Momentum Contrastive Pre-training for Question Answering

Minda Hu[1] **Muzhi Li**[1] **Yasheng Wang**[2] **Irwin King**[1]
[1]Dept. of Computer Science & Engineering, The Chinese University of Hong Kong
[2]Huawei Noah's Ark Lab
{mindahu21, mzli, king}@cse.cuhk.edu.hk,
wangyasheng@huawei.com

Abstract

Existing pre-training methods for extractive Question Answering (QA) generate cloze-like queries different from natural questions in syntax structure, which could overfit pre-trained models to simple keyword matching. In order to address this problem, we propose a novel **M**omentum **C**ontrastive p**R**e-training f**O**r que**S**tion an**S**wering (MCROSS) method for extractive QA. Specifically, MCROSS introduces a momentum contrastive learning framework to align the answer probability between cloze-like and natural query-passage sample pairs. Hence, the pre-trained models can better transfer the knowledge learned in cloze-like samples to answering natural questions. Experimental results on three benchmarking QA datasets show that our method achieves noticeable improvement compared with all baselines in both supervised and zero-shot scenarios.

1 Introduction

The task of extractive Question Answering (QA), which aims to select an answer span from a passage given a query, is a major focus of NLP research. Currently, deep learning systems (Huang et al., 2018; Devlin et al., 2019; Wu et al., 2021) have achieved competitive results with humans on large-scale QA datasets (Rajpurkar et al., 2016, 2018; Yang et al., 2018; Kwiatkowski et al., 2019). However, the collection of high-quality natural QA pairs is still a labor-intensive task, especially for the construction of domain-specific QA systems. To alleviate such data availability restrictions, pre-training methods have been drawing increasing attention (Dhingra et al., 2018; Ram et al., 2021).

Typically, pre-training methods for extractive QA generate cloze-like query-passage pairs with text-matching techniques. For instance, the Span Selection Pre-Training (SSPT) method (Glass et al., 2020a) generates these pairs with the Wikipedia corpus using BM25. Nevertheless, these methods have two major issues. Firstly, the format of cloze-like queries differs much from natural queries asked by humans (see Fig. 1). In addition, models trained by cloze-like queries overfit in capturing the lexical overlaps between queries and passages, which restricts the capability of high-level semantic reasoning (Li et al., 2020; Hu et al., 2021). Since answering natural questions requires a comprehensive understanding of queries and passages, models pre-trained with cloze-like queries are not well aligned with downstream QA tasks.

Cloze-like Query	In Jordan's second season of the playoffs against the **[BLANK]**, he even scored 63 points
Natural Query	What team did Jordan score 63 points against in the second season of the playoffs?
Answer Term	Boston Celtics

Figure 1: An example of the pre-training queries in SSPT (Glass et al., 2020b). The query is a sentence where a noun phrase or entity ("Boston Celtics") is replaced with the special token [BLANK], which differs syntactically from the natural one.

To solve these issues, we utilize Contrastive Learning (CL) techniques to align knowledge learned in cloze-like samples to answering natural language questions and circumvent overfitting. Specifically, CL aims to learn representations by contrasting augmentations of different input data, which has been successfully applied in various research fields such as bioinformatics, graph learning, computer vision, and NLP (Han et al., 2022; Zhang et al., 2022; He et al., 2020; Yang et al., 2021; Pan et al., 2021). Nevertheless, existing CL methods for QA focus on query-passage matching (Yang et al., 2021; Caciularu et al., 2022) and multilingual embedding alignment (Pan et al., 2021), which have not been tailored to deal with issues of query format inconsistency and overfitting in matching lexical overlaps.

In this paper, we propose a *Momentum Contrastive pRe-training fOr queStion AnSwer-*

Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, pages 4324–4330
December 7-11, 2022

ing (MCROSS) method for extractive QA. Specifically, MCROSS employs a momentum contrastive learning strategy along with the conventional answer prediction task to maximize the consistency of predicted answer distributions between cloze-like and natural query pairs and thus improves the performance of pre-trained models in answering natural language questions. We show the efficacy of our approach on standard English benchmark datasets. On SQuADv1.1 (Rajpurkar et al., 2016), MCROSS achieves about 2.7/3.5 percentage points gain on F1/Exact Match (EM) accuracy over BERT, and 1.4/1.8 percentage points improvement on the same metrics compared to SSPT (Glass et al., 2020a). MCROSS also consistently outperforms the baseline methods by a large margin on TriviaQA (Joshi et al., 2017) and NewsQA (Trischler et al., 2017) in supervised and zero-shot scenarios.

