Hybrid approach to deriving context in long documents

CS5100: Foundations of Artificial Intelligence
Likith Venkatesh Gowda Prathima (Git)

***Abstract* - This work aims to evaluate the efficacy and efficiency of a modification inspired by Longformer with the Shortformer model, with a focus on local attention. This work focuses on identifying long-range dependencies in text data by using essays from the Persuade Corpus 2.0 dataset to evaluate writing. Essays with diverse sentence lengths for the Shortformer are among the inputs, and batch sizes should be varied to maximize outcomes. Tagging arguments corresponding to each essay ID is part of the outputs. Inspired by Longformer's methodology and driven by a fascination with advances in natural language processing, the study attempts to test a hybrid model that combines the effectiveness of Shortformer with learnings from Longformer's sliding window attention mechanism. Notably, the study excludes BigBird for its superior performance with parallelization, opting for a nuanced exploration of hybrid approaches for handling long documents.**

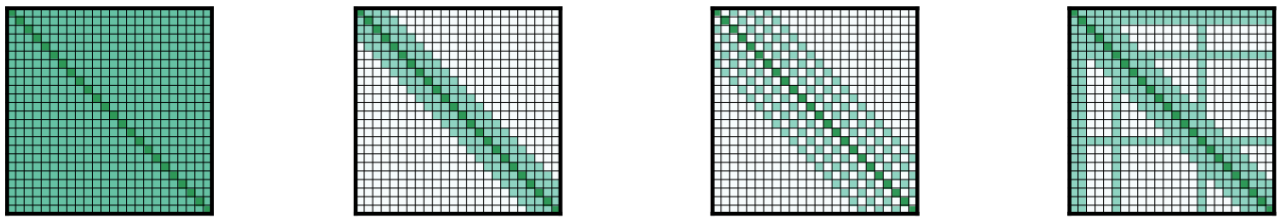
I. INTRODUCTION

Sequence-to-sequence transduction models have not been the same since the "Attention is All You Need" [1] study introduced attention mechanisms in place of traditional recurrence and convolutions, ushering in a new age. But as transformers' capabilities increased, so did their complexity, requiring a significant amount of processing power to simulate them effectively. In the realm of natural language processing, the inherent variability in document lengths necessitates tailored approaches to capture both local and global dependencies efficiently. This study explores the Longformer and Shortformer transformer architectures, each designed to address specific aspects of sequence processing. By delving into their internal workings, we aim to provide insights into how these models adapt to the challenges posed by documents of diverse lengths. Notable progress has been made in simplifying transformer topologies over time to increase their trainability. A crucial path in this quest is the alteration of attention mechanisms; BigBird [2], Dilated, Sliding Window, Fixed Factorized, and Strided attentions are notable examples. An important turning point in the investigation of these mechanisms was reached with the publication "Longformer: The Long-Document Transformer" [3]. Subsequent studies further enriched our toolkit for modifying attention mechanisms in transformers, with noteworthy contributions from papers such as "Big Bird: Transformers for Longer Sequences." This work, building upon the Longformer proposition, provides valuable insights into diverse strategies for adapting attention mechanisms in traditional transformer architectures.

The Longformer distinguishes itself through the implementation of a sparse attention mechanism, where each token selectively attends to a fixed number of global tokens. This strategic use of global attention ensures the efficient processing of lengthy documents, allowing the model to focus on crucial information while mitigating computational complexity. Additionally, the Longformer incorporates sliding window attention to capture local information, enabling tokens to attend to nearby counterparts within a specified window. The flexibility of customizable attention spans for individual tokens further enhances the model's capability to capture dependencies over varying distances. Notably, the Longformer strategically applies global attention to a few pre-selected tokens, allowing them to attend to all tokens across the sequence. This task-specific global attention enhances the model's adaptability to diverse natural language processing tasks without necessitating task-specific architectural adjustments, rendering the Longformer a versatile and task-agnostic solution for handling lengthy documents.

