSE-DeepCom, and the corresponding configurations are 35000 and 30000 in FunCom. All model is trained using NVIDIA Tesla A100 GPUs with a batch size of 64. We train all baselines including our models using AdamW optimizer with a multi_step learning rate scheduler, and set the initial learning rate to 0.0002 and 0.003 for Transformer-based and RNN-based models, respectively. And warmup strategy is used for stable training in the Transformer-based model. The training epochs for each method are listed in Table 2.

### 4.2 Main Results

**Comparisons with baselines.** The comparisons with baselines are listed in Table 2. Without bells and whistles, we find that our proposed PA-former substantially surpasses all baselines and achieves new state-of-the-art on both middle-size and large-size datasets. Compared with NeuralCodeSum which is a relative positional encoded Transformer, our method has a 5% improvement in BLEU on the middle-size dataset. For the current leading methods CAST and TPTrans, which both take all the nodes in ASTs and the sub-token sequence as input, our model only takes a subset of AST nodes and sub-tokens surpass both of them over 1.0+ BLEU scores on two datasets. And thanks to the efficient implementation and parameter sharing strategy, our model requires far less training time to converge, while TPTrans needs a very long time to converge as the large size pair-wise paths and the sparse scatter algorithm. The leading results demonstrate that our multi-scale formulation for the task of source code summarization is reasonable. To further evaluate the effectiveness of our method, we provide a case study in Appendix D. Moreover, we provide the results on RMSE-Deepcom dataset with error bars to show that the improvement achieved by our method is statistically significant in Table 11 at Appendix B.

**Human evaluation.** As a complement, human evaluation is conducted over the transformer-based methods following the previous works [12, 27, 15]. We invite 5 volunteers with over 5 years of programming experience and ask them to score (0 - 10, 10 is the best) the generated summaries over 100 examples sampled randomly from the testing sets (20 from RMSE-Deepcom and 80 from Funcom). Here we focus on two key points: **naturalness** (grammaticality and fluency of the generated summary) and **usefulness** (what extent the generated summary is useful to understand the code). We divide the 100 examples into 5 groups (each one has 20 examples) and assign each group into 3 different volunteers, and the final score of each example is the average of its 3 "review" scores. As shown in Table 3, our PA-former outperforms others in both naturalness and usefulness.

Table 3: Human evaluation results over 100 examples sampled randomly from the testing sets (20 from RMSE-Deepcom and 80 from Funcom).

| Methods | Naturalness | Usefulness |
|---|---|---|
| NeuralCodeSum | 5.12 | 4.98 |
| TPTrans | 5.51 | 5.23 |
| CAST | 5.55 | 5.19 |
| PA-former (ours) | **6.09** | **5.86** |

### 4.3 Ablation Study

In this part, we analyze the importance of designs in our proposed PA-former using a series of ablation studies on EMSE-DeepCom dataset. Unless otherwise specified, the training setup is the same as above.

**Multi-scale *v.s.* single-scale.** To demonstrate the effectiveness of multi-granularity formulation, we degenerate the pyramid input into a single large sequence by concatenating the three granularity sequences and directly feed the same sequence into a relative positional encoded Transformer with a copy mechanism. As shown in Tabel 4, more data in the input sequence obtains performance improvement compared to NeuralCodeSum which only takes a sub-token sequence as input. However, processing the same data using our proposed multi-scale way achieves higher performance.

Table 4: Evaluation results of multi-scale *v.s.* single-scale.

| Methods | BLEU | Rouge-L | Meteor | Precision | Recall | F1 |
|---|---|---|---|---|---|---|
| single-scale | 37.329 | 55.032 | 25.114 | 61.586 | 55.750 | 56.946 |
| multi-scale | **38.848** | **56.095** | **25.895** | **62.498** | **56.642** | **57.903** |

**Effectiveness of pyramid attention.** The core components of pyramid attention are the bottom-to-top aggregation and cross-granularity interaction. We show the necessity for both of these two components here. To demonstrate the effectiveness of bottom-to-top aggregation, we evaluate a variant of PA-former whose bottom-to-top pathway is disabled. As shown in Table 5, the model performance drops by about 1.1 BLEU scores without such a component. To demonstrate the effectiveness of cross-granularity interaction, we evaluate a variant of PA-former whose cross-granularity interaction is disabled which directly degenerates into NeuralCodeSum as we only use the fine-grained encoder output for decoding. As shown in Table 5, the model performance drops by about 1.7 BLEU scores without such a component. Such results reveal the importance of cross-granularities interaction. We further disable the granular-aware relative embedding to investigate its importance, and we observe a big drop in performance shown in Table 5, which shows its effectiveness.

