Deep Code Search

ABSTRACT

To implement a program functionality, developers can reuse previously written code snippets by searching through a large-scale codebase. Over the years, many code search tools have been proposed to help developers. The existing approaches often treat source code as textual documents and utilize information retrieval models to retrieve relevant code snippets that match a given query. These approaches mainly rely on the textual similarity between source code and natural language query. They lack a deep understanding of the semantics of queries and source code.

In this paper, we propose a novel deep neural network named CODEnn (Code-Description Embedding Neural Network). Instead of matching text similarity, CODEnn jointly embeds code snippets and natural language descriptions into a high-dimensional vector space, in such a way that code snippet and its corresponding description have similar vectors. Using the unified vector representation, code snippets related to a natural language query can be retrieved according to their vectors. Semantically related words can also be recognized and irrelevant/noisy keywords in queries can be handled.

As a proof-of-concept application, we implement a code search tool named DeepCS using the proposed CODEnn model. We empirically evaluate DeepCS on a large scale codebase collected from GitHub. The experimental results show that our approach can effectively retrieve relevant code snippets and outperforms previous techniques.

CODEnn没有匹配文本相似性，而是将代码片段和自然语言描述共同嵌入到高维向量空间中，使得代码片段及其对应的描述具有相似的向量。使用统一的向量表示，可以根据向量检索与自然语言查询相关的代码片段。语义相关的单词也可以识别，查询中不相关的/有噪声的关键字也可以处理。