MAFiD: Moving Average Equipped Fusion-in-Decoder for Question Answering over Tabular and Textual Data

Sung-Min Lee[†], **Eunhwan Park**[†], **Daeryong Seo**[‡], **Donghyeon Jeon**[‡], **Inho Kang**[‡], **Seung-Hoon Na**[†*]

[†]Jeonbuk National University [‡]NAVER Corporation

{cap1232, judepark, nash}@jbnu.ac.kr, {daeryong.seo, donghyeon.jeon, once.ihkang}@navercorp.com

Abstract

Transformer-based models for question answering (QA) over tables and texts confront a "long" hybrid sequence over tabular and textual elements, causing long-range reasoning problems. To handle long-range reasoning, we extensively employ a fusion-in-decoder (FiD) and exponential moving average (EMA), proposing a Moving Average Equipped Fusion-in-Decoder (**MAFiD**). With FiD as the backbone architecture, MAFiD combines various levels of reasoning: *independent encoding* of homogeneous data and *single-row* and *multi-row heterogeneous reasoning*, using a *gated cross attention layer* to effectively aggregate the three types of representations resulting from various reasonings. Experimental results on HybridQA indicate that MAFiD achieves state-of-the-art performance by increasing exact matching (EM) and F1 by 1.1 and 1.7, respectively, on the blind test set.

1 Introduction

While most studies have focused on text question answering (QA), where unimodal textual passages are provided as a source of evidence for an answer (Joshi et al., 2017; Yang et al., 2018; Rajpurkar et al., 2018; Welbl et al., 2018; Dua et al., 2019; Karpukhin et al., 2020; Zhu et al., 2021b; Pang et al., 2022), realistic questions often need to refer to "heterogeneous" evidences based on both tabular and textual contents to generate an answer, motivating researchers to address *table-and-text* QA (Chen et al., 2020; Wenhu Chen, 2021; Talmor et al., 2021; Zhu et al., 2021a; Nakamura et al., 2022).

Among the various tasks for table-and-text QA, we address the *multi-hop* table-and-text QA described in HybridQA (Chen et al., 2020), which is a large-scale table-and-text QA dataset focusing on the multi-hop reasoning across tabular and textual contents to extract an answer.

However, a table usually contains a nontrival number of rows and relevant passages; thus linearization of all relevant heterogeneous contents easily exceeds the maximum length limit for transformers, thereby causing *long range reasoning* problems.

To address long range reasoning, we present a novel encoder-decoder model that deploys fusion-in-decoder (FiD) (Izacard and Grave, 2021) and exponential moving average (EMA) (Ma et al., 2022), the Moving Average Equipped Fusion-in-Decoder (**MAFiD**). Armed with FiD as the backbone architecture, MAFiD combines various levels of reasoning:

- **Independent encoding of homogeneous data**, which independently encodes tabular and textual contents separately for each row, without being fused in the encoder step. Inherited from FiD, the resulting encoded representations are jointly fused in the decoder, which significantly reduces the computational time required for self-attention, thereby allowing us to use a longer sequence as an input for the encoder.
- **Single-row heterogeneous reasoning** (also referred to as *single-row reasoning*), which performs in-depth interaction between tabular and textual contents per row; it first concatenates the tabular and textual representations for each row and then applies the "self-attention" layer over the concatenated sequence. Thus, single-row heterogeneous reasoning is performed in a restricted manner only on heterogeneous contents within a specific row.
- **Multi-row heterogeneous reasoning** (also referred to as *multi-row reasoning*), which performs light interaction across tabular and textual contents of "multiple" rows based on the EMA layer; it concatenates the heteroge-

*Corresponding author

Findings of the Association for Computational Linguistics: EACL 2023, pages 2337–2344