Table 5: Evaluation results of pyramid attention ablation.

| Methods | BLEU | Rouge-L | Meteor | Precision | Recall | F1 |
|---|---|---|---|---|---|---|
| PA-former | **38.848** | **56.095** | **25.895** | **62.498** | **56.642** | **57.903** |
| *w/o bottom-to-top* | 37.532 | 55.153 | 25.131 | 61.815 | 55.590 | 56.977 |
| *w/o cross-granularity* | 37.133 | 54.800 | 25.051 | 61.356 | 55.413 | 56.682 |
| *w/o granular-aware rel.* | 36.905 | 54.690 | 24.896 | 61.443 | 55.224 | 56.560 |

**Decoding strategy.** The output of our PA-former encoder is a pyramid representation, thus there are various ways for decoding. Our PA-former directly uses the fine-grained sequence just for simplicity. Here, we study other implementations that utilize all three granularity of representation. The most intuitive method is concating the three granularity representations, and another way is using them in a serial strategy [18]. As Table 6 shows, using the fine-grained representation is good enough in this task to benefit from the design of information interaction across granularities. The results are a bit counterintuitive, we think a reasonable explanation is that the fine-grained representation already contains enough information benefited from the proposed pyramid attention. When all three

granularities are used, there is severe information redundancy between different granularities, which complicates the task.

Table 6: Evaluation results of decoding strategy.

| Decoding strategy | BLEU | Rouge-L | Meteor | Precision | Recall | F1 |
|---|---|---|---|---|---|---|
| fine-grained | **38.848** | **56.095** | **25.895** | **62.498** | **56.642** | **57.903** |
| serial | 38.729 | 55.894 | 25.803 | 62.223 | 56.425 | 57.669 |
| concat | 38.556 | 55.819 | 25.797 | 62.174 | 56.456 | 57.638 |

**Grammar input.** When constructing the pyramid input, we intentionally enforce the bias between different granularities by assigning different types of content. For example, we bias the semi-grained sequence into grammar features by assigning token grammar types to it. To evaluate the rationality of assigning different types of content, we compare the proposed pyramid input without grammar information, which set the initial content of semi-grained and coarse-grained sequences to empty. The results in Table 7 prove the rationality of pyramid input design.

Table 7: Evaluation results of grammar input.

| pyramid input content | BLEU | Rouge-L | Meteor | Precision | Recall | F1 |
|---|---|---|---|---|---|---|
| PA-former | **38.848** | **56.095** | **25.895** | **62.498** | **56.642** | **57.903** |
| *w/o grammar* | 38.394 | 55.744 | 25.706 | 62.198 | 56.458 | 57.637 |

## 5 Conclusion

In this paper, we present a multi-scale formulation for source code summarization. Given a code snippet, we construct a pyramidal input with three granularities. And a novel pyramid attention mechanism is proposed for efficient information interaction across granularities. Without bells and whistles, the proposed method obtains 5% improvement over the strong baseline and achieves new state-of-the-art results. Additional ablation studies show the effectiveness of the proposed modules. Furthermore, our multi-scale formulation for the source code can be easily extended to other code-understanding tasks. In future work, we would try to scale our model up and pre-train it on larger source code corpora to further improve its capability and generalizability on source code comprehension tasks.

## 6 Acknowledgement

This research was supported by NSFC (62076121, 61921006). The authors would like to thank Ying Li, Hao-Yuan He, Ren-Biao Liu, Yu-Xi Sun and Zhi-Yu Shen for helpful discussions. Finally, We are grateful to all reviewers for their insightful comments.

## References


[1] M. Allamanis, M. Brockschmidt, and M. Khademi. Learning to represent programs with graphs. In *Proceedings of the 6th International Conference on Learning Representations (ICLR)*, 2018.

[2] M. Allamanis, H. Peng, and C. Sutton. A convolutional attention network for extreme summarization of source code. In *International Conference on Machine Learning (ICML)*, 2016.

[3] U. Alon, S. Brody, O. Levy, and E. Yahav. code2seq: Generating sequences from structured representations of code. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2019.

[4] U. Alon, M. Zilberstein, O. Levy, and E. Yahav. code2vec: Learning distributed representations of code. In *Proceedings of the ACM on Programming Languages (POPL)*, 2019.

[5] S. Banerjee and A. 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*, 2005.

[6] H. Fan, B. Xiong, K. Mangalam, Y. Li, Z. Yan, J. Malik, and C. Feichtenhofer. Multiscale vision transformers. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2021.

[7] Z. Feng, D. Guo, D. Tang, N. Duan, X. Feng, M. Gong, L. Shou, B. Qin, T. Liu, D. Jiang, and M. Zhou. Codebert: A pre-trained model for programming and natural languages. In *International Conference on Learning Representations (ICLR)*, 2020.

