# Query-BERT:

CIS 522 Final Project Technical Report

November 12, 2022

**Team Members:**

- Xinyue Wang; wsinyue; Email: `wsinyue@seas.upenn.edu`
- Yifan Li; yfli; Email: `yfli@seas.upenn.edu`
- Yuxuan Wang; wangy49; Email: `wangy49@seas.upenn.edu`

**Abstract**

Semantic search is an essential topic in Natural Language Processing (NLP). Mainstream methods to conduct it are to adopt the same general large sentence encoder to acquire representations of queries and candidates and calculate their similarity. However, such methods are still suboptimal as they ignore the different characteristics of query and candidate corpus. To this end, we propose Query-BERT, a query embedding specific encoder to provide high-quality query embedding in a light model way. Inspired by the success of self-supervised learning (SSL) in computer vision, we leverage BYOL, a mature image contrastive learning approach without the need for negative samples, onto query embedding training to acquire a more robust query representation. Experiments show that under appropriate augmentation, our SSL method can reach competitive performance to the supervised method on Quora question duplicated pairs without big batch size and negative samples. However, the alignment between Query-BERT and candidates embedding is not satisfied and demands further optimization. We hope that this success of transfer learning could motivate people to rethink the appropriate effectiveness utilization of the light model and the role of domain-specific learning in the machine learning problem.

## 1 Introduction

Recently, semantic search has received increasing attention from academia and industry as neural-based Natural Language Processing (NLP) techniques advance. One crucial technology that realizes large-scale semantic document retrieval is Sentence Embedding, where we use a machine learning model to map texts to some metric vector spaces. We construct these vectors, i.e., the embeddings, so that a pair of documents are semantically similar if their corresponding

vector representations have a small distance, measured by a pre-defined metric, e.g., Euclidean Distance. Therefore, it is not hard to see that training the encoders that map from raw text to vectors is vital in the semantic search process.

Many robust sentence encoders have been proposed and proven effective in the past few years, such as InferSent, Universal Sentence Encoder, and Sentence-BERT [Con+17; Cer+18; RG19]. However, as their names suggest, these models are designed for domain-agnostic sentence encoding. They apply the same transformations (mapping) to all textual inputs regardless of their properties, e.g., length, topic, syntactic features, or downstream tasks. While these general settings are suitable and straightforward for most conditions, it is hard to acquire more accurate and robust sentence representations due to lacking specific domain consideration. For instance, query texts and answer documents are different in structure and content. Therefore, using the same encoder to encode them might not give us the optimal embeddings in terms of performance-efficiency trade-off. So, to build a better query search system, our work considers one particular subset of models corresponding to **Query Embedding** and **Ad-hoc Query Task.**

Two observations motivates our work. (1) Documents and queries are of different structures. For instance, documents are usually much longer than queries. Furthermore, documents are mostly declarative, where questions are generally interrogative. Intuitively, we need to encode the two families of texts differently. (2) Document embeddings are largely static after database indexing, but query embeddings are always calculated on the fly. Thus, we cannot use a powerful heavy universal embedder for every query given constrained computing power, which is the case on edge devices like Alexa Dot. However, since the questions have simpler structures, we assume that a more lenient model can still capture semantic information effectively.

To sum, we aim to find a better way for **Asymmetric Semantic Search:** different encoders for query and corpus. This work demonstrates that it is possible to train an extremely lightweight BERT encoder (as small as 16M) using a **Contrastive Representation Learning** method **without negative samples.** Further, our experiments show that we can align the query encodings to heavy-encoder-produced document encodings with a simple two-layer MLP projector. Lastly, we discuss the problems associated with our approach and potential directions for improvement.

