Exame de qualificação de mestrado

Aprendizagem de representação através do uso de redes neurais convolucionais na recuperação de trecho de código-fonte

Marcelo de Rezende Martins

sob orientação do Prof. Dr. Marco Aurélio Gerosa

Instituto de Pesquisas Tecnológicas do Estado de São Paulo - IPT

Intro

Definição

Recuperação de trecho de código-fonte consiste em recuperar um trecho de código a partir de um repositório de códigos-fontes, de modo a atender a intenção do desenvolvedor, expressa em linguagem natural [1][2].

[1] Jose Cambronero, Hongyu Li, Seohyun Kim, Koushik Sen, and Satish Chandra. 2019. When deep learning met code search.

[2] Xiaodong Gu, Hongyu Zhang, and Sunghun Kim. 2018. Deep code search.

Definição (ERRATA)

Code Retrieval: Dada uma questão em linguagem natural $q \in \mathbb{Q}$, um modelo F_r será treinado a recuperar os trechos $\mathbb{C}^+ \subset \mathbb{C}_a$ com a maior pontuação:

$$\mathbb{C}^+ = \underset{c \in \mathbb{C}_a}{argmax}\, F_r(q, c) \tag{1}$$

Abordagem

Joint Embedding

Sejam $\mathbb{Q}$ e $\mathbb{C}$ conjuntos de dados heterogêneos. *Joint embedding* pode ser formulado como:

$$f : q \rightarrow t_q \rightarrow h_\theta(t_q, t_c) \leftarrow t_c \leftarrow c : g \tag{2}$$

Joint Embedding

Como representar as palavras e os tokens das questões e trechos de código-fonte?	*Word2Vec*
Como representar as sentenças?	CNN
Como aproximá-los?	Função de custo *hinge*

