DC-BERT: Decoupling Question and Document for Efficient Contextual Encoding

Yuyu Zhang^{*,1}, **Ping Nie**^{*,2}, **Xiubo Geng**^{3}, **Arun Ramamurthy**^{4}, **Le Song**^{1} **& Daxin Jiang**^{3}
[1]Georgia Institute of Technology [2]Peking University
[3]Microsoft [4]Siemens Corporate Technology
{yuyu,lsong}@gatech.edu,ping.nie@pku.edu.cn
{xigeng,djiang}@microsoft.com,arun.ramamurthy@siemens.com

Abstract

Recent studies on open-domain question answering have achieved prominent performance improvement using pre-trained language models such as BERT. State-of-the-art approaches typically follow the "retrieve and read" pipeline and employ BERT-based reranker to filter retrieved documents before feeding them into the reader module. The BERT retriever takes as input the concatenation of question and each retrieved document. Despite the success of these approaches in terms of QA accuracy, due to the concatenation, they can barely handle high-throughput of incoming questions each with a large collection of retrieved documents. To address the efficiency problem, we propose DC-BERT, a decoupled contextual encoding framework that has dual BERT models: an *online* BERT which encodes the question only once, and an *offline* BERT which pre-encodes all the documents and caches their encodings. On SQuAD Open and Natural Questions Open datasets, DC-BERT achieves 10x speedup on document retrieval, while retaining most (about 98%) of the QA performance compared to state-of-the-art approaches for open-domain question answering.

1 Introduction

Open-domain question answering (QA) is an important and challenging task in natural language processing, which requires a machine to find the answer by referring to large unstructured text corpora without knowing which documents may contain the answer. Recently, pre-trained language models such as BERT (Devlin et al., 2019) have boosted up the performance of open-domain QA on several benchmark datasets, such as HotpotQA (Yang et al., 2018) and Natural Questions (Kwiatkowski et al., 2019). With high-quality contextual encodings, BERT-based approaches (Nie et al., 2019; Hu et al., 2019) have dominated the leaderboards and significantly outperformed previous RNN and CNN based approaches (Qi et al., 2019; Jiang & Bansal, 2019; Min et al., 2018; Yu et al., 2018; Wang et al., 2018).

Recent studies on open-domain QA typically follow the "retrieve and read" pipeline initiated by Chen et al. (2017), which combines information retrieval (IR) and machine reading comprehension (MRC) modules as a pipeline: the former retrieves the documents using off-the-shelf IR systems based on TF-IDF or BM25, and the latter reads the retrieved documents to extract answer. The IR systems are purely based on n-gram matching and have shallow understanding of the context. Thus, documents that contain the correct answer may not be ranked among the top by IR systems (Asai et al., 2020). If we simply feed more documents into the reader module to increase the chance of hitting under-ranked documents that contain the answer, it can be computationally expensive and bring more noise to the reader module, making it harder to find the answer. To alleviate this problem, state-of-the-art approaches (Nie et al., 2019; Hu et al., 2019; Lee et al., 2018) have proposed to train a BERT-based reranker, which is a binary classifier to filter retrieved documents before feeding them into the reader module. The input of the BERT retriever is the concatenation of a question and each retrieved document, formulated as:

$$[CLS]\ Question\ [SEP]\ Document\ [SEP],$$

[*]Equal contribution

where $[CLS]$ is the pooling token. Due to the high-quality contextual encodings generated by BERT, such methods significantly improve the retrieval performance over non-parameterized IR systems, and thus boost up the answer accuracy for open-domain QA. However, due to the concatenation, these approaches have to repeatedly encode a question with each of the retrieved document, which are hard to handle high-throughput incoming questions each with a large collection of retrieved documents. This severe efficiency problem prohibits existing approaches for open-domain QA from being deployed as real-time QA systems.

