Utilizing Evidence Spans via Sequence-Level Contrastive Learning for Long-Context Question Answering

Avi Caciularu[1*] **Ido Dagan**[1] **Jacob Goldberger**[2] **Arman Cohan**[3,4]

[1]Computer Science Department, Bar-Ilan University, Ramat-Gan, Israel
[2]Faculty of Engineering, Bar-Ilan University, Ramat-Gan, Israel
[3]Allen Institute for AI, Seattle, WA
[4]Paul G. Allen School of Computer Science, University of Washington, Seattle, WA

avi.c33@gmail.com, armanc@allenai.org
dagan@cs.biu.ac.il, jacob.goldberger@biu.ac.il

Abstract

Long-range transformer models have achieved encouraging results on long-context question answering (QA) tasks. Such tasks often require reasoning over a long document, and they benefit from identifying a set of evidence spans (e.g., sentences) that provide supporting evidence for addressing the question. In this work, we propose a novel method for equipping long-range transformers with an additional sequence-level objective for better identification of supporting evidence spans. We achieve this by proposing an additional contrastive supervision signal in finetuning, where the model is encouraged to explicitly discriminate supporting evidence sentences from negative ones by maximizing the question-evidence similarity. The proposed additional loss exhibits consistent improvements on three different strong long-context transformer models, across two challenging question answering benchmarks – HotpotQA and QAsper.[1]

1 Introduction

Answering questions that require reasoning over a long sequence (e.g., multiple paragraphs or a long document consisting of multiple sections) is a challenging task (Dasigi et al., 2021; Pang et al., 2021). Research in this domain mostly concerns tasks that involve multiple passages (paragraphs or documents), including benchmarks like HotpotQA (Yang et al., 2018b) and QAsper (Dasigi et al., 2021). HotpotQA is a multi-hop QA benchmark over multiple paragraphs from Wikipedia and QAsper involves reading comprehension from long academic papers, where relevant information on a question could be spread across the paper, often resulting in complex entailment challenges.

Recent long-context transformers (Tay et al., 2021) have been successfully applied to long-context QA by processing and contextualizing information across the entire input. Prominent examples of long-context transformer models are Longformer (Beltagy et al., 2020) and BigBird (Zaheer et al., 2020) where instead of a full self-attention operation, they perform sparse self-attention over the input by using a local sliding window attention and a global attention mechanism to few specific input locations. To answer questions from long inputs that require multi-hop reasoning, it is often critical to identify a set of evidence spans (e.g., sentences or paragraphs) that provide relevant information for resolving the question. Prior work shows that jointly training long-context transformer models to perform evidence span extraction in addition to answer generation is important for achieving high performance (Beltagy et al., 2020; Dasigi et al., 2021). To jointly perform evidence extraction and question answering, these models utilize special sentence markers in the input; the final layer representation corresponding to these markers is then passed through a classification layer and is optimized using the cross-entropy loss in conjunction with the answer extraction/generation loss. We hypothesize that this cross-entropy loss does not explicitly capture relationships between the question and the candidate evidence spans. In this work, we propose a method for *explicitly* capturing question-evidence relationship with contrastive learning (see Fig. 1).

Driven by the intuition that better QA generalization requires finding supporting evidences for answering a question, we propose a supervised contrastive learning objective for long-context QA tasks at the finetuning stage. Contrastive learning has been recently applied to a variety of deep learning models in computer vision (Khosla et al., 2020; Chen et al., 2020; Chen and He, 2021) and NLP (Gao et al., 2021a; Gunel et al., 2021). Unlike

[*] Work partly done as an intern at AI2.
[1]Code is available at github.com/aviclu/seq-contrast-qa

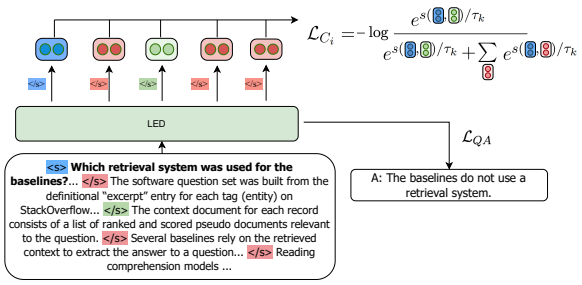

Figure 1: Demonstration of our method over an example instance taken from QAsper (Dasigi et al., 2021). A long sequence is fed to a sparse transformer model, producing representations for the marker special tokens. Then, these vectors are used to compute our contrastive objective. The token colored in blue represents the question, and the tokens colored in green (red) represent positive (negative) evidence sentences. The objective is to maximize the similarity between the blue vector and the green vectors.

the prior NLP related methods, we are the first to propose a novel model-agnostic contrastive loss for long-context transformers. Our proposed objective targets specifically the tokens that mark the question and evidence spans within input sequences, and unlike prior work, it is not based on individually encoding sentences or paragraphs. We show that our additional contrastive supervision provides consistent improvements on three different models and two long-context QA datasets, demonstrating its effectiveness and versatility.