2 Method

The structure of our proposed MCROSS method is shown in Fig. 2. Along with the cloze-like samples $s^{\text{cloze}} = (q_{\text{c}}, p, a)$ generated by SSPT (Glass et al., 2020a), we create positive natural samples $s^{\text{natural+}} = (q_{\text{n}}, p, a)$ containing natural queries q_{n} given passages and answers (p, a) from s^{cloze} with T5 answer-aware question generator (Raffel et al., 2020) which is fine-tuned with SQuADv1.1 training set. The details of pre-training datasets are illustrated in Appendix A.1. Though different in format, the cloze-like queries q_c and positive natural queries q_n are semantically similar since they have the same answers a in given supporting passages p. Therefore, we expect that the predicted probability distribution of answer span in positive pairs $(s^{\text{cloze}}, s^{\text{natural+}})$ should be closer. On the other hand, we generate negative query pairs $(s^{\text{cloze}}, s^{\text{natural}-})$ with different answers or supporting passages. Correspondingly, the predicted answer span distribution of $s^{\text{natural}-}$ is expected to be far from that of s^{cloze}. MCROSS achieves the goal with two pre-training tasks as follows.

2.1 Pre-Training Tasks

Answer Term Prediction Task. Given a training sample $s = (q, p, a)$, the answer term prediction task aims to predict the token span $a = (st_a, en_a)$ in the query q from the passage p[1], where st_a and en_a are the start and end positions of the answer term in the passage.

[1]We treat the first exact match of the token span in the passage as the answer term.

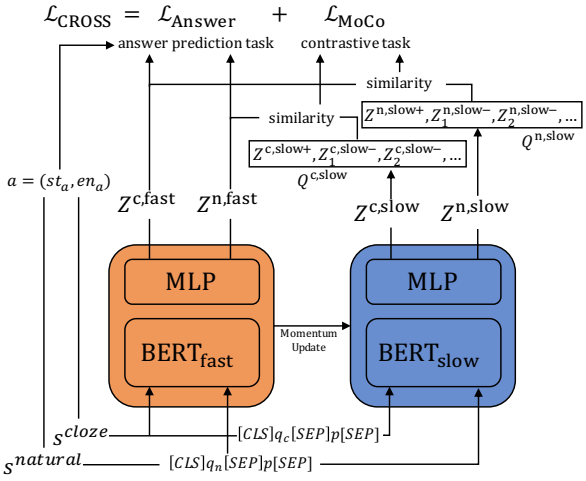

Figure 2: An overview of the proposed MCROSS method.

First, we prepare the concatenated token sequence $\mathrm{T} = [CLS]\, q\, [SEP]\, p\, [SEP]$ and leverage BERT (Devlin et al., 2019) to encode the context embeddings $\boldsymbol{H} \in \mathbb{R}^{[\text{len}(q)+\text{len}(p)+3]\times d}$ for each query-passage pair, where d is the dimension of embeddings. Then, the probability distribution $Z = (z^{st_a}, z^{en_a})$ of the start/end indices of answer span can be predicted as:

$$\begin{aligned} z_i^{st_a} &= p(i = st_a) = \text{softmax}(L_{\text{span}}^{\text{start}}(\boldsymbol{H}_i)), \\ z_i^{en_a} &= p(i = en_a) = \text{softmax}(L_{\text{span}}^{\text{end}}(\boldsymbol{H}_i)), \end{aligned} \tag{1}$$

where $L_{\text{span}}^{\text{start}}$ and $L_{\text{span}}^{\text{end}}$ are linear answer span prediction layers. The loss function of the answer term prediction task is defined below:

$$\begin{aligned} \mathcal{L}_{\text{Answer}}(Z) = -\frac{1}{2}\{ &\sum_{1\le i\le N} \mathbb{1}(i = st_a) \log z_i^{st_a} \\ &+ \sum_{1\le i\le N} \mathbb{1}(i = en_a) \log z_i^{en_a}\}, \end{aligned} \tag{2}$$

where $\mathbb{1}(\text{condition})$ is the indicator function that returns 1 if the condition is satisfied and 0 otherwise. N is the length of the sequence T.