Recent studies (Tenney et al., 2019; Hao et al., 2019) have probed and visualized BERT to understand its effectiveness, which show that the lower layers of BERT encode more local syntax information such as part-of-speech (POS) tags and constituents, while the higher layers tend to capture more complex semantics relying on wider contexts. Inspired by these observations, we propose DC-BERT, which decouples the lower layers of BERT into local contexts (question and document), and then applies Transformer layers on top of the independent encodings to enable question-document interactions. As illustrated in Figure 1, DC-BERT has two separate BERT models: an *online* BERT which encodes the question only once, and an *offline* BERT which pre-encodes all the documents and caches their encodings. With caching enabled, DC-BERT can instantly read out the encoding of any document. The decoupled encodings of question and document are then fed into the Transformer layers with global position and type embeddings for question-document interactions, which produces the contextual encodings of the (question, document) pair. DC-BERT can be applied to both document retriever and reader. In this work, we focus on speeding up the retriever, since the number of documents retrieved per question can be fairly large, while the number of documents fed into the reader module is much controlled. Therefore, it is more important to address the efficiency problem of the document retriever.

Speeding up BERT-based models for efficient inference is an active field of research. Previous works related to this direction mainly include: 1) model compression, where methods are proposed to reduce the model size through weight pruning or model quantization (Jacob et al., 2018); and 2) model distillation, where methods are proposed to train a small student network from a large teacher network, such as DistilBERT (Sanh et al., 2019). Both lines of research are actually orthogonal to our work, since the compressed / distilled BERT model can be combined with DC-BERT. In the experiments, we also compare with quantized BERT and DistilBERT, showing that DC-BERT has significant performance advantage over these methods.

Our main contributions are summarized as follows:

- *Decoupled QA encoding*: We propose to decouple question and document for efficient contextual encoding. To the best of our knowledge, our work is the first to explore a combination of local- and global-context encoding with BERT for open-domain QA.
- *Effective question-document interactions*: We propose an effective model architecture for question-document interactions, which employs trainable global embeddings with Transformer layers.
- *Fast document retrieval*: We successfully apply DC-BERT to document reranking, a key component in open-domain QA, by making it over 10x faster than the existing approaches, while retaining most (about 98%) of the QA performance on benchmark datasets.
- *New evaluation metrics*: We propose two symmetric new evaluation metrics to gauge the retriever's capability of discovering documents that have low TF-IDF scores but contain the answer.

2 METHODOLOGY

The overall architecture of DC-BERT (Figure 1) consists of a dual-BERT component for decoupled encoding, a Transformer component for question-document interactions, and a classifier component for document reranking.

Dual-BERT component. DC-BERT contains two BERT models to independently encode the question and each retrieved document. During training, the parameters of both BERT models are updated to optimize the learning objective, which is described later in the classifier component part. Once the model is trained, we pre-encode all the documents and store their encodings in an *offline* cache. During testing, we encode the question only once using the *online* BERT, and instantly read out the cached encodings of all the candidate documents retrieved by an off-the-shelf IR system. Com-