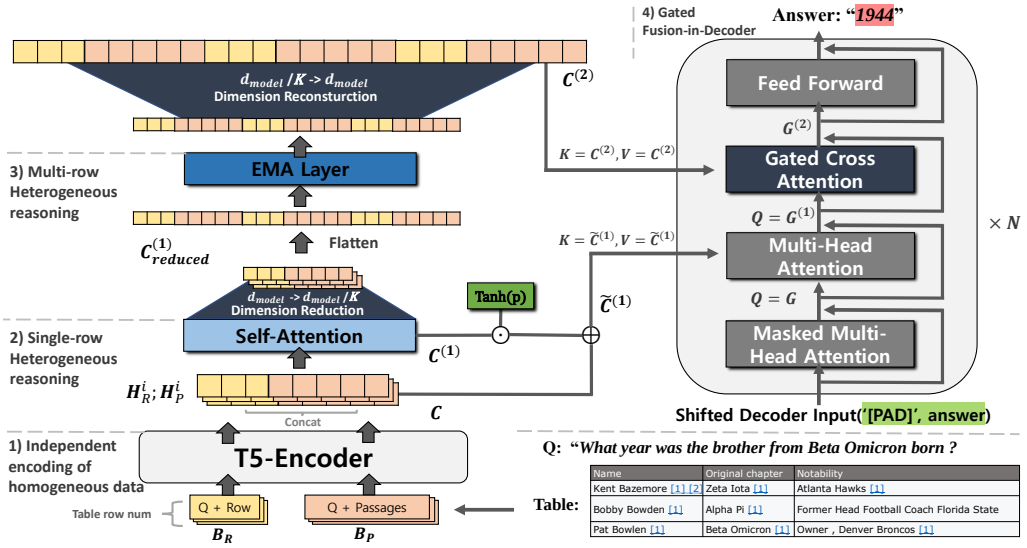

Figure 1: The overall neural architecture of the proposed MAFiD: 1) Independent encoding applies T5's encoder on the tabular and textual blocks in i-th row, separately (i.e., $b_R^i \in \mathcal{B}_R$ and $b_P^i \in \mathcal{B}_P$) and the resulting contextual representations are concatenated to obtain the i-th row's heterogeneous representation, $\boldsymbol{C}_i$ (Eq. (1)). 2) Single-row reasoning performs the row-specific cross-modal interaction by applying the single-head attention over $\boldsymbol{C}_i$ to generate $\boldsymbol{C}_i^{(1)}$ (Eq. (2)). 3) Multi-row reasoning preforms the between-row cross-modal interaction by applying the low-dimensional EMA over the long hybrid sequence $\boldsymbol{C}^{(1)}$ (Eq. (3)) to produce the $\boldsymbol{C}_i^{(2)}$ (Eq. (4)). 4) The gated FiD aggregates all types of representations of $\boldsymbol{C}$ (Eq. (4)), $\boldsymbol{C}^{(1)}$, and $\boldsymbol{C}^{(2)}$ using the gated cross-attention layer to finally yield the decoder's contextual representation $\boldsymbol{G}^{(2)}$ (Eq. (5)) which is fed to generate an output token.

neous contents of all rows in a table to obtain a "long" hybrid sequence and then applies the EMA layer over the resulting long sequence to produce aggregated representation. To process a long sequence more efficiently, we further propose a *low-dimensional* EMA, which additionally performs a *dimensionality reduction* and *reconstruction*.

In the decoder, we further propose the use of a *gated cross-attention layer* to effectively aggregates the aforementioned three representations resulting from various reasoning, motivated by the work of (Alayrac et al., 2022).

Our contributions are summarized as follows: 1) We propose MAFiD, which augments FiD with EMA and the gated cross-attention layer, thus effectively combining various types of reasoning. 2) We propose a low-dimensional EMA to efficiently process long sequences for table-and-text QA. 3) The proposed MAFiD achieves state-of-the-art performance on the HybridQA dataset.

2 Related Works

Recently, many datasets such as HybridQA (Chen et al., 2020), OTT-QA (Wenhu Chen, 2021), MultiModalQA (Talmor et al., 2021), HybriDialogue (Nakamura et al., 2022), and TAT-QA (Zhu et al., 2021a) have been presented for table-and-text QA. Various works on table-and-text QA have enhanced "pretraining" to strengthen the cross-modal matching and numerical reasoning, by learning on tables and texts jointly (Herzig et al., 2020; Yin et al., 2020) and exclusively on tables (Iida et al., 2021).

To handle the long range reasoning on table-and-text QA, early works employed "efficient" transformers based on *sparse attention* with selective attention masks, such as the LongFormer (Beltagy et al., 2020) in the work of (Huang et al., 2022) and ETC (Ainslie et al., 2020) in the work of (Wenhu Chen, 2021). MATE (Eisenschlos et al., 2021b) uses *structure*-based sparse attention that attends to either rows or columns in a given table. Recently, *truncation*-based approaches have been employed in MITQA (Kumar et al., 2021) where the *passage filter* module is additionally applied such that only the filtered passages are used as textual contents of a table's row.

Compared to these existing approaches, which rather limitedly reduce the computational cost in the encoder part, MAFiD significantly lightens the encoder part by minimizing the interaction between different rows and instead fuses the encoded representations in the decoder part under the framework of FiD. Equipped with the low-dimensional EMA,

MAFiD only performs the "shallow" interaction across rows, thus mostly maintaining the efficiency of the interaction-less encoder.

3 Moving Average Equipped Fusion-in-Decoder

Figure 1 shows the overall neural architecture of the proposed MAFiD model, which combines three types of representation. Here, we present the details of the MAFiD components.

