Secora: Semantic Code Retrieval Analysis

Datascience Project

Florian Hoelscher, Kai Glasenapp

Hochschule Bonn-Rhein-Sieg

October 2021

Outline I

Motivation: Semantic Code Search

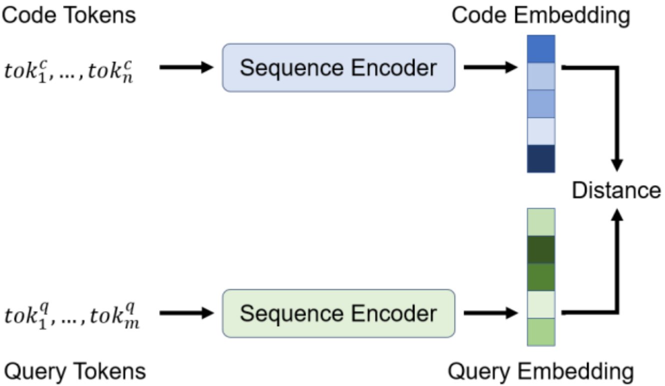
- Searching code is essential for Developers
- Semantic code search is about finding code with natural language queries
- Existing code is difficult to find and access
- There may be better approaches to solutions than term frequency

Goals

- Investigate a recent retrieval neural network on CodeSearchNet
- We want to surpass the baseline until the end of this project

CodeSearchNet Challenge [3]

- Information retrieval Task for source code
- Is now concluded and fully open to the public
- Succeeded by CodeXGlue, but still very relevant
- Premise: documentation serves as proxy for search queries
- 2 parts:
  1. CodeSearchNet Corpus, for training
  2. CodeSearchNet Challenge, for evaluation



[3]

Embedding distance of queries and the related code should be minimized

CodeSearchNet Corpus [3]

- Collected from Github based on stars and forks
- Languages: Go, Java, JavaScript, PHP, Python, Ruby
- Tests, unparseable and duplicate code is filtered out
- 2 million documentation and function pairs
- 4 million sole code samples
- 80-10-10 train/valid/test split of the whole dataset

Annotation for Validation [3]

- Set of 99 relevant natural language search queries from Bing
- Query results retrieved and filtered to top 10 candidates by an ensemble of the baseline models
- 4026 expert annotations for query - result pairs from 0 (totally irrelevant) to 3 (exact match)
- Normalized discounted cumulative gain as metric [7]

	Count by Relevance Score				Total
	0	1	2	3	Annotations
Go	62	64	29	11	166
Java	383	178	125	137	823
JavaScript	153	52	56	58	319
PHP	103	77	68	66	314
Python	498	511	537	543	2 089
Ruby	123	105	53	34	315

[3]

Distribution of annotations across the languages and scores, where 0 is totally irrelevant and 3 is an exact match

NDGC [1]

Normalized Discounted Cumulative Gain

- Used to evaluate information retrieval
- List of documents sorted by relevance
- CG - Sum of relevance:
- $\sum_{i=1}^{n} relevance_i = CumulativeGain$
- DCG - Takes order into account:
- $\sum_{i=1}^{n} \frac{relevance_i}{log_2(i+1)} = DiscountedCumulativeGain$
- Normalized DCG - found DCG value divided by ideal value to normalize
- $NDCG = \frac{DCG_f}{DCG_i}$

Related Work

Baseline Models:

- Self-Attention where multi-head attention is used to compute representations of each token in the sequence
- Neural Bag of Words where each (sub)token is embedded to a learnable embedding (vector representation)
- Bidirectional RNN models where we employ the GRU cell to summarize the input sequence
- 1D Convolutional Neural Network over the input sequence of tokens
- Elasticsearch with default tokenizer and parameters

Our Approach

- Investigate a recent unsupervised information retrieval model, SimCSE
- SimCSE performs well and is simple
- Build SimCSE based on a recent Bert variation, like Roberta [5], tbd.
- Test different configurations of natural language and code pretraining
- Finetune on CodeSearchNet corpus

SimCSE[2]

- Contrastive unsupervised and supervisedly trainable information retrieval model
- Simple idea, independently sample dropout for samples and minimize the distance to themself
- Dropout as data augmentation
- It optimizes alignment and uniformity, which is important to contrastive learning

