Retrieval Augmented Code Generation and Summarization

Md Rizwan Parvez[§], **Wasi Uddin Ahmad**[§], **Saikat Chakraborty**[†]
Baishakhi Ray[†], **Kai-Wei Chang**[§]
[§]University of California, Los Angeles, [†]Columbia University

EMNLP-Findings 2021

Motivation

Find the median of an array

```
def median(L):
    n = len(L)
    l = top_k(L, n/2 + 1)
    return max(l)
```

Motivation

- ✓ Diversity in identifiers

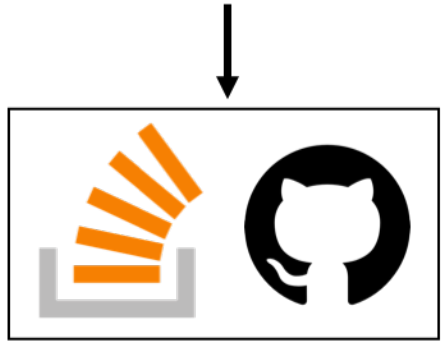
```
void function (Element arg0,
    Formula arg1) {

    arg0.addElement(
      "string").setText(
          arg1.getText());

}
```

Motivation

NL: Return the median of an unsorted list



RC: Retrieved Code

```
def median(L):
  n = len(L)
  l = top_k(L, n/2 + 1)
  return max(l)
```

```
def median(L):
  L = sorted(L)
  n = len(L)
  l = top_k(L, n/2 + 1)
  return max(l)
```

What Help Do Developers Seek, When and How?

How Developers Search for Code: A Case Study

Caitlin Sadowski
Google
Mountain View, CA, USA
supertri@google.com

Kathryn T. Stolee
Iowa State University
Ames, IA, USA
kstolee@iastate.edu

Sebastian Elbaum
University of Nebraska
Lincoln, NE, USA
elbaum@cse.unl.edu

Our approach (REDCODER)