Contrastive Learning Task. This task utilizes a contrastive loss function to guide the answer span distributions $(Z^{\text{cloze}}, Z^{\text{natural+}})$ of positive sample pairs $(s^{\text{cloze}}, s^{\text{natural+}})$ with same (p, a) to be closer, while keeping Z^{cloze} to be dissimilar from $Z^{\text{natural}-}$ of $s^{\text{natural}-}$ with different passages p^- or answers a^-.

Inspired by the **Mo**mentum **Co**nstrastive learning (MoCo) (He et al., 2020) strategy, we maintain a large pool of consistent negative samples

from several previous batch iterations to preserve more information on negative samples and obtain better pre-training objectives. Following MoCo, we employ a dual-encoder architecture $(\text{BERT}_{\text{fast}}, \text{BERT}_{\text{slow}})$ and maintain FIFO queues $(Q^{\text{slow}}_{st_a}, Q^{\text{slow}}_{en_a})$ containing a large set of answer span distributions $(z^{\text{slow}-}_{st_a}, z^{\text{slow}-}_{en_a})$ of negative samples predicted by $\text{BERT}_{\text{slow}}$. To maintain queue consistency, for each batch iteration, parameters θ_{slow} of $\text{BERT}_{\text{slow}}$ are only updated by the exponential moving average of θ_{fast} of $\text{BERT}_{\text{fast}}$. Here we have

$$\theta_{\text{slow}} = m\theta_{\text{slow}} + (1-m)\theta_{\text{fast}}, \tag{3}$$

where m is the momentum coefficient. We regulate the output of the answer-span classification layer with InfoNCE loss to bring positive pairs (s, s^+) closer to each other and push negative ones apart:

$$\mathcal{L}_{\text{MoCo}}(z^{\text{fast}}, z^{\text{slow}+}, Q) = -\log \frac{\exp\left(\mathcal{D}(z^{\text{fast}}, z^{\text{slow}+})/\tau\right)}{\sum_{z^{\text{slow}-}\in Q}\exp\left(\mathcal{D}(z^{\text{fast}}, z^{\text{slow}-})/\tau\right)}, \tag{4}$$

where z^{fast} and $z^{\text{slow}+}$ represent the start or end distributions of positive pairs, τ is the softmax temperature, and Q denotes queues containing distributions $z^{\text{slow}-}$ of negative samples s^- which have (p^-, a^-) being different from s^+. $\mathcal{D}$ is the similarity function measuring distances between answer span distributions. Cosine similarity is utilized in the original MoCo method (He et al., 2020). However, we argue that KL-Divergence is more suitable for the measurement of differences between two probability distributions, which is evaluated and validated in Appendix C.

2.2 Variants of the MCROSS Method

There are two variants of the MCROSS method with different loss functions, the first is involved with the unilateral loss function, and the second one with the bilateral loss.

MCROSS (UNI). In addition to the original SSPT training loss $\mathcal{L}_{\text{Answer}}(Z^{\text{c}})$, we add $\mathcal{L}_{\text{MoCo}}$ to maximize the consistency of prediction between s^{cloze} and s^{natural} with little overhead in the unilateral version of MCROSS. Here, we have the unilateral MoCo loss

$$\begin{aligned}\mathcal{L}_{\text{UNI}}(s^{\text{cloze}}, s^{\text{natural}}) &= \mathcal{L}_{\text{Answer}}(Z^{\text{c}}) \\ &+ \lambda_{\text{MoCo}}/2 * [\mathcal{L}_{\text{MoCo}}(z^{st_a,\text{c}}, z^{st_a,\text{n}}, Q^{\text{n,slow}}_{st_a}) \\ &+ \mathcal{L}_{\text{MoCo}}(z^{en_a,\text{c}}, z^{en_a,\text{n}}, Q^{\text{n,slow}}_{en_a})],\end{aligned} \tag{5}$$

Method	F1	EM	0-shot F1	0-shot EM
BERT	87.43	79.50	8.00	0.06
SSPT	88.75	81.25	22.88	15.42
SSPT†	89.25	82.13	64.41	44.44
MCROSS(UNI)	88.95	81.60	24.78	16.58
MCROSS(BI)	**90.11**	**83.03**	**65.68**	**45.40**
w/o $\mathcal{L}_{\text{MoCo}}$	89.92	82.96	65.20	44.82