SimCSE

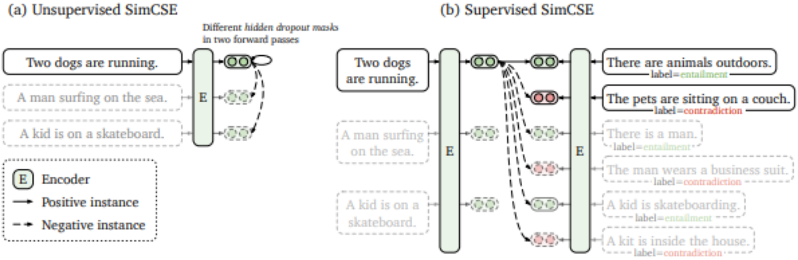


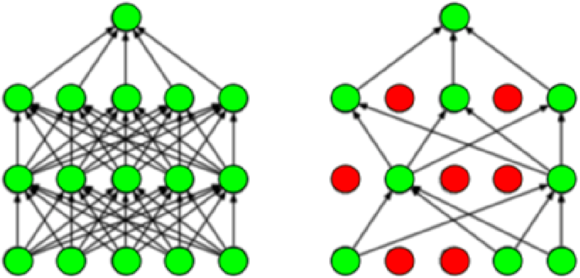
Figure of the SimCSE [2] contrastive architecture

Contrastive Self Supervised Learning

- SSL and Unsupervised are similar in meaning
- SSL use the parts of the data itself as labels for "self supervision"
- Contrastive learning reduces the amount of needed labeled data
- Contrastive learning uses similar positive tuples (x, x^+) and unrelated samples x^- [8]
- [4]

Dropout

Dropout [6]



(a) Standard Neural Net (b) After applying dropout.

SimCSE uses dropout with $p = 0.1$ to generate positive pairs [2]

SimCSE depends on dropout

Alignment [9]

A key properties for good contrastive learning

- Starting from a pretrained checkpoint is necessary because it provides good initial alignment
- Metric for embedding proximity of positive pairs should be minimized
- $\ell_{align} \triangleq \underset{(x,x^+)}{\mathbb{E}} \left[\, \|f(x) - f(x^+)\|^2 \, \right]$

Uniformity

- Metric for distribution of the (normalized) features on the vector space should be maximized
- $\ell_{uniform} \triangleq \; log\left[\mathop{\mathbb{E}}_{x,y\sim data} e^{-2\|f(x)-f(y)\|^2}\right]$

References I

Pranay Chandekar. *Evaluate your Recommendation Engine using NDCG*. 2020. URL: https://towardsdatascience.com/evaluate-your-recommendation-engine-using-ndcg-759a851452d1 (visited on 10/28/2021).

Tianyu Gao, Xingcheng Yao, and Danqi Chen. “SimCSE: Simple Contrastive Learning of Sentence Embeddings”. In: *Empirical Methods in Natural Language Processing (EMNLP)*. 2021.

Hamel Husain et al. “CodeSearchNet challenge: Evaluating the state of semantic code search”. In: *arXiv preprint arXiv:1909.09436* (2019).

References II

Yann LeCun and Ishan Misra. *Self-supervised learning: The dark matter of intelligence*. Facebook AI Research Blogpost. Mar. 2021 [Online]. URL: https://ai.facebook.com/blog/self-supervised-learning-the-dark-matter-of-intelligence?fileGuid=WyYwxqq8kWjKdWgd.

Yinhan Liu et al. "RoBERTa: A Robustly Optimized BERT Pretraining Approach". In: *ArXiv* abs/1907.11692 (2019).

Nisha McNealis. *A Simple Introduction to Dropout Regularization (With Code!)* 2020. URL: https://medium.com/analytics-vidhya/a-simple-introduction-to-dropout-regularization-with-code-5279489dda1e (visited on 10/29/2021).

References III

I. C. Mogotsi. "Christopher D. Manning, Prabhakar Raghavan, and Hinrich SchüTze: Introduction to Information Retrieval". In: *Inf. Retr.* 13.2 (Apr. 2010), pp. 192–195. ISSN: 1386-4564. DOI: 10.1007/s10791-009-9115-y. URL: https://doi.org/10.1007/s10791-009-9115-y.

Nikunj Saunshi et al. "A Theoretical Analysis of Contrastive Unsupervised Representation Learning". In: *Proceedings of the 36th International Conference on Machine Learning*. Ed. by Kamalika Chaudhuri and Ruslan Salakhutdinov. Vol. 97. Proceedings of Machine Learning Research. PMLR, Sept. 2019, pp. 5628–5637. URL: https://proceedings.mlr.press/v97/saunshi19a.html.

References IV

Tongzhou Wang and Phillip Isola. "Understanding Contrastive Representation Learning through Alignment and Uniformity on the Hypersphere". In: *CoRR* abs/2005.10242 (2020). arXiv: 2005.10242. URL: https://arxiv.org/abs/2005.10242.