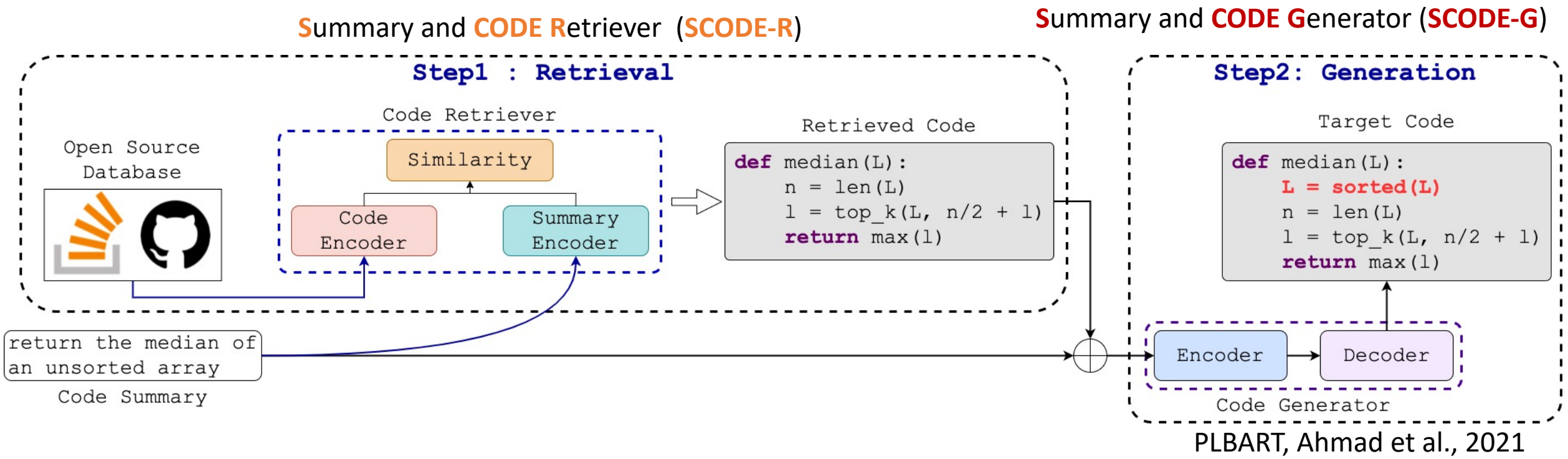
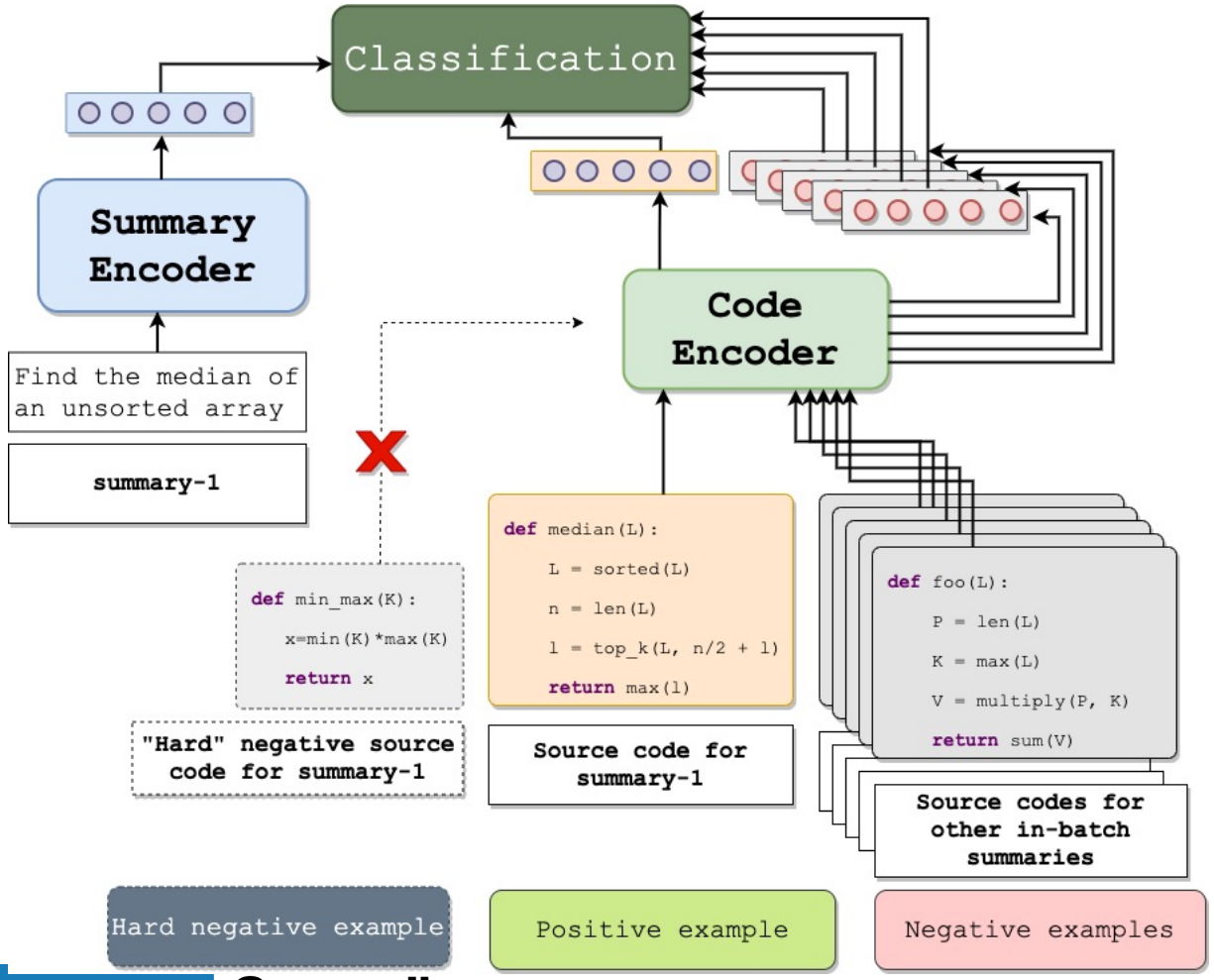


Fig: **R**etrieval augment**ED COD**e g**E**neration and summa**R**ization framework (**REDCODER**)

Sparse Vs Dense SCODE-R?