Table 1: Results on SQuADv1.1 dataset. (in domain)

where $Z^{\text{n}} = (z^{st_a,\text{n}}, z^{en_a,\text{n}})$ are the predicted distributions of start/end indices for s^{natural}, and $Z^{\text{c}} = (z^{st_a,\text{c}}, z^{en_a,\text{c}})$ are for s^{cloze}. $(Z^{\text{c}}, Z^{\text{n}})$ are predicted by $\text{BERT}_{\text{fast}}$ and , $\text{BERT}_{\text{slow}}$, respectively. Z^{n} are stored in $(Q^{\text{n,slow}}_{st_a}, Q^{\text{n,slow}}_{en_a})$ and will be reused in later training steps as negative examples. λ_{MoCo} denotes the ratio of MoCo loss.

MCROSS (BI). In MCROSS (UNI) method, only cloze-like samples s^{cloze} are considered in the calculation of answer term prediction loss $\mathcal{L}_{\text{Answer}}$, which makes the fast encoder $\text{BERT}_{\text{fast}}$ to be less sensitive to natural samples s^{natural}. Therefore, it will be difficult to properly maintain the consistency between the probability distributions generated from the two types of samples. To tackle this issue, following xMoCo (Yang et al., 2021), we leverage two $\mathcal{L}_{\text{UNI}}$ losses and jointly optimize the alignment of $(s^{\text{cloze}}, s^{\text{natural}})$ and $(s^{\text{natural}}, s^{\text{cloze}})$ in the following bilateral loss.

$$\begin{aligned}\mathcal{L}_{\text{BI}} &= \mathcal{L}_{\text{UNI}}(s^{\text{cloze}}, s^{\text{natural}}) \\ &+ \mathcal{L}_{\text{UNI}}(s^{\text{natural}}, s^{\text{cloze}}),\end{aligned} \tag{6}$$

where distributions of s^{natural} are obtained from $\text{BERT}_{\text{slow}}$ and stored in queue $(Q^{\text{n,slow}}_{st_a}, Q^{\text{n,slow}}_{en_a})$ for $\mathcal{L}_{\text{UNI}}(s^{\text{cloze}}, s^{\text{natural}})$. The same practices are applied to s^{cloze} for $\mathcal{L}_{\text{UNI}}(s^{\text{natural}}, s^{\text{cloze}})$.

We also evaluate MCROSS (w/o $\mathcal{L}_{\text{MoCo}}$), which utilizes multi-task answer prediction loss to predict similar span a given (q_c, p) and (q_n, p) with $\text{BERT}_{\text{fast}}$:

$$\mathcal{L}_{\text{MCROSS}} = \mathcal{L}_{\text{Answer}}(Z^{\text{n}}) + \mathcal{L}_{\text{Answer}}(Z^{\text{c}}). \tag{7}$$

3 Experiments

We use the following three English span-extraction QA datasets to evaluate pre-trained models on F1/EM metrics in both supervised and zero-shot scenarios. Specifically, TriviaQA (Joshi et al., 2017) and NewsQA (Trischler et al., 2017) are out-of-domain datasets from MRQA 2019 Shared

Task (Fisch et al., 2019). Since SQuADv1.1 (Rajpurkar et al., 2016) dataset is utilized to fine-tune the T5 question generator, it is also included to examine the in-domain QA performance.

The experimental settings, implementation details, and complexity analysis are presented in Appendix A, B, and D.

3.1 Baselines

We compare the MCROSS with the following four baselines:

- BERT: The 12-layer BERT model released by (Devlin et al., 2019).
- MRQA-BERT: The official multi-task baseline BERT-base model from MRQA[2].
- SSPT: The span selection pre-training method proposed by (Glass et al., 2020b). This method trains models with cloze-like samples s^{cloze} using answer term prediction loss $\mathcal{L}_{\text{Answer}}(Z^{\text{c}})$ in Eq. (7).
- SSPT†: SSPT method trained with only natural samples s^{natural} using answer term prediction loss $\mathcal{L}_{\text{Answer}}(Z^{\text{n}})$ in Eq. (7).

