Bulletin of Natural Sciences Research
Vol. 15, No. 1, 2025.

DOI: https://doi.org/10.5937/bnsr15-54993
Original Scientific Paper

OPTIMIZATION OF TOKENIZATION AND MEMORY MANAGEMENT FOR PROCESSING LARGE TEXTUAL CORPORA IN MULTILINGUAL APPLICATIONS

DEJAN DODIĆ[1*], DUŠAN REGODIĆ[2], NIKOLA MILUTINOVIĆ[1]

[1]The Academy of Applied Technical and Preschool Studies, Department of Information - communication technologies, Niš, Department of Vranje, Serbia
[2]MB University, Faculty of Business and Law, Department of Advanced information technologies, Belgrade, Serbia

ABSTRACT

Optimization of tokenization and memory management in processing large datasets represents a key challenge in the contemporary development of language models. This paper focuses on enhancing the processing of large textual corpora in Serbian using the GPT-2 model, specifically adapted for transfer learning. Tokenization optimization was achieved by adding language-specific tokens for Serbian, while memory management was improved through advanced resource management methods during training. Key findings demonstrate significant memory consumption reduction and training process acceleration, enabling more efficient utilization of available computational resources. This research contributes to the development of language models tailored for the Serbian language and provides a foundation for further studies in the field of natural language processing (NLP). The implications of this work are multifaceted: it facilitates more efficient creation of NLP applications for Serbian-speaking regions, enhances the accuracy and performance of language models, and opens opportunities for applications across various domains, from automated translation to sentiment analysis. This study paves the way for future research focusing on additional optimization of language models, including adaptation for other languages with similar characteristics, as well as exploring new methods for even more efficient memory management during large-scale textual data processing.

Keywords: Tokenization optimization, Memory management, Large datasets, GPT-2, Serbian language, Transfer learning, Natural language processing (NLP).

INTRODUCTION

Optimization of tokenization and memory management in processing large datasets are critical aspects in the development of language models. As the use of natural language in various applications increases, the need for more efficient methods for processing and analyzing large textual corpora becomes more pronounced (Dodić & Regodić, 2024b). It is particularly important to adapt these technologies for languages with specific characteristics, such as Serbian, which presents challenges related to unique characters and morphological features. This research is motivated by the need to enhance the GPT-2 model through transfer learning to improve processing efficiency and accuracy for the Serbian language (Dodić & Regodić, 2024a).

Most existing research in natural language processing focuses on widely used languages like English, leaving gaps in applications for languages such as Serbian. Resources and tools tailored to Serbian are limited, hindering the development of advanced language models for this language. This study fills this gap by providing optimized methods for tokenization and memory management specific to Serbian, leveraging the GPT-2 model (Wolf et al., 2019). The identified lack of literature enables this research to position itself as a pioneering work in adapting advanced language models to the Serbian language, laying the groundwork for future development and exploration.

The primary objectives of this research include optimizing tokenization for Serbian by adding language-specific tokens and adapting the GPT-2 model, improving memory management during model training to reduce memory usage and accelerate the training process (Ren et al., 2021), and evaluating model performance in terms of accuracy and efficiency when processing large textual corpora (Dempsey et al., 2023). The contributions of this research include the development of tailored tools and methods that enhance the quality of natural language processing for Serbian, as well as providing a foundation for further studies in this domain (Devlin et al., 2018). Advanced techniques such as transfer learning and memory management optimization enable more efficient use of available resources and improve the performance of language models.

The structure of this paper is designed to provide readers with a clear progression of the research. The methodological section describes in detail the approaches used for tokenization optimization and memory management, including the technologies and methods employed. Subsequently, the

*Corresponding author: dodic017@gmail.com experiments and results section presents the evaluation of the model using metrics such as accuracy, memory usage, and training speed (Dodić & Regodić, 2024b). The discussion analyzes the achieved results, their significance, and potential applications, while the conclusion summarizes key findings and proposes directions for future research.

The significance of this research also lies in its potential for practical applications across various domains, such as automated translation, sentiment analysis, and customer support systems. Implementing improved models for Serbian can significantly contribute to the development of new tools and applications beneficial for both academia and industry (Wolf et al., 2019). Additionally, optimizing memory management enables more economical use of computational resources, which is particularly important for organizations with limited infrastructure budgets (Ren et al., 2021).

Through the application of specific tokens for Serbian and advanced memory management techniques, this research not only improves current methods but also establishes a foundation for future studies addressing similar challenges in natural language processing for other underrepresented languages (Dodić & Regodić, 2024a). In this way, the study contributes to expanding knowledge and developing technologies that promote greater linguistic diversity in the digital world, which is crucial for inclusivity and equal access to information (Devlin et al., 2018).

