

Codex и GraphCodeBert

Кондратьев Захар

Введение

Как работать с кодом?

Можем просто взять модели из NLP и применить их к коду, как к тексту.

Подобный подход в статье Codex, то есть оригинальная модель GPT не меняется.

Можем подавать модели на вход информацию о коде.
Про это статья GraphCodeBert.

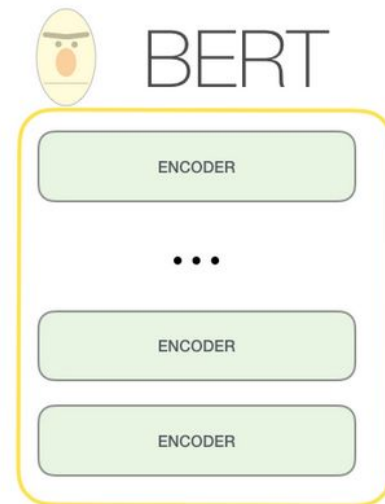
BERT и GPT

Выход:

последовательность vs токен

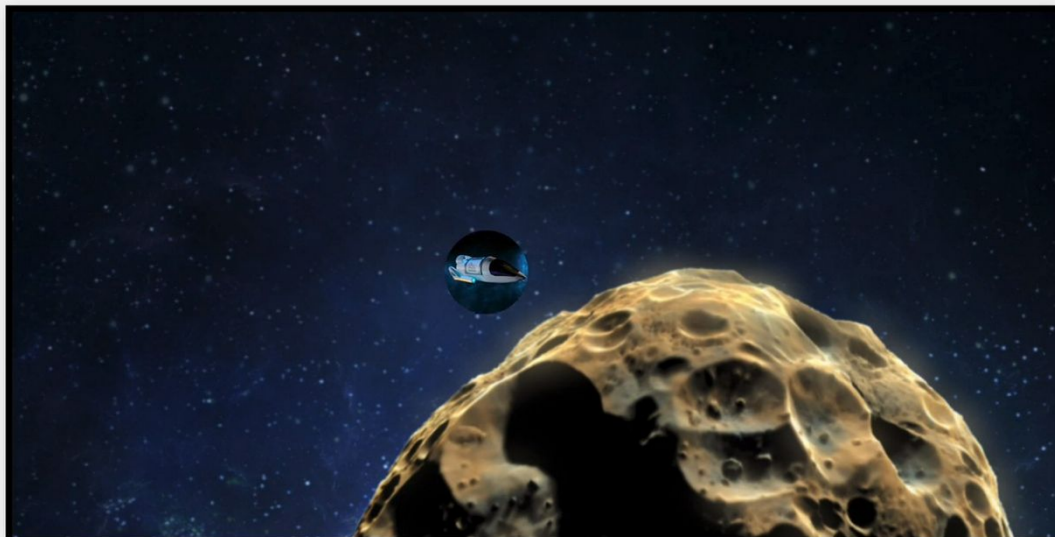
Задачи для обучения.

Решаемые задачи.



Codex

Что модель может?



Make it be the size of the rocketship times 0.75.

```
document.body.removeChild(text);
}, 250);
};
```

```
/* Now add an image of an
asteroid:
https://d.newsweek.com/en/full/1721338/asteroid.jpg?w=1600&h=1600&q=88&f=9d82d35c9de96a82b3fc9f7705eb325b */
var asteroid =
document.createElement('img');
asteroid.src =
'https://d.newsweek.com/en/full/1721338/asteroid.jpg?w=1600&h=1600&q=88&f=9d82d35c9de96a82b3fc9f7705eb325b';
document.body.appendChild(asteroid);
```


```
/* Make it be the size of the
rocketship times 0.75. */
asteroid.style.width =
rocketship.offsetWidth * 0.75
+ 'px';
asteroid.style.
```



Что модель может?

```
"""Now plot the results.
Label both axes (y axis is
max temperature), rotate
the x ticks, and add a
title."""
import matplotlib.pyplot
as plt

plt.plot(df['date'],
df['value'])
plt.xlabel('Date')
plt.ylabel('Max
Temperature (F)')
plt.xticks(rotation=90)
plt.title('Max Temperature
in San Francisco')
plt.show()
```



```
"""Jane has 9
balloons. 6 are
green and the rest
are blue. How many
balloons are
blue?"""
jane_balloons = 9
green_balloons = 6
blue_balloons =
jane_balloons -
green_balloons
print(blue_balloons)
```

Данные

Train

54 млн репозиторийев, 159 GB кода с GitHub

Test

Собрали HumanEval - 164 задачи по программированию

Данные: function signature, docstring, body, unit tests

Для проверки тестов есть песочница.

Метрика Pass@K

Генерируем k ответов модели для задачи.

Если хотя бы один код проходит все тесты, то задача решена.