## 2 Related Work

**Neural Matching Models** Information Retrieval (IR) researchers have recently introduced various BERT-based neural architectures for ranking. The broad spectrum of models broadly fall under two categories, *bi-encoder* versus *cross-encoder*. Information Retrieval (IR) researchers have recently introduced various BERT-based neural architectures for ranking. The broad spectrum of models broadly fall under two categories, *bi-encoder* versus *cross-encoder*. Cross-encoder models take query and document as inputs and produce a rel-

evance score for the query-document pair [Nog+19]. Though having slightly better IR performance, they suffer significantly from the high computational cost. Thus, pure cross-encoder architecture is rarely deployed in production. In contrast, bi-encoder models apply siamese networks or two separate encoders for query and document representation modeling, and then evaluate the similarities with a simple metric network or distance functions, e.g., cosine similarity [KZ20; Cho+21]. Therefore, it is possible to compute document representations offline for indexing. In previous studies, the query and document encoders are trained jointly to perform better. However, such a design brings an engineering disadvantage to the model update or iteration. For example, suppose we improve the architecture design of the document encoder; we need to re-train both encoders due to the coupling. So, in this work, we investigate the potential of elaborately training a query encoder and align it with an arbitrary large pre-trained document encoder with a simple two-layer MLP projector.

**Unsupervised Sentence Representation Learning** Unsupervised learning for sentence embedding is not a new idea; for instance, the Masked-Language-Modeling (MLM) task effectively learns meaningful representations [Dev+18]. However, we should notice that MLM is still a token-level modeling task instead of a sentence-level one. Moreover, recent research on sentence embedding also empirically shows the importance of learning sentence representations for various downstream tasks, and many such methods have been proposed, including TSDAE, SimCSE, CT [WRG21; GYC21; Car+21]. SimCSE and CT formulate the representation learning task as a contrastive learning task. We train the embedder to minimize the distance between anchor-positive sample pairs and maximize the distance between anchor-negative sample pairs. The two methods differ from each other in positive and negative sample generation. One caveat of contrastive learning is the definition of negative examples. i.e., whether two sentences are indeed semantically different can strongly affect final embedder performance.

Recent advancements in Computer Vision (CV) mainly inspire this work: negative samples may not be necessary for contrastive learning. Two notable training methods proposed are SimSiam [CH20] and BYOL [Gri+20], where researchers have empirically shown that predicting augmented samples alone can prevent the model from collapsing. By adopting this idea, BSL [Zha+21] investigates the possibility of learning sentence representation by way of BYOL. However, they still aim at a universal sentence encoder instead of a dedicated query encoder in our case.

**Textual Data Augmentation** Data augmentation is a vital component for contrastive learning without negative samples. In CV, augmentations can be performed simply by cropping, rotating, or flipping the images, and these approaches are almost guaranteed to be correct. However, since the input space is discrete, text data augmentation is significantly more difficult to perform. Removing a single word can completely revert the meaning of a sentence. Some

work has been done to try to address this issue and perform data augmentation for NLP tasks.

Back-translation has been widely adopted as a text augmentation method for various NLP tasks [Hay+18; SY19; Xie+20]. It refers to the procedure that, for a given text in language $A$, first translate it into another language $B$ and then back to $A$. The "back-translated" text will be used as an augmented sample.

Simple rule-based augmentations is another category of commonly used techniques [SKF21]. These methods relies on some simple rules only, without the need of an auxiliary neural network. Some example of rule-based augmentations are random swap, insertion, deletion, which, as indicated by the name, are just manipulating words at random. These approaches, however, are very likely to introduce grammatically incorrect or meaningless sentences. They are mostly used to add noise into the input to improve the robustness of the deep learning model. One exception is random synonym replacement, which are much more likely to generate synonymous sentences as an augmented sample.

Paraphrase generation [ZB21] is a widely researched natural language generation task that produces a paraphrase of the given text, which can be used as an augmented sample. This is a typical sequence-to-sequence problem, and many models including BART [Lew+19], Pegasus [Zha+20], and T5 [Raf+19] can be fine-tuned for paraphrase generation.

# 3 Dataset and Features

## 3.1 MS MARCO

MS MARCO [Baj+16] (MicroSoft MAchine Reading COmprehension) is a dataset designed for large-scale machine reading comprehension task such as Question Answering and Document Retrieval. The full dataset contains 1,010,916 anonymized user queries on the Bing search engine, each with a human generated answer, and 8,841,823 passages extracted from 3,563,535 web documents. The passages and documents provides information related to the queries.

In our practice, we utilized the queries and the passages data for the document retrieval task. The goal is to find the most relevant passage given a user query. The queries are also used as the source of augmentation to perform contrastive learning.