3.1 Problem Definition

Suppose that $\mathcal{B}_R$ and $\mathcal{B}_P$ are a set of tabular and textual blocks in a given table, where $b^i_R \in \mathcal{B}_R$ indicates the tabular block for the i-th row (i.e., a list of its cells), $b^i_P \in \mathcal{B}_P$ indicates the textual block for the i-th row (i.e., a set of its linked passages), and $L = |\mathcal{B}_R| = |\mathcal{B}_P|$ is the number of rows in a table. Given question q, the goal is to generate a correct answer by considering $\mathcal{B}_R$ and $\mathcal{B}_P$ as heterogeneous evidence.

3.2 Independent Encoding of Homogeneous Data: the Basic Encoder for FiD

Following the independent encoding in FiD (Izacard and Grave, 2021), independent encoding linearizes tabular and textual blocks into a sequence independently and concatenates each of them with q as follows:

$$\mathsf{row}^i = [q;\ [\mathsf{SEP}];\ b^i_R], \mathsf{psg}^i = [q;\ [\mathsf{SEP}];\ b^i_P]$$

where ; is the concatenation operator. The tabular and textual sequences are then fed into the encoder of T5 independently and concatenated as follows:

$$\begin{aligned} \boldsymbol{H}^i_R &= \mathsf{T5\text{-}enc}(\mathsf{row}^i) \in \mathbb{R}^{|\mathsf{row}^i| \times d_{model}} \\ \boldsymbol{H}^i_P &= \mathsf{T5\text{-}enc}(\mathsf{psg}^i) \in \mathbb{R}^{|\mathsf{psg}^i| \times d_{model}} \\ \boldsymbol{C}_i &= [\boldsymbol{H}^i_R; \boldsymbol{H}^i_P] \in \mathbb{R}^{(|\mathsf{row}^i|+|\mathsf{psg}^i|) \times d_{model}} \end{aligned} \quad (1)$$

where $|\mathsf{x}|$ is the length of sequence x and d_{model} is the dimensionality of the encoder of T5.

3.2.1 Single-row Heterogeneous Reasoning

In single-row reasoning, we perform an in-depth interaction between tabular and textual blocks for each row, b^i_R and b^i_P, using self-attention as follows:

$$\boldsymbol{C}^{(1)}_i = \mathsf{SHA}\left(\boldsymbol{C}_i, \boldsymbol{C}_i, \boldsymbol{C}_i\right) \quad (2)$$

where $\boldsymbol{C}^{(1)}_i \in \mathbb{R}^{(|\mathsf{row}^i|+|\mathsf{psg}^i|) \times d_{model}}$ and SHA is the *single*-head attention defined in Eq. (6) in Appendix D.

3.3 Multi-row Heterogeneous Reasoning by the Low-dimensional EMA

In multi-row reasoning, we first concatenate the contextual representations of all tabular and textual blocks as follows:

$$\boldsymbol{C}^{(1)} = [\boldsymbol{C}^{(1)}_1; \cdots; \boldsymbol{C}^{(1)}_L] \quad (3)$$

where $\boldsymbol{C}^{(1)} \in \mathbb{R}^{N \times d_{model}}$, provided $N = \sum_i \left(|\mathsf{row}^i| + |\mathsf{psg}^i|\right)$ for notational convenience.

We then adopt the low-dimensional EMA as a variant of EMA using dimensionality reduction and reconstruction based on linear layers as follows:

$$\begin{aligned} \boldsymbol{C}^{(1)}_{reduced} &= \mathsf{Linear}\left(\boldsymbol{C}^{(1)}\right) \\ \boldsymbol{C}^{(2)}_{reduced} &= \mathsf{EMA}\left(\boldsymbol{C}^{(1)}_{reduced}\right) \\ \boldsymbol{C}^{(2)} &= \mathsf{Linear}\left(\boldsymbol{C}^{(2)}_{reduced}\right) \end{aligned}$$

where $\boldsymbol{C}^{(1)}_{reduced}, \boldsymbol{C}^{(2)}_{reduced} \in \mathbb{R}^{N \times d_{model}/K}$, $\boldsymbol{C}^{(2)} \in \mathbb{R}^{N \times d_{model}}$, Linear is a linear layer, and EMA is the damped EMA of (Ma et al., 2022) defined in Appendix E.