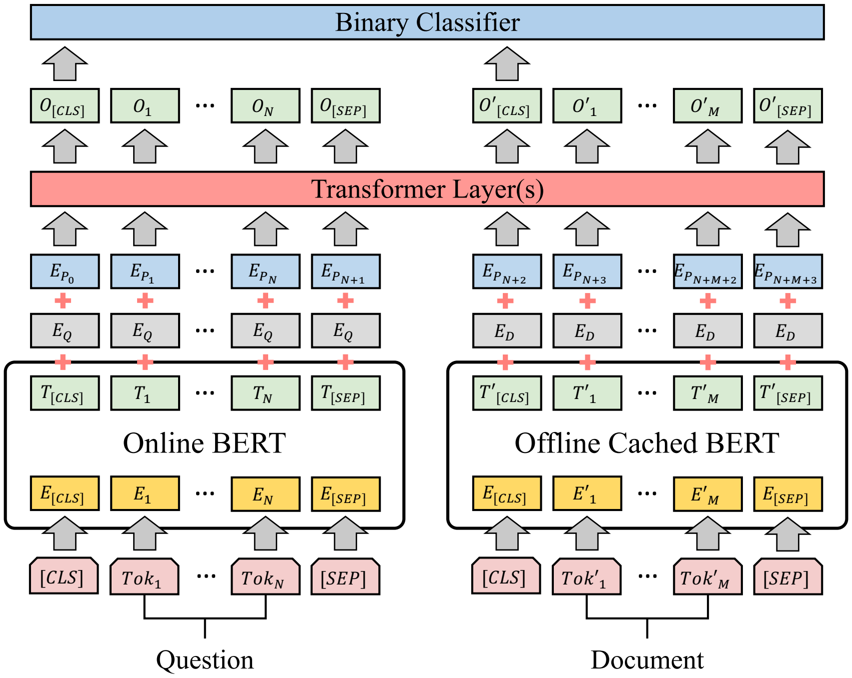


Figure 1: Overview architecture of DC-BERT, which decouples question and document for efficient contextual encoding. The binary classifier predicts whether the document is relevant to the question.

pared to the existing "concatenate and encode" approaches that concatenate the question and each retrieved document, DC-BERT only encodes a question for once, which reduces the computational cost of BERT lower layers from $O(N_q N_d (L_q + L_d)^2)$ to $O(N_q (L_q^2 + N_d L_d^2))$ where N_q denote the number of questions, N_d denote the number of retrieved documents for each question, L_q and L_d denote the average number of tokens of each question and document, respectively. Moreover, the decoupled BERT also enables caching of the document encodings offline, which further reduces the computational cost to $O(N_q L_q^2)$.

In fact, a number of previous RNN and CNN based QA approaches, such as BiDAF (Seo et al., 2017), DCN (Xiong et al., 2017) and QANet (Yu et al., 2018), also decouple the encoding of question and document, but their performance is dominated by recent BERT-based approaches. With self-attention Transformer layers, BERT is designed to perform contextual encoding by incorporating wide context from pre-training to fine-tuning. This explains why most state-of-the-art approaches for open-domain QA concatenate question and document for wide-context encoding and achieve prominent performance. Recent work (Lee et al., 2019) attempts to independently encode question and document with BERT, and compute the inner product of their encodings to retrieve documents, but performs worse than state-of-the-art approaches. We posit that the lack of interactions between question and document can significantly hurt the performance. Therefore, our method is designed to encapsulate question-document interactions with a Transformer component, as described below.

Transformer component. With the dual-BERT component, we obtain the question encoding $\mathbf{T} \in \mathbb{R}^{N \times d}$ and the document encoding $\mathbf{T}' \in \mathbb{R}^{M \times d}$, where d is the dimension of word embeddings, and N and M are the length of the question and the document, respectively. Since the document reranking task is to predict the relevance of the document for a question, we introduce a Transformer component with trainable global embeddings to model the question-document interactions.

More specifically, we have global position embeddings $\mathbf{E}_{P_i} \in \mathbb{R}^d$ to re-encode the token at position P_i in the concatenated question-document encoding sequence. We also have global type embeddings $\mathbf{E}_Q \in \mathbb{R}^d$ and $\mathbf{E}_D \in \mathbb{R}^d$ to differentiate whether the encoded token is from question or document. Both the global position and type embeddings are initialized by the position and sentence embeddings from pre-trained BERT, and will be updated during the training. These additional embeddings are added on top of the question and document encodings (with the encodings of $[CLS]$ and $[SEP]$), and then fed into the Transformer layers. The number of Transformer layers K is configurable to trade-off between the model capacity and efficiency. The Transformer layers are initialized by the last K layers of pre-trained BERT, and are updated during the training.