As can be expected from a search engine query log, the user queries cover a wide range of topics and are sometimes grammatically wrong or incomplete. Some examples of the queries are:

- what is a bank transit number
- how does a firefly light up
- work study average pay

The passages are paragraphs extracted from various web pages. An example of the passages are:

The Manhattan Project and its atomic bomb helped bring an end to World War II. Its legacy of peaceful uses of atomic energy continues to have an impact on history and science

### 3.2 Quora Question Pairs

Quora Question Paris is a dataset designed for identifying question pairs that have the same meaning. It contains 404,302 pairs of questions people raised on Quora, each labelled as whether they are duplicated or not by human.

Because of the question formatting requirements on Quora, the queries in this dataset are much more formal and grammatically correct. An example of the duplicated question pair is:

- How can I be a good geologist?
- What should I do to be a great geologist?

An example of the non-duplicated question pair is:

- What is the step by step guide to invest in share market in India?
- What is the step by step guide to invest in share market?

We mainly utilized this dataset as a source for natural language queries to perform query augmentation for contrastive learning. Moreover, we tested the model's ability to distinguish duplicated question pairs as an additional dimension to evaluate model performance.

## 4 Methodology

### 4.1 Query Augmentation

The goal of query augmentation is to generate a synonymous query, or equivalently, a paraphrase for each given query. To achieve this, we tested the following augmentation strategies:

**contextual** Randomly mask words from the query and input the surrounding words to a contextual language model, such as BERT, and use the predicted masked word as augmented query

**wordnet** Randomly replace words with their synonyms in the WordNet [Mil95]

**bart** Use the BART [Lew+19] model fine-tuned for paraphrasing

**t5** Use the T5 [Raf+19] model fine-tuned for paraphrasing

**pegasus** Use the Pegasus [Zha+20] model fine-tuned for paraphrasing [1]

[1] BART, T5, and Pegasus are all general purpose sequence-to-sequence models. We used their fine-tuned versions published on hugging face directly. They are not further tuned by us due to the limitation of our computation resources and dataset availability.

**google** Use the Google Translation API to translate the query to German and back to English

**facebook** Use Facebook FAIR's WMT19 News Translation Task solution model [Ng+19] to translate the query to German and back to English

**paa** Get similar queries by Google People Also Ask API, and use sentence bert models to get the one with the highest similarity

**shuffle** Split the tokens and shuffle them randomly

**styleformer** Use the Styleformer to perform style transfer on the input query between formal and informal

The following is the augmentation result of an example query, "what is a bank transit number":

| Strategy | Result |
|---|---|
| contextual | what not my bank transit number |
| wordnet | what is a bank theodolite number |
| bart | what is bank transit number? |
| t5 | what is a bank transaction number? |
| pegasus | what is the transit number for a bank? |
| google | what is a bank code |
| facebook | what is a bank code? |
| paa | what is a transit number in Canada? |
| shuffle | is number transit bank a what |
| styleformer | what is a bank transit number? |

Table 1: Augmentation Strategies and Results

We selected the strategies by reviewing the augmented queries manually, instead of evaluating them with some metrics. We made this choice mainly because the common evaluation metrics for paraphrase generation, such as BLEU and ROUGE, require a reference as ground truth, which is not available in our case.

## 4.2 Architecture

Our final query embeddings training architecture (Figure 1) is a modified version of BSL framework, which is shown to be more robust than SimSiam on NLP task by ablation study on Quora question duplicated pairs dataset. The architecture takes four augmented view of a query as input, processed them through two Bert encoders while one of which keeps gradient update off during training. Note that the updated one is final query embeddings trained and adopted, and the other encoder is only participated as assistance to maintain the whole contrastive learning process. On the updated side, after pooling by the Bert encoder, a

attention based projector is utilized to better filter augmentation noise and help attain important features in the augmented view. In addition, there is a MLP based predictor followed by projector to match the output of one view to the other view, and centralize the query embeddings in the upstream encoder. Note that bewtween the updated encoder and untrained encoder, there is a exponential moving average module to stabilize the difference between encoders and prevent collapsing happening in the contrastive learning without negative samples.

