UDC 004.89 https://doi.org/10.20535.2786-8729.5.2024/314724

SCIENTIFIC ARTICLE SUMMARIZATION MODEL WITH UNBOUNDED INPUT LENGTH

Oleksandr Steblianko
National Technical University of Ukraine
"Igor Sikorsky Kyiv Polytechnic Institute", Kyiv, Ukraine
https://orcid.org/0009-0006-5055-0934

Volodymyr Shymkovych *
National Technical University of Ukraine
"Igor Sikorsky Kyiv Polytechnic Institute", Kyiv, Ukraine
https://orcid.org/0000-0003-4014-2786

Peter Kravets
National Technical University of Ukraine
"Igor Sikorsky Kyiv Polytechnic Institute", Kyiv, Ukraine
https://orcid.org/0000-0003-4632-9832

Anatolii Novatskyi
National Technical University of Ukraine
"Igor Sikorsky Kyiv Polytechnic Institute", Kyiv, Ukraine
https://orcid.org/0009-0009-7457-7391

Lyubov Shymkovych
National Technical University of Ukraine
"Igor Sikorsky Kyiv Polytechnic Institute", Kyiv, Ukraine
https://orcid.org/0000-0002-1291-0373

*Corresponding author: v.shymkovych@kpi.ua

In recent years, the exponential growth of scientific literature has made it increasingly difficult for researchers and practitioners to keep up with new discoveries and developments in their fields. Thanks to this, text summarization has become one of the primary tasks of natural language processing. Abstractive summarization of long documents, such as scientific articles, requires large neural networks with high memory and computation requirements. Therefore, it is all the more important to find ways to increase the efficiency of long document summarization models.

The objects of this research are long document summarization transformer models and the Unlimiformer cross-attention modification. The article reviews the basic principles of transformer attention, which constitutes the primary computational expense in transformer models. More efficient self-attention approaches used for long document summarization models are described, such as the global+sliding window attention used by Longformer. The cross-attention mechanism of Unlimiformer, which allows a model to have unbounded input length, is described in detail. The objective of the study is the development and evaluation of a long document summarization model using the Unlimiformer modification. To achieve this goal, a Longformer Decoder-Encoder model pretrained on the arXiv dataset is modified with Unlimiformer cross-attention. This modification can be applied without additional model fine-tuning, avoiding the cost of further training a large sequence length model.

The developed model was evaluated on the arXiv dataset using the ROUGE-1, ROUGE-2 and ROUGE-L metrics. The developed model showed improved results compared to the baseline model, demonstrating the viability of using this approach to improve long document summarization models.

Key words: neural networks, transformers, text summarization, long document summarization, natural language processing, attention.

1. Introduction

As information is available in abundance for every topic on internet, condensing the important information in the form of summary would benefit a number of users. In recent years, the exponential growth of scientific literature has made it increasingly challenging for researchers and practitioners to keep up with new findings and developments in their fields. The sheer volume of published articles – spanning diverse disciplines and journals – poses significant barriers to effective information retrieval and comprehension. There is growing interest among the research community for developing new approaches to automatically summarize the text.

Automatic text summarization system generates a summary, i.e. short length text that includes all the important information of the document. Since the advent of text summarization in 1950s, researchers have been trying to improve techniques for generating summaries so that machine generated summary matches with the human made summary. Summary can be generated through extractive as well as abstractive methods. Abstractive methods are highly complex and require the use of machine learning.

Machine learning models are mathematical models that are capable of learning from a provided dataset. During training, these models adjust their own internal parameters in order to get closer to the correct answer, inferring hidden patterns in the dataset in the process. A properly trained model is capable of correctly answering previously unseen problems that match its training data, as it has successfully learned the associated rules.

Neural networks are a common class of machine learning models that imitate the structure of the human brain. They are made up of several layers of interconnected “neurons”, each of which performs a mathematical operation on the data it receives. Each of those operations is governed by internal variables, allowing the model to achieve a higher level of complexity and understanding. As a result, neural networks are able to learn a wide variety of rules, such as the rules behind human language.

Human language understanding and human language generation are two aspects of natural language processing (NLP) [1]. Understanding human language is more difficult due to the ambiguity of natural language. Speech recognition, summarizing documents, answering questions, speech synthesis, machine translation and other programs use NLP technology [2, 3].