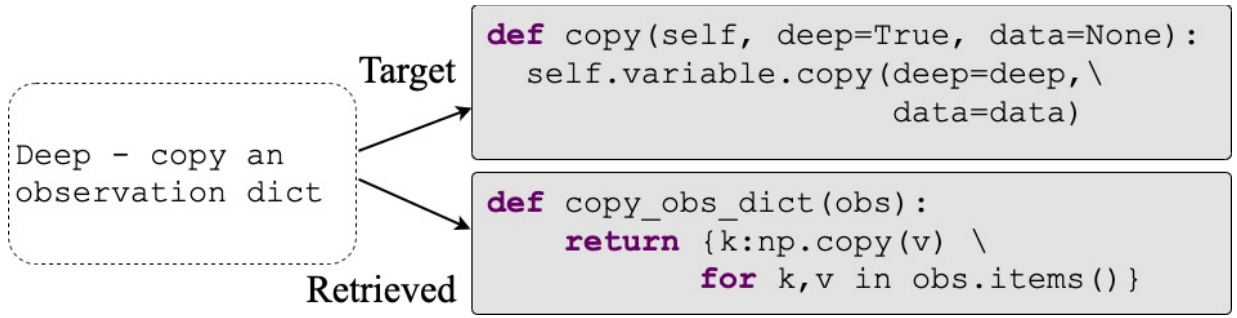
- ✓ Must be fast
- ✓ Need Understanding both syntactically and semantically