3.4 Gated Fusion-in-Decoder

In the decoder, we first concatenate the row-wise representations of independent encoding before feeding them to the FiD as follows:

$$\boldsymbol{C} = [\boldsymbol{C}_1; \cdots; \boldsymbol{C}_L] \quad (4)$$

In the decoder, we aggregate all representations of $\boldsymbol{C}$ (Eq. (1) and (4)), $\boldsymbol{C}^{(1)}$ (Eq. (2) and (3)) $\boldsymbol{C}^{(2)}$ (Eq. (4) using a gating mechanism similar to that of (Alayrac et al., 2022) as follows:

$$\begin{aligned} \tilde{\boldsymbol{C}}^{(1)} &= \boldsymbol{C} + \tanh(p) \odot \boldsymbol{C}^{(1)} \\ \boldsymbol{G}^{(1)} &= \mathsf{MHA}(\boldsymbol{G}, \tilde{\boldsymbol{C}}^{(1)}, \tilde{\boldsymbol{C}}^{(1)}) \\ \boldsymbol{G}^{(2)} &= \boldsymbol{G}^{(1)} + \\ &\tanh(q) \odot \mathsf{MHA}(\boldsymbol{G}^{(1)}, \boldsymbol{C}^{(2)}, \boldsymbol{C}^{(2)}) \end{aligned} \quad (5)$$

where $\boldsymbol{G} \in \mathbb{R}^{|N^{(dec)}| \times d_{model}}$ is the output of the masked multi-head attention in the decoder part, $|N^{(dec)}|$ is the sequence length of the decoder input, $\tanh(\cdot)$ is the tanh function, p and q are learnable parameters, and $\boldsymbol{G}^{(1)}, \boldsymbol{G}^{(2)} \in \mathbb{R}^{|N^{(dec)}| \times d_{model}}$.

4 Experiments

4.1 Experimental Setup

The details of the implementation and experiment setup is presented in Appendix A.

	Table				Passage				Total			
	Dev		Test		Dev		Test		Dev		Test	
	EM	F1	EM	F1	EM	F1	EM	F1	EM	F1	EM	F1
HYBRIDER	51.5	58.6	52.1	59.3	40.5	47.9	38.1	46.3	43.7	50.9	42.5	50.2
HYBRIDER-Large	54.3	61.4	56.2	63.3	39.1	45.7	37.5	44.4	44.0	50.7	43.8	50.6
DocHopper	-	-	-	-	-	-	-	-	47.7	55.0	46.3	53.3
POINTR + TAPAS	68.1	73.9	67.8	73.2	62.9	72.0	62.0	70.9	63.3	70.8	62.7	70.0
POINTR + MATE	68.6	74.2	66.9	72.3	62.8	71.9	62.8	71.9	63.4	71.0	62.8	70.2
MITQA	68.1	73.3	68.5	74.4	**66.7**	**75.6**	64.3	73.3	65.5	72.7	64.3	71.9
Ours	**69.4**	**75.2**	**68.5**	**74.9**	66.5	75.5	**65.7**	**75.3**	**66.2**	**74.1**	**65.4**	**73.6**
Human	-	-	-	-	-	-	-	-	-	-	88.2	93.5

Table 1: Comparison results on the dev and blind test dataset in HybridQA. The best is bolded text.

	Table		Passage		Total	
	EM	F1	EM	F1	EM	F1
Ours	**68.48**	**74.92**	**65.75**	**75.34**	**65.38**	**73.56**
w/o Multi-row reasoning	67.44	73.74	65.50	75.23	64.86	73.08
w/o Multi-row, Single-row reasoning	41.97	49.46	60.20	69.42	51.46	59.86
w/o Single-row tanh gate	67.21	73.44	64.86	74.82	64.45	72.75
w/o Multi-row tanh gate	67.58	73.96	66.43	75.47	65.46	73.29
w/o Single-row, Multi-row tanh gate	66.09	72.51	64.81	75.22	64.01	72.65

Table 2: Ablation study on blind test dataset in HybridQA. "w/o Single-row tanh gate" and "w/o Multi-row tanh gate" correspond to the runs of fixing $\tanh(p) = 1$ and $\tanh(q) = 1$ in Eq. (5), respectively.

4.2 Baselines

In the experiment, we compare MAFiD and other baseline systems on HybridQA as follows:

- **HYBRIDER** (Chen et al., 2020) employs a sparse passage retriever (i.e., TF-IDF and a longest-substring matching) to find relevant cells and performs the reasoning step consisting of the ranking, the hop, and the reading comprehension models to extract an answer.
- **DocHopper** (Sun et al., 2021) uses the "iterative hierarchical attention" to retrieve short or long contents in a multi-step navigational manner.
- **POINTR + (TAPAS or MATE)** (Herzig et al., 2020; Eisenschlos et al., 2021a). POINTR extends the cell with its entity description and performs a two-stage method that consists of "cell selection" and "passage reading" steps. Either TAPAS (Herzig et al., 2020) or MATE (Kumar et al., 2021) is considered as a transformer encoder.
- **MITQA** (Kumar et al., 2021) uses the pipelined module including a retriever, a reader, and a joint row+span reranker, etc., being trained using the multi-instance distant supervision approach.

