FOVEA TRANSFORMER: EFFICIENT LONG-CONTEXT MODELING WITH STRUCTURED FINE-TO-COARSE ATTENTION

*Ziwei He, Jian Yuan, Le Zhou, Jingwen Leng, Bo Jiang**

Shanghai Jiao Tong University
{ziwei.he, yuanjian, zhoule1217, leng-jw, bjiang}@sjtu.edu.cn

ABSTRACT

The quadratic complexity of self-attention in Transformers has hindered the processing of long text. To alleviate this problem, previous works have proposed to sparsify the attention matrix, taking advantage of the observation that crucial information about a token can be derived from its neighbors. These methods typically combine one or another form of local attention and global attention. Such combinations introduce abrupt changes in contextual granularity when going from local to global, which may be undesirable. We believe that a smoother transition could potentially enhance model's ability to capture long-context dependencies. In this study, we introduce Fovea Transformer, a long-context focused transformer that addresses the challenges of capturing global dependencies while maintaining computational efficiency. To achieve this, we construct a multi-scale tree from the input sequence, and use representations of context tokens with a progressively coarser granularity in the tree, as their distance to the query token increases. We evaluate our model on three long-context summarization tasks[1]. It achieves state-of-the-art performance on two of them, and competitive results on the third with mixed improvement and setback of the evaluation metrics.

Index Terms— Efficient Transformer, Long-Range Modeling, Structured Attention, Abstractive Summarization

1. INTRODUCTION

Transformers [1] have become the fundamental architecture in natural language processing (NLP). However, the quadratic time and space complexity of self-attention has hindered the application of mainstream pretrained transformer models [2, 3, 4, 5] to tasks requiring long texts. The past few years has witnessed considerable efforts to relieve this limitation. Existing works generally fall into two categories.

One line of research respects the length limitation by partitioning a long input into smaller segments and feeding them separately to models pretrained on short texts [6, 7]. These approaches are able to reuse various pretrained language models instead of training from scratch, albeit at the cost of breaking the integrity of long texts and hence hurting the performance.

The other line of research targets at sparsifying the attention matrix, by following some predefined patterns [8, 9, 10]. Most notable is local attention, where a token only attends to a small range of neighboring tokens [11, 12, 9, 10]. Although previous work has suggested that crucial information about a token can be mostly derived from its neighbors [11, 12], simply ignoring the other tokens may still hurt the performance on downstream tasks. Thus local attention is typically complemented by global attention, where input tokens also attend to shared global tokens that are each a coarse-grained representation of a long segment or even the entire sequence of input tokens [11, 12, 9, 10]. Such local-global combinations have successfully allowed transformer models to process inputs with up to 16k tokens and outperform methods from the first category [9, 10]. Nevertheless, the abrupt change in granularity when going from local to global may be undesirable, and a smoother transition could potentially enhance the ability to effectively handle long-context input.

Hence we propose *Fovea Transformer*, a long-context focused transformer that allows every token to attend to the entire sequence with structured fine-to-coarse granularities. To this end, we construct a multi-scale tree representation of the input sequence through a bottom-up process, where a leaf node corresponds to an input token, and an internal node is a coarser representation of its children. When computing the attention of a token, we use representations of context tokens with a progressively coarser granularity higher up in the tree, as their distance to the query token increases. This allows every token to attend to the entire context while minimizing computational requirements. As this attention mechanism draws inspiration from the acuity around the fovea of human eyes, we dub it Fovea Attention and the resulting model Fovea Transformer.

Note that BPT [13] shares a similar idea with Fovea Transformer. It constructs a multi-scale tree for each input sequence, but in a top-down fashion through recursive binary partitioning. The tree is then converted to a graph and node representations are updated by graph attention. From the perspective of token interactions, the conversion to graph attention introduces information staleness across graph layers, as a result of the two-hop separation between input tokens in the graph. We avoid this issue by a new design that shuns graph attention. In addition, we divide tokens into blocks, and apply our proposed attention pattern on top of it for better time and memory efficiency. Note also that BPT is not validated on long-context tasks.

Fovea Transformer does not introduce any new parameters into the original transformer. It provides an inexpensive drop-in replacement for the attention mechanism in existing transformer architectures. To avoid heavy and expensive pretraining, we warm-start our model on LongT5 [10] for all the experiments.

We test Fovea Transformer on three datasets with long context. It achieves state-of-the-art performances on two of them, and competitive results on the third with mixed improvement and setback of the evaluation metrics.