Text summarization – taking a given text and generating a short digest of the most important information in it – has emerged as one of the classic tasks in natural language processing in response to this [4, 5]. Approaches to text summarization have included extractive methods, where the most important parts of the text are extracted; abstractive methods, where a new text is generated on the basis of the given one, usually by an artificial intelligence; and hybrid approaches, such as extracting the most important sentences and writing a new summary based off of them or generating more specific abstractive summaries based off of user queries [6 – 9]. An example of text summarization using NLP shown on Fig. 1.

Abstractive text summarization has garnered a lot of interest with the rise of neural networks capable of processing text, replacing previous heuristic-based approaches [7, 10]. Summarizing scientific articles falls under the umbrella of long document summarization – a subset of text summarization focused on especially lengthy text sequences, such as articles, reports, or even whole books.

To process all that information, long document summarization models must either become much larger or use special methods to process that information more efficiently. Larger neural networks bring with them higher memory and computation requirements, increasing the time needed to train or use them. Additionally, such networks can be hosted on fewer devices, making them less accessible. Therefore, it is all the more important to find ways to increase the efficiency of long document summarization models.

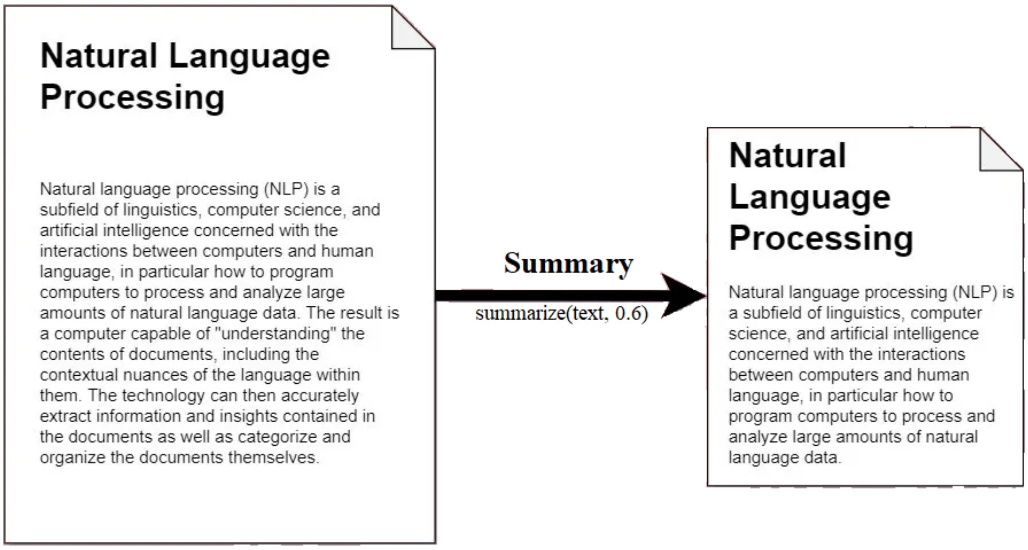

Fig. 1. Text summarization using NLP

The topic of improving the efficiency of long-document summarization models is relevant due to the rapid growth of textual information that needs to be processed across various fields. Efficient summarization approaches enable faster access to key information while reducing the burden on computational resources.

2. Literature review and problem statement

There have been many papers exploring special transformer architectures directed towards efficient long document NLP, largely focusing on changing the transformer's computationally expensive self-attention function.

DANCER (Divide-ANd-ConquER) [11] utilizes a divide-and-conquer approach to long document summarization, splitting the document into several sections that are summarized individually and then combined to form the final summary. This allows it to avoid the problems of long document summarization without having to change the attention structure of the model itself.

Longformer Encoder-Decoder (LED) [12] is a Transformer model that utilizes global+sliding window self-attention during encoding, and full cross-attention in the decoder. This sparseness allows LED to support much larger sequence lengths, offering 4k and 16k length versions of the model.

BigBird [13] is a Transformer model that further develops the attention mechanism – in addition to a sliding window and a set of global tokens, it also uses random attention to sparsely associate tokens that fall outside of the other two patterns. Additionally, BigBird has an input sequence length of 4,096 tokens.

Unlimiformer [14] is a transformer modification that instead alters the transformer's cross-attention function. Like in DANCER, inputs are processed in chunks, but the decoder performs an approximate k-nearest-neighbors search to select the most relevant embeddings from other parts of the input in addition to its own chunk.