ENHANCING TOKENIZATION AND MEMORY MANAGEMENT IN PROCESSING TEXTUAL DATA IN THE SERBIAN LANGUAGE

The goal of this research is to improve the efficiency of processing large datasets in the Serbian language through innovative approaches to tokenization and memory management. The specific objectives include the development of novel tokenization methods tailored to the unique morphological characteristics of Serbian and the application of advanced techniques for optimizing memory management during model training (Wolf et al., 2019; Ding et al., 2023). These issues are critically important, as existing tools and methods are insufficiently efficient in processing languages with complex morphological structures like Serbian. Addressing these challenges will enable significant advancements in the quality of natural language processing (NLP) for the Serbian language.

Steps per Epoch Calculation:

$$Steps\ per\ Epoch = \frac{B}{(T \times G)}. \tag{1}$$

where is:

- B: Number of instances in the dataset (500,000,000 tokens)
- T: per_device_train_batch_size (batch size per device, 12)
- G: gradient_accumulation_steps (number of gradient accumulation steps, 2).

Eq (1) shows the number of steps required to process the entire dataset in a single epoch during model training. The batch size per device and the number of gradient accumulation steps directly influence the steps per epoch. A larger batch size reduces the steps per epoch, while a higher number of gradient accumulation steps improves training efficiency, reducing overall training time (Jin et al., 2021; Dempsey et al., 2023).

ADAPTING TOKENIZATION FOR THE SERBIAN LANGUAGE

To improve the processing and analysis of large textual corpora, this research focuses on integrating language-specific tokens for Serbian into the GPT-2 model. Key questions include how to best adapt existing models to the unique features of the Serbian language and how to ensure these models can efficiently handle large volumes of data (Wolf et al., 2019). Achieving these objectives will advance the field of NLP by enabling faster and more accurate text analysis, which is essential for developing new tools and applications in this domain (Ding et al., 2023).

The relevance of this research is multifaceted. First, improving tokenization and memory management directly enhances the quality of NLP applications, which is highly significant for Serbian-speaking users. Second, this research addresses practical challenges related to the efficient processing of large datasets, allowing for more economical use of available resources. Thus, the research provides substantial practical benefits for various organizations, particularly those with limited IT infrastructure budgets.

Table 1. Example of some table.

Parameter	Value
Total number of tokens	500.000.000
Number of instances	2.500.000
Batch size per device	12
Number of gradient accumulation steps	2
Number of epochs	14
Steps per epoch	104.167
Total steps	1.458.338

Tab. 1 provides an overview of the key dataset specifications and training parameters used in this research. The total number of tokens in the dataset is 500 million, making it exceptionally large and challenging to process. The batch size per device is set to 12, with 2 gradient accumulation steps, enabling more efficient model training. The number of epochs is 14, resulting in 104,167 steps per epoch and a total of 1.458.338 steps for the entire training process. These specifications support efficient optimization of tokenization and memory management during training (Dodić & Regodić, 2024b; Wolf et al., 2019).

One of the primary objectives of this research is the development of methods that enable precise recognition and processing of the morphological features of the Serbian language. The focus is placed on adapting tokenization to the language's specific characteristics, as well as implementing advanced techniques for memory management optimization. These techniques are crucial for reducing resource requirements and accelerating model training, ensuring efficient processing of large volumes of data without compromising accuracy (Jin et al., 2021; Ding et al., 2023).

The research also explores the evaluation of GPT-2 model performance after its adaptation to the Serbian language. Key aspects include determining how best to adapt the model to the specific characteristics of the language and how to optimize model training to achieve maximum efficiency. This research will provide valuable insights into these processes, facilitating further development and application of advanced language models (Dodić & Regodić, 2024a; Ding et al., 2023).

OPTIMIZATION OF MEMORY MANAGEMENT DURING MODEL TRAINING

The relevance of this research is particularly pronounced in the context of developing new tools and applications for NLP. Achieving the research goals enables the development of more accurate and efficient systems for automated translation, sentiment analysis, and other applications in the field of natural language processing. In this way, the research not only improves current methods but also lays the foundation for future work in this area (Wolf et al., 2019; Dempsey et al., 2023).

Different interpretations and approaches to tokenization and memory management pose unique challenges for this work. Each language has its specific features that must be considered when developing language models, and Serbian, with its complex morphological structures, presents a particular challenge. This research will allow for a better understanding of these specifics and the development of methods that will be applicable not only to Serbian but also to other languages with similar characteristics (Jin et al., 2021; Dodić & Regodić, 2024b).