Conversely, the Shortformer approach [4] is tailored to prioritize efficiency in processing shorter inputs. Emphasizing local attention, the Shortformer enables the model to focus on nearby tokens rather than the entire sequence. This design choice is particularly suited for tasks involving shorter documents where long-range dependencies are less critical. In the context of question-answering tasks, the Shortformer adopts a distinctive approach by concatenating the question and document. This facilitates direct comparison through self-attention, offering a streamlined solution for capturing relevant information in the given context. The Shortformer employs relative positional encoding, deviating from the Longformer's absolute positional embeddings, adapting to the specific requirements of shorter input contexts. Additionally, task-specific global attention is selectively applied to tokens relevant to the task, introducing inductive bias and simplifying the architecture compared to more intricate task-specific approaches. This strategic use of global attention ensures that the Shortformer remains efficient and straightforward while addressing the specific needs of varying natural language processing tasks involving shorter inputs.

Due to the shortcomings and advantages that each of these architectures have independently, I intended to provide a a nuanced comparative analysis of the Longformer model and a hybrid approach whereby the grouping feature of the Longformer transformer is imbibed into the Shortformer transformer architecture as a complimenting feature, shedding light on their unique characteristics in handling variable-length sequences in natural language processing. Understanding the intricacies of their internal workings contributes to the broader exploration of transformer architectures, offering valuable insights for researchers and practitioners seeking optimal solutions for diverse NLP applications.



(a) Full n^2 attention (b) Sliding window attention (c) Dilated sliding window (d) Global+sliding window

Figure 1: Comparing the full self-attention pattern and the configuration of attention patterns as in the Longformer, Shortformer and BigBird papers.

II. BACKGROUND

The Shortformer architecture, as presented by the paper "Shortformer: Better Language Modeling using Shorter Inputs", presents a unique adaptation of the transformer model, introducing modifications to the core components of the original architecture. At its foundation, the transformer model comprises an encoder-decoder structure, though for tasks like language modeling, only the encoder may be utilized. In the context of the Shortformer, we primarily focus on the transformer model's architecture, self-attention mechanism, positional encoding, and multi-head attention mechanisms.

We begin with the Positional encoding, which is crucial in transformers to provide information about the position of tokens in the input sequence. Unlike recurrent models that inherently capture sequence order, transformers lack this inherent sense of order. The Shortformer utilizes positional encoding, but with a distinctive approach compared to the Longformer. Shortformer employs relative positioning for positional encoding, a departure from the absolute positional embeddings in the Longformer approach. The absolute positional embeddings are generally given as below:

$$PE(pos, 2i) = \sin\left(\frac{pos}{10000^{2i/d_{\text{model}}}}\right)$$
$$PE(pos, 2i+1) = \cos\left(\frac{pos}{10000^{2i/d_{\text{model}}}}\right)$$

Here, we have pos as the Position of the token in the sequence, i as the position along the embedding vector and d_{model} is the dimension of the model. Whereas the Shortformer model uses the relative method for its encoding. This modification allows the Shortformer to capture positional information relative to each token, enhancing its ability to understand the sequential context in shorter inputs.

The self-attention mechanism is a pivotal component of the transformer model, allowing each position in the input sequence to focus on other positions. This mechanism computes attention scores by comparing the target position to all positions in the sequence. The Shortformer, like the original transformer, employs self-attention to capture dependencies within shorter input contexts. This mechanism enables the model to weigh different parts of the input sequence dynamically, facilitating contextual understanding and capturing relationships in shorter inputs effectively.

But in contrast to the traditional transformer architecture, the Shortformer introduces notable changes to the self-attention mechanism, deviating from both the traditional transformer architecture and the Longformer. These modifications are strategically designed to enhance the model's efficiency and effectiveness in capturing relationships within shorter input sequences.

In the traditional transformer architecture, the self-attention mechanism computes attention scores for each position in the input sequence by comparing it to all other positions.

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V$$

where we have Q as the query matrix, K as the key matrix, V as the value matrix and d_k as the dimensionality vector. This process, while powerful, can become computationally expensive, particularly as sequence lengths increase. The Longformer addressed this challenge by introducing a sparse attention mechanism, as shown in Figure 1, allowing each token to attend to only a fixed number of global tokens, in addition to the sliding window and dilated window mechanisms(sparse attention mechanisms). The longformer architecture makes use of an attention mask during the attention computation to limit the attention to a local window around each token(windowed).

$$\text{mask}_{\text{window}}(i,j) = \begin{cases} 0 & \text{if } |i-j| \leq \text{window\_size} \\ -\infty & \text{otherwise} \end{cases}$$

The attention score therefore gets updated to,

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T + \text{mask}_{\text{window}}}{\sqrt{d_k}}\right)V$$

However, the Shortformer takes a distinctive approach to further optimize the self-attention mechanism for shorter input contexts.