Unlimiformer is directed towards tasks with input sequences vastly exceeding the maximum input length of available models, necessitating the usage of chunking to process the data. However, Unlimiformer's approach to selecting the most relevant embeddings may also be beneficial towards models with large sequence lengths while also increasing their effectiveness on unexpectedly long inputs. The effect of Unlimiformer on large sequence length models is underexplored and deserves further study.

3. The aim and objectives of the study

The aim of this study is to develop a large sequence length transformer model in order to explore the effects of the Unlimiformer modification. The usage of this modification on large sequence length models may lead to improvements in model effectiveness without requiring additional fine-tuning.

The objective of the study is the development and evaluation of a long document summarization model using the Unlimiformer modification.

To achieve this goal, the following tasks are set:

– to develop a single-document summarization transformer using LED's 16k length model and Unlimiformer's cross-attention mechanism.

– to evaluate the effectiveness of the developed model on the ROGUE-1, ROGUE-2 and ROUGE-L metrics; to compare it to the base LED 16k model in order to determine the effects of Unlimiformer.

4. The study materials and methods of summarization model development

4.1. Summarization dataset

One of the most important stages of training a neural network is selecting the appropriate dataset. This section details the process of preparing the dataset for the model.

arXiv [15] is a dataset assembled from scientific papers available on the arXiv.org repository, covering a wide variety of topics and containing 228,789 entries. The abstract is used as the expected output, and the rest of the article is given as the input sequence, which avoids the need for human-written summaries in training. The arXiv dataset was made specifically for long document summarization, and is widely used for evaluating the effectiveness of long summarization models.

PubMed is another widely-used scientific article dataset, consisting of articles available on the PubMed.com site. Like arXiv, it is dedicated to generating abstracts based on the contents of the articles. Unlike arXiv, it focuses on biomedical and life science publications, making it less representative of scientific papers as a whole than arXiv. Additionally, PubMed articles are shorter - using a SentencePiece tokenizer, the 90% percentile token length of Arxiv articles is 20,170 and the equivalent number for Pubmed is 8,883. PubMed itself is also smaller than arXiv, containing 133,215 entries [15, 16].

arXiv-Long [17] is a subset of the arXiv dataset created by selecting only the entries with an abstract of at least 350 tokens. This filters out articles with short summaries, allowing training to focus on extended summaries. However, this is not a widely used dataset and thus does not allow for effective comparison with existing models.

SciSummNet, also known as the Yale Scientific Article Summarization Dataset [18] is dataset intended for citation-aware summarization. It combines an article's abstract with its 'citances' – sentences in other papers that cite that paper – as input. Human-written summaries are used as the expected output. Instead of just using the article itself, models trained on this dataset are meant to more accurately reflect an article's impact on the field. An article's impact may be different from what is highlighted in the abstract. However, the process of making hand-written summaries and compiling citations is labour-intensive. SciSummNet only contains 1000 entries.

The Semantic Scholar Network dataset (SSN) [19] expands on SciSummNet's concept by adding on a full citation graph, showing both incoming and outgoing citations for a given article. This is particularly helpful for models meant to help in writing an abstract for a paper, as the yet-to-be-published paper won't have any articles citing it, making SciSummNet's approach pointless. Additionally, SSN is much larger than the Yale dataset, being made up of 140,799 papers and 660,908 citation relationships, with every paper being connected to at least two others.

This paper is concerned with summarization within the bounds of a single article, so the citation-based approaches of SciSummNet and SSN are not fully applicable to the task. The arXiv dataset is widely used in long document summarization research and represents a large number of potential topics, making it the optimal choice for training our network.

Each entry in the arXiv dataset consists of three parts - the abstract of the article, its body, and the section headers. All three are given in the string format. Images and tables are not included in the body string, though their annotations still remain. Citations and formulae are replaced with @xciteN and @xmathN sequences respectively, where N indicates the number of the citation or formula.

The section headers are discarded, and the abstracts and bodies are used unaltered. The article body is given as input to the model, and the abstract is used as the expected output.

4.2. Neural network architecture

An article summarization model requires the ability to process very large amounts of input text, which requires special consideration.

The commonly-used architecture for text processing is the Transformer. Previous recurrent neural network structures have needed to process inputs sequentially and store information in a hidden state, requiring repeated computations to get the forward and backward context. Transformer processes the entire input at once, instead using positional embeddings and attention in order to capture relationships and dependencies between tokens in the sequence. The structure of attention is shown in Fig.2 [20].