In image contrastive learning, heavy augmentation is common and prone to disrupt representation learning, which is frequently happened when adopting natural language augmentation methods like shuffle, style rewritten. Considering that heavy augmentation like shuffle and style rewritten are common in query search scenario, to alleviate this effect and fully unleash the role of augmentation in query representation learning, we transfer the directional contrastive loss from image representation learning task to our query embeddings learning inspired by [Bai+21]. Here we adopt in total four augmentations including two light augmentations (Pegasus back translation, Facebook back translation) and two heavy augmentations (Word-level shuffle, Styleformer rewritten) as our training input. Denoting that the output of projector are $z_1$, $z_2$, $z_3$, $z_4$, and the output of predictor are $p_1$, $p_2$, $p_3$, $p_4$, respectively referring to the output of Pegasus, Facebook, Shuffle, Styleformer augmentation mentioned above. In the training process, our main goal is to minimize the negative cosine similarity:

$$D(p, z) = -\frac{p}{\|p\|} \cdot \frac{stop\_grad(z)}{\|stop\_grad(z)\|} \tag{1}$$

$$Loss = \frac{1}{4}D(p_1, z_2) + \frac{1}{4}D(p_2, z_1) + \frac{1}{4}D(p_3, z_1) + \frac{1}{4}D(p_4, z_2) \tag{2}$$

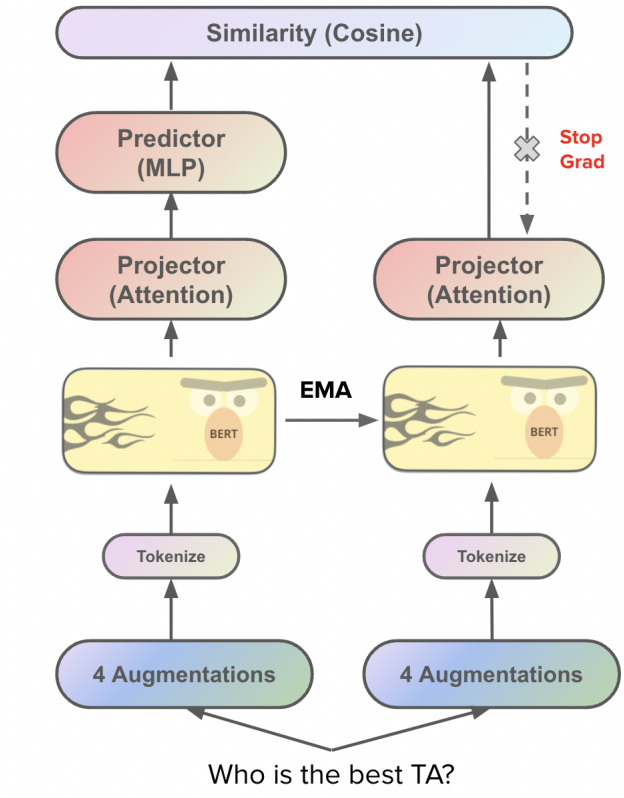


Figure 1: Query embeddings training framework

### 4.3 Training Details

In practice, we did ablation study and finalize the optimal baseline setting for query embedding learning as follows:

- *Encoder*. To fulfill our target of building a both light and effective query encoder, we adopt ALBERT-Small-V2[Lan+19] as base encoder due to its rich capacity (model width and depth) and superior distill quality (general performance on benchmark) compared to other light language model.
- *Projector*. The projector head is consisted of four self-attention layer with batch normalization, which has much less parameters than MLP based projector.
- *Predictor*. The predictor is a simple three layer MLP with batch normalization applied.
- *Optimization*. Benefit from contrastive learning without negative samples, large batch size training would not be needed in our development. We

utilize AdamW as optimizer and set 1e-2 as weight decay to prevent overfitting and alleviate forgetting effect, and conduct 3 epochs training with 64 and 1e-4 as batch size and learning rate. In addition, we introduced the early-stopping, batch normalization and linear scheduler techniques to further prevent overfitting and achieve the best generalization.