Our main contributions are summarized as follows:

- We propose Fovea Transformer, a long-context focused transformer that addresses the challenges of capturing global de-

*Bo Jiang is the corresponding author.

This work is supported in part by the National Natural Science Foundation of China under Grant No. 62072302.

[1]Our code is publicly available at: *https://github.com/ZiweiHe/Fovea-Transformer*

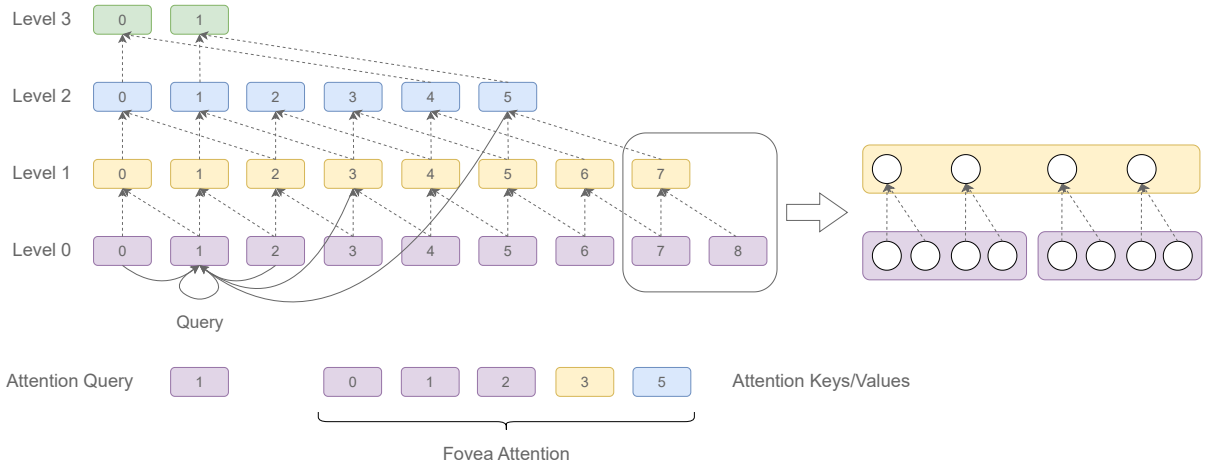


Fig. 1: Illustration for tree construction and fovea attention.

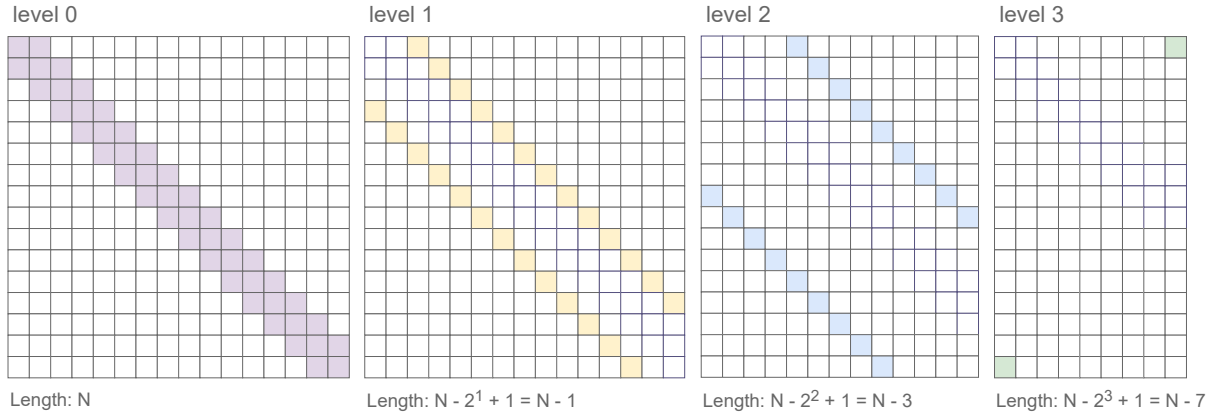


Fig. 2: Examples of building blocks for fovea attention. Each subplot indicates the attention matrix masks between query and key for each level of the tree (the colors correspond). Suppose there are originally N blocks in the input, the number of blocks from higher level decreases through the tree merging. Colored entries means active of attention, white entries indicates absence instead.

pendencies while maintaining computational efficiency.