2 Setup

Assume that we are given a question q_i and a context $\mathcal{S}_i{=}\langle s_1, ..., s_M\rangle$ consisting of M sentences (s_j can be also a document/paragraph/passage, depending on the dataset). From C_i, the task is to identify the correct answer a_i and a set of K evidence spans $\mathcal{S}_i^+{=}\{s_{i_1}, ..., s_{i_K}\}$ where i_j are indices of the sentences that are the supporting evidence for answering the question Q_i.

As common in the input setup of long-context transformer models (Beltagy et al., 2020; Zaheer et al., 2020; Caciularu et al., 2021), the question and context sentences are concatenated in a single long sequence with special tokens specifying sentence boundaries. Then the input is passed to the long-context transformer, and is trained to jointly identify the evidence sentences and extract/generate the answer. Concretely, for each example, we prepare the following concatenated input sequence:

$$\texttt{Input}_i{=}[\langle \texttt{s}\rangle, Q_i, \langle \texttt{/s}\rangle, s_1, \langle \texttt{/s}\rangle, s_2, ..., \langle \texttt{/s}\rangle, s_M]$$

where "," is the string concatenation operation, q_i and s_j are sequences of tokens corresponding to the question and the jth sentence in the input context, and $\langle \texttt{s}\rangle$ and $\langle \texttt{/s}\rangle$ are special tokens representing the question and a context sentence, respectively (See Fig. 1 for an example). Then, a QA loss, which we denote by $\mathcal{L}_{QA}$ is applied over the contextualized representation of each sentence token, and is optimized using supervision. $\mathcal{L}_{QA}$ depends on the dataset and can take the form of a multi-task objective, representing multiple tasks in the context of QA (Dasigi et al., 2021) (e.g., evidence extraction, answer generation, etc.).

3 Sequence-level contrastive loss

To encourage the long-context transformer model to explicitly capture relationships between the question and evidence sentences, while performing the QA task, we introduce an additional sequence-level loss that compares and contrasts the question with context sentences. The additional proposed loss $\mathcal{L}_C$ is based on the InfoNCE loss (Oord et al., 2018), and is applied over a triplet of vectors; it encourages the question and correct evidence representations to become closer to each other, while pushing the question and distracting target representations away. Formally, the contrastive loss is defined as the sum of negative log-likelihood losses over each example, where each loss term discriminates the positive units from negative ones, as follows:

$$\mathcal{L}_{C_i} = -\log \sum_{x^+ \in \mathcal{S}_i^+} \left(\frac{e^{sim(s^+, q_i)/\tau_k}}{\sum\limits_{s \in \mathcal{S}_i} e^{sim(s, q_i)/\tau_k}} \right), \quad (1)$$

where s^+ is the positive supporting evidence marker token representation, q_i is the vector representation corresponding to the question maker in the input ($\langle \texttt{s}\rangle$ in Fig. 1), τ_k is the configurable temperature hyperparameter, and $sim(\cdot)$ is a similarity metric, e.g., dot product ($sim\,(s,q){=}s^\top q$) or cosine similarity ($sim\,(s,q){=}\frac{s^\top q}{\|s\|\|q\|}$). Then the final aggregated loss is obtained by averaging over all the examples $\mathcal{L}_C{=}\frac{1}{n}\sum_{i=1}^{n}\mathcal{L}_{C_i}$.

We incorporate our additional contrastive loss into the main QA span extraction/generation loss

$\mathcal{L}_{QA}$ as follows:

$$\mathcal{L} = (1-\lambda)\cdot\mathcal{L}_{QA} + \lambda\cdot\mathcal{L}_{C}, \quad (2)$$

where $\lambda \in [0,1]$ is a hyper-parameter determining how much weight is assigned to each component of the total loss.