SCODE-R Training



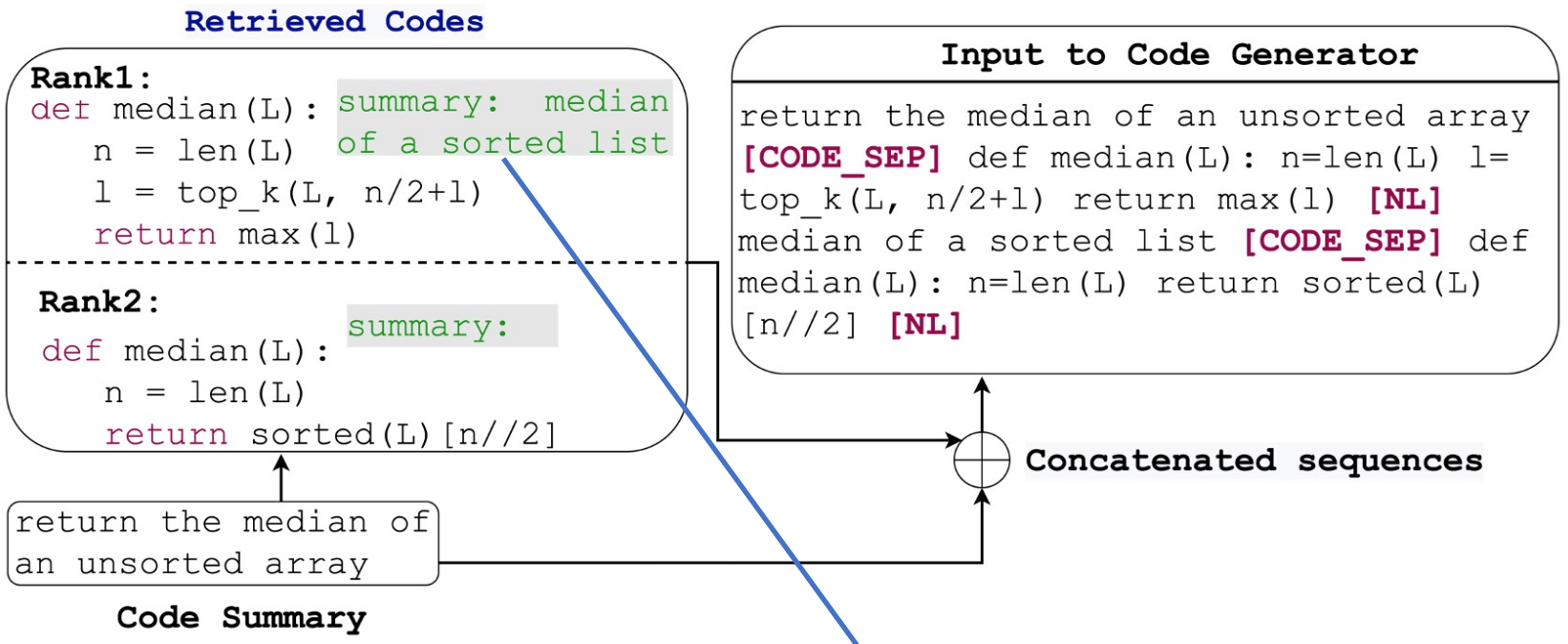
SCODE-R Training

SCODE-R is based on DPR (Karpukhin et al., 2020)