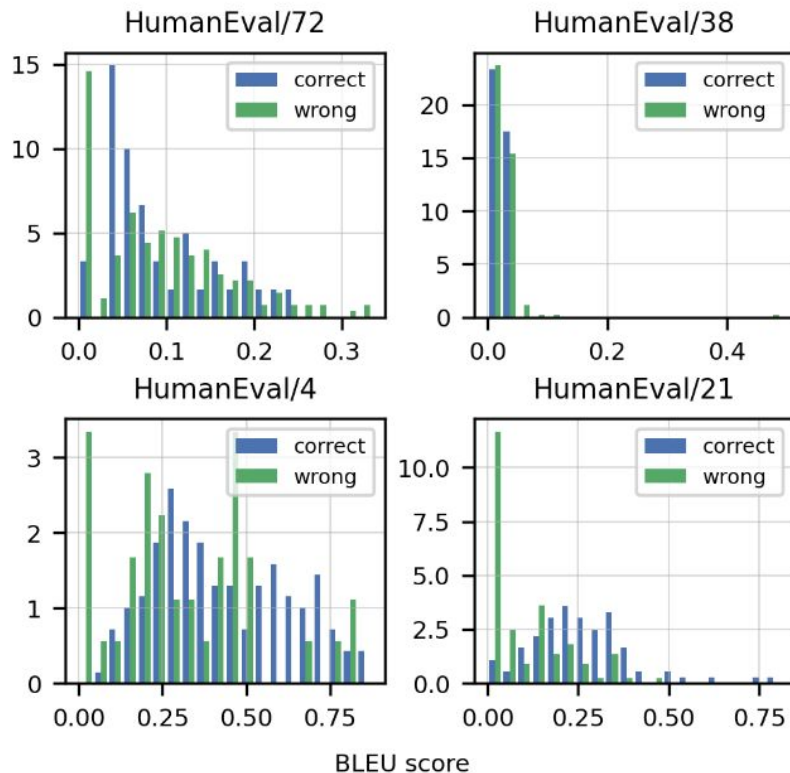
Метрика не очень стабильная, поэтому используют немного другой вариант:

Запусков будет n ($n > k$), а затем усредняются значения, чтобы получить ту же метрику.

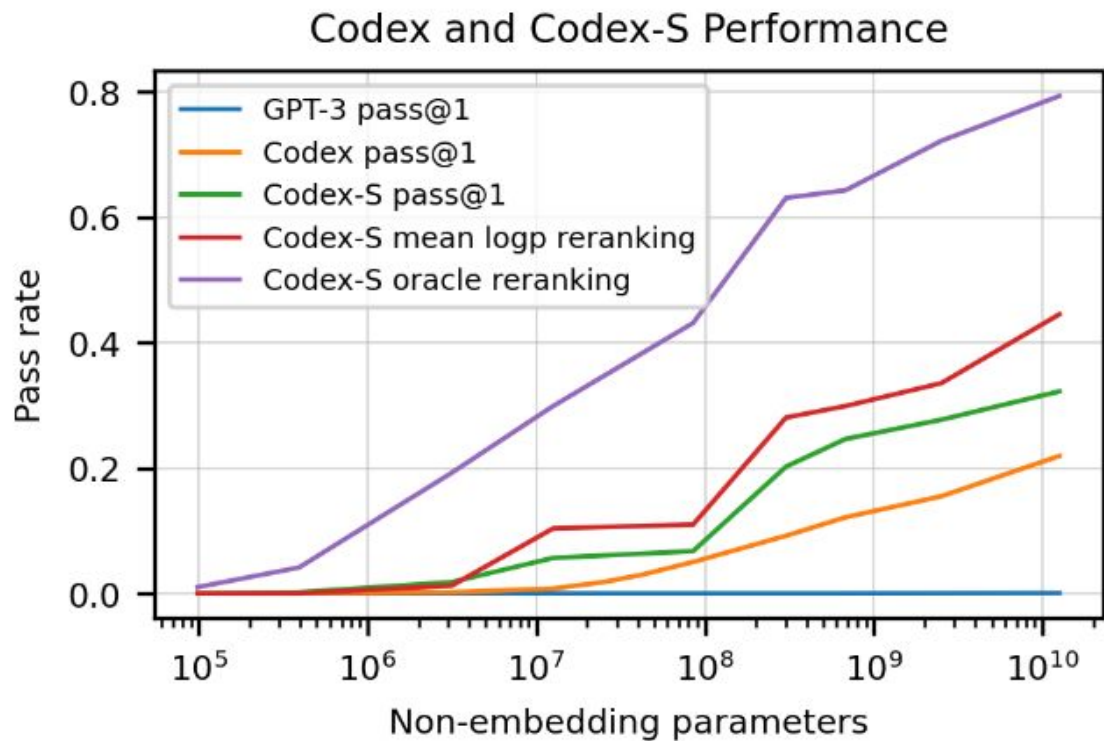
Что с BLEU не так?

Метрика хорошая для задач с текстами.

В случае кода есть проблемы.
Распределения BLEU для
правильных и неправильных
программ сложно разделить,
поэтому оптимизация BLEU
не даёт гарантий на код.



Результаты



Как можно улучшить?

1) Стартовая модель

Пробовали обучать с нуля и начиная с предобученной на тексте модели (GPT).

Улучшается только скорость сходимости, качество практически не меняется.