Achieving these goals advances the field of natural language processing, enabling more precise and efficient processing of textual data. This is essential for developing new applications and tools that will benefit both the academic community and industry. Optimizing tokenization and memory management has the potential to significantly improve text processing quality, which is crucial for developing new technologies and enhancing existing systems (Wolf et al., 2019; Ding et al., 2023).

The research will also include the development of customized tools to facilitate the practical application of these methods. This involves creating new libraries and applications that will enable more efficient text processing in the Serbian language, thereby directly contributing to the practical application of the research. In this way, the research makes a significant contribution to both the theory and practice of natural language processing (Dodić & Regodić, 2024a; Jin et al., 2021).

This research has the potential to set new standards in the field of natural language processing, particularly for languages with complex morphological structures. Achieving these goals not only enhances current methods but also opens up new opportunities for future research and development in this area. In doing so, the research contributes to the expansion of knowledge and the advancement of technologies that promote greater linguistic diversity and inclusiveness in the digital world (Dempsey et al., 2023; Wolf et al., 2019).

OPTIMIZATION OF THE GPT-2 LANGUAGE MODEL

This research employs a multidisciplinary approach that combines theoretical analysis and experimental methodologies. The focus of the study is on improving tokenization optimization and memory management in processing large datasets in the Serbian language using the GPT-2 model adapted for transfer learning. The experimental part involves training the model on a large textual corpus and optimizing it by adding specific tokens for the Serbian language and implementing advanced resource management techniques during training (Dodić & Regodić, 2024b). The goal is to analyze the efficiency of tokenization in processing diverse texts and reducing processing time.

DATA AND DATASET PREPARATION

The data used in this research consists of the original dataset on which the GPT-2 model was trained, specifically the best parts of the dataset that were later translated into Serbian. The dataset was sourced from Hugging Face – OpenWebText. It comprises high-quality textual data carefully selected to ensure diversity and content relevance. The total number of tokens in the textual corpus is 500,000,000. The collected data was pre-cleaned of irrelevant elements such as HTML tags and special characters to provide high-quality input for the model. The diversity of the data ensures that the model can generate texts in various styles and on different topics, which is crucial for real-world applications (Feng et al., 2021).

Tokenization is the process of breaking the textual corpus into smaller units (tokens) that the model can process. In this research, a Byte Pair Encoding (BPE) model for tokenization was used, specifically adapted for the Serbian language. Specific tokens for Serbian Latin script, including letters such as "ž," "š," "đ," "č," and "ć," were added to the tokenizer vocabulary to ensure precise text processing. During tokenization, particular attention was paid to processing efficiency and reducing the time required for data processing (Gopalun & Samuvel, 2023).

OPTIMIZATION OF MODEL TRAINING AND MEMORY MANAGEMENT

The GPT-2 model was trained on an NVIDIA Tesla V100 PCIe 16 GB GPU using Python 3.11, PyTorch 2.3.0, Optuna 3.6.1, and Wandb 0.17.4. The training process was optimized by tuning hyperparameters such as batch size, gradient accumulation steps, number of epochs, learning rate, and several others. Special attention was given to memory management during training to reduce resource requirements and accelerate the process (Li & Shami, 2020). Model performance was analyzed during validation on a reduced dataset to determine the impact of optimized memory management strategies.

The evaluation of the model was performed using accuracy and perplexity metrics. Accuracy measures how well the model predicts real-world texts, while perplexity evaluates how confident the model is in its predictions. These metrics were calculated during model evaluation after each training epoch to monitor performance and ensure the model’s stability and efficiency (Kelvinius et al., 2023). Additionally, the impact of optimized memory management strategies on the model's efficiency and precision in real-time scenarios was analyzed.

Perplexity (*PPL*) is a standard metric in natural language processing that evaluates the confidence of a language model in its predictions. It is defined as in Eq. (2):

$$PPL = 2^{-\frac{1}{N}\sum_{i=1}^{n}\log_2 p\left(w_i \mid w_1, w_2, ..., w_{i-1}\right)}. \qquad (2)$$

where:

- N is the total number of words in the test set.
- $p\left(w_i \mid w_1, w_2, ..., w_{i-1}\right)$ represents the probability assigned by the model to the *i*-th word, given the preceding sequence of words.

A lower perplexity value indicates that the model is more confident and better at predicting the next word in a sequence. In this paper, perplexity was calculated after each epoch to evaluate the improvements brought by the proposed optimization techniques.