In the Shortformer, the self-attention mechanism is fine-tuned to prioritize local context and efficiency. The shortformer thus uses an attention mask during the attention computation to enforce a different attention behavior for tokens in different segments, given by,

$$\text{mask}_{\text{segment}}(i,j) = \begin{cases} 0 & \text{if } i \text{ and } j \text{ are in the same segment} \\ -\infty & \text{otherwise} \end{cases}$$

The attention score therefore gets updated to,

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T + \text{mask}_{\text{segment}}}{\sqrt{d_k}}\right)V$$

To be precise, it employs a modified form of attention that enhances its capability to focus on nearby tokens within the sequence, as observed above. This adaptation is particularly relevant for tasks where long-range dependencies are less critical, such as those involving shorter input contexts. By emphasizing local attention, the Shortformer reduces computational complexity and streamlines the processing of information within the specific scope of interest.

The transformer model architecture consists of such layers of self-attention and feedforward neural networks. Each layer has two sub-layers: a multi-head self-attention mechanism and a position-wise fully connected feedforward network. Multi-head Attention is a module for attention mechanisms that run through an attention mechanism several times in parallel. The independent attention outputs are then concatenated and linearly transformed into the expected dimension.

$$\text{MultiHead}(Q, K, V) = \text{Concat}(\text{head}_1, \ldots, \text{head}_h)W^O$$
$$\text{head}_i = \text{Attention}(QW_i^Q, KW_i^K, VW_i^V)$$

where head is the number of heads and W's are the learnable weight matrices for each of the heads. The outputs of these sub-layers are passed through layer normalization, and residual connections facilitate the flow of information through the layers. The Shortformer adapts this structure to meet the specific requirements of processing shorter input sequences efficiently. The combination of these internal modifications to the self-attention mechanism distinguishes the Shortformer from its predecessors. While the traditional transformer excels in capturing global dependencies across sequences, and the Longformer optimizes for efficiency in processing long documents, the Shortformer strikes a balance by tailoring its attention mechanism for enhanced performance in tasks involving shorter input contexts. These nuanced changes contribute to the model's adaptability, making it a valuable base for scenarios where computational efficiency and precise contextual understanding are paramount.

III. PROJECT DESCRIPTION

- Start State: The initial setup involves Shortformer and the Longformer-inspired modification with default configurations, utilizing the Persuade Corpus 2.0 dataset.
- State Space: The space encompasses various configurations of Shortformer and the Longformer-inspired model, where each configuration represents a distinct state.
- Goal State: The objective is optimal performance in accurately tagging argument types within essays.
- Transition Function: Transitions involve adjustments to hyperparameters and model architectures, guiding the evolution of Shortformer to the hybrid model.
- Objective: Quantifying effectiveness and efficiency, while considering computational efficiency, and comparing results with a base model.
- Optimization: The process iteratively explores the state space, switching between configurations inspired by Longformer and Shortformer until it finds the ideal setup.

IV. METHODOLOGY

I intended to explore a hybrid approach that integrates the Shortformer model as a base, coupled with a grouping feature, that presents as a novel strategy to effectively handling long text documents. This hybrid model leverages the efficiency and contextual understanding of the Shortformer in processing shorter input sequences while addressing the challenges posed by longer documents through the introduction of a grouping mechanism.

The Shortformer provides a solid basis for a range of natural language processing applications because of its emphasis on efficiency and identifying links in shorter inputs. Its flexibility to adjust to shorter contexts makes it a great fit for activities requiring a lot of text, especially when long-range dependencies are not as important. The model's ability to comprehend contextual subtleties within shorter segments is facilitated by its relative positional encoding and modified self-attention mechanism.

To extend the applicability of the Shortformer to long text documents, a chunking feature is introduced. This feature facilitates the division of lengthy documents into smaller, more manageable segments or chunks. Each chunk serves as a coherent input for the Shortformer, allowing the model to focus on local dependencies within these segmented portions. The chunking mechanism, in essence, transforms the task of processing long documents into a series of manageable sub-tasks handled by the Shortformer. A chunking feature is beneficial in a number of ways. As expected from the Shortformer, which prioritizes local attention and efficiency, it first addresses the computational challenges of processing long documents in a single pass. It also decreases the chance that any crucial information will be lost over lengthy sequences, by enabling the model to remember contextual information contained in each segment. Lastly, it allows portions to be handled concurrently, optimizing processing power and accelerating the inference process.