3.2 Experiment Results

SQuADv1.1. Table 1 shows the performance of all models on SQuADv1.1. Compared with the BERT baseline without extended pre-training, three variants of the proposed MCROSS method increase the F1/EM metrics by at least 1.5/2.1 percentage points. In addition, MCROSS(w/o $\mathcal{L}_{\text{MoCo}}$) achieves discernible boosts among all metrics compared to baseline SSPT. It proves that the combination of cloze-like and natural samples in the QA task can endow the pre-trained model with a better understanding on supporting passages. Moreover, the gap between MCROSS(BI) and MCROSS(UNI) indicates that models trained only with cloze-like samples will be insensitive to natural questions. It should be also noticed that there exists a performance gap between SSPT† and MCROSS(UNI) among all metrics. This is attributed to the fact that the natural samples used in SSPT† are generated by the T5 question generator (Raffel et al., 2020). Since the T5 generator is fine-tuned on the training set of SQuADv1.1, it is well fitted in the domain of SQuAD. Compared to MCROSS(UNI) using cloze-like samples, SSPT† is trained with the natural samples containing more domain knowledge of the dataset, thus surpassing MCROSS(UNI).

[2] https://github.com/mrqa/MRQA-Shared-Task-2019/tree/master/baseline

Method	F1	EM	0-shot F1	0-shot EM
BERT	64.28	49.17	3.55	0.02
SSPT	65.86	50.62	9.58	6.05
SSPT†	65.60	50.52	34.03	15.67
MRQA-BERT	66.80	50.80	N/A	N/A
MCROSS(UNI)	66.45	51.33	10.90	6.55
MCROSS(BI)	**67.90**	**52.18**	34.82	14.98
w/o $\mathcal{L}_{\text{MoCo}}$	66.95	52.14	**35.68**	**16.50**

Table 2: Results on NewsQA dataset. (out of domain)

Method	F1	EM	0-shot F1	0-shot EM
BERT	62.86	57.39	3.25	0.04
SSPT	70.93	65.27	26.89	22.47
SSPT†	69.09	63.24	42.01	33.99
MRQA-BERT	71.60	65.60	N/A	N/A
MCROSS(UNI)	72.06	66.41	28.28	23.48
MCROSS(BI)	**73.77**	**68.05**	**47.79**	**39.53**
w/o $\mathcal{L}_{\text{MoCo}}$	73.30	67.72	46.63	38.54

Table 3: Results on TriviaQA dataset. (out of domain)

NewsQA and TriviaQA. Table 2 and Table 3 show the results on NewsQA and TriviaQA dataset from MRQA 2019 Shared Task. It is noticeable that MCROSS(BI) performs the best among all methods on F1 metrics. Compared with the state-of-the-art baseline SSPT, it achieves an improvement of 2.0 F1 score and 1.6 EM accuracy on NewsQA. On the TriviaQA dataset, MCROSS(BI) also surpasses SSPT by 2.8 percentage points on F1 and EM accuracy. Furthermore, the improvement on TriviaQA of MCROSS(BI) over MCROSS(w/o $\mathcal{L}_{\text{MoCo}}$) demonstrates the effectiveness of $\mathcal{L}_{\text{MoCo}}$ in the out-of-domain setting. In both datasets, the great improvement on zero-shot F1/EM of SSPT† over SSPT exhibits that SSPT† has gained the domain knowledge in understanding natural questions from SQuAD. In comparison with SSPT†, MCROSS(BI) significantly boosts zero-shot QA performance by 5.6/6.2 percentage points of F1/EM accuracy.

In contrast to the SQuAD dataset with in-domain settings, MCROSS(UNI) has gained a noticeable advantage over SSPT† on both NewsQA and TriviaQA datasets, hinting that SSPT† is overfitted to the natural samples of SQuAD with extra domain knowledge. Interestingly, MCROSS(BI) performs worse in zero-shot scenarios on NewsQA dataset than MCROSS(w/o $\mathcal{L}_{\text{MoCo}}$). This may be due to the fact that NewsQA has 32.7% samples that can be easily answered by simple text-matching (Trischler et al., 2017), into which MCROSS(w/o $\mathcal{L}_{\text{MoCo}}$) is overfitted to capture lexical overlaps.

4 Conclusion

This paper presents a novel pre-training method MCROSS for extractive QA which contains two tasks: 1) contrastive learning and 2) answer term prediction. Specifically, MCROSS adapts MoCo frameworks to maintain consistency in answering cloze-like and natural questions, enabling pre-trained models to have a more comprehensive understanding of supporting passages. The empirical experiments on three public datasets demonstrate that our approach can obtain noticeable improvements in extractive QA tasks in supervised and zero-shot scenarios.