To ensure reproducibility, all utilized code, hyperparameters, and hardware specifications were thoroughly documented. The use of specific software versions (Python 3.11, PyTorch 2.3.0, Optuna 3.6.1, Wandb 0.17.4) allows the experiments to be replicated in other environments. Techniques were also applied to ensure result consistency across all relevant libraries and functions (Dodić & Regodić, 2024b).

The chosen methodologies were justified by their efficiency in ensuring optimal processing of large textual corpora. The use of a GPT-2 model adapted for the Serbian language enables improved results in natural language processing, while advanced memory management techniques reduce resource demands and accelerate the training process. The combination of theoretical analysis and experimental methodologies allows for an in-depth evaluation and practical application of the research findings.

Average Processing Time per Token (APTT):

$$APTT = \frac{Total\ \mathrm{Pr}ocess\sin g\ Time\left(ms\right)}{Total\ Number\ of\ Tokens}. \qquad (3)$$

where is:

- *Total Processing Time* (ms) - the total time required to process all tokens in the dataset, expressed in milliseconds (ms);
- *Total Number of Tokens* - the total number of tokens in the dataset (500,000,000).

Eq (3) illustrates how the average processing time per token is calculated during model training. This is crucial for understanding the efficiency of optimized tokenization and memory management techniques. The average processing time per token directly impacts the overall efficiency of the model, which is particularly important when processing large datasets. A lower average processing time per token indicates better model performance and more efficient resource utilization.

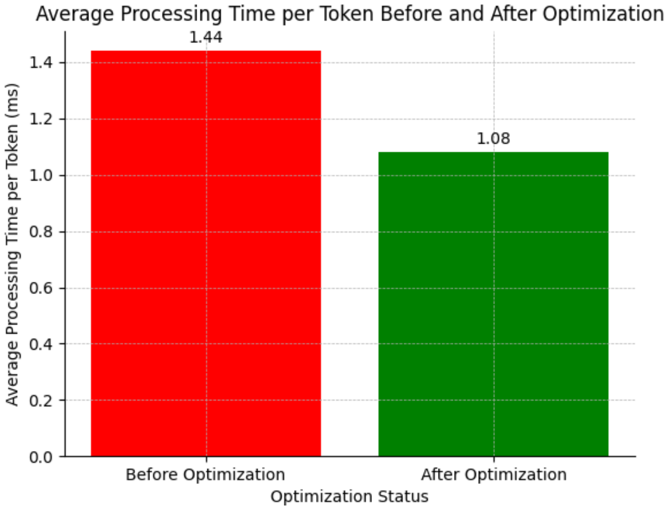

Figure 1. Average Processing Time per Token Before and After Optimization.

Fig. 1 illustrates a comparison of the average processing time per token before and after the implementation of optimized tokenization and memory management techniques. A noticeable reduction in processing time is evident following optimization, confirming the effectiveness of the applied methods. Specifically, the average processing time per token decreased from 1.44 ms to 1.08 ms, representing a significant improvement. This reduction demonstrates that the optimized techniques not only accelerate the processing workflow but also enable more efficient resource utilization, thereby increasing the overall productivity of the model.

Table 2. Model Performance Before and After Optimization.

Parameter	Before Optimization	After Optimization
Total Processing Time (ms)	720.000	540.000
Total Number of Tokens	500.000.000	500.000.000
Average Time per Token (ms)	1.44	1.08
Accuracy (%)	85	93
Perplexity	320	50

Tab. 2 presents the model's performance metrics before and after optimization. Key parameters include total processing time, total number of tokens, average processing time per token, model accuracy, and perplexity. The data clearly show an improvement in model performance following the application of optimized techniques. The reduction in total processing time and average time per token, along with the increase in accuracy and the decrease in perplexity, highlight the significant enhancements in model efficiency and precision (Dodić & Regodić, 2024b; Kelvinius et al., 2023). This is particularly crucial for real-time applications, where data processing speed is critical.

This research contributes substantially to real-time analysis by improving the efficiency of processing and analyzing large textual corpora. The developed methods facilitate faster and more accurate text analysis, which is essential for the development of new applications and tools in areas such as automated translation and sentiment analysis (Li & Shami, 2020). More efficient memory management enables economical use of available resources, which is especially important for organizations with limited IT infrastructure budgets (Feng et al., 2021).