Classifier component. After the Transformer layers, DC-BERT treats the document reranking task as a binary classification problem to predict whether the retrieved document is relevant to the question. Following previous work (Das et al., 2019; Htut et al., 2018; Lin et al., 2018), we employ paragraph-level distant supervision to gather labels for training the classifier, where a paragraph that contains the exact ground truth answer span is labeled as a positive example. We parameterize the binary classifier as a MLP layer on top of the Transformer layers:

$$p(Q_i, D_j) = \sigma(\text{MLP}([o_{[CLS]}; o'_{[CLS]}])), \tag{1}$$

where (Q_i, D_j) is a pair of question and retrieved document, and $o_{[CLS]}$ and $o'_{[CLS]}$ are the Transformer output encodings of the $[CLS]$ token of the question and the document, respectively. The MLP parameters are updated by minimizing the cross-entropy loss:

$$\mathcal{J} = -\sum_{(Q_i, D_j)} \Big(y \log{(p)} + (1 - y) \log{(1 - p)} \Big), \tag{2}$$

where $y = y(Q_i, D_j)$ is the distantly supervised label, and $p = p(Q_i, D_j)$ as defined in Eq. equation 1.

3 Experiments

Benchmark datasets. We evaluate DC-BERT and other baseline methods on two popular benchmark datasets: 1) SQuAD Open (Chen et al., 2017), which is composed of questions from the original crowdsourced SQuAD dataset (Rajpurkar et al., 2016); 2) Natural Questions Open (Min et al., 2019), which is composed of questions from the original Natural Questions dataset (Kwiatkowski et al., 2019). The questions are created from real user queries issued to Google Search engine. For all our experiments, we use the standard splits provided with the datasets, and report the performance on the development split.

Evaluation metrics. To evaluate the retriever speed, we compare the wall-clock time running on a single GPU. To evaluate the retriever ranking performance, we use the following metrics: 1) P@N, which is defined in previous work (Chen et al., 2017) as the percentage of questions for which the answer span appears in one of the top N documents; 2) PBT@N, a new evaluation metric that we propose to gauge the semantic retrieval capability of the document reranker, which is the percentage of questions for which at least one of the top N documents that contains the answer span is *not* in the TF-IDF top N documents. In other words, this new metric measures the retriever's capability beyond the TF-IDF retriever (the higher the better); 3) PTB@N, our proposed metric that is symmetric to PBT@N, which is the percentage of questions for which at least one of the top N TF-IDF retrieved documents that contains the answer span *not* in the retriever's top N documents. This metric measures the retriever's capability of retaining the relevant documents returned by TF-IDF retriever (the lower the better). To evaluate the downstream QA performance, we follow previous works (Chen et al., 2017; Nie et al., 2019; Das et al., 2019) and use the standard answer exact match (EM) score.

Implementation details. We use pre-trained BERT-base model (Devlin et al., 2019) for the document reranker and pre-trained BERT-wwm (whole word masking) model for the downstream QA model. For the standard TF-IDF retrieval, we use the released retrieval data from Min et al. (2019) for Natural Questions Open, and use the DrQA (Chen et al., 2017) TF-IDF retriever to collect 80 documents for SQuAD Open. We select top 10 documents ranked by retriever to feed into the reader module. For our method, we enable caching of the document encodings. We set $K = 1$ as the number of Transformer layers for all the experiments and vary this choice in the ablation study. We use $4e - 5$ as the initial learning rate. Our model is trained with Adam optimizer (Kingma & Ba, 2014). For the binary classifier, we use a two-layer MLP with the $\tanh(\cdot)$ activation function for nonlinear transformation.

Baseline methods. 1) BERT-base: a well-trained reranker using the BERT-base model (Devlin et al., 2019), which is a very strong baseline and represents state-of-the-art performance on both

Table 1: Performance comparison on two benchmark datasets. DC-BERT uses one Transformer layer for question-document interactions. Quantized BERT is a 8bit-Integer model. DistilBERT is a compact BERT model with 2 Transformer layers.

Retriever Model	SQuAD Open				Natural Questions Open			
	Retriever P@10	Retriever Speedup	Answer EM	EM Drop (%)	Retriever P@10	Retriever Speedup	Answer EM	EM Drop (%)
BERT-base	71.5	1.0x	40.1	0.0	65.0	1.0x	28.0	0.0
Quantized BERT	68.0	1.1x	39.5	1.5	64.3	1.1x	27.5	1.8
DistilBERT	56.4	5.7x	34.6	13.7	60.6	5.7x	25.1	9.7
DC-BERT	70.1	10.3x	39.2	2.1	63.5	10.3x	27.4	2.0

Table 2: Retriever performance in PBT@10 and PTB@10.