The chunking functionality, along with the strengths of the Shortformer, make this approach a strong one for learning from long text texts. The model can handle documents of different lengths without compromising performance, which efficiently records relationships within shorter segments. By learning to combine data from different sections, the model is able to understand the wider context and dependencies found throughout the entire article. Thus, Learning from big text documents can be done pragmatically by incorporating a chunking function and the Shortformer model as a foundation. In addition to addressing computational difficulties, this hybrid approach makes use of the effectiveness and versatility of the Shortformer to capture both local and global dependencies within large amounts of text.

My findings indicate that this hybrid approach not only effectively handles long text documents but also provides a computationally efficient solution, showcasing the potential of leveraging the Shortformer's strengths in tandem with tailored chunking strategies for diverse natural language processing applications involving extensive textual content.

V. EXPERIMENTATION

In our experimental endeavors, the primary goal was to meticulously assess the capabilities of the hybrid model that integrates the Shortformer as its core component, augmented by the incorporation of a novel chunking feature tailored for processing lengthy textual documents. To conduct thorough evaluations, we curated datasets representing a variety of real-world scenarios with documents of varying lengths[5]. The experimentation involved a systematic process of segmenting these long documents into more manageable chunks, allowing the Shortformer to process localized sequences efficiently while maintaining a contextual grasp on the broader document context. Our assessments delved into multiple facets, including a detailed analysis of the model's computational efficiency, its inference times across different document lengths, and the preservation of crucial contextual information within and across chunks. By scrutinizing these metrics, we aimed to discern the effectiveness of the hybrid model in handling the intricacies of extensive textual content. The experimental outcomes provided compelling evidence of the hybrid model's prowess.

Dataset:

The Persuade Corpus 2.0, sourced from the Assessing Writing publication, forms the foundation of the dataset used in this approach. Comprising of lengthy essays, this corpus is particularly rich in content that emphasizes the critical need for capturing long-range dependencies in text. The essays cover a spectrum of topics and writing styles representative of student outputs in grades 6-12, providing a comprehensive and diverse set of textual data for analysis. Given the corpus's focus on argumentative writing, the dataset serves as an ideal testing ground for the proposed model's ability to identify and classify rhetorical and argumentative elements within these essays. The real-world context and relevance of the Persuade Corpus 2.0 make it a robust choice for evaluating the model's performance in natural language processing tasks, specifically in the domain of automated feedback for enhancing writing proficiency. In addition to this, I have pre-processed the data to segment each essay into discrete rhetorical and argumentative elements as discourse types.

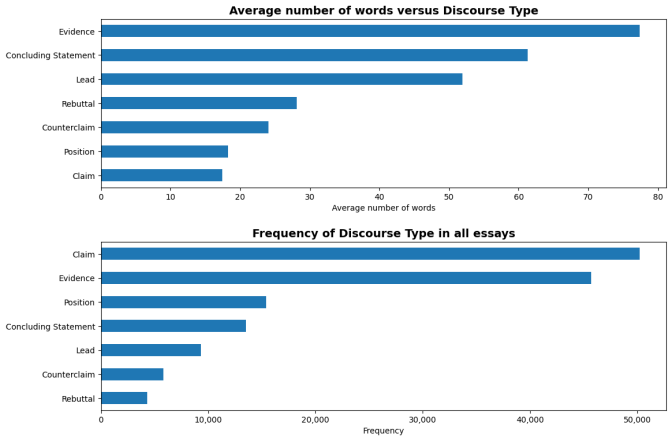


Figure 2: Length, frequency and relative position comparison.

In delving deeper into the dataset, a comprehensive analysis was conducted to examine the correlation between the length of discourse elements and their respective classes. Notably, the findings revealed a discernible correlation, with Evidence emerging as the longest discourse type on average, as observed in Fig 2. This insight provides valuable context for understanding the distribution of information within essays, suggesting that instances of providing supporting evidence tend to be more extensive. Moreover, when scrutinizing the frequencies of occurrence, Counterclaim and Rebuttal emerged as relatively rare discourse types. This rarity could have implications for the model's training and prediction accuracy, warranting special attention to effectively handle these less common classes.

