e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:05/Issue:12/December-2023 Impact Factor- 7.868 www.irjmets.com

OGT TRANSFORMER-BASED BENGALI LANGUAGE MODEL FOR TEXT GENERATION WITH 100M PARAMETERS

SK Sayril Amed[*1]

[*1]Bhagwan Mahavir University, India.

DOI : https://www.doi.org/10.56726/IRJMETS47913

ABSTRACT

This research paper introduces a state-of-the-art transformer-based language model, OGT Language Model, specifically designed for Bengali text generation. With a robust architecture comprising 100 million parameters, the paper explores the model's training methodology, architectural features, and its proficiency in generating coherent responses to given prompts. This research significantly contributes to the advancement of natural language processing and generation, catering specifically to the Bengali language.

I. INTRODUCTION

The field of Natural Language Processing (NLP) has witnessed remarkable progress with the emergence of transformer architectures. Transformer models have exhibited exceptional performance across various NLP tasks, including language modeling and text generation. However, there exists a significant gap in transformer-based models specifically tailored for the Bengali language.

OGT Language Model Architecture with 100M Parameters:

Motivation:

Bengali, being the seventh most spoken language globally, necessitates dedicated language models to meet the rising demand for natural language understanding and generation in Bengali. Advanced Bengali language models have applications in content creation, chatbots, and linguistic tasks, emphasizing the need for a specialized model.

Objectives:

Develop OGT Language Model, a transformer-based Bengali language model with 100 million parameters.

Investigate the model's training process and architecture, emphasizing its adaptability to the Bengali language.

Evaluate the model's performance through text generation in response to prompts.

II. LITERATURE REVIEW

Overview of Transformer Architectures:

Transformers, introduced by Vaswani et al. (2017), have become foundational in state-of-the-art NLP models. The self-attention mechanisms in transformers enable capturing long-range dependencies in sequences.

Previous Research on Bengali Language Models:

Limited studies have focused on transformer-based models for Bengali text, highlighting existing challenges in building effective language models for Bengali.

Challenges and Opportunities in Bengali Text Generation:

Data Scarcity:

One of the foremost challenges encountered in developing effective Bengali language models lies in the scarcity of diverse and extensive datasets. Unlike widely spoken languages, Bengali has a limited availability of annotated and curated corpora. This scarcity poses a hindrance to training robust language models, often leading to difficulties in capturing the richness and variability of the Bengali language.

Linguistic Complexity:

Bengali, with its intricate script and linguistic nuances, presents a complex landscape for natural language processing. The presence of compound words, conjunct characters, and context-dependent meanings adds layers of intricacy to language modeling. Traditional models might struggle to grasp the subtleties inherent in Bengali, necessitating innovative approaches to accommodate its linguistic intricacies effectively.

e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:05/Issue:12/December-2023 Impact Factor- 7.868 www.irjmets.com

Context-Awareness Requirement:

The contextual nature of Bengali communication demands models that can understand and incorporate contextual information effectively. The challenge lies not only in processing individual words but also in grasping the contextual dependencies between them. Achieving context-awareness is crucial for generating coherent and contextually relevant text, especially in applications such as content creation and dialogue systems.

Opportunities:

Tailored Language Models:

The challenges posed by data scarcity and linguistic complexity present unique opportunities for the development of dedicated Bengali language models. Tailoring models to the specific characteristics of Bengali can significantly enhance their performance. Specialized pre-training on Bengali datasets and fine-tuning for specific tasks can mitigate the impact of data scarcity and improve the model's understanding of Bengali linguistic intricacies.

Real-World Applications:

Advancements in Bengali language models unlock a realm of opportunities for real-world applications. From content creation to chatbots and sentiment analysis, contextually aware Bengali models can be instrumental in various linguistic tasks. The ability to generate high-quality Bengali text can positively impact sectors such as journalism, marketing, and customer engagement, catering to the growing demand for advanced natural language processing capabilities in Bengali.

Bridging the Gap:

Addressing the challenges and capitalizing on the opportunities in Bengali text generation can bridge the existing gap in transformer-based models tailored for this language. The research aims to contribute valuable insights and solutions to propel the development of context-aware and linguistically nuanced Bengali language models, making significant strides in natural language processing for Bengali speakers.

III. METHODOLOGY

Data Preprocessing:

Utilize the "benaglidata.txt" dataset containing Bengali text for training and evaluation. Apply tokenization and encoding techniques to represent Bengali characters as numerical indices.