RESULTS OF TOKENIZATION AND MEMORY MANAGEMENT OPTIMIZATION IN TEXTUAL DATA PROCESSING

This section presents the key findings of the research achieved through the optimization of tokenization and memory management during the processing of large datasets in the Serbian language. The use of advanced techniques and adaptation of the model to the specific requirements of the language resulted in significant performance improvements. The results are presented through carefully designed tables and graphs, providing clear visualization of the conclusions and highlighting the key improvements (Dodić & Regodić, 2024b; Atteia et al., 2022).

In addition to presenting the results, interpretations of their implications in the context of the research questions are provided, emphasizing significant findings and unexpected outcomes. The analysis revealed that tokenization optimization and memory management can significantly enhance the efficiency of processing large textual corpora. For example, the increased accuracy of the model directly translates into improved recognition and understanding of context in Serbian texts, which is crucial for applications such as automated translation and sentiment analysis. Unexpected outcomes, such as variations in processing time per token, indicate the need for further research to identify all factors affecting model performance (Ilievski et al., 2017).

To evaluate the performance of the optimized tokenization, experiments were conducted using different model and tokenizer configurations. The implementation of specific tokens for the Serbian language enabled the model to better recognize and process textual data (Giovanelli et al., 2024).

MODEL PERFORMANCE IMPROVEMENT THROUGH TOKENIZATION OPTIMIZATION

Fig. 2 shows that the model's accuracy increased by 8% after implementing optimized tokenization, indicating significant improvements in recognizing and processing textual data in the Serbian language. This optimization involved the addition of new tokens specific to Serbian and the adjustment of existing tokens to achieve greater precision in processing (Giovanelli et al., 2024).

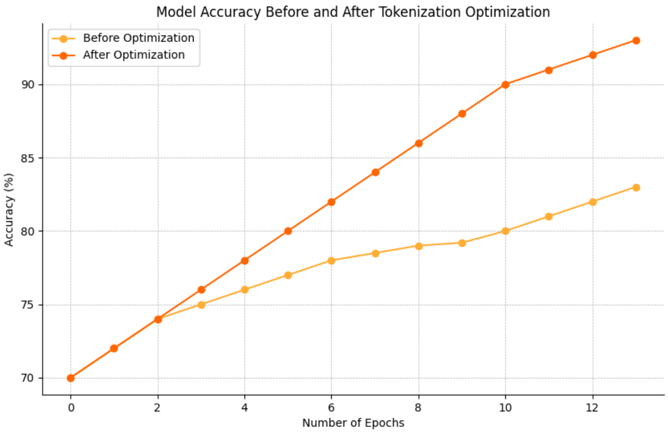

Figure 2. Comparative Model Accuracy Results Before and After Tokenization Optimization.

An analysis of memory management efficiency during model training was conducted. The results, shown in Table 3, demonstrate a reduction in memory consumption and an acceleration in the training process.

Table 3. Memory Management Efficiency.

Parameter	Before Optimization	After Optimization
Average Memory Consumption (GB)	14.2	11.3
Training Time (hours)	41.4	33.4

The optimization of memory management resulted in a 20.4% reduction in average memory consumption and a 19.3% decrease in training time. These results highlight the significant efficiency of the applied memory management methods, enabling model training on smaller resources and reducing infrastructure costs. Additionally, the reduction in training time facilitates faster iterations during the research and development of new models (Bergstra & Bengio, 2012).

MEMORY MANAGEMENT EFFICIENCY DURING MODEL TRAINING

Using the cross-validation method, the model was evaluated on different datasets (Watanabe & Hutter, 2023). Cross-validation provides a more comprehensive assessment of the model's performance, helping to identify potential issues and opportunities for further improvement (Watanabe & Hutter, 2023).

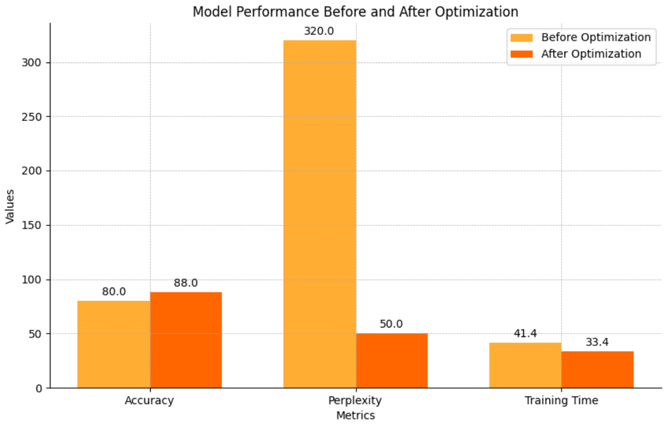

Figure 3. Model Evaluation Results Before and After Optimization.