Incorporating question types Question types, if available in the dataset, can further guide our contrastive term and potentially improve the results by inserting more inductive bias to the model. Following Iter et al. (2020), we can also define the similarity function as the bilinear distance according to different question types (e.g., yes/no, generative, non answerable, etc).

$$sim_k(s,q) = s^\top W_k q, \quad (3)$$

where k is the expected question type (which is given as a label in the training set), and W_k is the corresponding allocated learnable matrix. For example, in the HotpotQA dataset, we are given three question types: $k \in \{\text{yes}, \text{no}, \text{span}\}$. Such learnable linear projections ensure that a specific subspace per question type exists. We additionally incorporate different temperature hyperparameters τ_k based on each question type in Equation 1 (see the ablation Table 3 for the effect of using different temperature per question type). The model then can leverage this information as solving QA tasks might require knowledge about question-type relations.

The dimensions of the proposed W_k tend to be large (because of the dimensions of the transformer's hidden-layers).[2] Hence, following Barkan et al. (2020), we introduce new non-square linear projections instead of using W_k:

$$sim_k(s,q) = \frac{s_k^\top q_k}{\|s_k\|\|q_k\|}, \quad (4)$$

where we set $s_k{=}W_k^S s, q_k{=}W_k^Q q$, and W_k^S (or W_k^Q) is the learnable matrix that projects s (or q) into a lower dimension, in the k^{th} question-type sub-space. Note that a correct answer projected into the sub-space of the wrong question type is considered as wrong. We can further integrate this into the denominator of the term inside $\mathcal{L}_C$, namely, the representations of s and q under the incorrect question types.

While contrastive learning has recently emerged as a useful tool for improving NLP tasks, in both finetuning and pretraining phases (Gunel et al., 2021), prior work applied this technique primarily for *inter-sequence* representation enhancement, e.g., better sentence representations in a Siamese setup (Iter et al., 2020; Luo et al., 2020; Gao et al., 2021a), whereas we apply this learning method in an *intra-sequence* manner, aiming at improving mutual contextual representations, considering the same sequence.

4 Experimental Setup and Results

In this section, we provide details about the experiments that we conducted and their outcomes.

4.1 Experimental Setup

As previously stated, we conjecture that our additional contrastive loss term can be useful for learning meaningful representations for modeling long-range relationships, which are typically required solving tasks that involve long-document or multi-document processing.

Accordingly, in order to demonstrate the advantage of our method, we performed an extensive evaluation using the recent QAsper dataset (Dasigi et al., 2021) and the well-known HotpoQA dataset (Yang et al., 2018a), which share the input form (see Section 2).

QAsper (Dasigi et al., 2021) is a long-document QA dataset which was built over academic papers, where NLP practitioners were recruited to generate questions following the title and the abstract of a particular paper, as well as creating the the correct evidence and answers to those questions out of the entire paper. More than half of the examples in QAsper require collecting evidences from multiple paragraphs in the given paper. For this benchmark, $\mathcal{L}_{QA}$, represents the sum of the teacher-forced text generation and paragraph classification loss functions, in a multi-task training setup.

We replicated the experiments described in Dasigi et al. (2021); we finetuned the LED-base model,[3] and evaluated it on the question answering and evidence selection tasks. For further details about the structure of the dataset and finetuning, see Appendix A.1.

HotpotQA (Yang et al., 2018a) introduced the task of multihop question answering (in the reading comprehension setting), where the inputs are

[2] We end up with $W_k \in \mathbb{R}^{768\times768}$ for base-sized models and $W_k \in \mathbb{R}^{1024\times1024}$ for large-sized models.

[3] According to Dasigi et al. (2021), LED-base outperforms LED-large over QAsper.

Input	Extractive		Abstractive		Yes/No		Unanswerable		Evidence		Overall	
	Dev.	Test	Dev.	Test	Dev.	Test	Dev.	Test	Dev.	Test	Dev.	Test
LED (2020)	26.1	31.0	16.6	15.8	67.5	70.3	28.6	26.2	23.9	29.9	29.1	32.8
+ $\mathcal{L}_T$ (Δ)	+0.2	+0.2	+0.8	+1.0	+1.6	+0.2	+1.5	+1.9	+1.0	+0.7	+0.9	+0.7

Table 1: Performance change when applying our additional loss $\mathcal{L}_T$ to the LED SOTA model on QAsper (the metric is Answer-F_1).