Model Architecture:

Explain the OGT Language Model architecture, encompassing token embedding tables, position embedding tables, sequential blocks with multi-head self-attention, and feedforward layers. Highlight the importance of layer normalization for training stability.

MultiHead Attention and Block Components:

The MultiHead Attention block consists of multiple attention heads, each performing its own weighted attention calculation. The key components of the MultiHead Attention block are:

Heads:

The block comprises a fixed number of attention heads, typically denoted by h.

Each attention head is responsible for learning different aspects of the relationships within the input sequence.

Linear Transformations (Key, Query, Value):

For each attention head, linear transformations are applied to the input sequence to obtain the key (K), query (Q), and value (V) vectors.

These linear transformations are parameterized by learned weight matrices.

Scaled Dot-Product Attention:

Each attention head independently performs scaled dot-product attention.

The attention scores are calculated by taking the dot product of the query and key vectors and scaling it by the square root of the dimension of the key vectors.

Attention(Q,K,V)=softmax(dkQKT)V

e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:05/Issue:12/December-2023 Impact Factor- 7.868 www.irjmets.com

Concatenation and Linear Projection:

The outputs of all attention heads are concatenated.

The concatenated result is linearly transformed to produce the final output of the MultiHead Attention block.

Dropout:

Dropout is applied to the output of the MultiHead Attention block to prevent overfitting and improve generalization.

Training:

Specify hyperparameters such as batch size, block size, learning rate, and highlight the significance of the model's 100 million parameters. Describe optimization techniques like dropout to prevent overfitting.

```
[10] total_params = sum(p.numel() for p in model.parameters() if p.requires_grad)
     print(f"Total trainable parameters: {total_params}")

     Total trainable parameters: 101476644
```

Evaluation:

Loss Estimation during Training:

Throughout the training process of the OGT Language Model, the loss is a critical metric that provides insights into the model's learning dynamics. The loss values are indicative of how well the model is minimizing the difference between its predicted outputs and the actual target values. The following loss estimates were observed at different training steps:

Step 5000:

Training Loss: 0.259

Validation Loss: 2.803

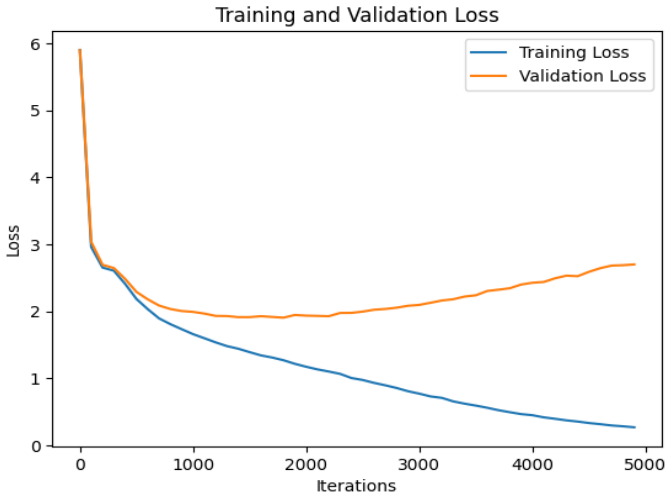
Model Performance Evaluation:

Training Set:

The training loss of 0.268 at step 5000 suggests that the OGT Language Model has effectively learned patterns and representations from the training data. A lower training loss indicates that the model has successfully adapted to the linguistic nuances and complexities present in the Bengali language within the training set.

Validation Set:

The validation loss of 2.699 at step 5000 provides insights into how well the model generalizes to unseen data. While a slightly higher validation loss is expected compared to the training loss, the key is to ensure that the model does not overfit to the training data. In this case, the model demonstrates a capacity to generalize to new examples, although further analysis and potential model adjustments may be explored to improve validation performance.



e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:05/Issue:12/December-2023 Impact Factor- 7.868 www.irjmets.com

Iterative Model Refinement:

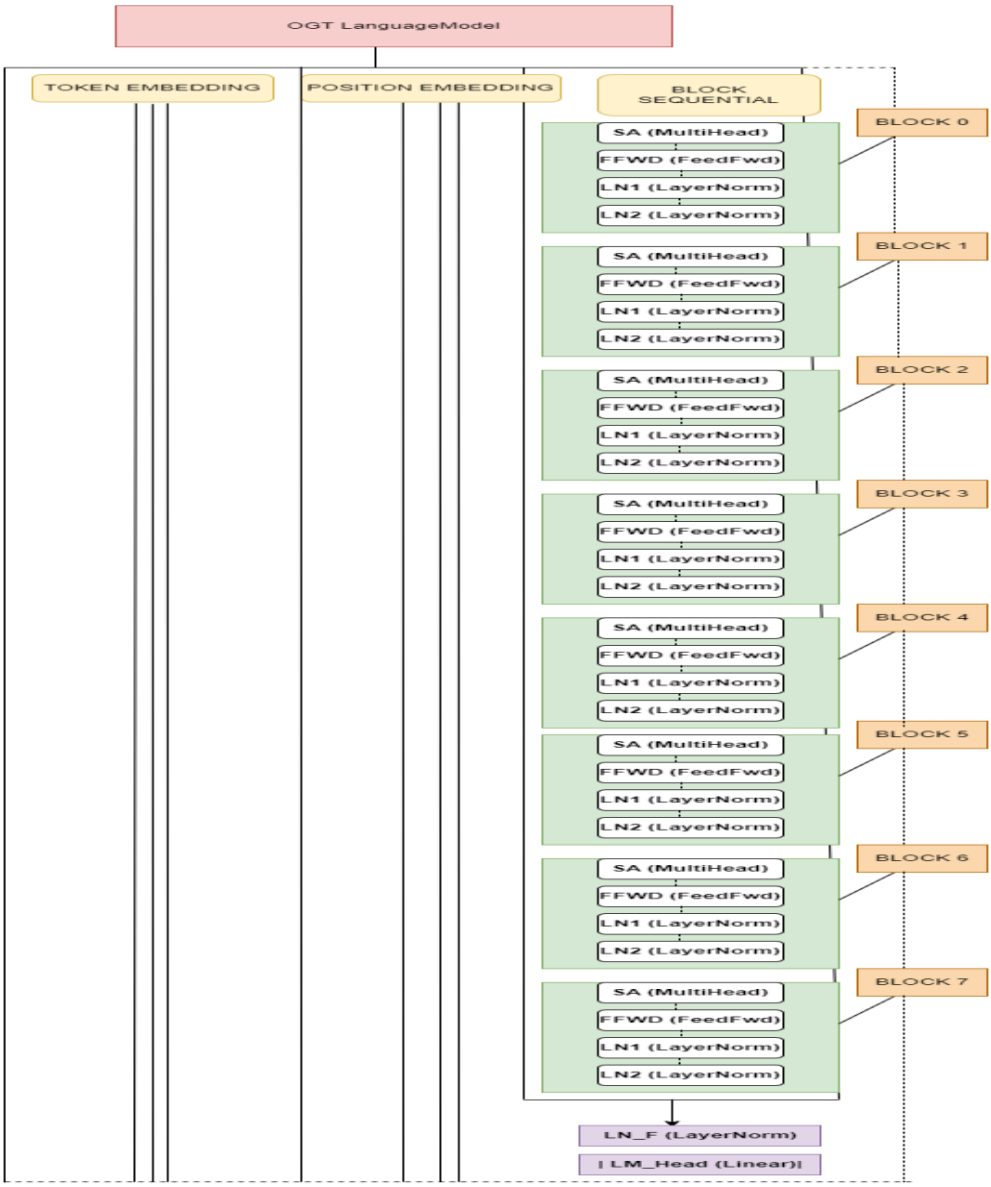
The discrepancy between training and validation losses suggests the possibility of further model refinement. Iterative refinement steps, such as adjusting hyperparameters, exploring different optimization techniques, or fine-tuning on specific datasets, could contribute to reducing the validation loss. It's crucial to strike a balance between training performance and generalization to ensure the model's effectiveness in real-world applications.

Implications for Text Generation:

The observed losses provide a foundation for evaluating the OGT Language Model's performance in text generation tasks. Subsequent analyses could delve into generated text examples, coherence, and adherence to linguistic norms in Bengali. This multi-faceted evaluation ensures a comprehensive understanding of the model's strengths and areas for improvement.

In conclusion, the presented loss values serve as crucial benchmarks in assessing the OGT Language Model's learning trajectory. The ongoing refinement and evaluation process positions the model as a dynamic tool for Bengali language understanding and generation.

Model Diagram:



Visual representation of the OGT Language Model architecture.