Retriever Model	SQuAD		Natural Questions	
	PBT@10	PTB@10	PBT@10	PTB@10
BERT-base (Devlin et al., 2019)	19.8	2.1	14.6	4.0
Quantized BERT (Jacob et al., 2018)	18.4	3.4	14.1	4.5
DistilBERT (Sanh et al., 2019)	14.3	10.2	11.5	7.1
DC-BERT	18.1	5.7	13.8	6.8

datasets. Recent work (Asai et al., 2020) uses external data such as hyperlinks and data augmentation to further improve the performance, which is out of the scope of this paper; 2) Quantized BERT: a recent work (Zafrir et al., 2019) that employs quantization techniques (Jacob et al., 2018) to compress BERT-base into a 8bit-Integer model. We use the official open-sourced code for experiments; 3) DistilBERT: a recent approach (Sanh et al., 2019) that leverages knowledge distillation techniques (Hinton et al., 2015) to train a smaller and compact student BERT model to reproduce the behavior of the teacher BERT-base model. We use the official open-sourced code for experiments. To achieve decent speedup, we set 2 Transformer layers for the student BERT model.

Retriever speed. The main experimental results are summarized in Table 1. We first compare the retriever speed. DC-BERT achieves over 10x speedup over the BERT-base retriever, which demonstrates the efficiency of our method. Quantized BERT has the same model architecture as BERT-base, leading to the minimal speedup. DistilBERT achieves about 6x speedup with only 2 Transformer layers, while BERT-base uses 12 Transformer layers.

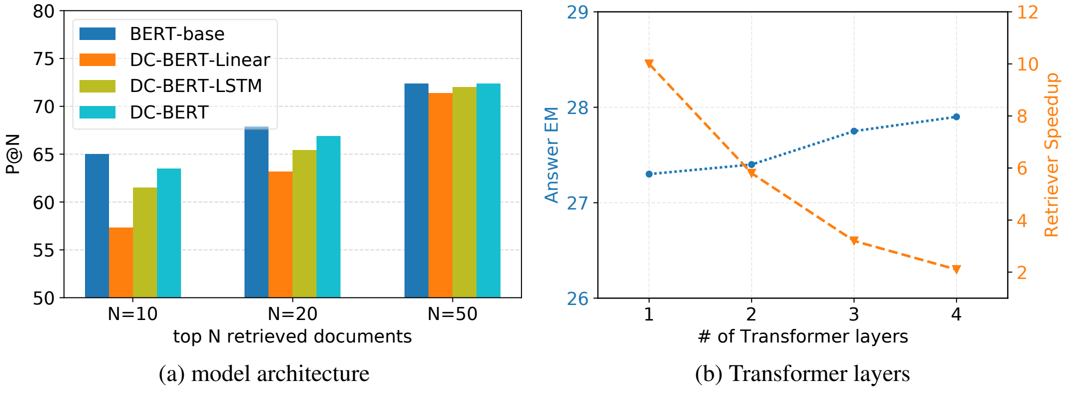
Retriever ranking performance. We evaluate the ranking metrics in terms of P@10 in Table 1. With a 10x speedup, DC-BERT still achieves similar retrieval performance compared to BERT-base on both datasets. At the cost of little speedup, Quantized BERT also works well in ranking documents. DistilBERT performs significantly worse than BERT-base, which shows the limitation of the distilled BERT model. We also report the proposed PBT@10 and PTB@10 metrics in Table 2. As discussed, PBT@10 is the higher the better, and PTB@10 is the lower the better. DC-BERT and Quantized BERT achieves similar performance compared to BERT-base, while DistilBERT is inferior in both metrics.

