# Fourth Milestone Report: Non-parametric Language Models for Natural Language to Code Generation

Alex Xie
https://axie66.github.io/07400-project/

Mentored by Vincent Hellendoorn
Institute for Software Research

February 28, 2022

## 1 Major Changes

There are no significant changes since the last milestone.

## 2 Progress Report

### 2.1 Accomplishments

My main accomplishment since the last checkpoint has been fine-tuning CodeT5 for the CodeSearchNet dataset. As CodeSearchNet is several times larger than the datasets previously used, training also took significantly longer, on the order of several hours per epoch. The best model I trained achieves 2.4 BLEU and 13 CodeBLEU - this is very poor, but consistent with other parametric models trained on this dataset [2]. This leaves a great deal of room for improvement and bodes well for retrieval based approaches such as $k$NN. I have not yet tested $k$NN retrieval on this dataset - this will be done by the next milestone.

Further, I experimented with an adaptive $k$NN retrieval mechanism that selectively performs retrieval for only a subset of generated tokens. Interestingly, I found that this performed extremely poorly - at train time, the loss plateaued around the value prior to training the retrieval network. Further, at test time, using adaptive retrieval greatly reduced accuracy by around 20 BLEU.

### 2.2 Previous Milestone Goals

While I have trained the base parametric model on CodeSearchNet, due to the lengthy training time, I have not yet been able to run $k$NN experiments on the dataset. Further, I have tested the adaptive retrieval approach as described in my last milestone.

### 2.3 Surprises

The main surprise is the poor performance of the adaptive retrieval mechanism on the CoNaLa dataset.

# 3 Next Steps

## 3.1 Looking Ahead

By the next milestone, I aim to complete $k$NN experiments on CodeSearchNet, as well as to have results for $k$NN code generation on the much larger CoDesc dataset [1], which may hopefully give a reflection of how $k$NN performs at a large scale.

## 3.2 Revisions to Future Milestones

No future milestones need to be revised at this time.

# References

[1] Masum Hasan, Tanveer Muttaqueen, Abdullah Al Ishtiaq, Kazi Sajeed Mehrab, Md. Mahim Anjum Haque, Tahmid Hasan, Wasi Ahmad, Anindya Iqbal, and Rifat Shahriyar. CoDesc: A large code–description parallel dataset. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 210–218, Online, August 2021. Association for Computational Linguistics.

[2] Md Rizwan Parvez, Wasi Uddin Ahmad, Saikat Chakraborty, Baishakhi Ray, and Kai-Wei Chang. Retrieval augmented code generation and summarization. In *EMNLP-Findings*, 2021.