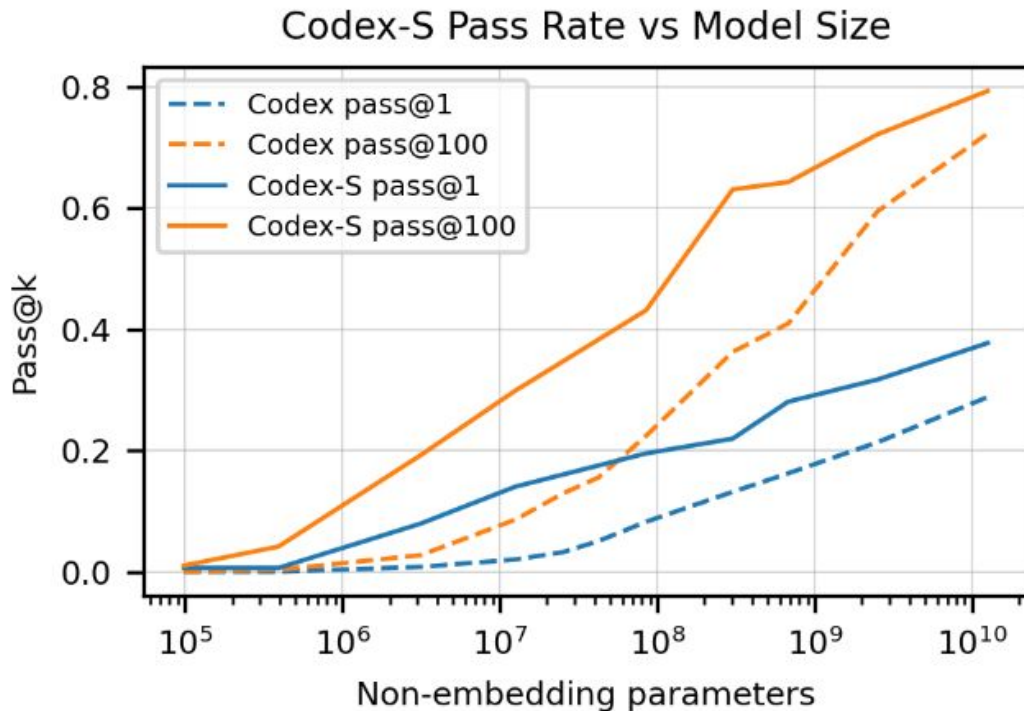
Возможное объяснение в том, что данных с кодом много.

Как можно улучшить?

2) Fine-tuning

- 10,000 задач из популярных контестов
- 40,000 функций с тестами из open source

Дообучили Codex-S



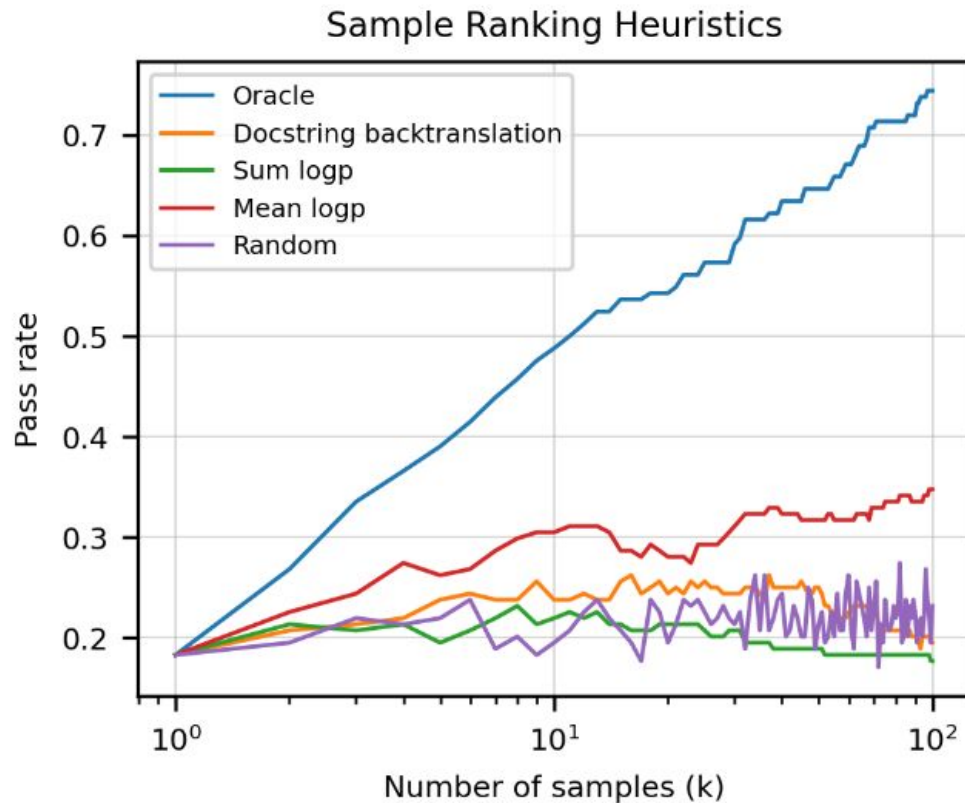
Как можно улучшить?

3) Размен времени работы на качество

Запускаем модель k раз и выбираем лучший результат.

Как выбрать лучший, если мы не можем вычислить метрику?

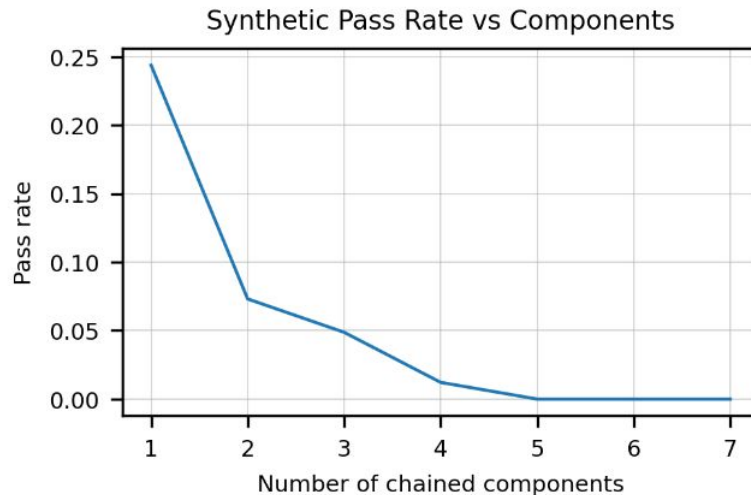
Mean token log probability



Слабости модели

- Неэффективное обучение. Человеку достаточно пары курсов по программированию, чтобы решить больше задач.
- Не справляется с последовательностью простых задач.

```
def do_work(x, y, z, w):  
    """ Add 3 to y, then subtract 4  
    from both x and w. Return the  
    product of the four numbers. """  
    t = y + 3  
    u = x - 4  
    v = z * w  
    return v
```