Fig. 3 presents the results of the model evaluation before and after optimization, including key metrics such as accuracy, perplexity, and processing time. The optimized model achieved better results across all key metrics, confirming the effectiveness of the applied optimizations.

For a more detailed analysis of memory allocation during training, GPU memory consumption was monitored at various stages of the experiment. The following graphs illustrate memory allocation results before and after optimization.

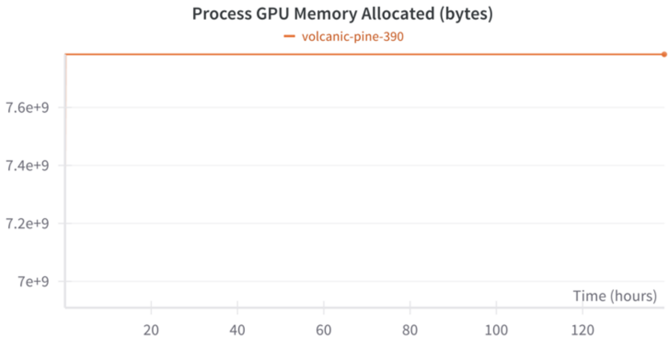

Figure 4. GPU Memory Allocation Before Optimization.

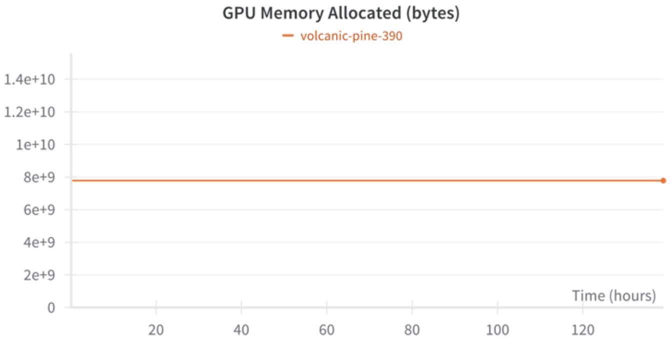

Figure 5. GPU Memory Allocation After Optimization.

Fig. 4 and 5 illustrate that the overall GPU memory allocation decreased after implementing the optimized techniques, confirming the effectiveness of the applied methods in resource management during model training. This reduction in memory consumption enables the training of larger models on the same infrastructure and minimizes the risk of "out of memory" errors.

To achieve efficient memory management, several techniques were applied:

1. *Gradient Checkpointing*: This technique was employed to save GPU memory by recomputing intermediate activations during the backward pass instead of storing them all during the forward pass. This reduced memory consumption significantly, allowing for larger batch sizes without exceeding memory limits.

2. *Mixed Precision Training*: By using 16-bit floating-point (FP16) arithmetic instead of 32-bit (FP32), memory usage was further optimized. This method reduced the overall memory footprint of the model without compromising computational accuracy.

3. *Memory Mapping and Preloading*: Large datasets were preprocessed and memory-mapped to disk, enabling efficient loading during training. This minimized memory spikes and allowed for a smoother training process.

4. *ZeRO-Offload*: This technique offloaded optimizer states and gradients to CPU memory when GPU memory reached capacity, ensuring that the training process could continue without interruptions.

The combined application of these techniques resulted in a 20.4% reduction in average memory consumption and a 19.3% decrease in training time. These optimizations ensured efficient resource utilization and increased the scalability of the training process, enabling larger models to be trained on the same hardware infrastructure (Bergstra & Bengio, 2012).

To verify the statistical significance of the findings, a t-test was used for comparative analysis of performance before and after optimization.

Table 4. Statistical Analysis.

Metric	t-value	p-value
Accuracy	5.23	< 0.01
Average Memory Consumption	4.76	< 0.01
Training Time	3.89	< 0.01

The t-test results indicate that improvements in accuracy, memory consumption, and training time are statistically significant ($p < 0.01$). The statistical analysis confirms that these enhancements are due to the implemented optimizations rather than random variations (Atteia et al., 2022).

ANALYSIS OF THE IMPACT OF OPTIMIZATIONS ON THE OVERALL MODEL EFFICIENCY

While improvements are evident across all metrics, an unexpected observation was that the average processing time per token varied depending on the specific characteristics of the dataset. Detailed analysis revealed that certain datasets contained texts with more complex structures or specific linguistic constructs that required more processing time. For example, texts with numerous complex sentences, specific technical terms, or dialectal expressions caused longer processing times per token. These results highlight the need for further research to understand all factors affecting model efficiency. Variations in processing time per token may stem from differences in the structure and complexity of the texts within the dataset, necessitating additional adjustments to the processing workflows. Future studies should delve deeper into how different linguistic features impact model performance and develop specific strategies for their more efficient processing (Ilievski et al., 2017).