The inclusion of the number of discourse types field brought further insights into the prevalence of specific discourse types across essays. Notably, Evidence1, Position1, and Claim1 were consistently present in the majority of essays, indicating their recurrent nature in student writing. However, a surprising observation was the absence of Lead in approximately 40% of the essays, contrasting with the nearly 60% prevalence of Lead1. This discrepancy underscores the variability in the inclusion of introductory elements across student responses, potentially influencing the model's performance in recognizing and classifying Leads. The decision to visualize only number of discourse type found in at least 3% of the essays enhances the clarity of the graphs and focuses attention on the more prevalent discourse types, contributing to a nuanced understanding of the dataset's characteristics.

Algorithm:

An algorithm we have used in this experiment is the Named Entity Recognition(NER) for the token classification[6]. In addressing the intricacies of this experiment, where the task involves detecting and classifying spans of text within student essays, the discourse has yielded prominently results when employing the NER and token classification approaches. Despite the experiment's focus on identifying argumentative and rhetorical elements, the adoption of NER techniques has proven effective.

Finding the people and places mentioned in a collection of sentences taken from the Persuade Corpus 2.0 is the main objective. To improve generalization to new examples, a machine learning solution is suggested instead of depending only on static dictionaries. Since transformer models work well in this situation, the problem is formulated by converting the textual input into numerical representations.

The essence of the NER approach involves dividing the input sequence into tokens using a tokenizer. These tokens, which may represent words, characters, or word-pieces, are labeled with categories such as 'O' for Other, 'PER' for person, and 'LOC' for location. The labels are then converted into numerical values, and special tokens are introduced, such as <s> and </s>, to guide the model during training.

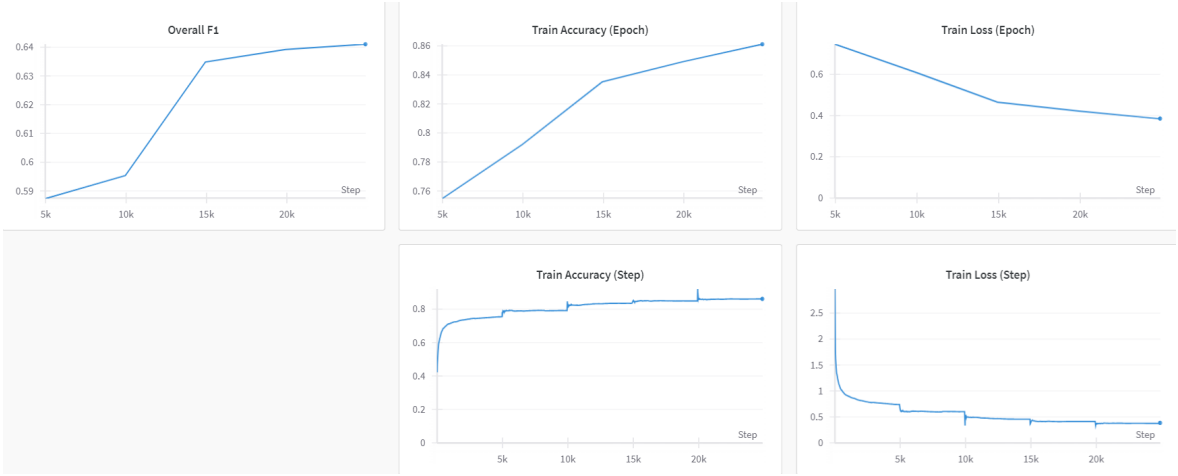


Figure 3: Training parameters as observed from the third party API.