Example: A relevant yet not same retrieved code

SCODE-G

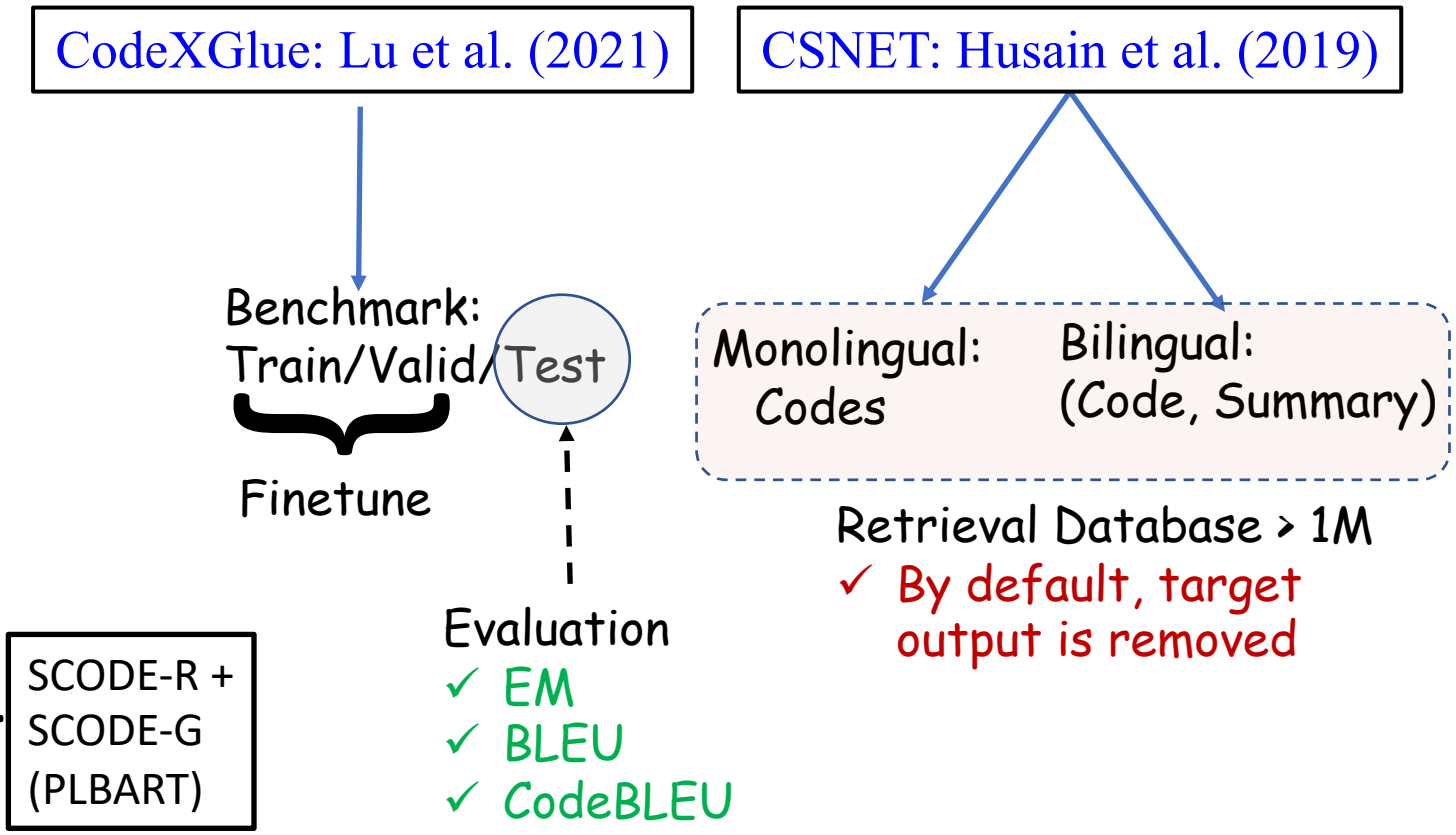

Retrieved Codes
Rank1:
def median(L):
n = len(L)
l = top_k(L, n/2+1)
return max(l)
summary: median of a sorted list
Rank2:
def median(L):
n = len(L)
return sorted(L)[n//2]
summary:
return the median of an unsorted array
Code Summary
Input to Code Generator
return the median of an unsorted array [CODE_SEP] def median(L): n=len(L) l= top_k(L, n/2+1) return max(l) [NL] median of a sorted list [CODE_SEP] def median(L): n=len(L) return sorted(L) [n//2] [NL]
Concatenated sequences
Example SCODE-G input (REDCODER-EXT)
Never used
SCODE-G input (REDCODER)
Leveraged when available
SCODE-G input (REDCODER-EXT)

Evaluation settings

Method	
Type	Name
Retrieval Based	BM25
	SCODE-R
Generative	CodeBERT
	GraphCodeBERT
	CodeGPT-adapted
	PLBART
Retrieval Augmented Generative	BM25 + PLBART
	REDCODER
	REDCODER-EXT

SCODE-R + SCODE-G (PLBART)

Baselines



Evaluation

Method		Java			Python		
Type	Name	EM	BLEU	CodeBLEU	EM	BLEU	CodeBLEU
Retrieval Based	BM25	0.00	4.90	16.00	0.00	6.63	13.49
	SCODE-R	0.00	25.34	26.68	0.00	22.75	23.92
Generative	CodeBERT	0.00	8.38	14.52	0.00	4.06	10.42
	GraphCodeBERT	0.00	7.86	14.53	0.00	3.97	10.55
	CodeGPT-adapted	0.00	7.10	14.90	0.01	3.11	11.31
	PLBART	0.00	10.10	14.96	0.00	4.89	12.01
Retrieval Augmented Generative	BM25 + PLBART	0.10	11.37	15.52	0.03	6.99	13.89
	REDCODER	8.95	26.92	31.15	8.88	22.74	28.93
	REDCODER-EXT	**10.21**	**28.98**	**33.18**	**9.61**	**24.43**	**30.21**

Table: Code gen. performances

Qualitative examples

Reference (Gold Output)

Input summary: Get the MuxerStream at the given position.

```
1 public MuxerStream getMuxer (int streamIndex)
2   throws java.lang.InterruptedException, java.io.IOException {
3   long cPtr = VideoJNI.MuxerStream_getMuxer(swigCPtr, this, streamIndex);
4   return (cPtr == 0) ? null : new MuxerStream (cPtr, false);
5 }
```

PLBART Prediction **[BLEU: 0.1439]**

```
1 public MuxerStream getMuxerStream (int position) {
2   if (muxerStream == null) {
3     muxerStream = new MuxerStream (this, position)
4   }
5   return muxerStream;
6 }
```

PLBART fails to predict the diverse identifiers (in red color) whereas REDCODER succeeds

Redcoder-ext Prediction **BLEU: 80.6**

```
1 public MuxerStream getMuxer (int streamIndex)
2   throws java.lang.InterruptedException, java.io.IOException {
3   long cPtr = VideoJNI.MuxerStream_getMuxer(swigCPtr, this, streamIndex);
4   return (cPtr == 0) ? null : new MuxerStream (cPtr, false);
5 }
```

Thank You!

Questions?

https://github.com/rizwan09/REDCODER