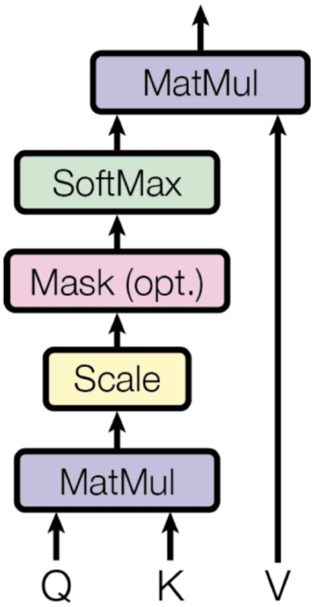

Fig.2. Transformer attention [20]

A transformer includes two kinds of attention:

– Self-attention, used in the encoder layer to associate parts of the input sequence with each other;

– Cross-attention, used to associate the encoded input with the output text being generated.

Attention is one of the primary elements of the transformer, and one of its most computationally expensive. The number of operations required to calculate attention rises quadratically with the length of the input sequence. This gives it a computational complexity of $O(n^2)$. It becomes a serious problem when models have to process thousands of input tokens. Therefore, one of the primary directions in long document summarization has been the development of new, more economical attention mechanisms.

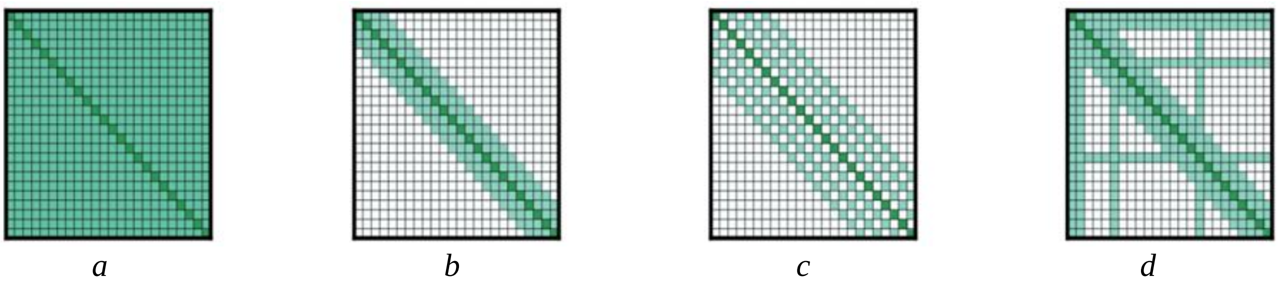

Fig. 3 – Self-attention patterns: *a* – Full n^2 attention; *b* – Sliding window attention; *c* – Dilated sliding window attention; *d* – Global+sliding window attention [12]

In addition to full attention, there exist a variety of sparse attention patterns that allow self-attention to have linearly scaling complexity, shown in Fig.3. Longformer utilizes global+sliding window attention: Each token only attends to the nearest w tokens, and certain predetermined tokens receive global attention - every token in the sequence can be associated with them, and they can be associated with every other token. Therefore, the final complexity is $O((w+g)n)$, where w is the width of the window, g is the amount of global tokens and n is the total size of the input. While g may be dependent on n, it is still a significantly smaller value, yielding nearly linear complexity [12].

This paper will explore the modification of led-large-16384-arxiv [21] – a version of LED's 16,384-token model fine-tuned for summarization on the arXiv dataset.

4.3. Modifying the cross-attention function

Led-large-16384-arxiv is already capable of processing a very large amount of input tokens – but it is possible for the model to encounter input longer than it can support. For instance, the arXiv training dataset itself contains entries that exceed 20k tokens when using LED's tokenizer. Extremely long documents represent a higher workload during human summarization, making the ability to effectively and thoroughly summarize them all the more important.

