Recent Advances in Long Documents Classification Using Deep-Learning

Muhammad Al-Qurishi
Elm Company,
Research Department,
Riyadh 12382, Saudi Arabia,
mualqurishi@elm.sa

Abstract

Long document classification is one of the most challenging linguistic processing tasks. Recently, Recent deep-learning models such as the transformers have proven to perform this task with a high-level of success. Such models typically include the self-attention mechanism, which makes the calculations extremely complex as document length increases. In order to unlock the use of the most accurate document classification tools on a wider range of document types and make the use of methods based on self-attention practically feasible, it's necessary to introduce some innovations that facilitate better scaling. In this work we provide a quick and concise survey of recent research work in the area of long documents classification using deep-learning techniques. The advantages and disadvantages of these methods have been discussed along with some directions that may be useful in future research.

1 Introduction

Modern deep learning models for semantic analysis can achieve impressive results after they are trained on very large datasets, gaining the ability to generate highly accurate predictions about content they haven't seen before. However, their capacity to capture the relationships between words and sentences is dependent on statistical operations that grow more complex as the length of the text sequence is increased (Tay et al., 2020). Effectively, this renders many of the best methods nearly useless from a practical standpoint and necessitates development of tools that can realistically process long documents and return reliable results within a reasonable timeframe. In particular, methods that rely on the attention mechanism (BERT (Devlin et al., 2018) and its derivatives) tend to become computationally demanding when working with long text documents (Vaswani et al., 2017).

Resolving this problem would allow for much wider use of document classification algorithms and would potentially allow large organizations in many different sectors to manage their administrative burden more efficiently. This is why researchers are looking into different possibilities for updating the existing algorithms, implementing changes aimed specifically at improving performance with long text, and testing hybrid approaches that offer better ratio between training/inference time and accuracy of prediction (Wagh et al., 2021; Tay et al., 2020). Their efforts are going in different directions and at present time it's unclear which approach might offer the best chances to overcome the current limitations, but there are already some promising findings that could hint towards sustainable solutions.

In this research paper we are trying to introduce a quick and concise summaries of the recent research papers published in the area of deep learning that trying to solve the problem of long document classification. Then we compare them from several perspectives such as: used method, authors objectives, performance and datasets. Finally, we conclude this review with a discussion of some ideas and suggestion that might become the basis for a new research in this area.

2 Long Document Classification Techniques

In this section we categorize the techniques and methods that have been used for long document classification in the most relevant existing research works into two main categories; Transformer-based technique such as BERT and hybrid techniques which combines two or more deep neural network (e.g., CNN, RNN, transformers, etc.), to improve the performance of the classification model. All of these techniques are aiming to identify procedures that could remove some of the well-known limitations of working with very long documents.

2.1 Transformer-based

Transformer architecture was first proposed in 2017 by (Vaswani et al., 2017), and it quickly lead to several successful implementations, most notably BERT by (Devlin et al., 2018) and XLNet by (Yang et al., 2019). Those models rely on bidirectional transformations of input that allow for superior tracking of semantic relationships. One of the most important properties of a Transformer such as BERT is that it can be pre-trained and fine-tuned for specific tasks, languages, and subject matters. One major problem with BERT is its limit to sequence length of 510 tokens. Several approaches to this problem were considered, including chunk selection, efficient self-attention, document truncation, and key sentence selection, with one existing method chosen as a representative of each theoretical direction.

(Sun et al., 2019) show how to train and fine-tune BERT for text classification. They tested their model using standardized hyper parameters such as batch size and sequence length, with several different datasets suitable for question classification, topic classification, and sentiment analysis. The best option was determined based on experimental results, for example, in this way it was found that multiple strategies can be used to get around BERT's limit to the sequence length, but the 'head & tails' strategy where only the first 128 and the last 382 tokens are kept performs the best (Pappagari et al., 2019).

On the other hand (Ding et al., 2020) presents a very creative solution to get around BERT's limit to sequence length. This solution inspired by a cognition theory of working memory proposed by Baddeley in 1992. The solution is named CogLTX, and it starts from the logical assumption that a majority of semantically important information is concentrated within specific sentences inside of a longer text, making it unnecessary to check for connections between all words in a document. Instead, they propose training a judge model that can recognize high-relevance sentences and pass them as input to the reasoning model that can complete the classification task.