- We proposed fovea attention mechanism, an inexpensive drop-in replacement for vanilla attention in existing transformer architectures.
- We validate the effectiveness of Fovea Transformer, positioning proposed approach as a state-of-the-art solution for long-context summarization tasks.

2. Fovea Transformer

We propose Fovea Transformer with fovea attention, which is a special self-attention that attends to further tokens with a progressively coarser granularity. To achieve this, we first construct a tree-structured, multi-scale sequence representation from the input tokens through a bottom-up process. Tokens are iteratively grouped together so that higher-level nodes in the tree represent coarser granularity (Figure 1). This constructed tree essentially provides a mapping between all input tokens and their corresponding multi-resolution representations. Next, for each query token, our proposed fovea attention collects nodes from the tree to form the key and value components. The constructed components then participate in the calculation of self-attention, allowing for smooth transitions in context granularity between short- and long-range context.

2.1. Constructing the Multi-scale Representation Tree

To generate the multi-scale representations as keys and values for fovea attention, we first organize a tree-structured representation of the sequence by iteratively averaging the embeddings of a consecutive set of input tokens. As the node in the tree climbs higher (indicated by a higher *level* in Figure 1), it represents an increasingly larger number of tokens.

Formally, for a certain i-th node at level q, noted as $u_{q,i}$, it is an average of tokens from the i-th to the $(i + 2^q - 1)$-th.

$$u_{q,i} = \frac{1}{2^q} \sum_{k=i}^{i+2^q-1} e_k \tag{1}$$

where e_k stands for the embedding of the k-th token in the input sequence. We can figure that $u_{q,i}$ depends on the embeddings of leaf tokens in the range of $[i, i + 2^q - 1]$, which we call the *receptive*

field of the node. Since we are computing the nodes iteratively in a bottom-up manner, the average operation can be accelerated by averaging over two of its children, i.e.,

$$u_{q,i} = \frac{1}{2}(u_{q-1,i} + u_{q-1,i+2^{q-1}}) \tag{2}$$

It is worth noting the distinction between this resulting tree and a standard binary tree. A standard binary tree has significantly fewer internal nodes compared to the one we present here. This is due to the large overlap of perceptive field between neighboring internal nodes in our tree. Our design can effectively ensure that the subsequent fovea attention can accurately attend to the specific range of tokens for every query token position.

In addition, to accelerate the tree construction process, in practice we divide the input sequence into equal-size blocks of tokens before we organize them into the tree. As a result, each leaf node in the tree stands for a *block* of neighboring tokens, and computing the mean value in equation 2 involves computing a block of mean values in parallel. The right part of Figure 1 illustrates how the block-wise averaging is performed with an example block size of 4. This practical modification makes the algorithm more hardware-friendly since coalesced memory transactions are much more efficient in many hardware accelerators.

2.2. Fovea Attention

For each query token, instead of attending to every token in the sequence, the fovea attention selects as key and values a set of nodes in the tree we built in Section 2.1 in a way that the receptive fields of the nodes concatenate back-to-back till covering the whole sequence without overlaps. In addition, the fovea attention selects nodes by following the principle that it attends to lower-level nodes while its corresponding receptive field is closer to the query, and progressively climbs up the tree as the nodes draw further away.

To elaborate, at level q, starting next to the endpoint of the receptive field of previous level $q-1$, the fovea attention selects k nodes to represent a fragment of non-overlapping actual input tokens that are of length $2^q k$. Then, it ascends to the upper level for nodes that are farther away, continuing until it reaches the end of the sequence. This selection process stretches symmetrically on both sides of the query token (see Figure 1).

Formally, for the i-th query token, on its right-hand side, the fovea attention selects the following k nodes in level q into the tree:

$$\{u_{q,i+k(2^q-1)+1+j2^q} \mid j \in [0, k-1]\} \tag{3}$$

where $j \in \mathbb{N}$. Note that the fovea attention also selects k nodes on the left side of the query token in the same manner.