Unlimiformer solves this problem by altering the decoder's cross-attention function. Instead of attending to every token at once, inputs are processed in chunks, which are then used to form an index of the entire input sequence. During decoding, the cross-attention function performs an approximate k-nearest-neighbors search to select the most relevant embeddings from all parts of the input sequence. This allows the model to more effectively account for valuable global context while scaling sublinearly with model size. The process is shown in Fig.4. [14]

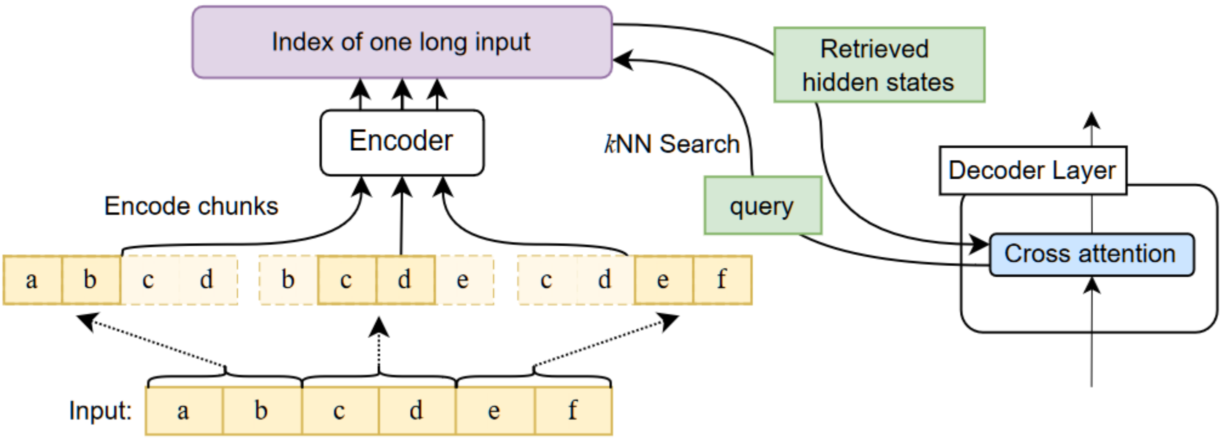

Fig. 4. Unlimiformer architecture for a model with a 2-token maximum input length. [14]

Another benefit of Unlimiformer is that it does not require training new attention weights or positional embeddings, allowing it to be applied to existing models without additional training. Fine-tuning can further improve performance, but it is not required as part of the augmentation process.

We augment our led-large-16384-arxiv model with Unlimiformer cross-attention. No further training is performed, as the process of fine-tuning a large sequence length model is very computationally expensive.

5. Results of investigating summarization effectiveness

PC characteristics on which the system was tested:

1. The clock frequency of the processor is 3.2 GHz;
2. 4 cores, 4 logical processors;
3. 14 GB of RAM;
4. Available NVIDIA GeForce GTX 1060 video card.

Both the base and modified led-large-16384-arxiv models were evaluated on a sample of the arXiv test dataset. Results are presented in Table 2.

Table 2. Model evaluation results

Model	ROUGE-1	ROUGE-2	ROUGE-L
led-large-16384-arxiv	46.42	23.43	30.86
led-large-16384-arxiv-unlim	**47.68**	**23.94**	**31.22**

Model effectiveness was evaluated using the ROUGE-1, ROUGE-2 and ROUGE-L metrics. ROUGE metrics offer a way to calculate Precision and Recall for summarization tasks, which can then be used to calculate the F-Measure (given as the ROUGE score):

$$Recall = \frac{X}{N_{expected}}, \tag{1}$$

$$Precision = X/N_{generated}, \tag{2}$$

$$FMeasure = \frac{2}{Recall^{-1} + Precision^{-1}}, \tag{3}$$

where *N_expected* is the total number of words in the expected summary, *N_generated* is the total number of words in the generated summary, and X depends on the ROUGE metric used:

For ROUGE-1, X is the amount of shared words - words that feature in both the expected and generated summaries.

For ROUGE-2, X is the amount of shared bigrams - two-word sequences that feature in both the expected and generated summaries.

For ROUGE_L, X is the length of the longest unbroken sequence of shared words.

6. Analysis of the obtained results

The augmented Unlimiformer model shows small but noticeable improvements across all ROUGE metrics compared to the base model, demonstrating that the addition of Unlimiformer cross-attention can help enhance model effectiveness. As the evaluation was performed without any further training, fine-tuning the modified model would likely lead to a further increase in summarization effectiveness.

Fine-tuning of long summarization models is often very computationally expensive, as the task requires the ability to process vast amounts of input tokens. The ability to refine such models through architectural changes is invaluable, as it allows us to bypass the expensive retraining process. Additionally, cross-attention augmentation can be used to adapt much smaller models to the long document summarization task, allowing for faster training and smaller, more accessible summarization models.

7. Conclusion

An augmented model for the summarization of scientific articles has been developed without the use of additional training. The resulting model shows improved ROUGE-1, ROUGE-2 and ROUGE-L scores compared to the base model, and is additionally able to process inputs beyond its size, having an unbounded input capacity.