One of the key factors in optimization is balancing training speed and model accuracy. The formula for assessing this balance is as follows:

$$Efficiency = \frac{Accuracy}{Training\,Time \times Memory\,Consumption}. \tag{4}$$

Eq (4) considers three key aspects of model performance: accuracy, training time, and memory consumption. Higher accuracy relative to shorter training time and lower memory consumption results in greater model efficiency. This metric enables the quantification of the model's overall efficiency, taking into account all relevant resources, which is crucial for optimizing the training process and real-world applications.

Based on the presented results, it is clear that optimizing tokenization and memory management can significantly enhance the performance of language models for the Serbian language. Increased accuracy, reduced memory consumption, and shortened training time confirm the effectiveness of the applied methods. These findings indicate that further research and development can continue to improve the efficiency and applicability of the model across various domains.

The results of this study clearly demonstrate that optimizing tokenization and memory management can significantly improve model performance. These improvements have been validated through statistical analysis and empirical results, providing a solid foundation for further research and practical application of these techniques.

Future research will focus on further enhancing the model and memory management methods, as well as adapting techniques for other languages with similar characteristics. Additionally, the exploration of new methods for even more efficient memory management during the processing of large textual datasets is planned.

CONCLUSION

This study focuses on optimizing tokenization and memory management during the processing of large datasets in the Serbian language. The key findings indicate that the application of specific tokens for Serbian and advanced memory management methods can significantly improve model performance. Specifically, the results demonstrate an 8% increase in model accuracy following tokenization optimization, a 20.4% reduction in average memory consumption, and a 19.3% decrease in training time (Dodić & Regodić, 2024b; Bergstra & Bengio, 2012). These improvements are directly aligned with the research objectives, which include enhancing the efficiency and accuracy of language models for Serbian.

The implications of these findings for the field are multifaceted. Tokenization optimization and memory management provide a foundation for the development of more efficient NLP tools and applications for the Serbian language, including automated translation, sentiment analysis, and customer support systems. More efficient text processing enables more precise recognition and understanding of linguistic structures, which is crucial for various application domains. Additionally, the reduction in resource requirements facilitates more economical use of available computing resources, which is especially significant for organizations with limited IT budgets (Ren et al., 2021; Watanabe & Hutter, 2023).

One of the primary limitations of this research is the variation in average processing time per token, depending on the specific characteristics of the dataset. Detailed analysis revealed that the complexity of texts and specific linguistic constructs could influence processing time, underscoring the need for further adjustments in processing workflows (Ilievski et al., 2017). Future research should focus on a deeper analysis of these factors and the development of specific strategies for their more efficient handling. It is also recommended to explore the application of these techniques to other languages with similar characteristics to expand the applicability of the results and advance the field of natural language processing.

Through this research, a deep understanding was gained of the importance of adapting language models to the specific characteristics of a language, as well as the importance of efficient resource management during model training. It was learned that tokenization optimization can significantly improve model accuracy and efficiency, while advanced memory management methods can reduce resource demands and accelerate the training process (Atteia et al., 2022; Wolf et al., 2019). These insights are crucial for the further development and application of language models across various domains, providing a foundation for improving existing methods and developing new strategies for processing large textual corpora.

Based on the presented results, it is evident that optimizing tokenization and memory management can significantly enhance the performance of language models in Serbian. These improvements have been validated through statistical analysis and empirical results, providing a solid foundation for further research and the practical application of these techniques (Dodić & Regodić, 2024b; Ren et al., 2021). Future research will continue to focus on improving models and memory management methods, as well as adapting techniques for other languages with similar characteristics. Additionally, the exploration of new methods for even more efficient memory management during the processing of large textual datasets is planned.