[8] D. Guo, S. Lu, N. Duan, Y. Wang, M. Zhou, and J. Yin. Unixcoder: Unified cross-modal pre-training for code representation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL)*, 2021.

[9] V. J. Hellendoorn, P. Maniatis, R. Singh, C. Sutton, and D. Bieber. Global relational models of source code. In *Proceedings of the 8th International Conference on Learning Representations (ICLR)*, 2020.

[10] X. Hu, G. Li, X. Xia, D. Lo, and Z. Jin. Deep code comment generation. In *Proceedings of IEEE/ACM International Conference on Program Comprehension (ICPC)*, 2018.

[11] X. Hu, G. Li, X. Xia, D. Lo, and Z. Jin. Deep code comment generation with hybrid lexical and syntactical information. In *Empirical Software Engineering*, 2019.

[12] S. Iyer, I. Konstas, A. Cheung, and L. Zettlemoyer. Summarizing source code using a neural attention model. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (ACL)*, 2016.

[13] A. LeClair, S. Haque, L. Wu, and C. McMillan. Improved code summarization via a graph neural network. In *Proceedings of International Conference on Program Comprehension (ICPC)*, 2020.

[14] A. LeClair and C. McMillan. Recommendations for datasets for source code summarization. In *Proceedings of NAACL-HLT*, 2019.

[15] J. Li, Y. Li, G. Li, X. Hu, X. Xia, and Z. Jin. Editsum: A retrieve-and-edit framework for source code summarization. In *Proceedings of the 36th IEEE/ACM International Conference on Automated Software Engineering (ASE)*, 2021.

[16] Y. Li, D. Tarlow, M. Brockschmidt, and R. Zemel. Gated graph sequence neural networks. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2016.

[17] Y. Liang and K. Q. Zhu. Automatic generation of text descriptive comments for code blocks. In *Proceedings of the 32nd AAAI Conference on Artificial Intelligence (AAAI)*, 2018.

[18] J. Libovický, J. Helcl, and D. Marecek. Input combination strategies for multi-source transformer decoder. In *Proceedings of the Third Conference on Machine Translation (WMT)*, 2018.

[19] C. Lin, Z. Ouyang, J. Zhuang, J. Chen, H. Li, and R. Wu. Improving code summarization with block-wise abstract syntax tree splitting. In *Proceedings of 29th IEEE/ACM International Conference on Program Comprehension (ICPC)*, 2021.

[20] C.-Y. Lin. Rouge: A package for automatic evaluation of summaries. In *Proceedings of the 42th Annual Meeting of the Association for Computational Linguistics (ACL)*, 2004.

[21] T.-Y. Lin, P. Dollt'ar1, R. Girshick, K. He, B. Hariharan, and S. Belongie. Feature pyramid networks for object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[22] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2021.

[23] L. Mou, G. Li, L. Zhang, T. Wang, and Z. Jin. Convolutional neural networks over tree structures for programming language processing. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, 2016.

[24] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL)*, 2002.

[25] H. Peng, G. Li, W. Wang, Y. Zhao, and Z. Jin. Integrating tree path in transformer for code representation. In *Advances of 35th Conference on Neural Information Processing Systems (NeurIPS)*, 2021.

[26] P. Shaw, J. Uszkoreit, and A. Vaswani. Self-attention with relative position representations. In *In Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*, 2018.

[27] E. Shi, Y. Wang, L. Du, H. Zhang, S. Han, D. Zhang, and H. Sun. Cast: Enhancing code summarization with hierarchical splitting and reconstruction of abstract syntax trees. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2021.

[28] I. Sutskever, O. Vinyals, and Q. V. Le. Sequence to sequence learning with neural networks. In *Advances of the Conference on Neural Information Processing Systems (NeurIPS)*, 2014.

[29] K. S. Tai, R. Socher, and C. D. Manning. Improved semantic representations from tree-structured long short-term memory networks. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, 2015.

[30] A. W. Uddin, C. Saikat, R. Baishakhi, and C. Kai-Wei. A transformer-based approach for source code summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL)*, 2020.

[31] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, and ukasz Kaiser. Attention is all you need. In *Advances of the 31st Conference on Neural Information Processing Systems (NeurIPS)*, 2017.

[32] B. Wei, G. Li, X. Xia, Z. Fu, , and Z. Jin. Code generation as a dual task of code summarization. In *Advances of 33th Conference on Neural Information Processing Systems (NeurIPS)*, 2019.