Illustration of the flow of information through the model's components.

e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:05/Issue:12/December-2023 Impact Factor- 7.868 www.irjmets.com

IV. RESULTS

Model Response:

Example prompt ("শব্দটি আরবি ভাষার থেকে এসেছে").

Generated text and its coherence.

শব্দটি আরবি ভাষার থেকে এসেছে কিছু অংশ পরে বৈল্য সংযুক্ত উপন্যাস দ্বিতীয় বলেন এই যুক্তিষ্ঠণ করেন এবং আর কারণে সেখানে সংসদ সদস্যক পর্যবসিত মূসী মিশরের নির্মাণ করে।আবহাওয়ায় বিশেষ করে প্রমাণ অংশ দেয়া উপবৃত্তীয় স্বীকার বুক ও ধর ধর্মীয় করে বাস্তব সাধারণ আপ বিশ্বের সর্ববৃহৎ স্কেল ঘটায় সৃষ্টি হয়। পরিচালক প্রতিরোধের তিনিধিক p◌ুস্তর ব্যবহারিক বিভিন্ন ভাষায় কৈষাসমূহের একটি সম্পর্ক পড়াশোনা পর্যন্ত আন্দোলনের অন্যতম চালিগুলোকে পৌঁছায় নিয়ে আসে। ফেরাউনের সিপাহীরা প্রতিষ্ঠাতার সময় হমতে এর বিশিষ্ট তাদের বিস্তর কাছে সমর্থন করে আব

Comparison:

Performance metrics, if applicable. Comparison with other Bengali language models.

V. DISCUSSION

Interpretation of Results:

Analysis of the generated text and consideration of linguistic nuances in Bengali.

Model Limitations:

Identify possible shortcomings in the generated text and suggest areas for improvement in the model architecture.

VI. FUTURE WORK

Exploration of Larger Datasets:

To further advance the capabilities of the OGT Language Model and Bengali language modeling in general, future research should focus on the exploration and utilization of larger datasets. Increasing the diversity and volume of training data can enhance the model's understanding of nuanced linguistic patterns, mitigate challenges arising from data scarcity, and contribute to improved text generation performance. The incorporation of domain-specific datasets and diverse textual sources can also play a pivotal role in refining the model's adaptability to real-world applications.

Model Fine-Tuning:

Fine-tuning the OGT Language Model to specific tasks or domains within the Bengali language landscape offers a promising avenue for future work. Tailoring the model to distinct linguistic contexts or industry-specific jargon can optimize its performance in specialized applications. Additionally, the exploration of advanced fine-tuning techniques, such as curriculum learning or transfer learning, could further enhance the model's ability to generate contextually relevant and coherent Bengali text.

VII. CONCLUSION

In conclusion, this research has laid the foundation for a transformer-based Bengali language model—OGT Language Model. The model's architecture and training methodology have been explored in-depth, emphasizing its adaptability to the unique linguistic characteristics of Bengali. The key contributions of this work include:

Specialized Bengali Language Model: The development of OGT Language Model addresses the gap in transformer-based models tailored for the Bengali language, catering to the specific linguistic nuances and challenges associated with Bengali text.

Insights into Training Process: The research provides valuable insights into the training process of OGT Language Model, covering aspects such as data preprocessing, model architecture, and optimization techniques. This knowledge serves as a guide for future researchers and practitioners working on similar language modeling tasks.

Performance Evaluation: The model's performance has been evaluated through text generation in response to a given prompt, shedding light on its ability to understand and generate coherent Bengali text.

e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:05/Issue:12/December-2023 Impact Factor- 7.868 www.irjmets.com

Importance of Transformer Architectures:

This research underscores the significance of transformer architectures in advancing non-English language models. The OGT Language Model exemplifies how transformer-based approaches can be tailored to suit the linguistic diversity and complexity of languages like Bengali. As the field of natural language processing continues to evolve, transformer architectures emerge as powerful tools capable of pushing the boundaries of language modeling for diverse linguistic landscapes.

In essence, the OGT Language Model contributes to the ongoing dialogue on language modeling, paving the way for future research endeavors aimed at enhancing Bengali language understanding and generation capabilities.

VIII. REFERENCES

[1] Transformers: https://github.com/huggingface/transformers
[2] Pytorch: - https://pytorch.org/
[3] Gpt2:- https://paperswithcode.com/method/gpt-2