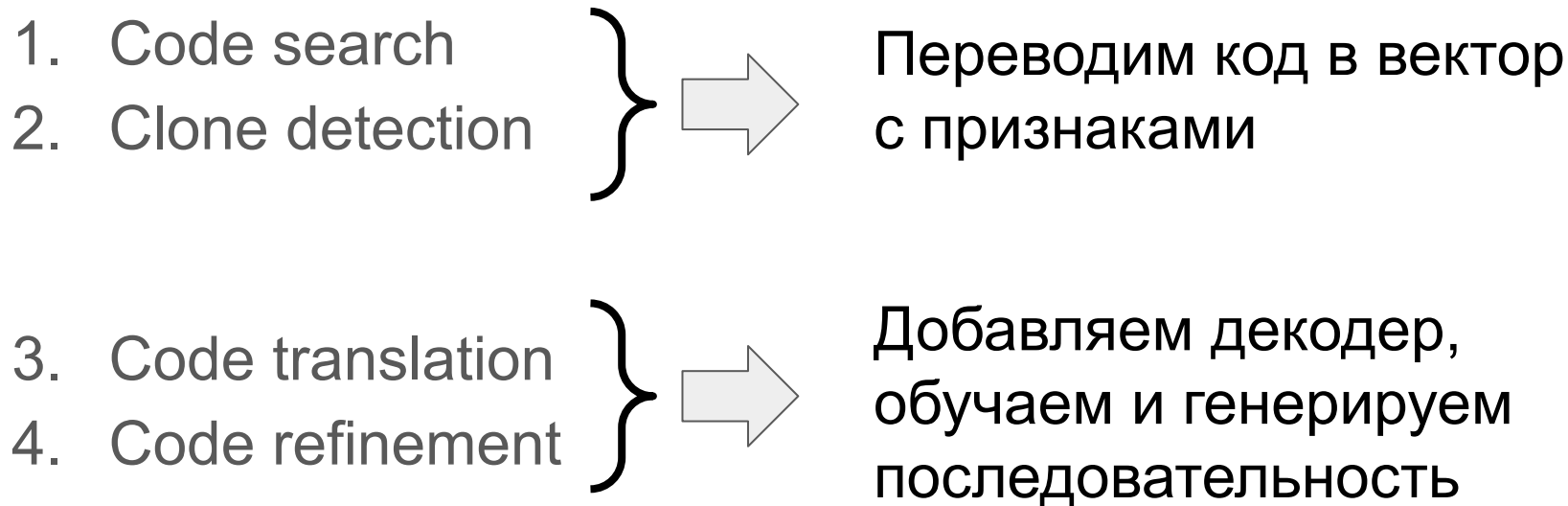


GraphCodeBert

В чём отличие?

- Другая модель
- Модель решает другие задачи
- Обучается тоже на других задачах
- Использует не только текстовое представление кода

Какие задачи решает модель?



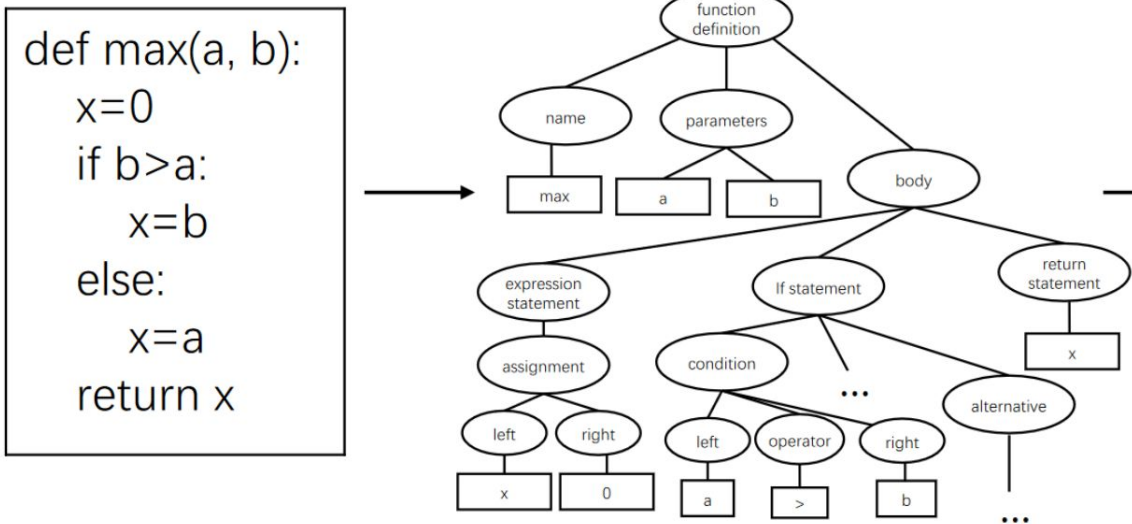
Извлечение информации из кода

Зачем?

`v = max_value - min_value` - сложно работать с `v`

Почему не AST?