References

[1] D. Khurana, A. Koli, K. Khatter, and S. Singh, “Natural language processing: state of the art, current trends and challenges,” Multimedia Tools and Applications, 2022, vol. 82, no. 3, pp. 3713–3744. https://doi.org/10.1007/s11042-022-13428-4.

[2] J. Sawicki, M. Ganzha, and M. Paprzycki, “The State of the Art of Natural Language Processing – A Systematic Automated Review of NLP Literature Using NLP Techniques,” Data Intelligence, 2023, vol. 5, no. 3, pp. 707–749. https://doi.org/10.1162/dint_a_00213.

[3]S. Kusal, S. Patil, J. Choudrie, K. Kotecha, D. Vora, and I. Pappas, “A systematic review of applications of natural language processing and future challenges with special emphasis in text-based emotion detection,” Artificial Intelligence Review, 2023, vol. 56, no. 12, pp. 15129–15215. https://doi.org/10.1007/s10462-023-10509-0.

[4]M. Gambhir and V. Gupta, “Recent automatic text summarization techniques: a survey,” Artificial Intelligence Review, 2016, vol. 47, no. 1, pp. 1-66. https://doi.org/10.1007/s10462-016-9475-9.

[5]S. Gupta and S. K. Gupta, “Abstractive summarization: An overview of the state of the art,” Expert Systems with Applications, 2019, vol. 121, pp. 49-65. https://doi.org/10.1016/j.eswa.2018.12.011.

[6]A. P. Widyassari et al., “Review of automatic text summarization techniques & methods,” Journal of King Saud University - Computer and Information Sciences, 2022, vol. 34, no. 4, pp. 1029–1046. https://doi.org/10.1016/j.jksuci.2020.05.006.

[7]M. M. Saiyyad and N. N. Patil, “Text Summarization Using Deep Learning Techniques: A Review,” RAiSE-2023. MDPI, 2024, p. 194. https://doi.org/10.3390/engproc2023059194

[8]D. O. Cajueiro et al., “A comprehensive review of automatic text summarization techniques: method, data, evaluation and coding,” 2023, arXiv. https://doi.org/10.48550/ARXIV.2301.03403.

[9]O. Klymenko, D. Braun, and F. Matthes, “Automatic Text Summarization: A State-of-the-Art Review,” Proceedings of the 22nd International Conference on Enterprise Information Systems. SCITEPRESS - Science and Technology Publications, 2020. https://doi.org/10.5220/0009723306480655.

[10]V. Patel and N. Tabrizi, “An Automatic Text Summarization: A Systematic Review,” Computación y Sistemas, vol. 26, no. 3. Instituto Politecnico Nacional/Centro de Investigacion en Computacion, Sep. 05, 2022. https://doi.org/10.13053/cys-26-3-4347.

[11]A. Gidiotis and G. Tsoumakas, “A Divide-and-Conquer Approach to the Summarization of Long Documents,” IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 28. Institute of Electrical and Electronics Engineers (IEEE), pp. 3029–3040, 2020. https://doi.org/10.1109/taslp.2020.3037401.

[12]I. Beltagi, M. E. Peters, and A. Cohan, "Longformer: The Long-Document Transformer," 2020, arXiv. doi: 10.48550/arXiv.2004.05150.

[13]M. Zaheer et al., “Big Bird: Transformers for Longer Sequences,”2020, arXiv. https://doi.org/10.48550/arXiv.2007.14062.

[14]A. Bertsch, U. Alon, G. Neubig, and M. R. Gormley, “Unlimiformer: Long-Range Transformers with Unlimited Length Input,” 2023, arXiv. https://doi.org/10.48550/arXiv.2305.01625.

[15]A. Cohan et al., “A Discourse-Aware Attention Model for Abstractive Summarization of Long Documents,” Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers). Association for Computational Linguistics, 2018. https://doi.org/10.18653/v1/n18-2097.

[16]M. Guo et al., “LongT5: Efficient Text-To-Text Transformer for Long Sequences,” Findings of the Association for Computational Linguistics: NAACL 2022. Association for Computational Linguistics, pp. 724–736, 2022. https://doi.org/10.18653/v1/2022.findings-naacl.55.

[17]S. Sotudeh, A. Cohan, andN. Goharian,“On Generating Extended Summaries of Long Documents,” 2020. arXiv. https://doi.org/10.48550/arXiv.2012.14136.