4.3 Main Results

As summarized in Table 1, MAFiD shows the state-of-the-art performance by increasing EM and F1 by 1.1 and 1.7 over MITQA (Kumar et al., 2021) on the blind test set. It is observed that MAFiD outperforms POINTR + (TAPAS or MATE) (Herzig et al., 2020; Eisenschlos et al., 2021a) that replies on the pretrained TAPAS, likely indicating that the long-range reasoning needs to be importantly handled on HybridQA, thus motivating the literature to go towards "reasoning"-enhanced pretraining in addition to the existing self-supervised tasks.

4.4 Ablation Studies

Single-row & Multi-row Heterogeneous Reasoning. To examine the effect of single-row and multi-row reasoning, we further evaluate MAFiD by removing either or both reasonings. As shown in Table 2, MAFiD without multi-row reasoning slightly decreases EM and F1 by 1.04 and 1.18, respectively. Importantly, MAFiD without both reasonings significantlly deteriorates the performance of EM and F1 by 13.92 and 13.7, respectively. The results confirm that the cross-modal interaction should be performed at least within a specific row, whereas the between-row interaction is somehow effectively proceeded by the proposed EMA module, although its effect is not large.

Single-row & Multi-row Tanh Gating. We further examine the effect of using the gated flows

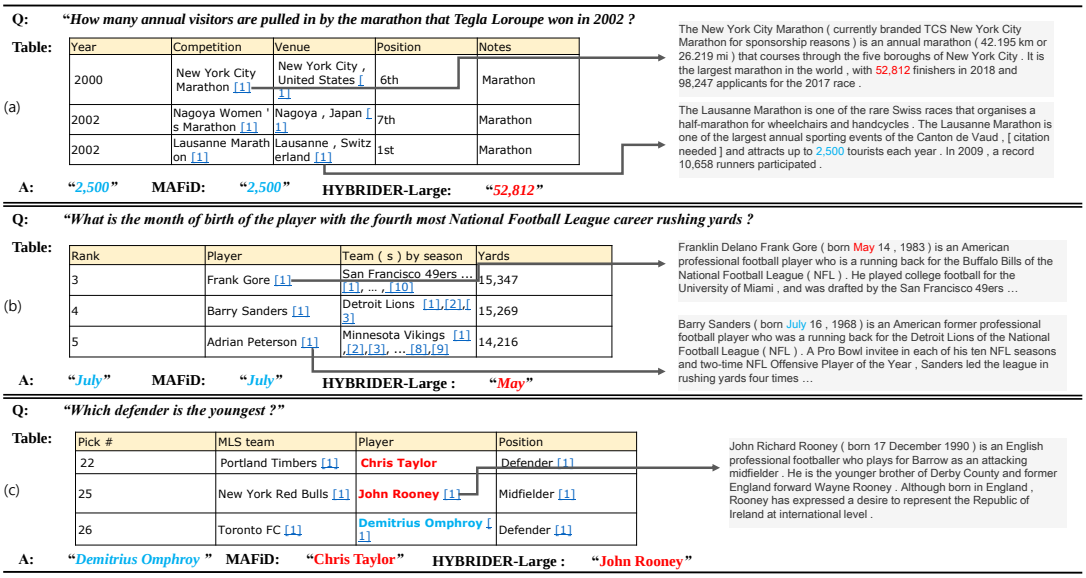

Figure 2: Illustrating examples of HYBRIDER-Large (Chen et al., 2020) and MAFiD in HybridQA.

	Total			
	Dev		Test	
	EM	F1	EM	F1
EMA	**66.2**	**74.1**	**65.4**	**73.6**
sliding window attention	65.7	73.3	65.3	73.1
Human	-	-	88.2	93.5

Table 3: Comparison results on thd dev and blind test sets in HybridQA between EMA and the sliding window attention of (Beltagy et al., 2020) for long-range reasoning.

	Total			
	Dev		Test	
	EM	F1	EM	F1
original rows	**66.2**	**74.1**	**65.4**	**73.6**
permuted rows	51.5	59.4	51.1	59.2
Human	-	-	88.2	93.5

Table 4: Comparison results of MAFiD on HybridQA between the case using original rows and that with permuted rows for tabular contents.

by evaluating MAFiD by fixing $\tanh(p) = 1$ and $\tanh(q) = 1$ without being learned in Eq. (5. In particular, MAFiD without the single-row tanh gate ($\tanh(p) = 1$) slightly decreases EM and F1 by approximately 1$\tilde{1}$.5, indicating that the gated FiD is helpful for further improvements.