Слишком сложная и избыточная структура.

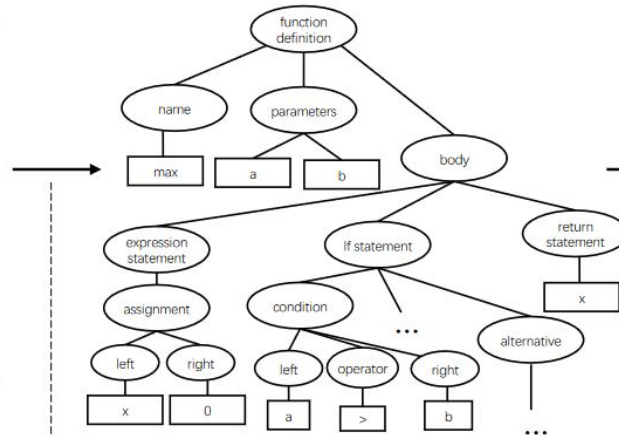


Замена AST - Data flow

Source code

```
def max(a, b):  
    x=0  
    if b>a:  
        x=b  
    else:  
        x=a  
    return x
```

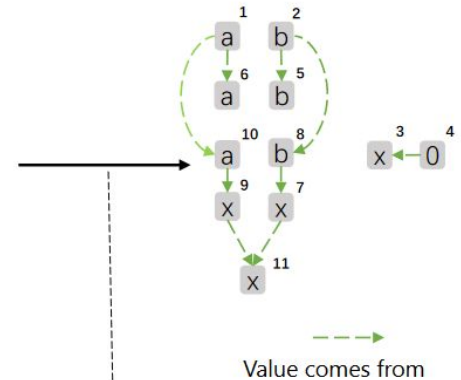
Parse into AST



Identify variable sequence

```
def max(a1, b2):  
    x3=04  
    if b5>a6:  
        x7=b8  
    else:  
        x9=a10  
    return x11
```