	Model	Ans	Sup	Joint
base size	Longformer-base (2020)	74.5	83.9	64.5
	+$\mathcal{L}_T$ (Δ)	+0.9	+0.2	+1.0
	CDLM-base (2021)	74.7	86.3	66.3
	+$\mathcal{L}_T$ (Δ)	+0.9	+0.3	+0.2
large size	Longformer-large (2020)	81.3	88.3	73.2
	+$\mathcal{L}_T$ (Δ)	+0.3	+0.1	+0.9
	CDLM-large (2021)	81.3	89.1	73.8
	+$\mathcal{L}_T$ (Δ)	+0.3	+0.6	+0.8

Table 2: HotpotQA-distractor results (F_1) for the dev set. We use the "base" and "large" model size results of CDLM and the Longformer for direct comparison. Ans: answer span, Sup: Supporting facts.

	Joint	Δ
CDLM-large + $\mathcal{L}_T$ (Eq. 4)	74.6	
– using a single τ parameter for all the question types	74.2	-0.4
– $sim(\cdot)$ is the dot product	73.1	-1.5
– $sim(\cdot)$ is the cosine similarity	74.0	-0.6
– $sim_k(\cdot)$ is the bilinear distance (Eq. 3)	73.7	-0.9
– w/o soft negatives (incorrect question type negatives)	74.2	-0.4

Table 3: Similarity function ablation results (Joint F_1) of CDLM and our loss term on the HotpotQA-distractor dev set.

a question and multiple paragraphs from various related and non-related documents. A model is queried to extract answer spans and evidence sentences, where it should handle challenging questions, that answering them requires finding and reasoning over multiple supporting documents. For this benchmark, $\mathcal{L}_{QA}$, represents the the standard cross-entropy answer extraction loss.

We replicated the experiment described in Beltagy et al. (2020); Caciularu et al. (2021), where we used the Longformer model and CDLM[4] as the backbone language models for this task, as they support long-sequence inputs. Since the CDLM work provided a base-sized model only, we pretrained[5] a larger version of the CDLM model, and hence we used both the base and large versions of the Longformer and CDLM. For further details about the structure of the dataset and finetuning, see Appendix A.2.

For both QAsper and HotpoQA, we performed a similar grid search for determining the hyperparameters of the contrastive loss (see more details in Appendix B).

4.2 Results

Main results We adopted the same evaluation metrics of original works, from Dasigi et al. (2021) and Beltagy et al. (2020) (where the reader is refered for more details). Tables 1 and 2 illustrate the evaluation results over the QAsper and HotpoQA datasets, respectively. We show the performance difference of adding our additional loss term with "+$\mathcal{L}_C$". The addition of our contrastive loss term exhibits the best performance among all the models and benchmarks, clearly demonstrating its consistent advantage. Note that maximizing the question-evidence similarity resulted also in the evidence detection improvement – see the "Evidence" and the "Sup" metrics in Table 1 and in Table 2, respectively.

Ablations Table 3 demonstrates the ablation study results for evaluating the effectivness of our design decisions. Using a constant temperature parameter for all question types, as well as using a different, degenerated similarity functions exhibits a lower performance results. Notably, by the last row in Table 3 we see that treating correct answers that are projected to the wrong question type as negatives seems to slightly improve the results. Overall, the ablation study shows the advantage of using Eq. 4 as a similarity function that provides an fine-grained, expressive modeling for question types, since it handles each question type in its own subspace. An additional theoretical justification to our contrastive learning is provided in (Gao et al., 2021b), where we can imply that our loss term improves the uniformity and therefore the expressiveness of the question and evidence representations.

[4]CDLM was shown to be an effective long-range encoder model for HotpotQA.

[5]we used the code from https://github.com/aviclu/CDLM/tree/main/pretraining

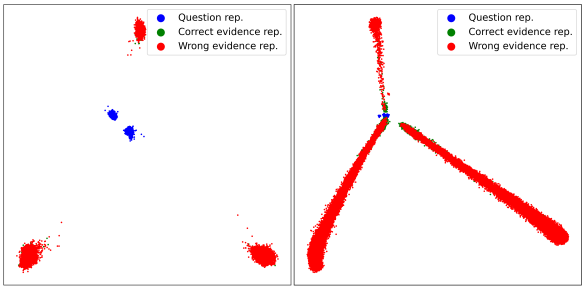

Figure 2: PCA plots of the learned question and answer token embeddings on the HotpotQA validation set, comparing early training epochs results (left) and results after model convergence (right). The wrong evidence representations correspond to both wrong evidences, or correct evidences using the wrong question type projections (our soft negatives).