If we put everything together, the whole set $\mathcal{S}_i$ of nodes in all levels that are selected by fovea attention for the i-th query token can be written as the union of all the nodes selected at every level:

$$\begin{aligned}
&\mathcal{S}_i = \\
&level\ 0: \quad \{u_{0,i}\} \cup \{u_{0,i+j} \mid j \in [1,k] \cup [-k,-1]\} \cup && (4)\\
&level\ 1: \quad \{u_{1,i+k+1+2j} \mid j \in [0,k-1]\} \cup \\
&\qquad\qquad \{u_{1,i-k-2-2j} \mid j \in [0,k-1]\} \cup && (5)\\
&\qquad\qquad \cdots \\
&level\ q: \quad \{u_{q,i+k(2^q-1)+1+j2^q} \mid j \in [0,k-1]\} \cup \\
&\qquad\qquad \{u_{q,i-k(2^q-1)-2^q-j2^q} \mid j \in [0,k-1]\} && (6)
\end{aligned}$$

where $j \in \mathbb{N}$. Note that the node $u_{a,b}$ should exist in the tree, so all nodes in the above set should subjects to the following constraint:

$$s.t. \quad \forall u_{a,b}, \quad a,b \in \mathbb{N} \quad and \quad 0 \le b \le N - 2^a \tag{7}$$

where N is the number of blocks in the sequence.

Despite the scatter in notations, it is pretty straightforward to understand the fovea attention by visualizing the attention matrix mask. In Figure 2, we vertically list the N queries and horizontally concatenate all nodes in the tree for every query. The nodes being attended on by the corresponding query are filled in color and the white entries are not selected by fovea attention. Despite the large size of the attention weight matrix, only a small portion of entries need to be computed. It enjoys a complexity of $O(N(\log N))$.

In addition, our proposed fovea attention does not introduce any new parameters, so it is friendly to pretrained models and could act as a drop-in replacement for most existing pretrained transformers.

3. EXPERIMENTS

In this section, we evaluate Fovea Transformer on text summarization tasks that involves extreme long sequences.

3.1. Implementation Details

We build Fovea Transformer with the Huggingface library, instead of pretraining from scratch, we warm-start finetune Fovea Transformer from the publicly released LongT5-xl ($\sim$3B) [2] checkpoint. We follow the same configurations of LongT5-xl including the hidden size, number of layers, block size, etc.. The results use input length 16384 and output length 512 for all datasets. For simplicity, we set $k = 1$. We use batch size of 128, learning rate of $\sim$0.001 with polynomial scheduler for all the experiments. We evaluate Fovea Transformer on 8 Nvidia A800 GPUs.

3.2. Datasets

We evaluate Fovea Transformer on three abstractive summarization datasets, we only use publicly available datasets from Huggingface[3], to make sure the reproducibility of our work. Table. 1 provides a statistical analysis about the dataset size and input/output length. More details are listed as follows.

Dataset	#Source	#Target	#Examples		
			Train	Valid	Test
Multi-News	2103	264	44972	5622	5622
WCEP-10	3866	28	8158	1020	1022
PubMed	3224	214	119924	6633	6658

Table 1: Dataset statistics.

Multi-News [14] A large-scale news dataset which summarizing multiple news documents into a human-written summary, additionally, each summary is professionally written by editors.
WCEP [15] A dataset for multi-document summarization, the input are collected leveraging the Wikipedia Current Events Portal (WCEP), the output are neutral human-written summaries of news events. There are at most 100 documents within each cluster in the original dataset, in our paper, we use the WCEP-10 [16], a version

[2]https://huggingface.co/google/long-t5-tglobal-xl
[3]https://huggingface.co/datasets

that have all the duplicates removed and only keep up 10 most relevant documents for each cluster.

PubMed [17] The task consists of scientific papers collected from *PubMed.com* where the long-form document content is used as input and their abstracts are ground-truth summaries.

3.3. Results

We evaluate model performances in terms of the Rouge scores: Rouge-1(R1), Rouge-2(R2) and Rouge-L(RL) [18] for all the datasets. We compare proposed model with various approaches which achieve significant results on Multi-News, WCEP and PubMed: BigBird [11], Longformer [9], LongT5 [10], PRIMER [16], GoSum [19], BART-LS [20], BART-Long-Graph [21], SPADE [22], UPER [23] and LSG [24].

The quantitative results are summarized in Tables 2, 4 and 3, which indicate that Fovea Transformer is able to effectively model long-term dependencies from up to 16k tokens. Fovea Transformer achieves state-of-the-art performance on Multi-News and WCEP. On PubMed, Fovea Transformer gives competitive results, with better R1 score but worse R2 and RL scores than the competing methods.