### 4.4 Baseline Model

We use averaging Word2Vec [Mik+13] token embeddings as the baseline query embedding. For the non-DL baseline, we use a linear projector to align the query embedding with document embedding, which is produced by pre-trained Sentence-BERT model. For the simple-DL baseline, we instead use a four-layer similar to DAN [Iyy+15] as the projector. We evaluate both models on the MS MARCO hard-negative dataset provided by Sentence-BERT team [2].

## 5 Results

Start with a sentence or two on what you want to show. Describe how you showed that and what you learned. This section should include your results, i.e. the performance of your models with regards to your chosen performance metric, as well as tables/visualizations of these results, the training process, confusion matrices for classification projects... (whatever works for your project, make a sensible choice here!). Please report the loss function that you've minimized and additional measures of performance/quality you looked at, too.

### 5.1 Quora Question Duplicated Pairs

In this experiment, our experiments can be divided into two parts, pseudo self-supervised learning and self-supervised learning.

In the first part, our main goal is to evaluate different settings and finalize the best architecture of query embedding learning based on pseudo self-supervised learning, as shown in the Table 2. A main point here we adopt this data set is that all the duplicated pairs are golden positive pairs and our main goal is to develop the final architecture used to do contrastive learning on large augmentation data set. Doing empirical study on it allows us not care the the error caused by mistaken augmentation. Here we selected 80% of positive pairs in the whole data set as training set, and sampled in total one-fourth of training set size as test set, which is mixed with positive samples and negative samples in a 1:1 ratio. However, it is also a pseudo supervised based since all the training pairs are ground truth positive samples. Therefore the encoders were trained and tested on the 5 fold division of Quora question duplicated pairs data set.

In the framework comparison, BSL demonstrated better learning ability than SimSiam. Even the smallest encoder BSL trained could outperform a lot than well-tuned ALBERT trained by SimSiam, which could be induced to richer

[2] https://huggingface.co/datasets/sentence-transformers/msmarco-hard-negatives

learning capacity brought by bi-encoder structure. When substituting MLP projector with attention projector proved, both encoders trained on BSL and SimSiam were improved, which indicated that attention projector can more effectively filter augmentation noise and exact important features from encoding. In addition, it appeared that exponential moving average can better prevent collapsing and enhance the stability and performance of encoder. Our experiments show that transfering these computer vision contrastive learning to natural language processing task is feasible and successful. Without negative samples participated, our query encoder did not collapse and exhibited high quality query representation.

| Framework | Encoder | Projector | EMA | F1 | Encoder Size |
|---|---|---|---|---|---|
| SimSiam | BERT-Tiny | MLP | | 0.608 | 12MB |
| SimSiam | ALBERT | MLP | | 0.614 | 43MB |
| SimSiam | ALBERT | Attention | | 0.623 | 43MB |
| BSL | BERT-Tiny | MLP | | 0.668 | 12MB |
| BSL | ALBERT | MLP | | 0.694 | 43MB |
| BSL | ALBERT | Attention | | 0.704 | 43MB |
| BSL (Ours) | ALBERT | Attention | ✓ | 0.712 | 43MB |

Table 2: Experiments on Quora question dulicated pairs. *AP: Attention projector instead of MLP projector

In the second part, we trained query encoders on the augmentation data of MS MARCO Passage Ranking data set to fully unleash the potential of contrastive learning without negative sampels, and tested them on the same Quora test set as the first part, whose result can be found in Table 3. With all augmented data, our finalized architecture in the first part with directional contrastive loss outperformed selected light augmentation and demonstrated competitive representation quality to the one trained by supervised mode in part one. In addition, increasing model size is shown to be a promising way to acquire better query embedding. Note that straightly using all augmentations led to worse performance than selected light augmentations, while it performed better when adopting directional self-supervised learning. Based on that, we argue that appropriately utilizing all kinds of augmentations in self-supervised learning can boost model performance than only using light augmentations.

| Augmentation | Loss | Encoder | F1 | Encoder Size |
|---|---|---|---|---|
| Pegasus&Facebook | Regular | BERT-Tiny | 0.606 | 12MB |
| Pegasus&Facebook | Regular | ALBERT | 0.648 | 43MB |
| All | Regular | ALBERT | 0.635 | 43MB |
| All | DSSL | ALBERT | 0.662 | 43MB |

Table 3: Models trained with MS MARCO queries and evaluated on Quora question dulicated pairs.

