Can CodeT5 embeddings be adapted for efficient Code Clone Detection and Retrieval?

Chinmay Rane, Master of Science in Computer Science

University of Dublin, Trinity College, 2022

Supervisor: Professor David Gregg

With the increase in the number of programs written every day, it becomes challenging to maintain a large software repository. Owing to this, large-scale code clone detection and retrieval have become a necessity. There exist several works offering solutions to solve the problem. However, most of the works have trouble maintaining a balance between accuracy and scalability. Classical approaches have high scalability but lower precision whereas recent neural network-based models have high precision but suffer from scalability. In this work, we show how CodeT5 a recent neural network-based model could be modified to reduce its usage during clone retrieval. Particularly, we fine-tune the architecture to extract code embeddings rich in semantic and syntactic information. Through experiments on the BigCloneBench dataset, we assess the efficacy of the generated code embeddings and show how our proposed Nearest Neighbor-based retrieval approach fetches clones in real-time while achieving comparable accuracy to the original CodeT5 architecture.

Keywords: Clone Detection, Clone Retrieval, Deep Learning, Neural Networks, Code Embeddings, K-Nearest Neighbor Search