Arquitetura

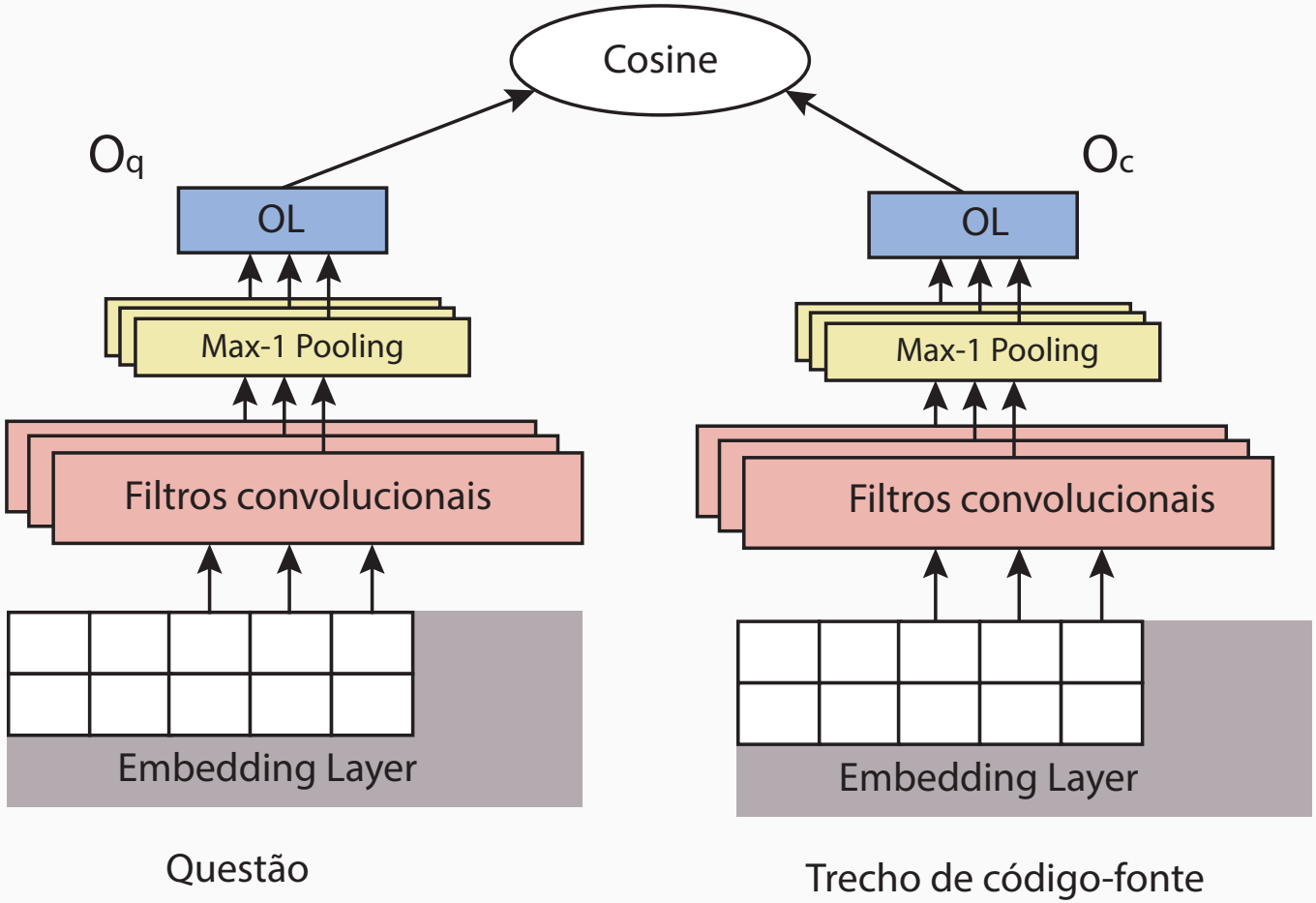


Figura 1: Arquitetura CNN proposta para recuperação de trecho de código-fonte.