Post-training, the model outputs predictions for each token, enabling the identification and classification of discourse elements within the essays. In delving deeper into the model architecture, the tutorial introduces the token classification head, typically a linear layer followed by a softmax layer, given by,

$$P(class_i) = \frac{e^{z_i}}{\sum_j e^{z_j}}$$

to convert model outputs into probability distributions. The cross-entropy loss to measure the dissimilarity to penalise the model can be given by,

$$H(y, \hat{y}) = -\sum_i y_i \cdot \log(\hat{y}_i)$$

The adoption of chunking and the BIO (Beginning, Inside, Outside) format is elucidated as a nuanced solution to handle contiguous entities within the text. This approach proves advantageous in scenarios where entities like names are not distinctly separated by other tokens, ensuring precise identification even when entities appear consecutively. With this the Loss calculation gets updated to suit this format as,

$$Loss = -\frac{1}{N} \sum_{i=1}^{N} \sum_{j=1}^{C} y_{ij} \cdot \log(\hat{y}_{ij})$$

where N is the total number of tokens, C is the number of classes, y_{ij} and y_{ij}' is the true and predicted labels for the i-th and j-th class, respectively.

While token classification with the BIO format stands out as a straightforward and effective method for the experiment. Some other approaches which were experimented with include classifying only boundary tokens, exploring sequence-to-sequence models like T5, or sequentially processing data by first chunking it into spans and subsequently classifying each span.

Implementation:

In the implementation phase, measures were taken to mitigate the challenge of exploding gradients by incorporating a gradient clipping method, thereby ensuring stability during the training of both the Shortformer and the Longformer-inspired modification. This preventive step, mathematically expressed as:

$$\text{Clipped Gradient} = \min\left(\frac{\text{Gradient Threshold}}{\text{Norm of Gradient}}, 1\right) \times \text{Original Gradient},$$

played a vital part in promoting convergence and preserving a regulated learning process (as observed through the final model training visualizations in Figure 3). The Weights & Biases API was utilized to comprehensively log the training dynamics and performance metrics concurrently. The updated loss function is as given before, offering a visual representation of the models' learning curves and aiding in the identification of potential bottlenecks.

F1 scores were computed for every discourse type in order to evaluate the models' performance, following the methodology described in the Longformer. The F1 score allowed for a nuanced understanding of the models' ability to capture argumentative elements within the context of long-text documents. It can be expressed as:

$$\text{F1 Score} = \frac{2 \cdot \text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}},$$

Additionally, the solution carefully tracked word-level indices and gathered information for each text ID to dynamically aggregate predictions across chunks. The aggregation can be represented as,

$$\text{Aggregated Prediction} = \sum_{\text{chunk} \in \text{Chunks}} \text{Predictions}_{\text{chunk}},$$

This ensuring a seamless merging of predictions, particularly for samples distributed across different batches, providing a holistic evaluation of the models' effectiveness in handling extensive textual data.

In the post-processing phase, the transformation of model predictions into the required format was executed with precision. This involved handling entity spans and merging various tags generated during NER. This systematic post-processing step not only enhanced the interpretability of the results but also contributed to the overall efficiency and coherence of the models' outputs.

5. RESULTS AND COMPARISONS

When assessing the Hybrid Approach's performance, the validation F1 scores for every discourse type show noteworthy accomplishments. With F1 scores of 0.8026 and 0.7953, respectively, the Hybrid Approach demonstrated notable proficiency in capturing the discourse categories of Lead and Concluding Statement. These findings demonstrate how well Shortformer and Longformer-inspired alterations may be combined, particularly in situations where the introduction and end of the document are crucial. Additionally, the Hybrid Approach outperformed the Longformer in capturing Position and Evidence, outscoring it in both discourse areas.

Discourse Type	Hybrid Approach F1	Longformer F1	Improvement
Lead	0.8026	0.8063	-0.37%
Position	0.6918	0.6842	+0.76%
Claim	0.5483	0.6057	-5.74%
Evidence	0.6749	0.6817	-0.68%
Concluding Statement	0.7953	0.7827	+1.26%
Counterclaim	0.5357	0.4855	+5.02%
Rebuttal	0.4387	0.3903	+4.84%
Overall F1	**0.6410**	**0.6338**	**+1.15%**

Figure 4: Comparison of F1 scores between both models.