## 5.2 MS MARCO Passage Ranking

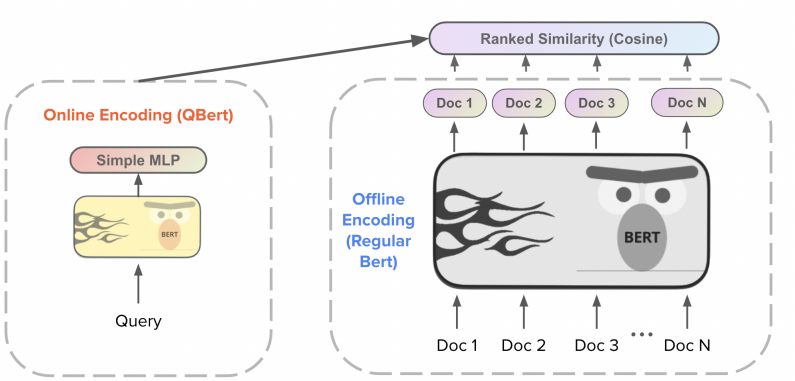


Figure 2: Model Inference Workflow on MS MARCO

During inference time, we first pre-compute all document embeddings with large pre-trained Sentence-BERT [3]. Next, we use the QueryBERT and a two-layer MLP to produce query embeddings. Then the rankings of passages are determined by the cosine similarity between query-document embeddings. Instead of evaluating the document retrieval performance, we calculate the average anchor-positive-negative margins to give a more direct quantitative evaluation of the model ranking capabilities. The margin is defined as,

$$margin = d(\text{query, document relevant}) - d(\text{query, document irrelevant})$$

where $d$ denotes the cosine similarity. We randomly sample $2^{15}$ anchor-positive-negative triples to perform the evaluation, and the result is shown below. A larger margin indicates better ranking capability. We see that the performance is positively correlated with model size. But, although the ALBERT-based query encoder only achieves half of the margin results from the original jointly-trained Sentence-BERT model, the model size is less than one-fifth.

[3] https://www.sbert.net/docs/pretrained-models/msmarco-v3.html

| Encoder | Margin | Encoder Size |
|---|---|---|
| Word2Vec-Linear | 0.062 | - |
| Word2Vec-MLP | 0.080 | - |
| BERT-Tiny | 0.109 | 12MB |
| ALBERT | 0.136 | 43MB |
| DistilBERT (Ori.) | 0.234 | 290MB |

Table 4: Anchor-positive-negative distance margin evaluated over MS MARCO hard-negative dataset.

# 6 Discussion

Our experiment demonstrates that we can adopt SimSiam and BYOL from CV to the NLP domain for representation learning. And it is possible to align the query encoder to a large pre-trained model using simple projection layers. At the same time, we identify the challenges of applying these methods, particularly concerning data augmentation, which can be the potential cause of inferior performance.

## 6.1 Findings

Query embedding should be different from documents embedding, which leads us to the motivation of this project, as well as the representation of the results. On the Quora question duplicated pairs test set, our positive-sample-only contrastive learning method demonstrates its superiority and robustness, which can reach competitive performance to pseudo supervised learning. However, when aligned with documents embedding, the performance is not satisfied and robust as it on the query pair test. What indicates here is that our query encoder is so good at learning query embedding that it is too domain specific to be harmoniously projected into the document representation space. Here we suggest that carefully think and develop the use of bi-encoder structure system whether in NLP or other fields, since reaching expectation respectively does not mean they could work well as a unity.

Having a much smaller query encoder that can be tuned with unsupervised learning can be a great advantage to user privacy. We can deploy the small model on the user devices and update parameters with Federated Learning [Kai+19], which is a distributed machine learning paradigm that mathematically ensures user privacy.

## 6.2 Limitations and Ethical Considerations

Though we confirm the effectiveness of BSL in preventing the model from collapsing, the model inference performance is still far from the current SOTA. Moreover, our training method heavily relies on the data augmentation strategy. In our perspective, back-translation results in the highest-quality aug-

mentations. However, since back-translation depends on statistical language modeling, it is hard to generalize our training strategy to languages with fewer resources due to the questionable translation quality. Further, the query encoders will inherit the bias from the translators we used. Thus, the deployment of our query encoder may widen the social gap.