Model	R1	R2	RL
BART-Long-Graph	49.24	18.99	23.97
LongT5-xl	48.20	19.40	24.90
PRIMER	49.90	21.10	25.90
SPADE	-	19.63	23.70
Fovea Trans. (ours)	**50.32**	**21.50**	**26.62**

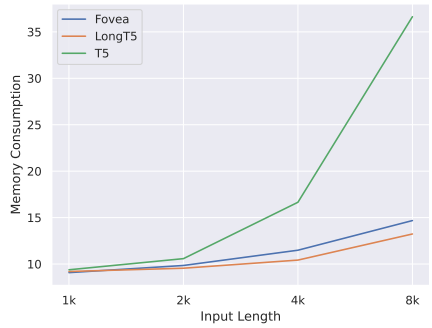
Table 2: Results for Multi-News. The Rouge scores are taken from their respective papers.

Model	R1	R2	RL
PRIMER	**46.1**	25.2	37.9
UPER	41.4	18.7	33.8
LSG-BART	46.0	24.2	37.4
Fovea Trans.	**46.1**	**25.3**	**38.1**

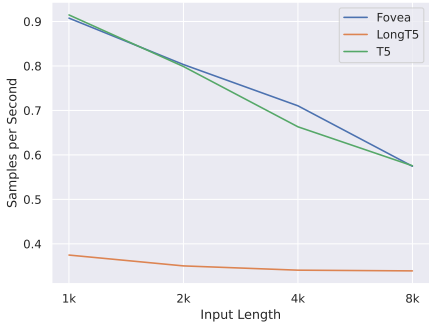
Table 3: Results for WCEP-10. The scores of UPER and LSG-BART are taken from [23], scores of PRIMER are from [16].

Model	R1	R2	RL
BigBird	46.32	20.65	42.33
Longformer	47.00	20.20	42.90
GoSum	49.83	23.56	45.10
LongT5-xl	50.23	**24.76**	**46.67**
BART-LS	50.30	24.30	46.30
Fovea Trans. (ours)	**50.41**	24.65	46.08

Table 4: Results for PubMed. The Rouge scores of BigBird, LongT5-xl are taken from [25], the rest are taken from their respective papers.



(a) Memory Consumption



(b) Samples per Second

Fig. 3: The training speed and GPU memory consumptions of Fovea Transformer, LongT5 and T5. All the models are in *large* size with input length of 1k, 2k, 4k and 8k. Measurements taken with batch size 1 on 1×4 A100-40 GPUs.

3.4. Analysis

We quantitatively evaluate the training speed and memory consumption of Fovea Transformer, LongT5 and T5 on Multi-News dataset, considering various input length. The results are summarized in Figure. 3, we can see that the memory usage is comparable among the models for shorter lengths, but the difference becomes significant as we increase the sequence length. Generally, LongT5 and Fovea Transformer have a much smaller memory footprint compared to the regular transformer T5. However, Fovea Transformer trains significantly faster than LongT5. We did not extend the input length any further as the trends are easily observable.

4. CONCLUSION

In this work, we propose Fovea Transformer, a long-context focused transformer that addresses the challenges of capturing global dependencies while maintaining computational efficiency. The experimental results and analysis validate the effectiveness of our proposed approach, positioning Fovea Transformer as a state-of-the-art solution for long-context abstractive summarization tasks.

5. REFERENCES

[1] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia

Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[2] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[3] Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer, "Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension," *arXiv preprint arXiv:1910.13461*, 2019.

[4] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov, "Roberta: A robustly optimized bert pretraining approach," *arXiv preprint arXiv:1907.11692*, 2019.

[5] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu, "Exploring the limits of transfer learning with a unified text-to-text transformer," *The Journal of Machine Learning Research*, vol. 21, no. 1, pp. 5485–5551, 2020.

[6] Chuhan Wu, Fangzhao Wu, Tao Qi, and Yongfeng Huang, "Hi-transformer: hierarchical interactive transformer for efficient and effective long document modeling," *arXiv preprint arXiv:2106.01040*, 2021.

[7] Maor Ivgi, Uri Shaham, and Jonathan Berant, "Efficient long-text understanding with short-text models," *Transactions of the Association for Computational Linguistics*, vol. 11, pp. 284–299, 2023.

[8] Aurko Roy, Mohammad Saffar, Ashish Vaswani, and David Grangier, "Efficient content-based sparse attention with routing transformers," *Transactions of the Association for Computational Linguistics*, vol. 9, pp. 53–68, 2021.