Another work by (Adhikari et al., 2019) attempt to develop a computationally more efficient version of BERT that would be better suited for classification of long documents. Knowledge distillation is used where knowledge is transferred from a large version of BERT to a much smaller BiLSTM network, which is then used to perform the classification task on new examples. Likewise, (Beltagy et al., 2020) attempt to adjust the successful Transformer architecture and make it better suited for analyzing long documents. Due to exponential increase in computational complexity, with models of this kind it becomes nearly impossible to handle documents whose length exceeds a certain arbitrary threshold. To remedy this issue, the authors propose an altered model they named Longformer, which reduces the complexity of the self-attention matrix.

2.2 Hybrid Methods

Two main approaches that we investigate are algorithms with sparse attention and hierarchical models, each of which has inspired a number of attempts to adjust the DNN architecture for the long document classification tasks. Some works use CNN and utilizing local convolutional feature aggregation to obtain the predictions of the document class (Liu et al., 2018). They proposed a RNN component to the architecture described in the first method, and uses it to track the semantic order for each of the selected sequences. Another work by (He et al., 2019) present a hybrid solution developed specifically with the idea to perform long document classification more efficiently. It combines a RNN-based controller component with a CNN that extracts discriminative features from linguistic content. The controller contextualizes the attention and localizes the extracted bits into a coarse representation of the document, before grouping the extracted features to acquire the overall structure.

Some other ideas are to avoid the exponential growth of complexity along with text sequence length that is typical for architectures of this kind (Huang et al., 2021; Park et al., 2022). For example, (Khandve et al., 2022) propose a hybrid solution with transfer learning as the central principle and a hierarchical architecture that reduces the number of necessary operations. The input data was divided into parts and processed with Universal Sentence Encoder and BERT, both of which were pre-trained. After this, the results were passed onto a shallow network based on either LSTM or CNN concept, which was used in the role of a classifier. Different combinations were explored to determine whether an improvement can be observed, and in case of USE it was possible to improve accuracy in this manner while for BERT the addition of

CNN classifier resulted in similar level of accuracy but with much lower computational requirements thanks to the hierarchical structure.

Another example was developed for the medical domain, and is characterized by strong performance with document of extraordinary length that are typical within this sector (Hu et al., 2021; Si and Roberts, 2021). (Hu et al., 2021) describes a hybrid model that includes several components, including bi-directional recurrent neural network (RNN), convolutional neural network (CNN) and the attention mechanism. Words are embedded as vectors in the initialization phase, before n-grams are extracted from sentences to capture more of the semantic context. A matrix of features is constructed with a combination of CNN output with the ReLu unit.

3 Comparison of the proposed methods

Most of the recent works addressing the problem of long document classification start from similar principles common to all deep learning methods. They also diverge in many aspects, as the authors explore different avenues for leveraging the power of the learning algorithms and overcoming the most significant obstacles (Dai et al., 2022). Since the authors are essentially attempting to solve the same problem, namely how to maintain high accuracy of semantic predictions while keeping the computing demands reasonable, it would be fair to describe the papers as belonging to the same family despite the considerable differences in approach.

To provide an impartial comparison of the proposed models and evaluate the degree of innovation they introduce, it's necessary to look at several elements present in each work, including:

- Basic methodological blueprint used to construct the model
- Priority objectives the authors are trying to achieve
- Statistical operations and training procedures designed to improve accuracy and/or efficiency
- Datasets and standards used for quantitative evaluation of the model
- Potential for real-world applications and follow-up work

In terms of methodological choices, practically all works from this group acknowledge the unmatched power of the attention mechanism for analyzing semantic relationships, and incorporate it in some way into the proposed architecture. There is a division between works that mostly (or completely) embrace an existing architecture and perform only minor operations such as fine-tuning or knowledge transfer in order to reduce the computational demands (Adhikari et al., 2019; Sun et al., 2019). On a different end of the spectrum, there are works that propose innovative hybrid solutions in which the attention mechanism and/or Transformer architecture are combined with elements of different deep learning paradigms, such as RNNs and CNNs. In particular, a common strategy is to adopt a hierarchical structure for the overall solution and use the attention mechanism only in a limited role, thus avoiding the exponential growth of complexity (Huang et al., 2021; Si and Roberts, 2021).