[18]M. Yasunaga et al., “ScisummNet: A Large Annotated Corpus and Content-Impact Models for Scientific Paper Summarization with Citation Networks,” Proceedings of the AAAI Conference on Artificial Intelligence, vol. 33, no. 01. Association for the Advancement of Artificial Intelligence (AAAI), pp. 7386–7393, Jul. 17, 2019. https://doi.org/10.1609/aaai.v33i01.33017386.

[19]C. An, M. Zhong, Y. Chen, D. Wang, X. Qiu, and X. Huang, “Enhancing Scientific Papers Summarization with Citation Graph,” Proceedings of the AAAI Conference on Artificial Intelligence, vol. 35, no. 14. Association for the Advancement of Artificial Intelligence (AAAI), pp. 12498–12506, May 18, 2021. https://doi.org/10.1609/aaai.v35i14.17482.

[20]A. Vaswani et al., “Attention Is All You Need,” 2017, arXiv. https://doi.org/10.48550/arXiv.1706.03762.

[21]Huggingface, “allenai/led-large-16384-arxiv.” Accessed: Nov. 1, 2024. [Online]. Available: https://huggingface.co/allenai/led-large-16384-arxiv.

УДК 004.89

МОДЕЛЬ ДЛЯ КОНСПЕКТУВАННЯ НАУКОВИХ СТАТЕЙ З НЕОБМЕЖЕНОЮ ДОВЖИНОЮ ВХІДНИХ ДАНИХ

Олександр Стеблянко
Національний технічний університет України
«Київський політехнічний інститут імені Ігоря Сікорського», Київ, Україна
https://orcid.org/0009-0006-5055-0934

Володимир Шимкович
Національний технічний університет України
«Київський політехнічний інститут імені Ігоря Сікорського», Київ, Україна
https://orcid.org/0000-0003-4014-2786

Петро Кравець
Національний технічний університет України
«Київський політехнічний інститут імені Ігоря Сікорського», Київ, Україна
https://orcid.org/0000-0003-4632-9832

Анатолій Новацький
Національний технічний університет України
«Київський політехнічний інститут імені Ігоря Сікорського», Київ, Україна
https://orcid.org/0009-0009-7457-7391

Любов Шимкович
Національний технічний університет України
«Київський політехнічний інститут імені Ігоря Сікорського», Київ, Україна
https://orcid.org/0000-0002-1291-0373

Останніми роками експоненційне зростання наукової літератури зробило для дослідників і практиків все більш складним завдання встигати за новими відкриттями та розробками у своїх галузях. Завдяки цьому, конспектування тексту стало одним із основних завдань обробки природної мови. Абстрактивне конспектування довгих документів, наприклад наукових статей, вимагає великих нейронних мереж із високими вимогами до пам'яті та обчислень. Відповідно, все більш важливо знайти шляхи підвищення ефективності моделей конспектування довгих документів.

Об'єктами дослідження є трансформерні моделі для конспектування довгих документів та модифікація перехресної уваги *Unlimiformer*. У статті розглянуто основні принципи уваги трансформеру, що складає більшість обчислювальних витрат в трансформерних моделях. Описано більш ефективні підходи до самоуваги, які використовуються в моделях конспектування довгих документів, наприклад увага глобального + ковзаючого вікна, що використовується в *Longformer*. Детально описано механізм перехресної уваги Unlimiformer, який дозволяє моделі мати необмежену вхідну довжину.

Метою дослідження є розробка та оцінка моделі конспектування довгих документів за допомогою модифікації *Unlimiformer*. Для досягнення цієї мети модель *Longformer Decoder-Encoder*, попередньо навчену на наборі даних *arXiv*, модифікується за допомогою перехресної уваги *Unlimiformer*. Цю модифікацію можна застосувати без додаткового тонкого настроювання моделі, уникаючи витрат на подальше навчання моделі з великою довжиною послідовності.

Розроблена модель була оцінена на наборі даних *arXiv* використовуючи показники *ROUGE*-1, *ROUGE*-2 і *ROUGE*-L. Розроблена модель показала покращені результати порівняно з базовою моделлю, демонструючи життєздатність використання цього підходу для вдосконалення моделей конспектування довгих документів.

Key words: нейронні мережі, трансформери, конспектування тексту, конспектування довгих документів, обробка природньої мови, увага.