Comparing Natural Language Embeddings for Libc Functions as Rich Labels

Bachelor defense

Ruben Triwari

Ludwig Maximilian University Munich

19, February 2025

Outline

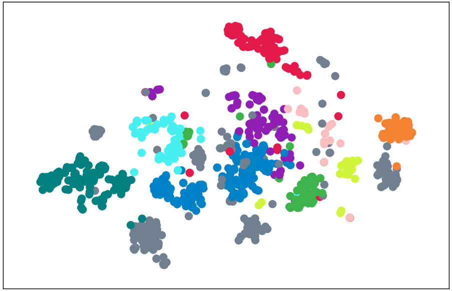
Motivation & Research Objective

Methodology

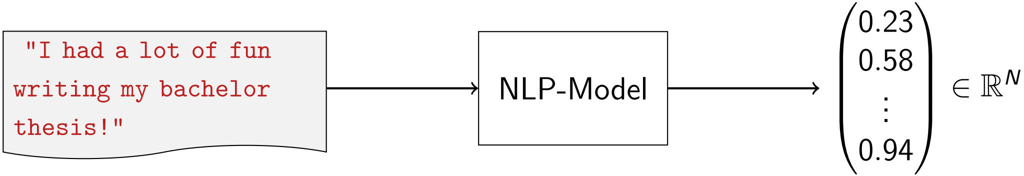
Results

Limitations

Conclusion & Future Work



Motivation



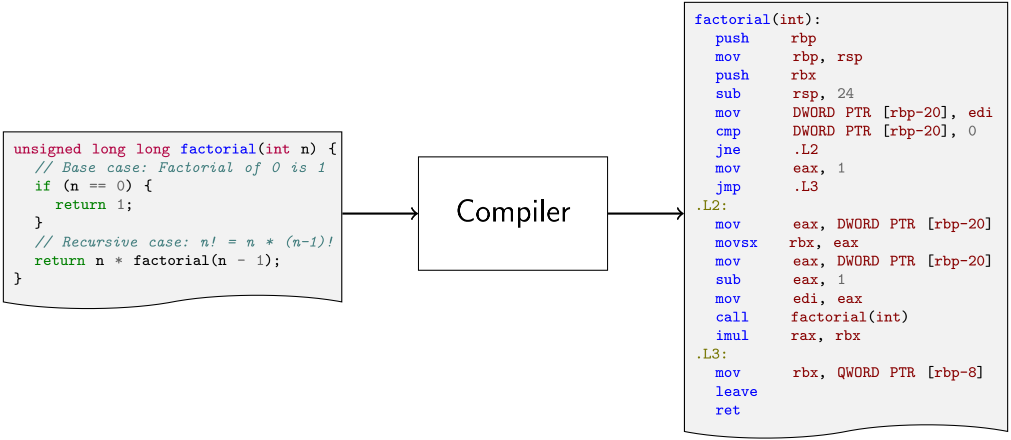
"I had a lot of fun writing my bachelor thesis!"

NLP-Model

$$\begin{pmatrix} 0.23 \\ 0.58 \\ \vdots \\ 0.94 \end{pmatrix} \in \mathbb{R}^N$$

- ⇝ Encoding natural language was a huge factor in recent nlp advancements
- ⇝ Information described as a vector can be used in many downstream task
- ⇝ That motivates encoding binary code and describing them as a vector
- ⇝ That motivates using NLP tools to encode binary code

Motivation



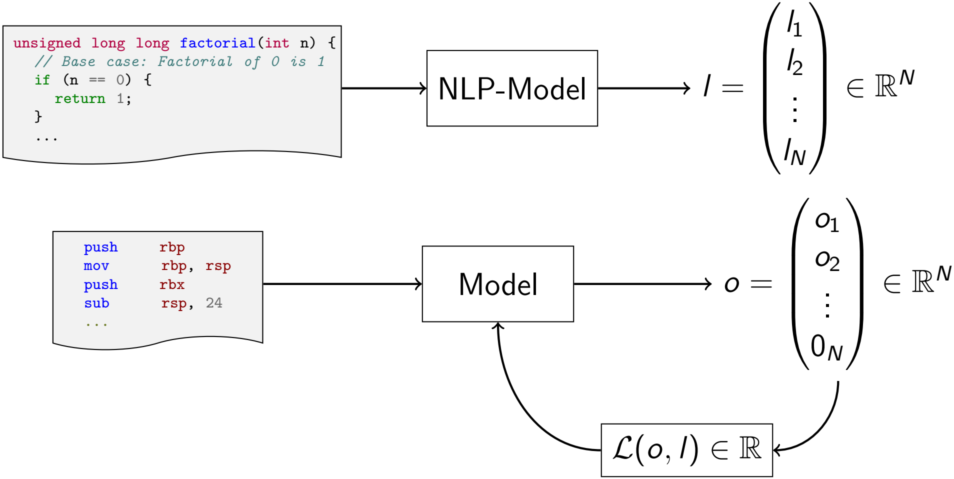
```
unsigned long long factorial(int n) {
  // Base case: Factorial of 0 is 1
  if (n == 0) {
    return 1;
  }
  // Recursive case: n! = n * (n-1)!
  return n * factorial(n - 1);
}
```