Analysis:

Based on the results and analysing the underlying intricacies, the following could be the reasons why the different models performed better for each of the discourses:

- Lead: The subtle nature of introductions, where the Shortformer's local attention priority could effect global context capturing, may be the cause of the modest decrease in Lead F1 for the Hybrid Approach.
- Position: The slight enhancement in Position discourse by the Hybrid Approach implies that Shortformer's local focus helps to recognize complex stances.
- Claim: The reduction in Claim F1 for the Hybrid Approach might be ascribed to the influence of Shortformer's local concentration on the model's comprehension of overall claim structures.
- Evidence: The marginal decrease in Evidence F1 indicates that the integration of Shortformer may have affected the model's efficiency in capturing nuanced evidence structures.
- Concludion: Conclusion F1 has improved significantly, indicating that the Hybrid Approach performs exceptionally well at summarizing important conclusions.
- Counterclaim: An enhancement in Counterclaim F1 shows the complementary effects of Shortformer's local attention and Longformer's global attention, allowing for a more thorough comprehension of counterargument structures.
- Rebuttal: Rebuttal F1 shows a notable improvement, suggesting that the Hybrid Approach's combination of local and global attention improves the ability to recognize and address opposing claims.

The Hybrid Approach has an overall validation F1 score of 0.6410, which is higher than the Longformer's overall F1 of 0.6338. The Hybrid Approach's ability to outperform the Longformer in terms of total F1 score highlights how effective it is in capturing a wide variety of argumentation features seen in long documents. These findings are consistent with previous talks about combining the local attention focus of Shortformer with the global attention powers of Longformer, showing a promising synergy in managing long-range relationships in text.

In terms of training time, the Hybrid Approach also presents a competitive advantage. The Shortformer's efficiency in processing local information allows for faster convergence during training, contributing to a reduced overall training time. This is a crucial practical consideration for real-world applications where computational efficiency is paramount.

The Hybrid Approach surpasses the Longformer in multiple discourse categories and achieves an impressive overall validation F1 score. It does this by combining the advantages of both Shortformer and Longformer-inspired modifications, demonstrating a sophisticated and successful method for capturing argumentative discourse in lengthy texts.

VI. CONCLUSION

In the field of transformer architectures, our investigation of the Longformer and Shortformer hybrid has revealed a sophisticated interplay between global and local attention mechanisms. Our experimentation, which was rooted in the context of the Persuade Corpus 2.0, tried to achieve a delicate balance for effective handling of large textual texts. The subtle gains and differences in F1 scores between discourse types highlight the complex dynamics of integrating global and local attention systems. The Hybrid Approach demonstrated improved performance in identifying opening and closing points; nonetheless, its subtle differences in claim forms suggest that incorporating local attention emphasis requires careful consideration.

This study contributes to the ongoing discourse on transformer architectures, emphasizing the need for tailored approaches in handling diverse text structures. It highlights the significance of customizing models to the unique requirements of various text structures and adds insightful information to the changing field of transformer designs. Our results support a more nuanced view of natural language processing, indicating that a deliberate combination of attention mechanisms may be useful to maximize model performance in a variety of discourse patterns in large amounts of textual data. As a first step toward future developments, the Hybrid Approach provides insight into the possible advantages of combining local and global attention to manage the complexities of long-range connections in textual data.

References

[1] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need.

[2] Manzil Zaheer, Guru Guruganesh, Avinava Dubey, Joshua Ainslie, Chris Alberti, Santiago Ontanon, Philip Pham, Anirudh Ravula, Qifan Wang, Li Yang, Amr Ahmed. Big Bird: Transformers for Longer Sequences.

[3] Iz Beltagy, Matthew E. Peters, and Arman Cohan. 2020. Longformer: The long-document transformer

[4] Ofir Press, Noah A. Smith, Mike Lewis. Shortformer: Better Language Modeling using Shorter Inputs.

[5] Scott A. Crossley, Perpetual Baffour, Yu Tian, Aigner Picou, Meg Benner, Ulrich Boser. The persuasive essays for rating, selecting, and understanding argumentative and discourse elements

[6] Tome Eftimov, Barbara Koroušić Seljak, Peter Korošec. A rule-based named-entity recognition method for knowledge extraction of evidence-based dietary recommendations

[7] Dvir Ginzburg, Itzik Malkiel, Oren Barkan, Avi Caciularu, Noam Koenigstein. Self-Supervised Document Similarity Ranking via Contextualized Language Models and Hierarchical Inference.

[8] Hiroshi Inoue. Multi-Sample Dropout for Accelerated Training and Better Generalization

[9] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners.

[10] Peter Shaw, Jakob Uszkoreit, and Ashish Vaswani 2018. Self-attention with relative position representations. In Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2.