The aforementioned methodological differences stem largely from the expectations for each paper, which range from proving a theoretical point to attempting to develop specialized model for long document classification. Works with a narrower scope tend to stay closer to the original BERT model design (Beltagy et al., 2020), while more ambitious efforts that aim to create new tools are more inclined to experiment with previously untested combinations of elements. In some papers, the scope of intended applications is limited to long documents from a certain domain (i.e. medical) (Si and Roberts, 2021), while others are approaching the problem in more general terms. Finally, there is an important distinction between works that aim for greater accuracy, and those that primarily attempt to improve computational efficiency and shorten the inference time (Park et al., 2022).

It's a fair assessment that practically all works from this group are grappling with the same problem – the tendency of attention-based models to become prohibitively complex as the length of the analyzed text is increased. In response, the authors tried a variety of ideas that rely on vastly different mechanisms to decrease complexity. From fine-tuning and knowledge distillation to introduction of hierarchical architectures and restrictive elements such as fixed-length sliding window (Beltagy et al., 2020; Wang et al., 2020), the proposed techniques are quite innovative and typically leverage some known properties of deep learning models to affect

Table 1: Comparison of the reported performance of different long document classification methods

Author	Method	Reported accuracy
(Liu et al., 2018)	Local convolutional feature aggregation	From 88 to 95.4%
(Sun et al., 2019)	BERT + head and tail truncation	Error rates from 0.67 to 5.4
(He et al., 2019)	Recurrent attention learning	From 77.4 to 80.4%
(Adhikari et al., 2019)	DocBERT	From 54 to 91%
(Pappagari et al., 2019)	RoBERT/ToBERT	From 82 to 95.4%
(Wang et al., 2020)	Dynamic hierarchical topic graph	From 68.8% to 97.3%
(Beltagy et al., 2020)	LongFormer	F1 score from 64.4 to 94.8
(Ding et al., 2020)	CogLTX a BERT-based model + MemRecall algorithm	F1 score from 70% to 97%
(Si and Roberts, 2021)	Hierarchical Transformer	Up to 94.7%
(Huang et al., 2021)	Hybrid self-attention sparse network	From 73.2% to 95.7%
(Khandve et al., 2022)	Hierarchical Attention Network (HAN) + BERT + CNN/LSTM	From 80 to 100%

how the attention mechanism performs in a particular deployment. The diversity of ideas found in those papers illustrates that researchers are currently casting a wide net and searching for unconventional answers to a difficult problem, without a single dominant strategy. On the other hand, hybrid approaches hold a lot of promise and they combine some proven elements from different methodologies into new, potentially more optimal configurations (Huang et al., 2021; He et al., 2019).

Evaluation of the proposed changes to established algorithms is crucially important, and all of the reviewed works include some form of empirical confirmation of their premises. While the numbers seemingly validate that the proposed solutions achieve state-of-the-art results under the best possible conditions, those findings are self-reported and may often be too optimistic. All of the papers are interested in document classification tasks and use it to evaluate their solutions, but datasets used for testing may not be the same in terms of size, diversity, and content. When directly comparing different solutions, it's extremely important to keep in mind the particulars of the evaluation protocols. Studies aiming to provide evaluations with independently administered comparative testing of several different BERT-like algorithms for document classification are slowly emerging and reporting some interesting findings that often diverge from self-assessed results (Wagh et al., 2021; Dai et al., 2022; Park et al., 2022). Still, there are no widely accepted evaluation standards and every comparison suffers from 'apples-to-oranges' problem up to an extent.

When it comes to practical use of the proposed solutions, there is a general lack of field data and even discussions of use cases are rare. This is understandable considering the main focus is on discovering more efficient methods, but without real world testing it's difficult to predict whether any of the solutions can deliver similar results to their reported findings. Some works may be directed as specific niches such as legal (Wan et al., 2019; Bambroo and Awasthi, 2021) or medical (Si and Roberts, 2021), but even in this case little attention is paid to practicalities associated with real world application. This weakness may reflect the current state of the field, which is highly experimental and mostly built on data collected in a controlled environment.

4 Datasets and Reported Accuracy

As previously stated, all papers include an experimental evaluation of the proposed solution and present certain quantitative findings that underscore their methodological choices. In particular, they measure the ability of the model to correctly classify long documents by topic. The performance is typically reported in terms of accuracy with several different metrics, but other aspects may be tracked as well such as complexity or speed of inference. It's important to note that multiple versions of the algorithms are tested on several datasets in each study, which is why accuracy estimations are given as ranges as shown in Table 1.