Compiler

```
factorial(int):
   push     rbp
   mov      rbp, rsp
   push     rbx
   sub      rsp, 24
   mov      DWORD PTR [rbp-20], edi
   cmp      DWORD PTR [rbp-20], 0
   jne      .L2
   mov      eax, 1
   jmp      .L3
.L2:
   mov      eax, DWORD PTR [rbp-20]
   movsx    rbx, eax
   mov      eax, DWORD PTR [rbp-20]
   sub      eax, 1
   mov      edi, eax
   call     factorial(int)
   imul     rax, rbx
.L3:
   mov      rbx, QWORD PTR [rbp-8]
   leave
   ret
```

⇝ Compiler removes important inforamtion in natural language

Motivation



```
unsigned long long factorial(int n) {
  // Base case: Factorial of 0 is 1
  if (n == 0) {
    return 1;
  }
  ...
```

NLP-Model

$$l = \begin{pmatrix} l_1 \\ l_2 \\ \vdots \\ l_N \end{pmatrix} \in \mathbb{R}^N$$

```
    push      rbp
    mov       rbp, rsp
    push      rbx
    sub       rsp, 24
    ...
```

Model

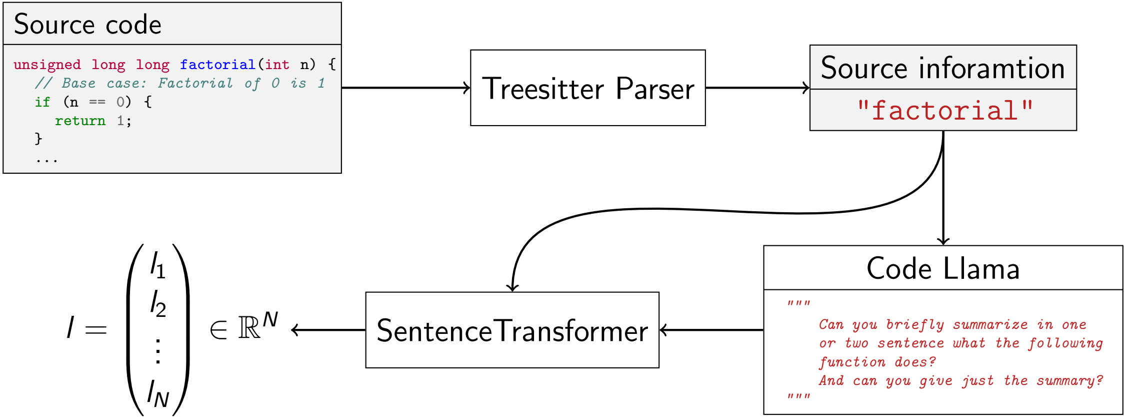
$$o = \begin{pmatrix} o_1 \\ o_2 \\ \vdots \\ 0_N \end{pmatrix} \in \mathbb{R}^N$$

$$\mathcal{L}(o, l) \in \mathbb{R}$$

Research Objectives

- Compare diffrent approaches generating an Embedding with NLP tools
  1. Embed function names with SentenceTransformer
  2. Embed function comments with SentenceTransformer
  3. Embed Code-Llama code summaries with SentenceTransformer
- Compare NLP approach to the existig Code2Vec Model
- Propose a new way comparing embedding spaces

Architecture



Source code

```
unsigned long long factorial(int n) {
  // Base case: Factorial of 0 is 1
  if (n == 0) {
    return 1;
  }
  ...
```

Treesitter Parser

Source inforamtion

"factorial"

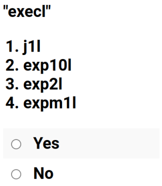
Code Llama

```
"""
    Can you briefly summarize in one
    or two sentence what the following
    function does?
    And can you give just the summary?
"""
```

SentenceTransformer

$$l = \begin{pmatrix} l_1 \\ l_2 \\ \vdots \\ l_N \end{pmatrix} \in \mathbb{R}^N$$

Expert Survey



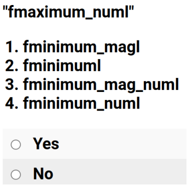
"execl"

1. j1l
2. exp10l
3. exp2l
4. expm1l

○ Yes

○ No

Figure: Positve exmaple



"fmaximum_numl"

1. fminimum_magl
2. fminimuml
3. fminimum_mag_numl
4. fminimum_numl

○ Yes

○ No

Figure: Negative exmaple

Ergebnisse der Expertenbefragung				
Strategie	Code-Llama-Erklärungen	Funktionsnamen	Funktionskommentare	*Code2Vec*
Score	0.596	0.532	0.433	0.321

Embeddings space comparison

Embeddings space comparison

Evaluation with T-SNE

HALLO

HALLO

Discussion