CNN

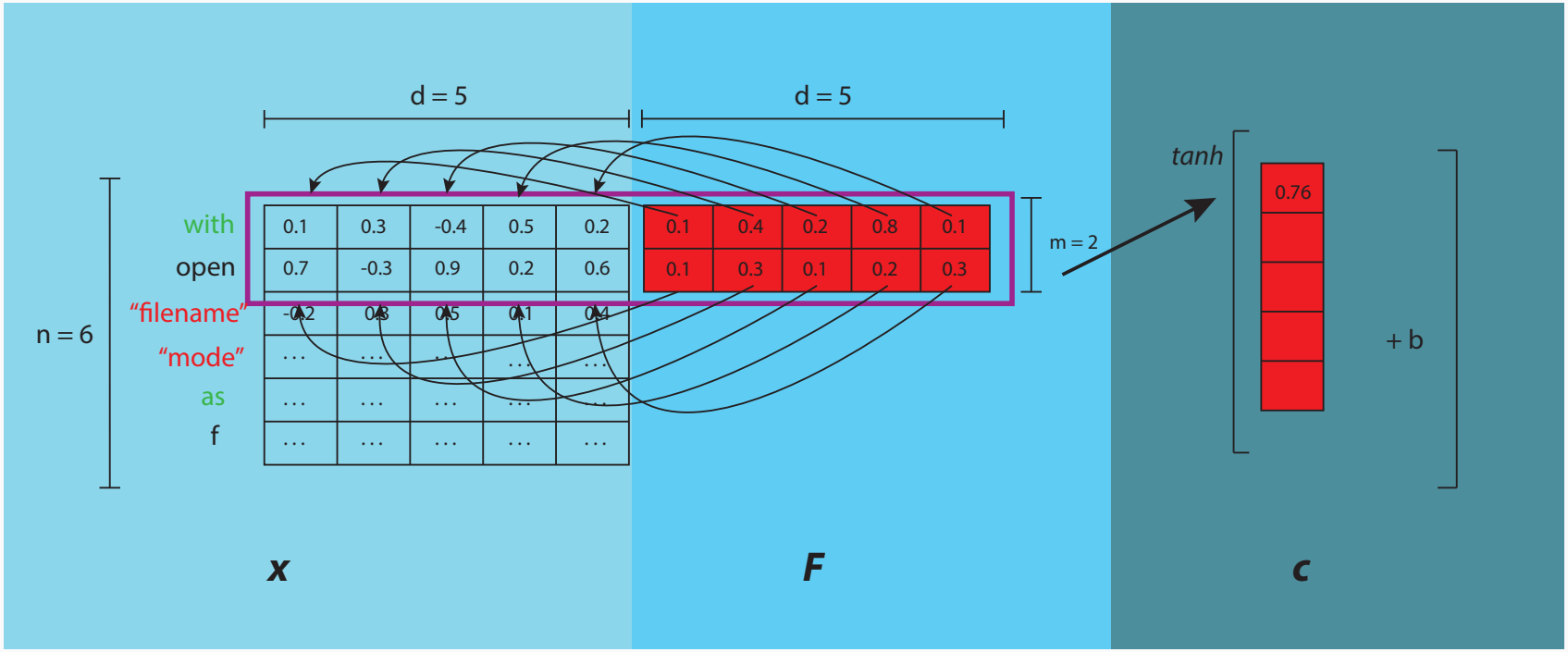


Figura 2: Primeiro passo da operação de convolução em um vetor de entrada ***x*** composto por vetores de representação distribuída de cada palavra da sentença.

Questões

Questões

- A aprendizagem de representação através do CNN auxilia na recuperação de trecho de código-fonte?
- O CNN é capaz de extrair as características mais relevantes de modo a facilitar o modelo a encontrar uma correlação entre as questões e os trechos de código-fonte?

Indiretamente:

- As interações locais auxiliam na aproximação das intenções aos trechos de código?

Avaliação

Dados de treinamento	Conjunto de pares de questões e trechos de código-fonte em Python coletados do Stack Overflow por Yao et al. (2018)[3]
Dados para avaliação final	Conjunto de dados anotados manualmente e disponibilizados por Yao et al. (2018)[3]
Métrica de desempenho	*MRR*

[3] Yao, Ziyu and Weld, Daniel S. and Chen, Wei-Peng and Sun, Huan. 2018. StaQC: A Systematically Mined Question-Code Dataset from Stack Overflow

Avaliação

Arquiteturas de referência para comparação	• Embedding • Rede neural com mecanismo de atenção proposto por Cambronero et al. (2019)[4]

[4] Jose Cambronero, Hongyu Li, Seohyun Kim, Koushik Sen, and Satish Chandra. 2019. When deep learning met code search.

Avaliação

Análise dos resultados	• Inspeção manual • Análise dos piores casos • Patologia das redes neurais (Feng et al. , 2018)[5]

[5] Shi Feng, Eric Wallace, Alvin Grissom II, Mohit Iyyer, Pedro Rodriguez, Jordan Boyd-Graber. 2018. Pathologies of Neural Models Make Interpretations Difficult.