### 6.3 Future Research Directions

Due to time and computing power constraints, we cannot evaluate the method on a more extensive set of base models. In this work, we only fine-tuned ALBERT [Lan+19] and BERT-tiny [Tur+19]. However, since the research community has recently proposed more advanced architectures, we believe it is necessary to explore the compatibility between our proposed method and these new architectures. Further, we need to experiment with more comprehensive combinations of hyper-parameters. The current setup may result in under-fitted models. Lastly, data is at the center of concern. On the one hand, we can further improve data augmentation and negative-sampling strategies. Unlike data augmentation in CV where inputs are more continuous (pixels are more fine-grained units), adding or subtracting word tokens can drastically change the meaning of sentences. On the other hand, we may train the models on more datasets to evaluate the effectiveness of our methodology.

## 7 Conclusions

This report introduces QueryBERT, an experimental query encoder designed for information retrieval and trained with contrastive representation learning methods inspired by CV research. Experiment results confirm that BSL and SimSiam training methods can prevent models from collapsing in the absence of negative samples. Further, we show that aligning the contrastively trained query encoder with a pre-trained document encoder by a simple two-layer MLP projector is possible. Meanwhile, we notice the performance gap between our model and supervised-trained models on the Quora duplicate-question-detection dataset and between jointly trained query-document-encoders. Further ablation study and improvement are still required.

## 8 References

## References

[Bai+21] Yalong Bai et al. "Directional Self-supervised Learning for Risky Image Augmentations". In: *CoRR* abs/2110.13555 (2021). arXiv: 2110.13555. URL: https://arxiv.org/abs/2110.13555.

[Baj+16] Payal Bajaj et al. "Ms marco: A human generated machine reading comprehension dataset". In: *arXiv preprint arXiv:1611.09268* (2016).

[Car+21] Fredrik Carlsson et al. "Semantic Re-tuning with Contrastive Tension". In: *ICLR*. 2021.

[Cer+18] Daniel Cer et al. "Universal Sentence Encoder". In: *CoRR* abs/1803.11175 (2018). arXiv: 1803.11175. URL: http://arxiv.org/abs/1803.11175.

[CH20] Xinlei Chen and Kaiming He. "Exploring Simple Siamese Representation Learning". In: *CoRR* abs/2011.10566 (2020). arXiv: 2011.10566. URL: https://arxiv.org/abs/2011.10566.

[Cho+21] Jaekeol Choi et al. "Improving Bi-encoder Document Ranking Models with Two Rankers and Multi-teacher Distillation". In: *CoRR* abs/2103.06523 (2021). arXiv: 2103.06523. URL: https://arxiv.org/abs/2103.06523.

[Con+17] Alexis Conneau et al. "Supervised Learning of Universal Sentence Representations from Natural Language Inference Data". In: *CoRR* abs/1705.02364 (2017). arXiv: 1705.02364. URL: http://arxiv.org/abs/1705.02364.

[Dev+18] Jacob Devlin et al. "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding". In: *CoRR* abs/1810.04805 (2018). arXiv: 1810.04805. URL: http://arxiv.org/abs/1810.04805.

[GYC21] Tianyu Gao, Xingcheng Yao, and Danqi Chen. "SimCSE: Simple Contrastive Learning of Sentence Embeddings". In: *CoRR* abs/2104.08821 (2021). arXiv: 2104.08821. URL: https://arxiv.org/abs/2104.08821.

[Gri+20] Jean-Bastien Grill et al. "Bootstrap Your Own Latent: A New Approach to Self-Supervised Learning". In: *CoRR* abs/2006.07733 (2020). arXiv: 2006.07733. URL: https://arxiv.org/abs/2006.07733.

[Hay+18] Tomoki Hayashi et al. "Back-Translation-Style Data Augmentation for end-to-end ASR". In: *2018 IEEE Spoken Language Technology Workshop (SLT)* (2018), pp. 426–433.