Impact of EMA. To examine the impact of EMA for multi-row reasoning, we evaluate the sliding window attention of (Beltagy et al., 2020) as the baseline to handle long-range reasoning. As shown in Table 3, the use of EMA increases F1 and EM by 0.1 and 0.5, respectively, suggesting that EMA is more helpful for promoting the enhanced local sequence representation.

Impact of Sequential Order. To examine the impact of using the sequential order of rows in tabular contents, Table 4 further shows the results of a variant of MAFiD by randomly permuting rows in tabular contents both for training and inference, referred to as "permuted row", comparing to the original case; the results strongly indicate that keeping original row orders is important for MAFiD.

4.5 Error Analysis

Figure 2 shows some illustrating QA examples in HybridQA dataset comparing the results of HYBRIDER-Large (Chen et al., 2020) and MAFiD; (a)-(b) require only keyword matching and numerical comparison, where HYBRIDER is failed; (c) requires sophisticated multi-hop reasoning across table rows and passages where both MAFID and HYBRIDER are incorrect.

5 Conclusion

In this paper, we address long range-reasoning for the multi-hop table-and-text QA and propose MAFiD, which extends FiD by equipping EMA and the gated cross-attention layer for the encoder and decoder parts, respectively, to design an effective way of combining various types of encoded representations. The experimental results on HybridQA showed that the proposed MAFiD achieved state-of-the-art performances in both the development and blind test sets. In future work, we will extend MAFiD to open-domain table-and-text QA and explore a unified approach that integrates single-row and multi-row reasoning.

Limitations

This paper proposes the use of EMA under FiD to tractably perform multi-row reasoning; however, EMA simply puts strong weights on nearby contexts, thus performing a restricted type of the long-range reasoning. Thus, our EMA-based method heavily depends on the sequential order of tables and texts, so hardly performing matching between long-distance but semantically related tokens in a long hybrid sequence. In using EMA, the current limitation of our method is that we only used the "damped EMA" of MEGA (Ma et al., 2022), which is only one of the basic components in MEGA. Because MEGA additionally combines the single-head attention unit over a long sequence, using MEGA could allow us to handle long-distance semantic matching. In the future work, it will be valuable to explore such extensions of EMA, such as MEGA, to strengthen the long-range reasoning.

In MAFiD, we show that EMA can be applied over a maximally long sequence in HybridQA (Chen et al., 2020). However, when moving to OTT-QA (Wenhu Chen, 2021), EMA cannot be naively applicable over retrieved long sequences without any truncation, because the size of a retrieved set of tables and texts is significantly larger than that of HybridQA. Given that OTT-QA more closely matches the real-world situation, the EMA-based reasoning should be extended further by incorporating retrieval and selection modules. Thus, our current framework needs to be extended further to handle open-domain table-and-text QA, under the retriever-reader framework.

Acknowledgements

This work was supported by NAVER Corporation. We would like to thank all anonymous reviewers for their valuable comments and suggestions.

References

Joshua Ainslie, Santiago Ontanon, Chris Alberti, Vaclav Cvicek, Zachary Fisher, Philip Pham, Anirudh Ravula, Sumit Sanghai, Qifan Wang, and Li Yang. 2020. ETC: Encoding long and structured inputs in transformers. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 268–284, Online. Association for Computational Linguistics.

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katie Millican, Malcolm Reynolds, Roman Ring, Eliza Rutherford, Serkan Cabi, Tengda Han, Zhitao Gong, Sina Samangooei, Marianne Monteiro, Jacob Menick, Sebastian Borgeaud, Andrew Brock, Aida Nematzadeh, Sahand Sharifzadeh, Mikolaj Binkowski, Ricardo Barreira, Oriol Vinyals, Andrew Zisserman, and Karen Simonyan. 2022. Flamingo: a visual language model for few-shot learning.

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

Wenhu Chen, Hanwen Zha, Zhiyu Chen, Wenhan Xiong, Hong Wang, and William Yang Wang. 2020. HybridQA: A dataset of multi-hop question answering over tabular and textual data. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1026–1036, Online. Association for Computational Linguistics.

Dheeru Dua, Yizhong Wang, Pradeep Dasigi, Gabriel Stanovsky, Sameer Singh, and Matt Gardner. 2019. DROP: A reading comprehension benchmark requiring discrete reasoning over paragraphs. 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)*, pages 2368–2378, Minneapolis, Minnesota. Association for Computational Linguistics.

Julian Eisenschlos, Maharshi Gor, Thomas Müller, and William Cohen. 2021a. MATE: Multi-view attention for table transformer efficiency. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7606–7619, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Julian Martin Eisenschlos, Maharshi Gor, Thomas Müller, and William W. Cohen. 2021b. Mate: Multi-view attention for table transformer efficiency.

