Improving Automated Commit Message Generation

Megha Chakraborty, MDS202022
meghac@cmi.ac.in

June 2021 - December 2021

During our summer internship as part of the IBM GRM Program, we had worked on commit message generation based on the CC2Vec model [2]. The model produced distributed representations (or feature vectors) for code changes using Hierarchical Attention Network. These feature vectors were then used to generate commit messages for a given code change.

However, we observed that the embedding space created by the model was "anisotropic" in nature, i.e. arbitrary vectors have high cosine similarities and thus the embedding space has a narrow cone representation. As a result, the model is unable to output similar commit messages for similar code changes, which would have been the case if the distributed representations were clustered in a better way. The narrow cone structure misleads the commit message generation process and gives undesirable results in some cases.

As an extension of the internship, for the Industry Project, I worked on solving the anisotropy problem in an attempt to get better distributed representations. The hypothesis is that creating more distinguishable clusters in the embedding space would improve the current commit message generation scores.

During the project, I thoroughly understood the concept of Isotropy of Contextual Embedding Spaces and concluded through CC2Vec is indeed highly anisotropic. I found a measure of this isotropy derived from previous work by [1], [3], [4].

Next, I applied a simple topic-modelling technique on the commit messages given in the code-patch based on "action phrases" in the commits. I extracted the "action phrases" in each commit in the form of a list and appended it to the corresponding code change. This gave us a new input data to work on- the new input being enriched by the topic words in the code-change.

Then, I performed the experiments advocated by the original authors of the CC2Vec paper [2] on the new input data. The results showed an improvement in the final Bleu score results- from 20.57(original) to 21.11(new). And we also saw an improvement in the isotropy of the new embedding space formed.

The experiment has shown that adding topics to the input has improved the isotropy of the embedding space formed as well as given a better results in average Bleu score.

- Meausuring Isotropy
- Average Bleu Scores

The desired outcome of this project was to improve the current embedding space (generated by CC2Vec) such that the vectors are not all closely clustered in a narrow cone but form distinguishable clusters based on their features. This resulted in an improvement of the commit messages generated.

Mentors:

- Dr. Monika Gupta, IBM Research
  Email id: `mongup20@in.ibm.com`
- Dr. Venkatesh Vinayakarao, CMI
  Email id: `venkateshv@cmi.ac.in`

References

[1] Sanjeev Arora, Yuanzhi Li, Yingyu Liang, Tengyu Ma, and Andrej Risteski. A latent variable model approach to pmi-based word embeddings. *Transactions of the Association for Computational Linguistics*, 4:385–399, 2016.

[2] Thong Hoang, Hong Jin Kang, David Lo, and Julia Lawall. Cc2vec: Distributed representations of code changes. *Proceedings of the ACM/IEEE 42nd International Conference on Software Engineering*, 2020.

[3] Jiaqi Mu, Suma Bhat, and Pramod Viswanath. All-but-the-top: Simple and effective postprocessing for word representations. *arXiv preprint arXiv:1702.01417*, 2017.

[4] Sara Rajaee and Mohammad Taher Pilehvar. A cluster-based approach for improving isotropy in contextual embedding space. *arXiv preprint arXiv:2106.01183*, 2021.