QA performance. As reported in Table 1, DC-BERT and Quantized BERT retain most of the QA performance (Answer EM) compared to BERT-base, on both SQuAD Open and Natural Questions Open datasets. Due to the inferior retrieval performance, DistilBERT also performs the worst in answer accuracy. With a 10x speedup, DC-BERT only has a performance drop of about 2%, which demonstrates the effectiveness of our method in the downstream QA task.

Ablation study. To further investigate the impact of our model architecture design, we compare the performance of DC-BERT and its variants, including 1) DC-BERT-Linear, which uses linear layers instead of Transformers for interaction; and 2) DC-BERT-LSTM, which uses LSTM and bi-linear layers for interactions following previous work (Min et al., 2018). We report the results in Table 3. Due to the simplistic architecture of the interaction layers, DC-BERT-Linear achieves the best speedup but has significant performance drop, while DC-BERT-LSTM achieves slightly worse

Table 3: Ablation study results on Natural Questions Open.

Retriever Model	Retriever P@10	Retriever Speedup	Answer EM
DC-BERT-Linear	57.3	43.6x	24.8
DC-BERT-LSTM	61.5	8.2x	26.5
DC-BERT	63.5	10.3x	27.4



(a) model architecture

(b) Transformer layers

Figure 2: Ablation study results on Natural Questions Open.

performance and speedup than DC-BERT. Figure 2a shows that DC-BERT consistently outperforms its variants for different number of top retrieved documents, and leads with a larger margin when retrieving less documents. We also investigate the impact of the number of Transformer layers for question-document interactions, and report the results in Figure 2b. When we increase the number of Transformer layers, the QA performance consistently improves, and the speedup decreases due to the increased computational cost. This shows the trade-off between the model capacity and efficiency of our method.

4 Conclusion

This paper introduces DC-BERT to decouple question and document for efficient contextual encoding. DC-BERT has been successfully applied to document retrieval, a key component in open-domain QA, achieving 10x speedup while retaining most of the QA performance. With the capability of processing high-throughput of questions each with a large collection of retrieved documents, DC-BERT brings open-domain QA one step closer to serving real-world applications.

References

Akari Asai, Kazuma Hashimoto, Hannaneh Hajishirzi, Richard Socher, and Caiming Xiong. Learning to retrieve reasoning paths over wikipedia graph for question answering. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=SJgVHkrYDH.

Danqi Chen, Adam Fisch, Jason Weston, and Antoine Bordes. Reading Wikipedia to answer open-domain questions. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1870–1879, Vancouver, Canada, July 2017. Association for Computational Linguistics. doi: 10.18653/v1/P17-1171. URL https://www.aclweb.org/anthology/P17-1171.

Rajarshi Das, Shehzaad Dhuliawala, Manzil Zaheer, and Andrew McCallum. Multi-step retriever-reader interaction for scalable open-domain question answering. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=HkfPSh05K7.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1423. URL https://www.aclweb.org/anthology/N19-1423.

Yaru Hao, Li Dong, Furu Wei, and Ke Xu. Visualizing and understanding the effectiveness of BERT. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 4143–4152, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1424. URL https://www.aclweb.org/anthology/D19-1424.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.

Phu Mon Htut, Samuel Bowman, and Kyunghyun Cho. Training a ranking function for open-domain question answering. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Student Research Workshop*, pp. 120–127, New Orleans, Louisiana, USA, June 2018. Association for Computational Linguistics. doi: 10.18653/v1/N18-4017. URL https://www.aclweb.org/anthology/N18-4017.

Minghao Hu, Yuxing Peng, Zhen Huang, and Dongsheng Li. Retrieve, read, rerank: Towards end-to-end multi-document reading comprehension. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 2285–2295, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1221. URL https://www.aclweb.org/anthology/P19-1221.

Benoit Jacob, Skirmantas Kligys, Bo Chen, Menglong Zhu, Matthew Tang, Andrew Howard, Hartwig Adam, and Dmitry Kalenichenko. Quantization and training of neural networks for efficient integer-arithmetic-only inference. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2704–2713, 2018.