Jonathan Herzig, Pawel Krzysztof Nowak, Thomas Müller, Francesco Piccinno, and Julian Eisenschlos. 2020. TaPas: Weakly supervised table parsing via pre-training. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4320–4333, Online. Association for Computational Linguistics.

Junjie Huang, Wanjun Zhong, Qian Liu, Ming Gong, Daxin Jiang, and Nan Duan. 2022. Mixed-modality representation learning and pre-training for joint table-and-text retrieval in openqa.

Hiroshi Iida, Dung Thai, Varun Manjunatha, and Mohit Iyyer. 2021. TABBIE: Pretrained representations of tabular data. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3446–3456, Online. Association for Computational Linguistics.

Gautier Izacard and Edouard Grave. 2021. Leveraging passage retrieval with generative models for open domain question answering. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 874–880, Online. Association for Computational Linguistics.

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.

Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense passage retrieval for open-domain question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6769–6781, Online. Association for Computational Linguistics.

Vishwajeet Kumar, Saneem A. Chemmengath, Yash Gupta, Jaydeep Sen, Samarth Bharadwaj, and Soumen Chakrabarti. 2021. Multi-instance training for question answering across table and linked text. *CoRR*, abs/2112.07337.

Xuezhe Ma, Chunting Zhou, Xiang Kong, Junxian He, Liangke Gui, Graham Neubig, Jonathan May, and Zettlemoyer Luke. 2022. Mega: Moving average equipped gated attention. *arXiv preprint arXiv:2209.10655*.

Kai Nakamura, Sharon Levy, Yi-Lin Tuan, Wenhu Chen, and William Yang Wang. 2022. HybriDialogue: An information-seeking dialogue dataset grounded on tabular and textual data. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 481–492, Dublin, Ireland. Association for Computational Linguistics.

Richard Yuanzhe Pang, Alicia Parrish, Nitish Joshi, Nikita Nangia, Jason Phang, Angelica Chen, Vishakh Padmakumar, Johnny Ma, Jana Thompson, He He, and Samuel Bowman. 2022. QuALITY: Question answering with long input texts, yes! In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5336–5358, Seattle, United States. Association for Computational Linguistics.

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, Melbourne, Australia. Association for Computational Linguistics.

Haitian Sun, William W. Cohen, and Ruslan Salakhutdinov. 2021. End-to-end multihop retrieval for compositional question answering over long documents. *CoRR*, abs/2106.00200.

Alon Talmor, Ori Yoran, Amnon Catav, Dan Lahav, Yizhong Wang, Akari Asai, Gabriel Ilharco, Hannaneh Hajishirzi, and Jonathan Berant. 2021. Multimodal{qa}: complex question answering over text, tables and images. In *International Conference on Learning Representations*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Johannes Welbl, Pontus Stenetorp, and Sebastian Riedel. 2018. Constructing datasets for multi-hop reading comprehension across documents. *Transactions of the Association for Computational Linguistics*, 6:287–302.

Eva Schlinger William Wang William Cohen Wenhu Chen, Ming-wei Chang. 2021. Open question answering over tables and text. *Proceedings of ICLR 2021*.

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, Brussels, Belgium. Association for Computational Linguistics.

Pengcheng Yin, Graham Neubig, Wen-tau Yih, and Sebastian Riedel. 2020. TaBERT: Pretraining for joint understanding of textual and tabular data. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8413–8426, Online. Association for Computational Linguistics.

Fengbin Zhu, Wenqiang Lei, Youcheng Huang, Chao Wang, Shuo Zhang, Jiancheng Lv, Fuli Feng, and Tat-Seng Chua. 2021a. TAT-QA: A question answering benchmark on a hybrid of tabular and textual content in finance. 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 3277–3287, Online. Association for Computational Linguistics.

Fengbin Zhu, Wenqiang Lei, Chao Wang, Jianming Zheng, Soujanya Poria, and Tat-Seng Chua. 2021b. Retrieving and reading: A comprehensive survey on open-domain question answering. *CoRR*, abs/2101.00774.