[33] H.-H. Wei and M. Li. Supervised deep features for software functional clone detection by exploiting lexical and syntactical information in source code. In *Proceedings of the 26th International Joint Conference on Artificial Intelligence (IJCAI)*, 2017.

[34] M. White, M. Tufano, C. Vendome, , and D. Poshyvanyk. Deep learning code fragments for code clone detection. In *Proceedings of the 31st IEEE/ACM International Conference on Automated Software Engineering (ASE)*, 2016.

[35] H. Wu, H. Zhao, and M. Zhang. Code summarization with structure-induced transformer. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP*, 2021.

[36] Y. Yuan, R. Fu, L. Huang, W. Lin, C. Zhang, X. Chen, and J. Wang. Hrformer: High-resolution transformer for dense prediction. In *Advances of 35th Conference on Neural Information Processing Systems (NeurIPS)*, 2021.

[37] J. Zhang, X. Wang, H. Zhang, H. Sun, K. Wang, and X. Liu. A novel neural source code representation based on abstract syntax tree. In *Proceedings of the IEEE/ACM International Conference on Software Engineering (ICSE)*, 2019.

[38] D. Zügner, T. Kirschstein, M. Catasta, J. Leskovec, and S. Günnemann. Language-agnostic representation learning of source code from structure and context. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2021.

## Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [No]
   (c) Did you discuss any potential negative societal impacts of your work? [No]
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [N/A]
   (b) Did you include complete proofs of all theoretical results? [N/A]
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes] See directory named `pa_former`, and we also provide training logs (`rmse_deepcom.log` and `funcom.log`) for our results in in the supplemental material.
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes]
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [Yes] In Table 11 at Appendix B
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes] See Section 4.1
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [Yes] See Section 4.1
   (b) Did you mention the license of the assets? [No]
   (c) Did you include any new assets either in the supplemental material or as a URL? [Yes]
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A]
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [No]
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]

# Appendix

We first provide the implementation details of PA-former and the compared baselines (Appendix A). Then, more supplementary experiments will be listed in Appendix B. Next, we introduce the data process for code (Appendix C). Finally, qualitative examples are provided in Appendix D.

## A Implementation Details

### A.1 PA-former

**Pyramid attention** Here we provide the pseudocode of our pyramid attention forward process in a PyTorch-like style in Algorithm 1. Compared with other methods, the algorithm implementation is straightforward and efficient. Please refer to the provided code in `pa_former` for more details.

**Algorithm 1** PyTorch-like pseudocode of pyramid attention

```
# self_attn: relative position encoded multi-head self-attention module
# ln_*: layer norm layer
# fc_*: linear layer
# fine_seq (BxLfxD): batched fine-grained sub-token sequence
# semi_seq (BxLsxD): batched semi-grained token sequence
# fine_seq (BxLcxD): batched coarse-grained statement sequence
# fine_to_semi (BxLsxLf): mappings between fine_seq and semi_seq
# fine_to_coarse (BxLcxLf): mappings between fine_seq and coarse_seq
# rel_pos_map (BxLxL): relative position map

# normalize mappings
fine_to_semi /= (fine_to_semi.sum(dim=-1, keepdim=True) + 1e-8)
fine_to_coarse /= (fine_to_coarse.sum(dim=-1, keepdim=True) + 1e-8)

# bottom-to-top aggregation process
# here, @: batch matrix multiplication
semi_seq = ln_s(fc_s(fine_to_semi @ fine_seq) + semi_agg)
coarse_seq = ln_c(fc_c(fine_to_coarse @ fine_seq) + coarse_seq)

# cross-granularity interaction
# here, cat: concatenation
x = cat((fine_seq, semi_seq, coarse_seq), dim=1) # BxLxD
x_lens = fine_seq.size(1), semi_seq.size(1), coarse_seq.size(1)
x = self_attn(x, rel_pos_map) # BxLxD
fine_seq, semi_seq, coarse_seq = split(x, x_lens, dim=1)
```

### A.2 Baselines

In this section, we provide the implementation details of baselines.

CODE-NN [12] is a basic attention-based seq2seq model which takes a token sequence as input. For the encoder, we use a 2-layer bi-directional LSTM and set the hidden size to 256. For decoder, we use a 2-layer single-directional LSTM and set hidden size to 512. Following [30], the hidden and cell vectors of the bi-directional encoder is concated separately to form 512-D hidden vectors which are used as the initial hidden states for decoder. And the copy mechanism is used over the encoder-decoder framework. This method directly uses the token sequence, which leads to a big risk of out-of-vocabulary and limits the performance of this method.