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

[10] Mandy Guo, Joshua Ainslie, David Uthus, Santiago Ontanon, Jianmo Ni, Yun-Hsuan Sung, and Yinfei Yang, "LongT5: Efficient text-to-text transformer for long sequences," in *Findings of the Association for Computational Linguistics: NAACL 2022*, Seattle, United States, July 2022, pp. 724–736, Association for Computational Linguistics.

[11] Manzil Zaheer, Guru Guruganesh, Kumar Avinava Dubey, Joshua Ainslie, Chris Alberti, Santiago Ontanon, Philip Pham, Anirudh Ravula, Qifan Wang, Li Yang, et al., "Big bird: Transformers for longer sequences," *Advances in neural information processing systems*, vol. 33, pp. 17283–17297, 2020.

[12] Joshua Ainslie, Santiago Ontanon, Chris Alberti, Vaclav Cvicek, Zachary Fisher, Philip Pham, Anirudh Ravula, Sumit Sanghai, Qifan Wang, and Li Yang, "Etc: Encoding long and structured inputs in transformers," *arXiv preprint arXiv:2004.08483*, 2020.

[13] Zihao Ye, Qipeng Guo, Quan Gan, Xipeng Qiu, and Zheng Zhang, "Bp-transformer: Modelling long-range context via binary partitioning," *arXiv preprint arXiv:1911.04070*, 2019.

[14] Alexander Fabbri, Irene Li, Tianwei She, Suyi Li, and Dragomir Radev, "Multi-news: A large-scale multi-document summarization dataset and abstractive hierarchical model," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, Florence, Italy, July 2019, pp. 1074–1084, Association for Computational Linguistics.

[15] Demian Gholipour Ghalandari, Chris Hokamp, Nghia The Pham, John Glover, and Georgiana Ifrim, "A large-scale multi-document summarization dataset from the wikipedia current events portal," *arXiv preprint arXiv:2005.10070*, 2020.

[16] Wen Xiao, Iz Beltagy, Giuseppe Carenini, and Arman Cohan, "Primera: Pyramid-based masked sentence pre-training for multi-document summarization," *arXiv preprint arXiv:2110.08499*, 2021.

[17] Arman Cohan, Franck Dernoncourt, Doo Soon Kim, Trung Bui, Seokhwan Kim, Walter Chang, and Nazli Goharian, "A discourse-aware attention model for abstractive summarization of long documents," in *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, New Orleans, Louisiana, June 2018, pp. 615–621, Association for Computational Linguistics.

[18] Chin-Yew Lin, "ROUGE: A package for automatic evaluation of summaries," in *Text Summarization Branches Out*, Barcelona, Spain, July 2004, pp. 74–81, Association for Computational Linguistics.

[19] Junyi Bian, Xiaodi Huang, Hong Zhou, and Shanfeng Zhu, "Gosum: Extractive summarization of long documents by reinforcement learning and graph organized discourse state," *arXiv preprint arXiv:2211.10247*, 2022.

[20] Wenhan Xiong, Anchit Gupta, Shubham Toshniwal, Yashar Mehdad, and Wen-tau Yih, "Adapting pretrained text-to-text models for long text sequences," *arXiv preprint arXiv:2209.10052*, 2022.

[21] Ramakanth Pasunuru, Mengwen Liu, Mohit Bansal, Sujith Ravi, and Markus Dreyer, "Efficiently summarizing text and graph encodings of multi-document clusters," in *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2021, pp. 4768–4779.

[22] Simiao Zuo, Xiaodong Liu, Jian Jiao, Denis Charles, Eren Manavoglu, Tuo Zhao, and Jianfeng Gao, "Efficient long sequence modeling via state space augmented transformer," *arXiv preprint arXiv:2212.08136*, 2022.

[23] Shangqing Tu, Jifan Yu, Fangwei Zhu, Juanzi Li, Lei Hou, and Jian-Yun Nie, "Uper: boosting multi-document summarization with an unsupervised prompt-based extractor," in *Proceedings of the 29th International Conference on Computational Linguistics*, 2022, pp. 6315–6326.

[24] Charles Condevaux and Sébastien Harispe, "Lsg attention: Extrapolation of pretrained transformers to long sequences," in *Pacific-Asia Conference on Knowledge Discovery and Data Mining*. Springer, 2023, pp. 443–454.

[25] Mandy Guo, Joshua Ainslie, David Uthus, Santiago Ontanon, Jianmo Ni, Yun-Hsuan Sung, and Yinfei Yang, "Longt5: Efficient text-to-text transformer for long sequences," *arXiv preprint arXiv:2112.07916*, 2021.