5 Limitations

Although MCROSS can already obtain satisfactory QA performance, due to limited time and computational resources, we only use 5 million cloze-like samples for pre-training, which is one-twentieth of the scale of original SSPT experiments.

References

Avi Caciularu, Ido Dagan, Jacob Goldberger, and Arman Cohan. 2022. Long context question answering via supervised contrastive learning. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2872–2879, Seattle, United States. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *NAACL-HLT (1)*.

Bhuwan Dhingra, Danish Danish, and Dheeraj Rajagopal. 2018. Simple and effective semi-supervised question answering. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 582–587.

Adam Fisch, Alon Talmor, Robin Jia, Minjoon Seo, Eunsol Choi, and Danqi Chen. 2019. MRQA 2019 shared task: Evaluating generalization in reading comprehension. In *Proceedings of 2nd Machine Reading for Reading Comprehension (MRQA) Workshop at EMNLP*.

Michael Glass, Alfio Gliozzo, Rishav Chakravarti, Anthony Ferritto, Lin Pan, G P Shrivatsa Bhargav, Dinesh Garg, and Avi Sil. 2020a. Span selection pre-training for question answering. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2773–2782, Online. Association for Computational Linguistics.

Michael Glass, Alfio Gliozzo, Rishav Chakravarti, Anthony Ferritto, Lin Pan, G P Shrivatsa Bhargav, Dinesh Garg, and Avi Sil. 2020b. Span selection pre-training for question answering. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2773–2782, Online. Association for Computational Linguistics.

Wenkai Han, Yuqi Cheng, Jiayang Chen, Huawen Zhong, Zhihang Hu, Siyuan Chen, Licheng Zong, Liang Hong, Ting-Fung Chan, Irwin King, Xin Gao, and Yu Li. 2022. Self-supervised contrastive learning for integrative single cell rna-seq data analysis. *Briefings Bioinform.*, 23(5).

Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. 2020. Momentum contrast for unsupervised visual representation learning. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9726–9735.

Ziniu Hu, Yizhou Sun, and Kai-Wei Chang. 2021. Relation-guided pre-training for open-domain question answering. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 3431–3448, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Hsin-Yuan Huang, Chenguang Zhu, Yelong Shen, and Weizhu Chen. 2018. Fusionnet: Fusing via fully-aware attention with application to machine comprehension. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net.

Mandar Joshi, Eunsol Choi, Daniel Weld, and Luke Zettlemoyer. 2017. TriviaQA: A large scale distantly supervised challenge dataset for reading comprehension. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1601–1611, Vancouver, Canada. Association for Computational Linguistics.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, et al. 2019. Natural questions: A benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:453–466.

Zhongli Li, Wenhui Wang, Li Dong, Furu Wei, and Ke Xu. 2020. Harvesting and refining question-answer pairs for unsupervised QA. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6719–6728, Online. Association for Computational Linguistics.

Lin Pan, Chung-Wei Hang, Haode Qi, Abhishek Shah, Saloni Potdar, and Mo Yu. 2021. Multilingual BERT post-pretraining alignment. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 210–219. Association for Computational Linguistics.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67.

Pranav Rajpurkar, Robin Jia, and Percy Liang. 2018. Know what you don't know: Unanswerable questions for squad. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 784–789.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. Squad: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392.

Ori Ram, Yuval Kirstain, Jonathan Berant, Amir Globerson, and Omer Levy. 2021. Few-shot question answering by pretraining span selection. 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)*, pages 3066–3079, Online. Association for Computational Linguistics.

Adam Trischler, Tong Wang, Xingdi Yuan, Justin Harris, Alessandro Sordoni, Philip Bachman, and Kaheer Suleman. 2017. Newsqa: A machine comprehension dataset. *ACL 2017*, page 191.

Jeff Wu, Long Ouyang, Daniel M Ziegler, Nisan Stiennon, Ryan Lowe, Jan Leike, and Paul Christiano. 2021. Recursively summarizing books with human feedback. *arXiv preprint arXiv:2109.10862*.