TreeLSTM [29] is a attention-based tree2seq model which takes a parsed AST as input. For the encoder, we use child-sum tree-based LSTM with the hidden size of 512 and implement it using the excellent Deep Graph Library (DGL) [5]. For the decoder, we use a 2-layer single-directional LSTM and set the hidden size to 512. The hidden and cell vectors of the root node are used as the initial hidden states for decoder. And a copy mechanism is used over the encoder-decoder framework. As

[5] `https://www.dgl.ai/`

the large size of AST and the risk of out-of-vocabulary, such a structure-aware doesn't obtain large performance improvement.

HDeepCom [11] is a hybrid method which encoder the flattened AST sequence and sub-token sequence separately, and combines these two representation when decoding using the attention mechanism. We set the embedding size and hidden size to 512 for fair comparison.

ASTAttnGRU [13] is also a hybrid method that encodes the AST-based graph and sub-token sequence separately, and combines these two representations when decoding using the attention mechanism. We use the DGL to implement the tree encoder and set the embedding size and hidden size to 512 for a fair comparison.

SiT [35] uses the structure information (abstract syntax edges, control flow edges and data dependency edges) as attention mask to guide the attention computations. Following the paper, this method only considers the punctuation-removed token sequence. Copy mechanism is applied and the model configuration is consistent with the default Transformer.

GREAT [9] introduces various edge types following [1] and models the pairwise relations using the summed edge type embeddings between two terminal nodes (i.e., token) in AST. Then these relations are used as relative distance bias in the attention computation process as Equation 4. They only consider the token sequence and embed the tokens by averaging embeddings of its sub-token(s). We also add a copy mechanism over the encoder-decoder framework for a fair comparison.

NeuralCodeSum [30] uses the sub-token sequence as input and directly adapts relative position encoded Transformer. And copy mechanism is also introduced in this method.

CAST [27] is a hybrid method that encodes the sub-token sequence using the Transformer encoder and split the AST into statement subtrees for structure representation learning. Then use tree-based RvNN to encode each subtree and then a max-pooling is applied to get the statement-level structure representations. Next, these statement-level representations are combined into a tiny tree according to the AST to capture global structure information. For the decoder, a serial attention mechanism is used to combine the learned sequential and structural features. And copy mechanism is also used in this method.

TPTrans [25] integrates the absolute paths and relative paths between sub-tokens into Transformer by representing the paths as order non-terminal node sequences and encoding them using GRUs. The learned path embeddings are used as a relative relation bias in attention computations as Equation 4. And copy mechanism is also used in this method.

### A.3 Training configuration

We show the details of training configurations in Table 8. The "rnn-24e" is used for all LSTM/GRU-based methods on both middle-size and large-size datasets. The "Transformer-30e" is used for all Transformer-based methods on middle-size dataset. The "Transformer-36e" is used for all Transformer-based methods on large-size dataset. For the items of "lr scheduler", "MS-(12, 19)" means that MultiStep learning rate scheduler which update the `current_lr` into $0.1\times$ `current_lr` at 12th and 19th epoch.

Table 8: Training configurations.

| tag | #epochs | optimizer | initial lr | lr scheduler | warmup steps | clip grad |
|---|---|---|---|---|---|---|
| rnn-24e | 24 | AdamW | 0.003 | MS-(12, 19) | – | 8.0 |
| Transformer-30e | 30 | AdamW | 0.0002 | MS-(16, 24, 28) | 1500 | 5.0 |
| Transformer-36e | 36 | AdamW | 0.0002 | MS-(16, 24, 28) | 1500 | 5.0 |

## B More Experiment Results

**Improvement for Transformer.** we show the effect of training tricks on improving model performance in Table 9. The results shows the optimizer AdamW and warmup strategy really helps. Such a improvement caused by training tricks also shows the strong ability of Transformer. And we use this training settings to all compared baselines for fair comparison.

Table 9: Evaluation results of training tricks. "−" represents the base model trained using Adam optimizer and MultiStep lr_scheduler which is used in our baseline [30].

| Tricks | BLEU | Rouge-L | Meteor | Precision | Recall | F1 |
|---|---|---|---|---|---|---|
| − | 34.159 | 52.632 | 23.191 | 60.121 | 52.954 | 54.690 |
| + *adamW* | 36.383 | 53.656 | 24.345 | 59.825 | 54.588 | 55.579 |
| + *warmup* | 37.133 | 54.800 | 25.051 | 61.356 | 55.413 | 56.682 |

**Number of encoder layer.** We further perform the ablation study by varying the number of encoder layers and list the results in Table 10. As the results show, a deeper model helps improve performance as more cross-granularity aggregations and interactions bring richer semantic information.