Analysis We apply PCA over the relevant normalized token representations of the validation data of HotpotQA (i.e., the question and answers representations in Fig. 1), and depicted them in 2. The projected representations of the correct and wrong answers are equally distributed at the beginning of the training (left figure). After several epochs when the model converged (right figure), the answer representations' manifold got closer to the questions' representations (in terms of radial distance). Each beam in the figure corresponds to a different question type (there are 3 in HotpotQA). The correct evidence representations (green dots) are the closest among the whole answer representations, confirming that our additional contrastive loss term generalizes and maximizes the question-evidence similarity.

5 Conclusion

In this work, we propose an effective sequence-level contrastive loss for improving the performance of long-range transformers in solving QA tasks that require reasoning over long contexts. We demonstrate consistent improvement when using our approach on three different models over two different benchmarks. In future work, we would like to explore variations of our proposed supervised loss on other long-context tasks, such as long document summarization.

Acknowledgments

We thank the Semantic Scholar research team at AI2 for fruitful discussions and helpful feedback. The work described herein was supported in part by grants from Intel Labs, the Israel Science Foundation grant 1951/17, and the Israeli Ministry of Science and Technology.

References

Oren Barkan, Avi Caciularu, and Ido Dagan. 2020. Within-between lexical relation classification. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3521–3527, Online. Association for Computational Linguistics.

Iz Beltagy, Matthew E Peters, and Arman Cohan. 2020. Longformer: The long-document transformer. *arXiv preprint arXiv:2004.05150*.

Avi Caciularu, Arman Cohan, Iz Beltagy, Matthew Peters, Arie Cattan, and Ido Dagan. 2021. CDLM: Cross-document language modeling. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 2648–2662, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Tianqi Chen, Bing Xu, Chiyuan Zhang, and Carlos Guestrin. 2016. Training deep nets with sublinear memory cost. *arXiv preprint*, abs/1604.06174.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. 2020. A simple framework for contrastive learning of visual representations. In *International Conference on Machine Learning (ICML)*.

Xinlei Chen and Kaiming He. 2021. Exploring simple siamese representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15750–15758.

Pradeep Dasigi, Kyle Lo, Iz Beltagy, Arman Cohan, Noah A. Smith, and Matt Gardner. 2021. A dataset of information-seeking questions and answers anchored in research papers. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4599–4610, Online. Association for Computational Linguistics.

Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2021a. SimCSE: Simple contrastive learning of sentence embeddings. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6894–6910, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2021b. SimCSE: Simple contrastive learning of sentence embeddings. In *Empirical Methods in Natural Language Processing (EMNLP)*.

Beliz Gunel, Jingfei Du, Alexis Conneau, and Veselin Stoyanov. 2021. Supervised contrastive learning for

pre-trained language model fine-tuning. In *International Conference on Learning Representations (ICLR)*.

Jeremy Howard and Sebastian Ruder. 2018. Universal language model fine-tuning for text classification. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 328–339, Melbourne, Australia. Association for Computational Linguistics.

Dan Iter, Kelvin Guu, Larry Lansing, and Dan Jurafsky. 2020. Pretraining with contrastive sentence objectives improves discourse performance of language models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4859–4870, Online. Association for Computational Linguistics.

Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. 2020. Supervised contrastive learning. *Advances in Neural Information Processing Systems (NeurIPS)*.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *ArXiv*, abs/1412.6980.

Fuli Luo, Pengcheng Yang, S. Li, Xuancheng Ren, and X. Sun. 2020. Capt: Contrastive pre-training for learning denoised sequence representations. *ArXiv*, abs/2010.06351.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. 2018. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*.

Richard Yuanzhe Pang, Alicia Parrish, Nitish Joshi, Nikita Nangia, Jason Phang, Angelica Chen, Vishakh Padmakumar, Johnny Ma, Jana Thompson, He He, and Samuel R. Bowman. 2021. QuALITY: question answering with long input texts, yes!

Yi Tay, Mostafa Dehghani, Samira Abnar, Yikang Shen, Dara Bahri, Philip Pham, Jinfeng Rao, Liu Yang, Sebastian Ruder, and Donald Metzler. 2021. Long range arena : A benchmark for efficient transformers. In *International Conference on Learning Representations (ICLR)*.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. 2018a. 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, Brussels, Belgium. Association for Computational Linguistics.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William W. Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. 2018b. HotpotQA: A dataset for diverse, explainable multi-hop question answering. In *EMNLP*.