Variable relation

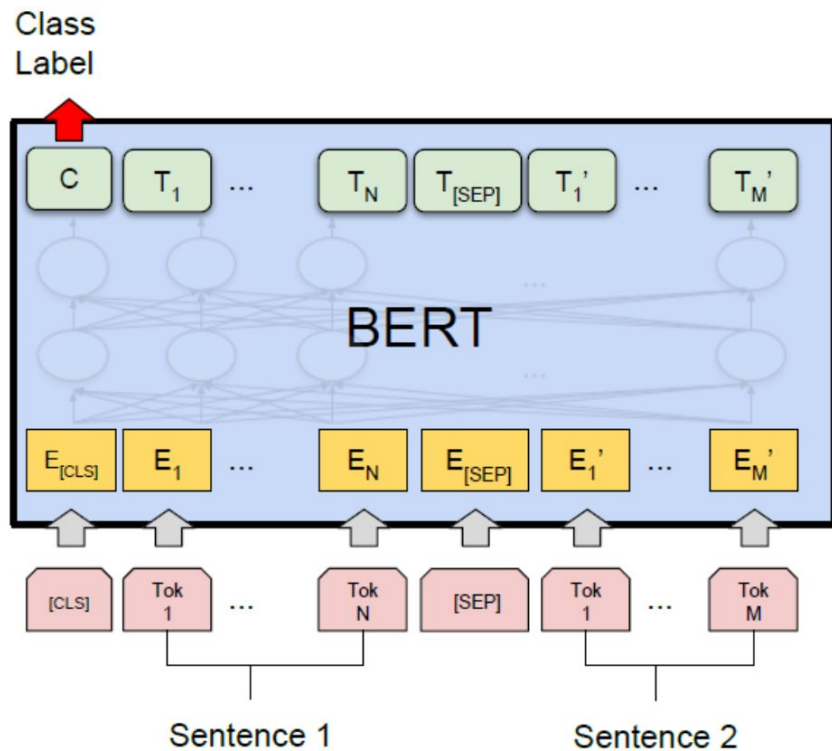


Compiler Tool

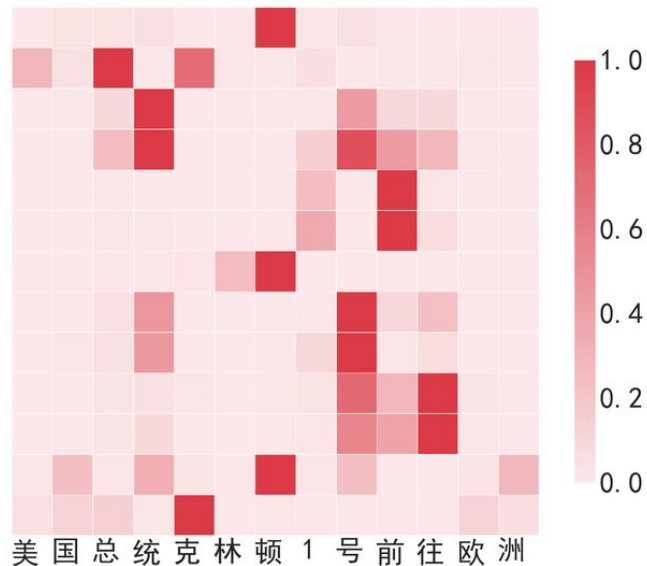
Identify variable sequence in AST

Extract variable relation from AST

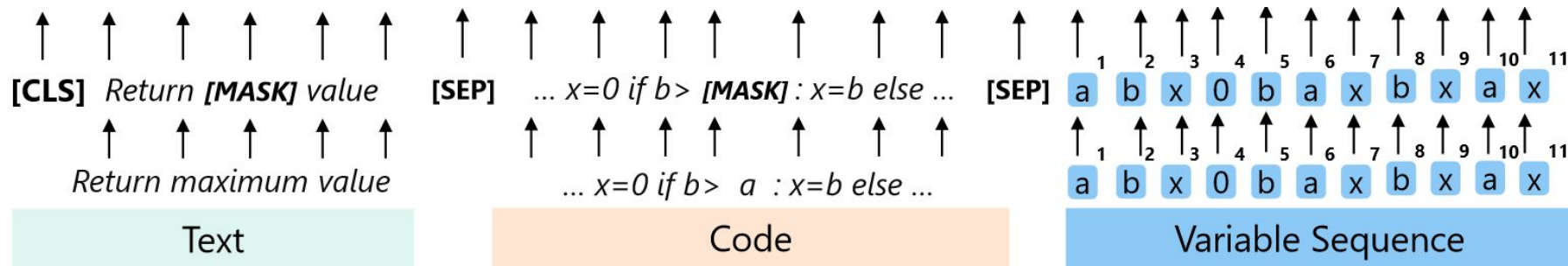
BERT



美国总统克林顿 1 号前往欧洲



Входные данные

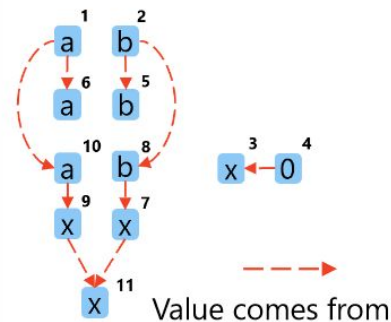


Comment
Return maximum value

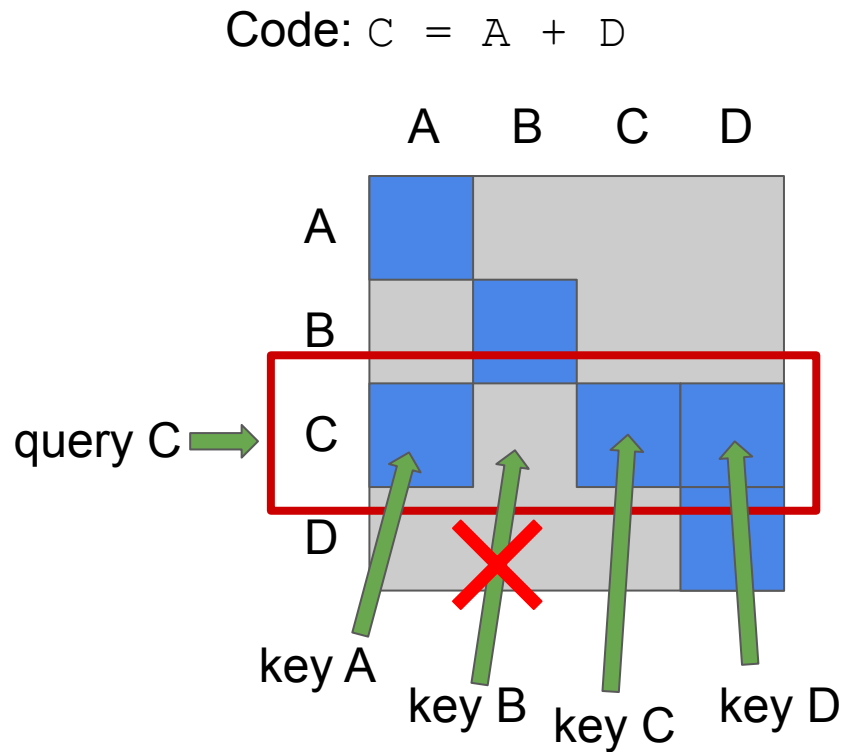
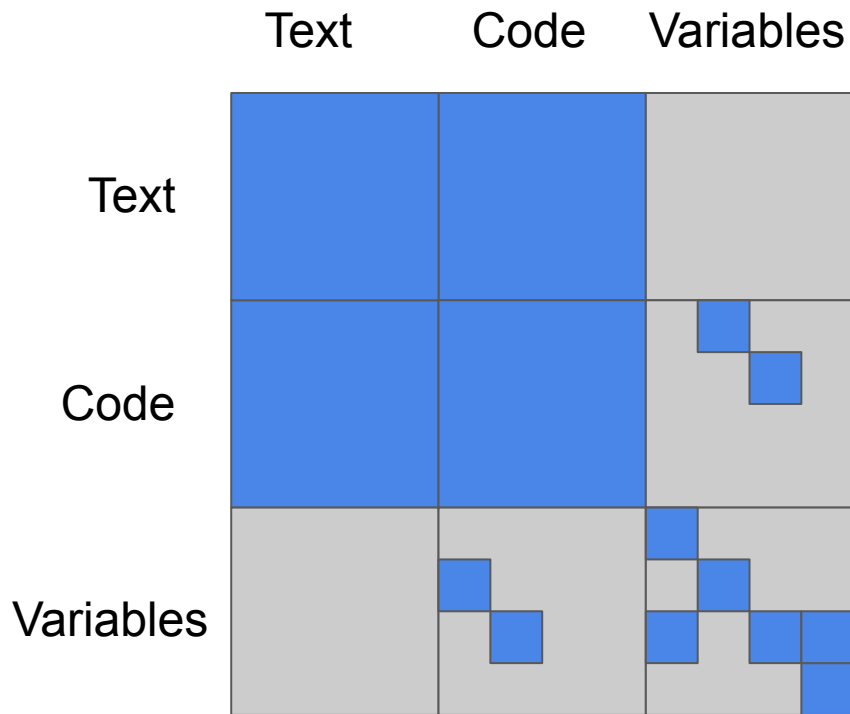
Source code

```
def max(a1, b2):  
    x3=04  
    if b5>a6:  
        x7=b8  
    else:  
        x9=a10  
    return x11
```