Experimento piloto

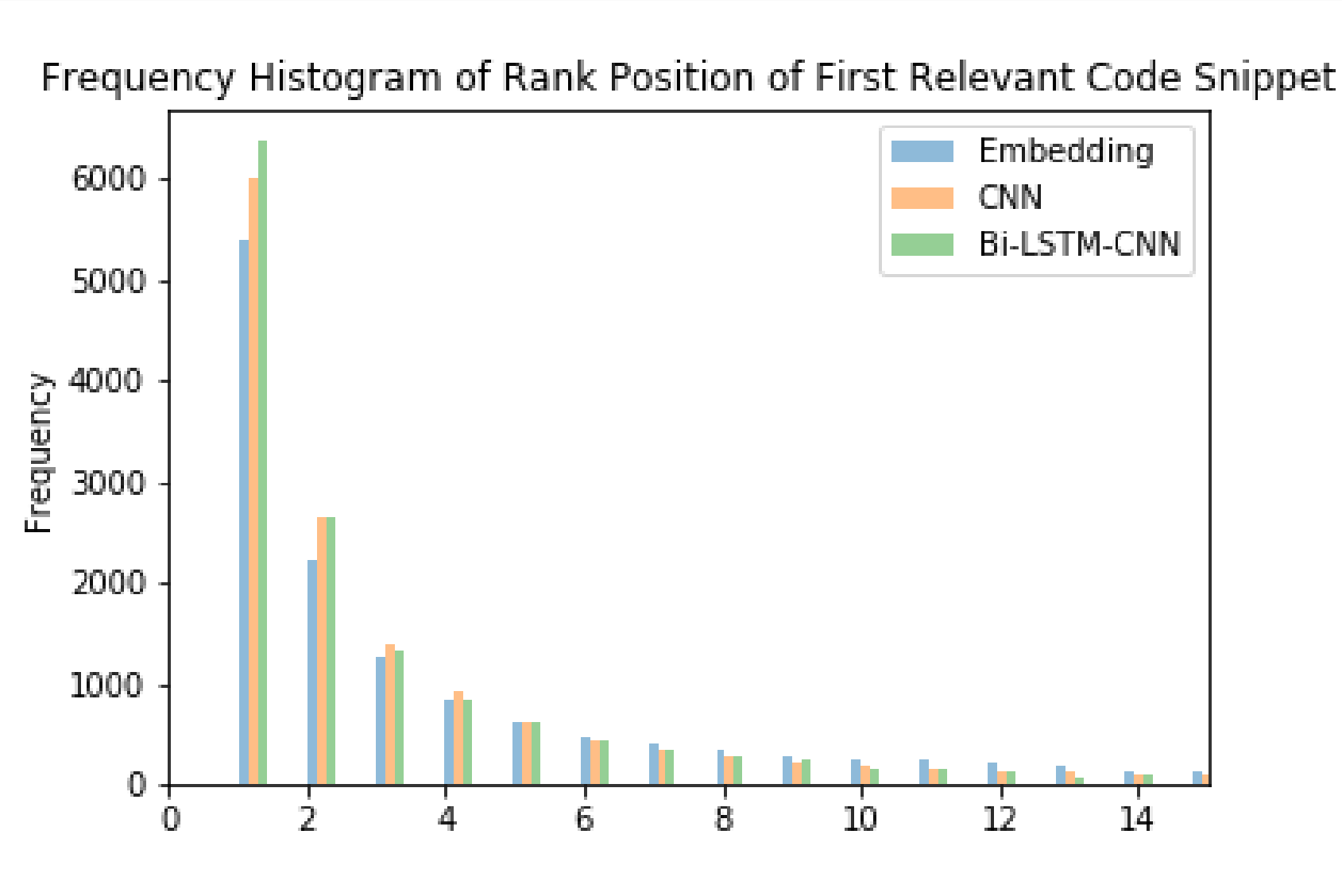
Resultados preliminares[6]

Modelos	Resultados (MRR)
Embedding	$0,52 \pm 0,01$
CNN	$\mathbf{0,58 \pm 0,01}$
bi-LSTM-CNN	$0,60 \pm 0,02$

Tabela 1: Resultado preliminar do modelo CNN em comparação com outras duas arquiteturas (bi-LSTM com CNN e Embedding). Estes resultados foram obtidos a partir da amostra EVAL.

[6] Marcelo de Rezende Martins e Marco Aurélio Gerosa. 2019. Um estudo preliminar sobre o uso de uma arquitetura deep learning para seleção de respostas no problema de recuperação de trecho de código-fonte.

Histograma das posições dos trechos de código-fonte relevantes



Exemplos

Python and appending items to text and excel file[7]

BiLSTM-CNN

```
Yvalues = [1, 2, 3, 4, 5]
file_out = open('file.csv','wb')
mywriter= csv.writer(file_out, delimiter = '\n')
mywriter.writerow(Yvalues)
file_out.close()
```

CNN

```
import csv

with open("output.csv", "wb") as f:
    writer = cwv.writer(f)
    writer.writerows(a)
```

[7] https://stackoverflow.com/questions/24593478/python-and-appending-items-to-text-and-excel-file

Próximos passos

Próximas tarefas

- Implementação da arquitetura proposta por Cambronero et al. (2019)[8]
- Adição de regularização aos modelos
- Coleta e análise dos resultados

[8] Jose Cambronero, Hongyu Li, Seohyun Kim, Koushik Sen, and Satish Chandra. 2019. When deep learning met code search.

Perguntas?