Manzil Zaheer, Guru Guruganesh, Kumar Avinava Dubey, Joshua Ainslie, Chris Alberti, Santiago Ontanon, Philip Pham, Anirudh Ravula, Qifan Wang, Li Yang, et al. 2020. Big Bird: Transformers for longer sequences. *Advances in Neural Information Processing Systems (NeurIPS)*.

A Datasets and Finetuning Details

In this section, we provide details, regrading finetuning and hyper-parameter configuration, over the benchmarks we used during our experiments.

A.1 QAsper

Since some of the questions included in QAsper are not answerable, we apply our contrastive loss only over examples that are answerable and contain at least one evidence sentence.

We train all models using the Adam optimizer (Kingma and Ba, 2014) and a triangular learning rate scheduler (Howard and Ruder, 2018) with 10% warmup. To determine number of epochs, peak learning rate, and batch size, we performed manual hyperparameter search on a subset of the training data. We searched over {1, 3, 5} epochs with learning rates {$1e^{-5}$, $3e^{-5}$, $5e^{-5}$, $9e^{-5}$}, and found that smaller batch sizes generally work better than larger ones. Our final configuration was 10 epochs, peak learning rate of $5e^{-5}$, and batch size of 2, which we used for all reported experimental settings. When handling full text, we use gradient checkpointing (Chen et al., 2016) to reduce memory consumption. We run our experiments on a single RTX 8000 GPU, and each experiment takes 30–60 minutes per epoch.

A.2 HotpotQA

We used the HotpotQA-distractor dataset (Yang et al., 2018a). Each example in the dataset is includes a question and 10 paragraphs from different documents, extracted from Wikipedia. Two gold paragraphs include the relevant information for properly answering the question, mixed and shuffled with eight distractor paragraphs (for the full dataset statistics, see Yang et al. (2018a)). There are two goals for this task: detecting the supporting facts, i.e., evidence sentences, as well as extraction of the correct answer span.

For preparing the data for training and evaluation, we follow the same finetuning scheme of the CDLM (Caciularu et al., 2021) and the Longformer (Beltagy et al., 2020); For each example, we concatenate the question

and all the 10 paragraphs in one long context. We particularly use the following input format with special tokens and our document separators: "[CLS] [q] question [/q] ⟨doc-s⟩⟨t⟩ title_1 ⟨/t⟩ ⟨s⟩ $\text{sent}_{1,1}$ ⟨/s⟩ ⟨s⟩ $\text{sent}_{1,2}$ ⟨/s⟩ ⟨/doc-s⟩ ... ⟨t⟩ ⟨doc-s⟩ title_2 ⟨/t⟩ $\text{sent}_{2,1}$ ⟨/s⟩ ⟨s⟩ $\text{sent}_{2,2}$ ⟨/s⟩ ⟨s⟩ ..." where [q], [/q], ⟨t⟩, ⟨/t⟩, ⟨s⟩, ⟨/s⟩, [p] are special tokens representing, question start and end, paragraph title start and end, and sentence start and end, respectively. The new special tokens were added to the models vocabulary and randomly initialized before task finetuning. We use global attention to question tokens, paragraph title start tokens as well as sentence tokens. The model's structure is taken from Beltagy et al. (2020).

AS in Beltagy et al. (2020); Caciularu et al. (2021), we finetune our models for 5 epochs, using a batch size of 32, learning rate of 1e-4, 100 warmup steps. Finetuning on our models took $\sim$6 hours per epoch, using four 48GB RTX8000 GPUs for finetuning our models.

B Contrastive Loss Details

In this section, we provide the details for reproducing our contrastive term, which is relevant for both QAsper and HotpotQA.

We searched over $d \times \{d, \frac{d}{2}, \frac{d}{4}, \frac{d}{8}\}$ to determine the linear projections' dimensions, where d is the model's hidden layer representation dimension (it depends on the size of the model).

In order to determine the temperature hyperparameter τ, we searched over $\{0.2, 0.4, 0.6, 0.8, 1.0\}$ per question type (if applicable).

We also applied dropout, with a rate of $p = 0.1$, over the linear projections, which consistently improved the results over all the benchmarks.