Table 10: Evaluation results of number of encoder layers. #layers represents the number of encoder layers.

| #layers | BLEU | Rouge-L | Meteor | Precision | Recall | F1 |
|---|---|---|---|---|---|---|
| 3 | 37.417 | 54.988 | 25.128 | 61.632 | 55.586 | 56.878 |
| 6 | 38.848 | 56.095 | 25.895 | 62.498 | 56.642 | 57.903 |
| 9 | **39.029** | **56.382** | **26.136** | **62.576** | **57.094** | **58.177** |

**Results with error bar.** We train the models 3 times with different random seeds on RMSE-Deepcom dataset and list the scores with error bars in Table 11. The results suggest that our method achieves significant improvement (e.g. 1.5 on BLEU) over the baselines with low standard deviation ($< 0.1$).

Table 11: Comparisons results with error bars.

| Methods | BLEU | Rouge-L | Meteor | Precision | Recall | F1 |
|---|---|---|---|---|---|---|
| RMSE-Deepcom (middle-size) | | | | | | |
| CODE-NN [12] | 28.414 ± 0.084 | 43.463 ± 0.210 | 17.869 ± 0.062 | 47.404 ± 0.168 | 44.922 ± 0.355 | 44.715 ± 0.207 |
| TreeLSTM [29] | 28.946 ± 0.157 | 44.002 ± 0.117 | 18.221 ± 0.086 | 48.167 ± 0.318 | 45.370 ± 0.232 | 45.281 ± 0.130 |
| HDeepCom [11] | 32.052 ± 0.180 | 48.890 ± 0.197 | 21.440 ± 0.124 | 54.236 ± 0.052 | 50.398 ± 0.262 | 50.618 ± 0.193 |
| ASTAttnGRU [13] | 32.865 ± 0.159 | 49.624 ± 0.119 | 22.095 ± 0.107 | 54.585 ± 0.232 | 51.452 ± 0.340 | 51.331 ± 0.118 |
| SiT [35] | 35.638 ± 0.048 | 53.666 ± 0.082 | 24.134 ± 0.065 | 60.566 ± 0.202 | 54.269 ± 0.073 | 55.637 ± 0.100 |
| GREAT [9] | 36.388 ± 0.086 | 53.665 ± 0.062 | 24.193 ± 0.060 | 60.176 ± 0.193 | 54.266 ± 0.092 | 55.545 ± 0.115 |
| NeuralCodeSum [30] | 37.044 ± 0.098 | 54.653 ± 0.179 | 24.918 ± 0.126 | 61.193 ± 0.140 | 55.301 ± 0.181 | 56.541 ± 0.154 |
| CAST [27] | 37.089 ± 0.106 | 54.794 ± 0.075 | 24.990 ± 0.079 | 61.452 ± 0.151 | 55.346 ± 0.023 | 56.668 ± 0.079 |
| TPTrans [25] | 37.225 ± 0.023 | 54.954 ± 0.043 | 24.993 ± 0.029 | 61.739 ± 0.283 | 55.433 ± 0.079 | 56.828 ± 0.056 |
| PA-former (ours) | **38.777** ± 0.093 | **56.071** ± 0.069 | **25.901** ± 0.016 | **62.489** ± 0.168 | **56.650** ± 0.031 | **57.908** ± 0.077 |

# C Data preprocessing

For the summaries of code which are natural language sentences, we preprocess them following the natural language process (NLP) pipeline. In this section, we mainly focus on the code preprocessing pipelines which are based on the parsed AST. Given a code snippet, we use Tree-sitter to get the parsed AST, an example is illustrated in Figure 5.

## C.1 Pyramid input

We show the details of our proposed pyramid input construction with the example in Figure 5.

The fine-grained sequence $\mathcal{T}$ is easy to handle. We first extract the ordered terminal nodes from AST, we get `[public, int, parseStringAsInt, (, String, s, ), ...]` for the given example. And then we split each element in the token sequence into sub-tokens and we get $\mathcal{T}$ = `[public, int, parse, String, As, Int, (, String, ...]`. Here, we ignore the grammatical information and just treat elements of $\mathcal{T}$ as natural words, so we name it as $\mathcal{T}$ext sequence.