Nan Yang, Furu Wei, Binxing Jiao, Daxing Jiang, and Linjun Yang. 2021. xMoCo: Cross momentum contrastive learning for open-domain question answering. 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)*, pages 6120–6129. Association for Computational Linguistics.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William Cohen, Ruslan Salakhutdinov, and Christopher D Manning. 2018. Hotpotqa: A dataset for diverse, explainable multi-hop question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2369–2380.

Yifei Zhang, Hao Zhu, Zixing Song, Piotr Koniusz, and Irwin King. 2022. COSTA: covariance-preserving feature augmentation for graph contrastive learning. In *KDD '22: The 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, Washington, DC, USA, August 14 - 18, 2022*, pages 2524–2534. ACM.

A Experimental Settings

A.1 Pre-Training Dataset

Cloze-like Samples. We follow the method proposed by SSPT (Glass et al., 2020b) to generate cloze-like pre-training instances. Specifically, we first choose a sentence in the Wikipedia corpus and randomly replace an entity or noun phrase with the special token [BLANK]. The new sentence will be considered as a query. Based on the query, we retrieve a paragraph that contains the masked term as its corresponding passage. Due to the time and computational resource limitations, we finally collect 5 million unique pre-training examples in English for our pre-training, which is one-sixth of the original SSPT method.

Natural Samples. For each $s^{\text{cloze}} = (q_{\text{c}}, p, a)$, we use a public T5-base model[3] to generate natural queries q_{n} given (p, a). It is fine-tuned for answer-aware question generation using the SQuADv1.1 train set. If answer a is empty, q_{n} is not generated and loss functions related to s^{natural} will not be included during training. We do not apply extra filtering or other quality control operations for generated questions.

A.2 Evaluation

We use the official model evaluation scripts of SQuADv1.1[4] and MRQA[5] to calculate the F1 and EM metrics between predictions and ground truth answers. Besides, we employ zero-shot F1 and zero-shot EM to quantify the performance of models without fine-tuning. Official training and development sets are used in all experiments.

B Implementation Details

All model architectures are based on the 12-layer BERT-base model with 110M parameters, and we employed Google's original implementation of the BERT model published in the Huggingface[6] to build up our encoder. The training batch size for pre-training and fine-tuning are 32 and 8. Total training steps for MCROSS and other baseline models are both kept to 156,250. The max sequence

[3] https://huggingface.co/valhalla/t5-base-qg-hl

[4] https://github.com/allenai/bi-att-flow/blob/master/squad/evaluate-v1.1.py

[5] https://github.com/mrqa/MRQA-Shared-Task-2019

[6] https://huggingface.co/docs/transformers/v4.21.2/en/model_doc/bert#transformers.BertForQuestionAnswering

$\mathcal{D}$	Dataset	F1	EM
$\mathcal{D}_{\text{KL}}$	SQuADv1.1	**88.95**	**81.60**
	NewsQA	**66.45**	**51.33**
	TriviaQA	**72.06**	**66.41**
$\mathcal{D}_{\text{Cosine}}$	SQuADv1.1	88.47	80.79
	NewsQA	65.98	50.40
	TriviaQA	69.73	64.24

Table 4: Comparison between choosing $\mathcal{D}_{\text{KL}}$ and $\mathcal{D}_{\text{Cosine}}$

length of the transformer encoder is 384. In all experiments, we use the Adam optimizer with a learning rate of 2e-5. The MoCo momentum m is set to 0.999, queue size 32,000, moco ratio λ_{MoCo} 1.0 and temperature τ 0.05. Models are implemented on PyTorch. When the passage length exceeds the max sequence length, we follow the sliding window strategy proposed for BERT transformers in both fast/slow encoder structures.

During QA prediction, we apply the constraint $st_a \leq en_a$ to filtering out implausible answer spans and then select the ones with the highest joint probability $p(st_a) \times p(en_a)$ as results.

C Design Choice: KL Divergence

The performance gain of the MCROSS(UNI) method choosing KL divergence $\mathcal{D}_{\text{KL}}$ over cosine similarity $\mathcal{D}_{\text{Cosine}}$ as similarity function $\mathcal{D}$ is shown in Table 4.

D Space and Time Complexity

The additional parameters of the answer prediction layer come from Eq. (1). The number of parameters is $(768 + 1) * 2 = 1,538$, which is negligible in front of the parameters in BERT-base (110M). Training of the SSPT baselines and the MCROSS method took approximately 40 hours each on 8 V-100 GPUs.