[Iyy+15] Mohit Iyyer et al. "Deep Unordered Composition Rivals Syntactic Methods for Text Classification". In: *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. Beijing, China: Association for Computational Linguistics, July 2015, pp. 1681–1691. DOI: 10.3115/v1/P15-1162. URL: https://aclanthology.org/P15-1162.

[Kai+19] Peter Kairouz et al. "Advances and Open Problems in Federated Learning". In: *CoRR* abs/1912.04977 (2019). arXiv: 1912.04977. URL: http://arxiv.org/abs/1912.04977.

[KZ20] Omar Khattab and Matei Zaharia. "ColBERT: Efficient and Effective Passage Search via Contextualized Late Interaction over BERT". In: *CoRR* abs/2004.12832 (2020). arXiv: 2004.12832. URL: https://arxiv.org/abs/2004.12832.

[Lan+19] Zhenzhong Lan et al. *ALBERT: A Lite BERT for Self-supervised Learning of Language Representations*. 2019. DOI: 10.48550/ARXIV.1909.11942. URL: https://arxiv.org/abs/1909.11942.

[Lew+19] Mike Lewis et al. "Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension". In: *arXiv preprint arXiv:1910.13461* (2019).

[Mik+13] Tomás Mikolov et al. "Distributed Representations of Words and Phrases and their Compositionality". In: *CoRR* abs/1310.4546 (2013). arXiv: 1310.4546. URL: http://arxiv.org/abs/1310.4546.

[Mil95] George A Miller. "WordNet: a lexical database for English". In: *Communications of the ACM* 38.11 (1995), pp. 39–41.

[Ng+19] Nathan Ng et al. "Facebook FAIR's WMT19 news translation task submission". In: *arXiv preprint arXiv:1907.06616* (2019).

[Nog+19] Rodrigo Nogueira et al. "Multi-Stage Document Ranking with BERT". In: *CoRR* abs/1910.14424 (2019). arXiv: 1910.14424. URL: http://arxiv.org/abs/1910.14424.

[Raf+19] Colin Raffel et al. "Exploring the limits of transfer learning with a unified text-to-text transformer". In: *arXiv preprint arXiv:1910.10683* (2019).

[RG19] Nils Reimers and Iryna Gurevych. "Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks". In: *CoRR* abs/1908.10084 (2019). arXiv: 1908.10084. URL: http://arxiv.org/abs/1908.10084.

[SKF21] Connor Shorten, Taghi M Khoshgoftaar, and Borko Furht. "Text data augmentation for deep learning". In: *Journal of big Data* 8.1 (2021), pp. 1–34.

[SY19] Amane Sugiyama and Naoki Yoshinaga. "Data augmentation using back-translation for context-aware neural machine translation". In: *EMNLP*. 2019.

[Tur+19] Iulia Turc et al. "Well-Read Students Learn Better: The Impact of Student Initialization on Knowledge Distillation". In: *CoRR* abs/1908.08962 (2019). arXiv: 1908.08962. URL: http://arxiv.org/abs/1908.08962.

[WRG21] Kexin Wang, Nils Reimers, and Iryna Gurevych. "TSDAE: Using Transformer-based Sequential Denoising Auto-Encoder for Unsupervised Sentence Embedding Learning". In: *CoRR* abs/2104.06979 (2021). arXiv: 2104.06979. URL: https://arxiv.org/abs/2104.06979.

[Xie+20] Qizhe Xie et al. “Unsupervised Data Augmentation for Consistency Training”. In: *arXiv: Learning* (2020).

[Zha+20] Jingqing Zhang et al. “Pegasus: Pre-training with extracted gap-sentences for abstractive summarization”. In: *International Conference on Machine Learning*. PMLR. 2020, pp. 11328–11339.

[Zha+21] Yan Zhang et al. “Bootstrapped Unsupervised Sentence Representation Learning”. In: *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. Online: Association for Computational Linguistics, Aug. 2021, pp. 5168–5180. DOI: 10.18653/v1/2021.acl-long.402. URL: https://aclanthology.org/2021.acl-long.402.

[ZB21] Jianing Zhou and S. Bhat. “Paraphrase Generation: A Survey of the State of the Art”. In: *EMNLP*. 2021.