COLUMBIA UNIVERSITY
IN THE CITY OF NEW YORK

Unsupervised Document Retrieval by Contrastive Learning

Ryan Anselm, Kuan-Yao Huang, Todd Morrill, Howard Yong
Department of Computer Science

Introduction

Objective: Perform document retrieval by self-supervised contrastive learning.

- Contrastive learning[3] is a deep-learning based representation learning approach to maximize the similarity of representations between similar data and minimize the similarities between dissimilar data.
- Self-supervised learning attempts to learn robust representations by training on pseudo-labels (e.g. pairs of sentences in the same document).
- The system encodes queries and keys and retrieves a ranked list of relevant documents using cosine similarity.

Datasets

Wikipedia[5] contains **6,583,470 articles**, from which we can generate positive and negative training examples.

Positive Example (same Wikipedia article)

- "A tornado outbreak occurred on Saint Patrick's Day in the Deep South"
- "Six people were injured by four different tornadoes across Alabama during the outbreak"

Negative Example (different Wikipedia articles)

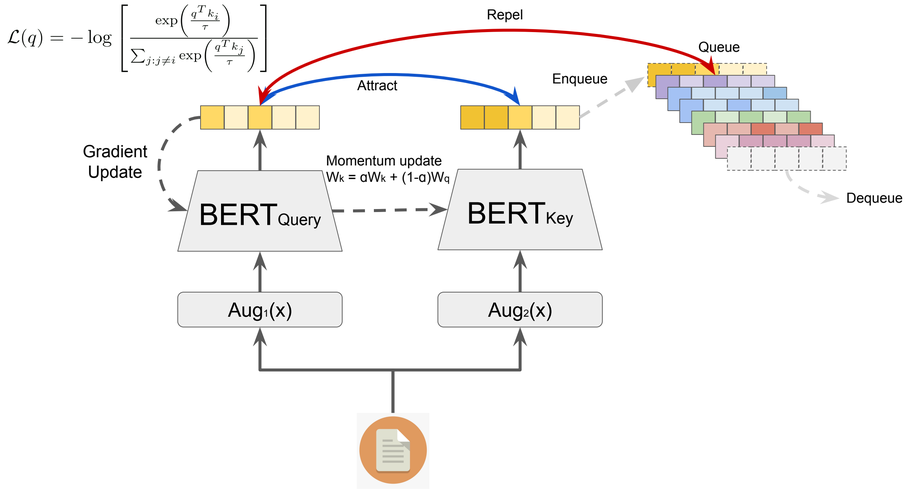
- "The reception room to the right from the foyer harmoniously mixes different periods and styles"
- "These taiga forests are almost completely intact apart from clearance around the city of Fairbanks, Alaska"

The **BEIR benchmark dataset**[4] is a collection of document retrieval datasets used to evaluate the performance of document retrieval systems. It includes MSMARCO, the Quora question-answer pairs dataset, and many more.

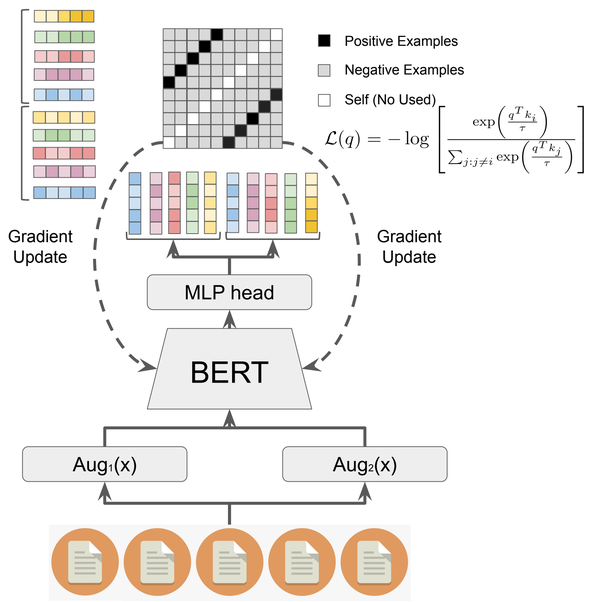
Architectures

Momentum contrastive learning (MoCo)[2]: We used random text deletion, masking, and back-translation to create the positive examples. 65,000 negative examples are evaluated against 1 positive example to improve the quality of contrastive learning. We used the InfoNCE loss as the loss function.

$$\mathcal{L}(q) = -\log\left[\frac{\exp\left(\frac{q^T k_i}{\tau}\right)}{\sum_{j:j\neq i}\exp\left(\frac{q^T k_j}{\tau}\right)}\right]$$



SimCLR[1]: The SimCLR model uses only one encoder but contains a dimension reduction head to make better representational alignment.



Architecture Details

Baseline: The baseline model was trained using the system in [3] for 5,000 optimization steps on 3 GPUs to establish baseline metrics.

Respect Document Boundaries: The original authors of [3] did not respect document boundaries when creating positive examples. This experiment patches that bug.

MoCo + SimCLR: This experiment combines the MoCo and SimCLR objectives with the following loss function.

$$\mathcal{L} = \mathcal{L}_{\text{MoCo}} + \lambda \cdot \mathcal{L}_{\text{SimCLR}}$$

Results

We evaluated the system on the BEIR benchmark[4] using standard metrics nDCG@10 and recall@100. nDCG@10 measures the quality of the top 10 results and recall@100 measures the portion of relevant results found in the top 100 hits.

nDCG@10

Dataset	Baseline	Respect Boundary	+SimCLR
Quora	2.315	3.35	4.2
ArguAna	7.732	3.136	35.1
SCIDOCS	0.238	0.393	0.4
Trec-COVID	0.317	0	0.3
FiQA-2018	0.143	0.925	0.4
NFCorpus	4.241	5.613	10
SciFact	9.797	10.933	35.4

Recall@100

Dataset	Baseline	Respect Boundary	+SimCLR
Quora	7.922	14.321	19.9
ArguAna	56.33	29.374	92.8
SCIDOCS	2.258	7.583	5.9
Trec-COVID	0.056	0.119	0.1
FiQA-2018	0.949	5.844	8.4
NFCorpus	9.207	8.72	14.2
SciFact	51.589	45.144	77.3

References

[1] Chen, Ting et al. "A Simple Framework for Contrastive Learning of Visual Representations." (2020). arxiv. https://arxiv.org/pdf/2002.05709.pdf
[2] He, Kaiming et al. "Momentum Contrast for Unsupervised Visual Representation Learning." 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2019): 9726-9735.
[3] Izacard, Gautier et al. "Unsupervised Dense Information Retrieval with Contrastive Learning." (2021). arxiv. https://arxiv.org/pdf/2112.09118.pdf
[4] Thakur, Nandan et al. "BEIR: A Heterogenous Benchmark for Zero-shot Evaluation of Information Retrieval Models." (2021). In NeurIPS 2021.
[5] Wikipedia Foundation, "Wikimedia Downloads." https://dumps.wikimedia.org