We use the semi-grained sequence $\mathcal{G}$ to represent the $\mathcal{G}$rammatical details of the code, which is a complement to $\mathcal{T}$. Each element in $\mathcal{G}$ corresponds to an element in the token sequence. For

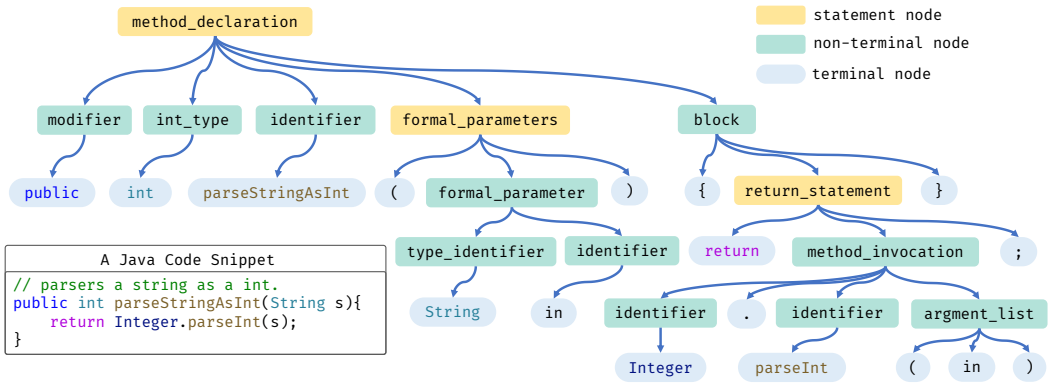


Figure 5: Parsed abstract syntax tree (AST) example using Tree-sitter.

each token, its parent node in AST represents its grammatical property, e.g., `parseStringAsInt` is an `identifier` and `public` is a `modifier`. However, we notice that all variable names and method names are marked as `identifier`. To further differentiate them, we instead use their grandparents' node in ASTs to mark the grammatical types, e.g. `parseStringAsInt` is marked as `method_declaration` and `parseInt` is marked as `method_invocation`. Finally, we get $\mathcal{G}$ = [`modifiers, int_type, method_declaration, (, type_identifier, ...`]. For the mapping relations $\mathcal{M}_{\mathcal{G}\to\mathcal{T}} \in \mathbb{R}^{|\mathcal{G}|\times|\mathcal{T}|}$, we set $\mathcal{M}_{\mathcal{G}\to\mathcal{T}}(i,j)$ to 1 if $\mathcal{T}[j]$ is split from the token corresponding to $\mathcal{G}[i]$. In this example, we set $\mathcal{M}_{\mathcal{G}\to\mathcal{T}}(3,3), \mathcal{M}_{\mathcal{G}\to\mathcal{T}}(3,4), \mathcal{M}_{\mathcal{G}\to\mathcal{T}}(3,5)$ and $\mathcal{M}_{\mathcal{G}\to\mathcal{T}}(3,6)$ to 1, as [`parse, String, As, Int`] is split from `parseStringAsInt`.

The $\mathcal{L}$ogical nature of the code is mostly in statement-level coarse-grained sequence $\mathcal{L}$. As the type of statement is limited, we first set the pre-defined statement set $\mathcal{S}$ shown in Table 12 for java dataset. In addition to the normal java statements, we also treat `method_declaration` and `formal_parameters` as statements. We treat the node in AST as statement node, if its tag is in $\mathcal{S}$. Thus we get $\mathcal{L}$ = [`method_declaration, formal_parameters, return_statement`] (yellow nodes in Figure 5). For the mapping relations $\mathcal{M}_{\mathcal{L}\to\mathcal{G}} \in \mathbb{R}^{|\mathcal{L}|\times|\mathcal{G}|}$, we set $\mathcal{M}_{\mathcal{L}\to\mathcal{G}}(i,j)$ to 1 if $\mathcal{L}[i]$ is the nearest ancester of $\mathcal{G}[j]$ in the scope of $\mathcal{L}$. In this example, we set $\mathcal{M}_{\mathcal{L}\to\mathcal{G}}(2,4), \mathcal{M}_{\mathcal{L}\to\mathcal{G}}(2,5), \mathcal{M}_{\mathcal{L}\to\mathcal{G}}(2,6)$ and $\mathcal{M}_{\mathcal{L}\to\mathcal{G}}(2,7)$ to 1, as `formal_parameters` is the nearest ancester of `(, String, in` and `)`.

Table 12: The pre-defined statement set.

| statement type | tag of statement node |
|---|---|
| expression statement | `expression_statement, explicit_constructor_invocation, local_variable_declaration` |
| exit statement | `return_statement, yield_statement, throw_statement` |
| conditional statement | `if_statement, switch_expression` |
| loop statement | `for_statement, enhanced_for_statement, while_statement do_statement, continue_statement, break_statement` |
| exception statement | `try_statement, catch_clause, finally_clause` |
| others | `method_declaration, formal_parameters` |

### C.2 Data preprocessing for baselines

Since some compared baselines don't provide available preprocessing code, we have to implement the preprocessing pipelines according to their papers.