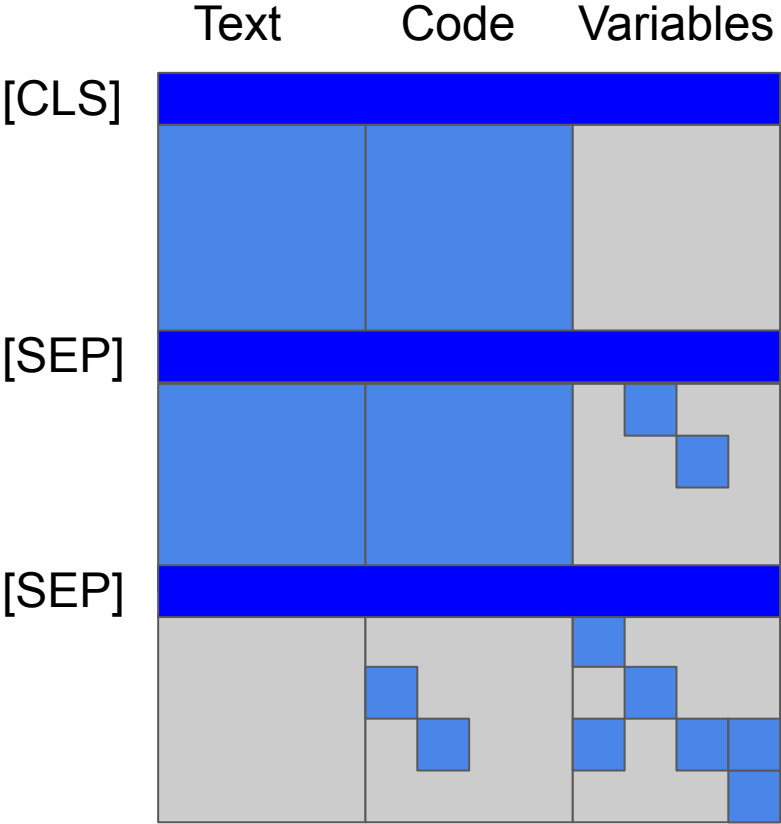
Data Flow



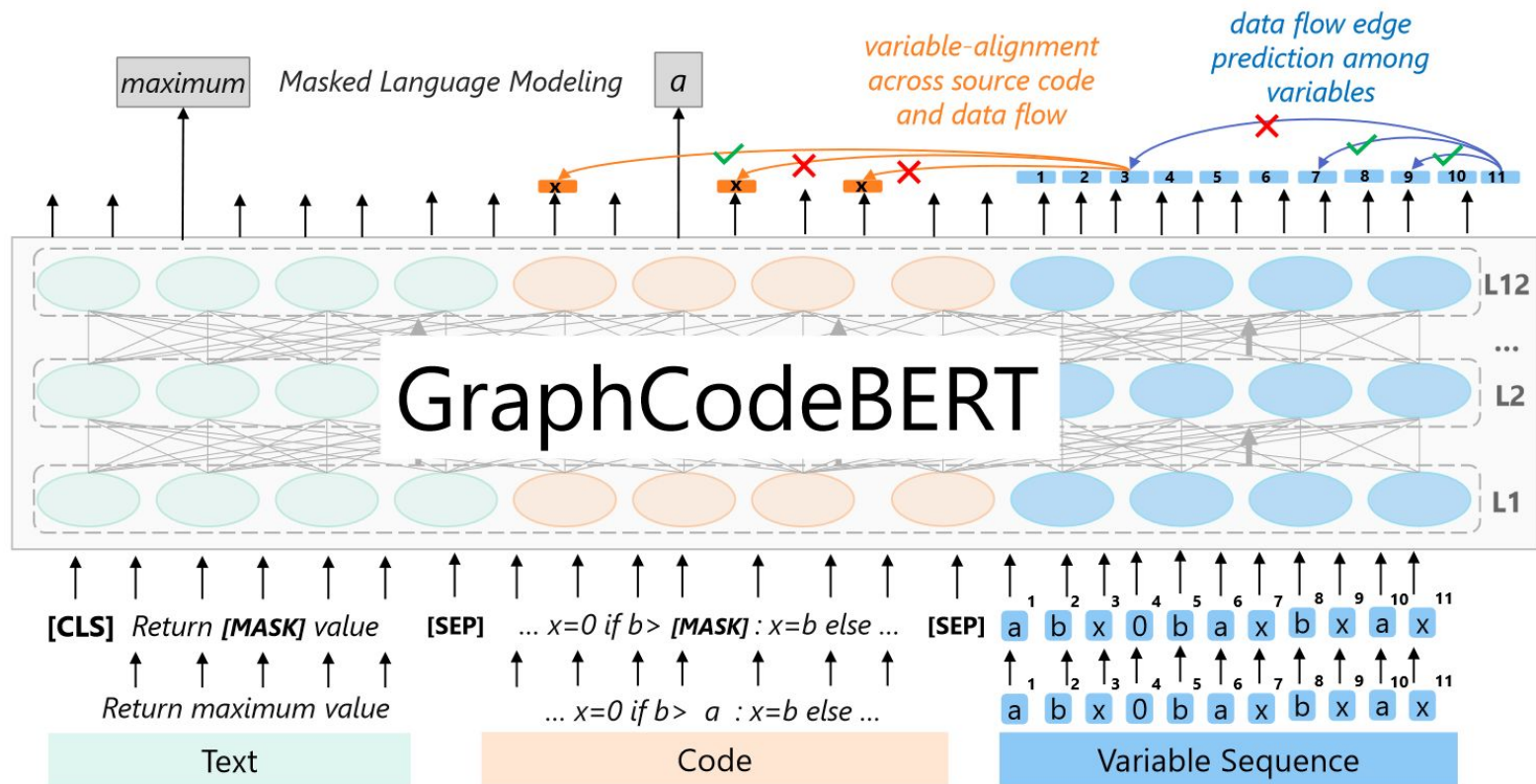
Masked attention



Masked attention



Задачи для обучения



Как предсказывать связи?

Edge Prediction

Случайно выбираем 20% переменных V

Удаляем связи переменных с V

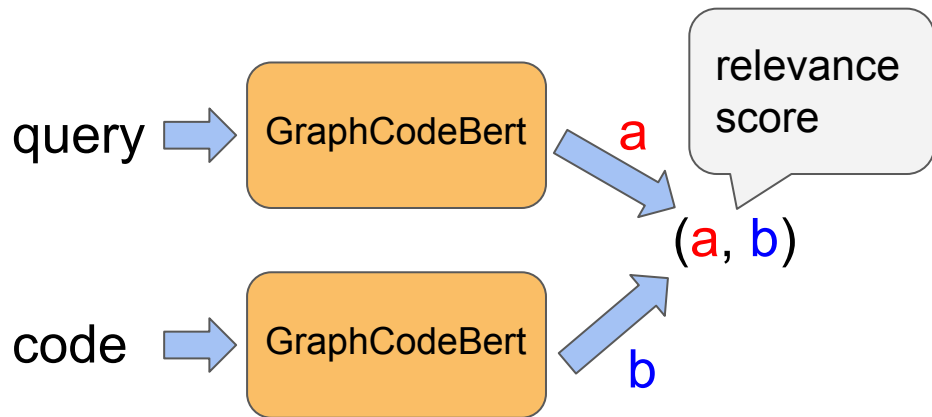
Считаем сигмоиду от скалярного произведения двух векторов и получаем вероятность

Node Alignment

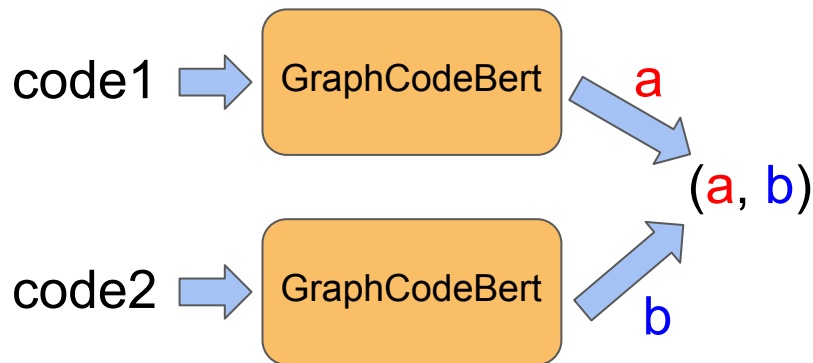
Удаляем связи токенов кода с V

Решение задач

Natural Language Code Search



Code Clone Detection



Code Clone Detection

Input: Two source codes

```
protected String downloadURLtoString(URL url) throws IOException
{
    BufferedReader in = new BufferedReader(new
        InputStreamReader(url.openStream()));
    StringBuffer sb = new StringBuffer(100 * 1024);
    String str;
    while ((str = in.readLine()) != null) {
        sb.append(str);
    }
    in.close();
    return sb.toString();
}
```

Output: Semantically similar (score: 0.983)

```
public static String fetchUrl(String urlString)
{
    try {
        URL url = new URL(urlString);
        BufferedReader reader = new BufferedReader(new