A Implementation Details

We used the HybridQA dataset, which is a large-scale multi-hop question answering dataset over tabular and textual data. Table 5 presents detailed

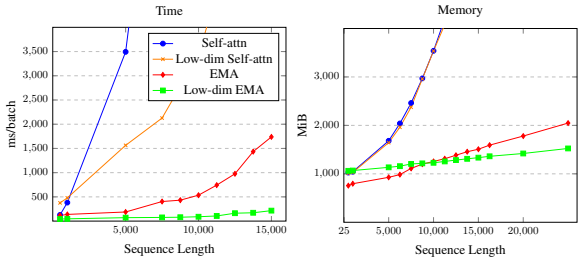

Figure 3: Comparison of memory and time complexities of self-attention, EMA, and their low-dimensional versions.

statistics of the HybridQA dataset. We used the T5-base[1] transformer encoder-decoder model as a pre-trained language model. Additional parameters were initialized from $\mathcal{N}(0, 0.2)$ and the bias was set to 0. To prepare the input of the encoder, we set the maximum sequence lengths of the tabular and textual blocks to 300 and 800, respectively. All the models were trained using the AdamW optimizer with a learning rate of $1e-4$. All trainings were conducted for 3 epochs, and the random seed number was fixed at 42 to reproduce the results. The batch size was four with two accumulation steps. Training was conducted for 1.5 days on 4 NVIDA Quadro RTX 8000. For answer generation, we employed a greedy decoding method. For the evaluation, we used exact matching (EM) and F1 metrics. Evaluations were conducted every 500 steps on the dev dataset and the best model with the highest EM score was chosen.

The maximum number of rows in a given table in HybridQA is 20, that is, $L \leq 20$. In our experiments, L was fixed at 20 by adding padding sequences when the number of rows was less than 20. The rate of the dimensionality reduction for the low-dimensional EMA, i.e., K, was fixed to 6.

B Dataset Statistics

Split	Train	Dev	Test	Total
In-Passage	35,215	2,025	20,45	39,285
In-Table	26,803	1,349	1,346	29,498
Missing	664	92	72	828
Total	62,682	3,466	3,463	69,611

Table 5: Hybrid QA dataset statistics. In-Passage and In-Table indicate that exact answer span is founded in a passage or table. Missing is the exact answer span not founded in given source.

[1] https://huggingface.co/t5-base

Split	min	mean	max	Count
Table	137	763	8,298	3,466
Row	14	48	1,454	55,036
Passages per row	2	656	10,797	55,036

Table 6: Statistics of length of tokenized sequence on dev dataset. *'Passages per row'* is the length of all concatenated passages in a row.

C An Example of Linearized Blocks

Specifically the i-th table row block is defined as follows:

$$b_R^i = [\mathsf{TITLE}]\ t\ [\mathsf{SECTITLE}]\ t_{(sec)}$$
$$[\mathsf{ROW}]\ h^1\ \text{'is'}\ v^{i,1}\ [\mathsf{SEP}]\ \cdots\ [\mathsf{SEP}]\ h^N\ \text{'is'}\ v^{i,N}$$

where h and v are the head and value, t and $t_{(sec)}$ are the title and section title, respectively.

Passage block is defined as follows:

$$b_P^i = [\mathsf{PSG}]\ psg_{linked}^{i,1}\ [\mathsf{PSG}]\ \cdots\ [\mathsf{PSG}]\ psg_{linked}^{i,N}$$

where psg is a linked passage at row.

D Single-head and Multi-head Attentions

The single-head and multi-head attentions (Vaswani et al., 2017) are defined as follows:

$$\begin{aligned}
&\mathsf{SHA}\left(\boldsymbol{Q}, \boldsymbol{K}, \boldsymbol{V}\right) = [head_1]\boldsymbol{W}^O, \\
&\mathsf{MHA}\left(\boldsymbol{Q}, \boldsymbol{K}, \boldsymbol{V}\right) = [head_1; \cdots; head_h]\boldsymbol{W}^O, \\
&head_i = \mathsf{Attn}\left(\boldsymbol{Q}\boldsymbol{W}_i^Q, \boldsymbol{K}\boldsymbol{W}_i^K, \boldsymbol{V}\boldsymbol{W}_i^V\right) \quad (6)
\end{aligned}$$

E EMA

EMA has widely been applied in time series and long range text modeling. Among variants of EMA presented in the work of (Ma et al., 2022), we employed the "damped EMA"[2] for proceeding long range sequences. The damped EMA is based on the recursive calculation for computing the output $\mathbf{Y}$ as follows:

$$\mathbf{y_t} = \alpha \odot \mathbf{x_t} + (1 - \alpha \odot \delta) \odot \mathbf{y_{t-1}} \quad (7)$$

where $\odot$ is the element-wise multiplication operator, $\alpha \in (0, 1)^d$ is a decaying factor for making exponentially decreasing effects from older tokens, $\delta \in (0, 1)^d$ is the damping factor, and α and δ are learnable weight parameters. This recursive computation of EMA can be efficiently implemented as the convolution and the fast Fourier transforms.

[2] This is different from the further extended multi-dimensional damped EMA (Ma et al., 2022)