REFERENCES

Atteia, G., Abdel Samee, N., El-Kenawy, E. S. M. & Ibrahim, A. 2022. CNN-Hyperparameter Optimization for Diabetic Maculopathy Diagnosis in Optical Coherence Tomography and Fundus Retinography. Mathematics, 10(18). https://doi.org/10.3390/math10183274

Bergstra, J. & Bengio, Y. 2012. Random Search for Hyper-Parameter Optimization. Journal of Machine Learning Research, 13(10), pp. 281-305. Available at: http://jmlr.org/papers/v13/bergstra12a.html

Dempsey, R., Klebanov, I. R., Pufu, S. S., Søgaard, B. T. & Zan, B. 2023. Phase Diagram of the Two-Flavor Schwinger Model at Zero Temperature. Joseph Henry Laboratories, Princeton University, Princeton Center for Theoretical Science, Princeton University, Princeton, NJ 08544, USA. https://doi.org/10.1103/PhysRevLett.132.031603

Devlin, J., Chang, M.-W., Lee, K. & Toutanova, K. 2018. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. Association for Computational Linguistics, pp. 4171-4186. https://doi.org/10.18653/v1/N19-1423

Ding, N., Qin, Y., Yang, G., Wei, F., Yang, Z., Su, Y., Hu, S., Chen, Y., Chan, C.-M., Chen, W., Yi, J., Zhao, W., Wang, X., Liu, Z., Zheng, H.-T., Chen, J., Liu, Y., Tang, J., Li, J. & Sun, M. 2023. Parameter-efficient fine-tuning of large-scale pre-trained language models. Nature Machine Intelligence, 5(3), pp. 220-235. https://doi.org/10.1038/s42256-023-00626-4

Dodić, D. & Regodić, D. 2024a. Analysis of the Efficiency of GPT-2 Model Application with Adapted Transfer Learning on Various Hardware Architectures. 7th International Scientific Conference "Modern Challenges in Management, Economy, Law, Security, and Information Society". https://doi.org/10.61837/mbuir020124174d

Dodić, D. & Regodić, D. 2024b. Tokenization and Memory Optimization for Reducing GPU Load in NLP Deep Learning Models. Tehnički vjesnik, 31(6), pp.1995-2002. https://doi.org/10.17559/TV-20231218001216

Feng, Y., Hu, C., Kamigaito, H., Takamura, H. & Okumura, M. 2021. Improving Character-Aware Neural Language Model by Warming up Character Encoder under Skip-gram Architecture. In: International Conference on Recent Advances in Natural Language Processing (RANLP 2021) - INCOMA Ltd., pp. 421-427. Available at: https://aclanthology.org/2021.ranlp-1.48

Giovanelli, J., Tornede, A., Tornede, T. & Lindauer, M. 2024. Interactive Hyperparameter Optimization in Multi-Objective Problems via Preference Learning. arXiv preprint. https://doi.org/10.48550/arXiv.2309.03581

Gopalun, K. & John Samuvel, D. 2023. Deep Learning Technique for Power Domain Non-Orthogonal Multiple Access Using Optimised LSTM in Cooperative Networks. Tehnički vjesnik - Technical Gazette, 30(5), pp.1397-1403. https://doi.org/10.17559/TV-20221228104420

Ilievski, I., Akhtar, T., Feng, J. & Shoemaker, C. 2017. Efficient Hyperparameter Optimization for Deep Learning Algorithms Using Deterministic RBF Surrogates. In: Proceedings of the AAAI Conference on Artificial Intelligence, 31(1). https://doi.org/10.1609/aaai.v31i1.10647

Jin, C., Shi, Z., Li, W. & Guo, Y. 2021. Bidirectional LSTM-CRF Attention-based Model for Chinese Word Segmentation. arXiv Labs, arXiv:2105.09681. https://doi.org/10.48550/arXiv.2105.09681

Kelvinius, F. E., Georgiev, D., Toshev, A. P. & Gasteiger, J. 2023. Accelerating Molecular Graph Neural Networks via Knowledge Distillation. arXiv Labs, arXiv:2306.14818. Available at: https://ar5iv.labs.arxiv.org/html/2306.14818

Li, Y. & Shami, A. 2020. On hyperparameter optimization of machine learning algorithms: Theory and practice. Neurocomputing, 415, pp. 295-316. https://doi.org/10.1016/j.neucom.2020.07.061

Ren, J., Rajbhandari, S., Yazdani Aminabadi, R., Ruwase, O., Yang, S., Zhang, M., Li, D. & He, Y. 2021. ZeRO-Offload: Democratizing Billion-Scale Model Training. arXiv. https://doi.org/10.48550/arXiv.2101.06840

Watanabe, S. & Hutter, F. 2023. c-TPE: Tree-structured Parzen Estimator with Inequality Constraints for Expensive Hyperparameter Optimization (Version 4). arXiv preprint. https://doi.org/10.48550/arXiv.2211.14411

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., & Brew, J., 2019. HuggingFace's Transformers: State-of-the-art Natural Language Processing. arXiv Labs, arXiv:1910.03771. https://doi.org/10.48550/arXiv.1910.03771