            InputStreamReader(url.openStream()));

        String line = null;
        StringBuilder builder = new StringBuilder();
        while ((line = reader.readLine()) != null) {
            builder.append(line);
        }
        reader.close();
        return builder.toString();
    } catch (MalformedURLException e) {
    } catch (IOException e) {
    }
    return "";
}
```

Решение задач

Code Translation

Input: A Java method

```
public void print(boolean b)
{
    print(String.valueOf(b));
}
```



Output: Its C# version

```
public void print(bool b)
{
    print(b.ToString());
}
```

Code Refinement

Input: A buggy Java method

```
public int add ( int a , int b )
{
    return a * b ;
}
```



Output: The fixed one

```
public int add ( int a , int b )
{
    return a + b ;
}
```

```
public void METHOD_1 ( TYPE_1 c )
{
    return VAR_1 . remove ( c ) ;
}
```



```
public void METHOD_1 ( TYPE_1 c )
{
    VAR_1 . remove ( c ) ;
}
```

ИСТОЧНИКИ

- <https://arxiv.org/pdf/2107.03374.pdf>
- <https://arxiv.org/pdf/2009.08366.pdf>
- <https://jalammar.github.io/illustrated-gpt2/>
- <https://arxiv.org/pdf/2002.08155.pdf>
-