For GREAT [9], we add the additional edges to AST according to construct the code graph representation according to [1]. The edges include `NextToken, LastRead, LastWrite, ComputeFrom`

`LexicalUse` and `LeafCFG`. The edge `LeafCFG` represents the edge of control flow graph (CFG) (see details in [9]) and please refer to [1] for details about other edges.

For SiT [35], we add `Flow` edge between each in-statement token pair, add adopt `LexicalUse`, `LastRead` and `LastWrite` in GREAT into the `Data dependency` edge.

## D Qualitative Examples

In this section, we provide several samples of code summarization with different models. We can see that PA-former generates most precise comment in most cases.

```
public boolean isOnPieChart(Point screenPoint){
    double sqValue = (MathUtil.pow(mCenterX - screenPoint.getX(), 2)\
        + MathUtil.pow(mCenterY - screenPoint.getY(), 2));
    double radiusSquared = mPieChartRadius * mPieChartRadius;
    boolean isOnPieChart = sqValue <= radiusSquared;
    return isOnPieChart;
}
```

Reference: checks if point falls within piechart.
PA-former (ours): checks if a point falls within a pie chart.
TPTrans: determines if a point is on a pie chart.
CAST: determines if a point is on the pie chart.
NeuralCodeSum: returns true if the specified point is on the pie chart.
GREAT: checks if point is within pie pie pie pie pie pie pie area.
SiT: checks if a point is on the chart.
ASTAttnGRU: checks if at the current position.

```
public XmlReporter(OutputStream out, String title){
    this.out = out;
    this.title = title;
}
```

Reference: create xmlreporter which writes to the given stream.
PA-former (ours): create an xml reporter which writes to the given stream.
TPTrans: creates a new xml reporter.
CAST: create an xml reporter to write to the given stream.
NeuralCodeSum: create an instance of this class.
GREAT: create an xml stream.
SiT: create a new reporter.
ASTAttnGRU: create an object decoding the values from the passed stream.

```
public final boolean isModified(boolean isEnable) {
    if(isStopped()) {
        return flase;
    }
    DependencyContainer dependencies = _dependencies;
    if (dependencies == null) {
        return false;
    }
    if (isEnable) {
        boolean isModified = dependencies.isModified();
        return isModified;
    } else {
        boolean isModified = isModified(getParent());
        return isModified;
    }
}
```

Reference: returns true if any of the classes have been modified.
PA-former (ours): returns true if any of the dependencies have been modified.
TPTrans: returns true if the underlying dependencies have changed.
CAST: returns true if the configuration of this object has changed.
NeuralCodeSum: returns true if the user has changed the dependency container.
GREAT: returns true if the entry has been modified.
SiT: returns true if the underlying dependencies have changed.
ASTAttnGRU: returns whether or not this property has been modified or not.

```java
public int consumeInteger()
    throws ParsingException , IllegalStateException {
    if (completed) {
        throwNoMoreContent();
    }
    String value = currentToken().value();
    try {
        int result = Integer.parseInt(value);
        moveToNextToken();
        return result;
    } catch ( NumberFormatException e ) {
        Position position = currentToken().position();
        throw new ParsingException(position, value);
    }
}
```

Reference: convert the value of this token to an integer return it and move to the next token.
PA-former (ours): convert the value of this token to an integer and move it to the next token.
TPTrans: consume an integer value from the input stream.
CAST: consume an integer value off the current token throwing an exception if it is not an integer.
NeuralCodeSum: gets the next token and remove it from the list.
GREAT: convert this token to an integer and move to an integer.
SiT: consume an integer value from the request.
ASTAttnGRU: gets the value of the int property.

```java
public void delete(String path, ParameterMap params,
    AsyncCallback callback) {
        HttpDelete req = new HttpDelete(path, params);
        executeAsync(req, callback);
}
```

Reference: execute a delete request and invoke the callback on completion.
PA-former (ours): execute a delete request and invoke the callback on completion.
TPTrans: execute a delete request.
CAST: execute a delete request.
NeuralCodeSum: execute a delete delete request.
GREAT: perform a delete on the specified path.
SiT: execute a delete request.
ASTAttnGRU: delete a parameter from a map.

```java
public FXDialog(final Stage parent , final Parent content, final
    Modality modality, final double width , final double height) {
    init(parent, modality, content);
    dialog.setWidth(width);
    dialog.setHeight(height);
    final Scene scene = new Scene(content, width, height);
    setScene (scene);
}
```

Reference: create a new dialog with a specified width and height.
PA-former (ours): create new dialog with the specified width and height.
TPTrans: creates a new dialog.
CAST: create dialog with parent content.
NeuralCodeSum: initializes the dialog with the given parent modality and content.
GREAT: creates new form <unk>.
SiT: creates a new dialog box.
ASTAttnGRU: create a new dialog box.