Yichen Jiang and Mohit Bansal. Self-assembling modular networks for interpretable multi-hop reasoning. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 4474–4484, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1455. URL https://www.aclweb.org/anthology/D19-1455.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

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

Jinhyuk Lee, Seongjun Yun, Hyunjae Kim, Miyoung Ko, and Jaewoo Kang. Ranking paragraphs for improving answer recall in open-domain question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 565–569, Brussels, Belgium, October-November 2018. Association for Computational Linguistics. doi: 10.18653/v1/D18-1053. URL https://www.aclweb.org/anthology/D18-1053.

Kenton Lee, Ming-Wei Chang, and Kristina Toutanova. Latent retrieval for weakly supervised open domain question answering. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 6086–6096, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1612. URL https://www.aclweb.org/anthology/P19-1612.

Yankai Lin, Haozhe Ji, Zhiyuan Liu, and Maosong Sun. Denoising distantly supervised open-domain question answering. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1736–1745, Melbourne, Australia, July 2018. Association for Computational Linguistics. doi: 10.18653/v1/P18-1161. URL https://www.aclweb.org/anthology/P18-1161.

Sewon Min, Victor Zhong, Richard Socher, and Caiming Xiong. Efficient and robust question answering from minimal context over documents. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1725–1735, Melbourne, Australia, July 2018. Association for Computational Linguistics. doi: 10.18653/v1/P18-1160. URL https://www.aclweb.org/anthology/P18-1160.

Sewon Min, Danqi Chen, Hannaneh Hajishirzi, and Luke Zettlemoyer. A discrete hard EM approach for weakly supervised question answering. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 2851–2864, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1284. URL https://www.aclweb.org/anthology/D19-1284.

Yixin Nie, Songhe Wang, and Mohit Bansal. Revealing the importance of semantic retrieval for machine reading at scale. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 2553–2566, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1258. URL https://www.aclweb.org/anthology/D19-1258.

Peng Qi, Xiaowen Lin, Leo Mehr, Zijian Wang, and Christopher D. Manning. Answering complex open-domain questions through iterative query generation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 2590–2602, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1261. URL https://www.aclweb.org/anthology/D19-1261.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. SQuAD: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pp. 2383–2392, Austin, Texas, November 2016. Association for Computational Linguistics. doi: 10.18653/v1/D16-1264. URL https://www.aclweb.org/anthology/D16-1264.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*, 2019.

Minjoon Seo, Aniruddha Kembhavi, Ali Farhadi, and Hannaneh Hajishirzi. Bidirectional attention flow for machine comprehension. In *International Conference on Learning Representations*, 2017. URL https://openreview.net/forum?id=HJ0UKP9ge.

Ian Tenney, Dipanjan Das, and Ellie Pavlick. BERT rediscovers the classical NLP pipeline. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 4593–4601, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1452. URL https://www.aclweb.org/anthology/P19-1452.

Shuohang Wang, Mo Yu, Xiaoxiao Guo, Zhiguo Wang, Tim Klinger, Wei Zhang, Shiyu Chang, Gerry Tesauro, Bowen Zhou, and Jing Jiang. R^3: Reinforced ranker-reader for open-domain question answering. In *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

Caiming Xiong, Victor Zhong, and Richard Socher. Dynamic coattention networks for question answering. In *International Conference on Learning Representations*, 2017. URL https://openreview.net/forum?id=rJeKjwvclx.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. HotpotQA: A dataset for diverse, explainable multi-hop question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 2369–2380, Brussels, Belgium, October-November 2018. Association for Computational Linguistics. doi: 10.18653/v1/D18-1259. URL https://www.aclweb.org/anthology/D18-1259.

Adams Wei Yu, David Dohan, Quoc Le, Thang Luong, Rui Zhao, and Kai Chen. Fast and accurate reading comprehension by combining self-attention and convolution. In *International Conference on Learning Representations*, 2018. URL https://openreview.net/forum?id=B14TlG-RW.

Ofir Zafrir, Guy Boudoukh, Peter Izsak, and Moshe Wasserblat. Q8bert: Quantized 8bit bert. *arXiv preprint arXiv:1910.06188*, 2019.