The choice of the training and testing datasets also carries great significance when analyzing the output of various deep learning algorithms. The same is true for the length of documents, as all of the reviewed papers state among their objectives the improvement of performance with long text sequences. Datasets may also differ by their volume, the number of classes, and other parameters as well, and some studies may include tasks other

Table 2: Overview of the datasets used for training and evaluation with average sequence length

Author	Datasets	Average document length
(Liu et al., 2018)	Custom set comprised of arXiv papers	6000 words
(Wang et al., 2020)	20NG, R8, R52, The Oshumed, MR,	From 39 to 389
(Sun et al., 2019)	IMDB, Yelp reviews, TREC, Yahoo answers, AG News, DBPedia, Sogou	10 – 740 words
(He et al., 2019)	Custom set comprised of arXiv papers	6300 words
(Adhikari et al., 2019)	Reuters, AAPD, IMBD, Yelp 2014	145 -390 words
(Beltagy et al., 2020)	WikiHop, TriviaQA, HotpotQA, OntoNotes, IMDB, Hyperpartisan	from 139 -2000
(Ding et al., 2020)	NewsQA, 20NewsGroups,Alibaba and HotpotQA	from 300-650 Limited by MemRecall block
(Pappagari et al., 2019)	CSAT, 20NG, Fisher Phase I	260 – 1790 words
(Si and Roberts, 2021)	MIMIC III	700 tokens
(Huang et al., 2021)	IMDB, Yelp 2018	From 100 to 500 words per review
(Khandve et al., 2022)	20NG, BBC News, AG News, BBC Sports, IMDB, R8	From 39 to 389 words per document

than document classification. The overall datasets used for training and evaluation with average sequence length are presented in Table 2.

5 Conclusion and Future Directions

It is beyond doubt that Transformer architecture changed the way linguistic analysis is performed, and in a very short time BERT has been widely accepted as the golden standard of semantic understanding. However, the greatest value of this concept may be tied to its flexibility, as it allows for extensive customization and specialization with only minimal modifications of the training procedure. While there have been numerous adaptations of successful Transformer models in the past, it's highly likely that the number and quality of derivative work will increase in the near future. Figuring out ways to improve an already impressive model is not easy, but growing presence of this topic in the online forums and greater availability of research papers dealing with some of the outstanding challenges could power the next wave of research in this direction. This process is already underway, and a breakthrough achieved with Transformers is being actively exploited by research teams from around the world.

Computational efficiency remains the central challenge, and developing models that can achieve elite accuracy on a wide range of tasks without requiring escalating amount of resources is a top priority for the next stage of research. Some of the ideas presented in the reviewed works will certainly be revisited and expanded in the coming years, and their cumulative contributions could eventually lead to a consensus solution. In parallel with the process of consolidation of knowledge and resolving practical difficulties, we can also expect to see a larger number of domain-specific applications that are designed and trained with real-world use in mind. Since in many domains there are long documents to be classified, solving the difficulties that current algorithms are having with long text sequences will stay a key objective. Localization is another issue that should be addressed in future work, as most of the current tools were never tested with non-English datasets. Given that the volume of non-English documents is enormous and growing very fast, it would be refreshing to see language-specific applications that match the quality of original BERT.

Hybridization of models remains an area that hasn't been sufficiently explored, in part due to huge potential for mixing and matching different elements. The advantages offered by older paradigms such as recurrent or convolutional neural networks shouldn't be ignored, and some very imaginative efforts to combine them with the attention mechanism were made. Hybrid approaches to long document classification are rapidly emerging over the last few years (Qin et al., 2022), and some of them deserve to be explored further. Balancing complexity of the model and compatibility of all components presents a unique challenge, and it may take several years before fully mature solutions of this type start appearing.

6 Acknowledgments

Author wants to express their gratitude towards the Research Department in Elm Company for funding and support this project.

References

Ashutosh Adhikari, Achyudh Ram, Raphael Tang, and Jimmy Lin. 2019. Docbert: Bert for document classification. *arXiv preprint arXiv:1904.08398*.

Purbid Bambroo and Aditi Awasthi. 2021. Legaldb: Long distilbert for legal document classification. In

2021 International Conference on Advances in Electrical, Computing, Communication and Sustainable Technologies (ICAECT), pages 1–4. IEEE.

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

Xiang Dai, Ilias Chalkidis, Sune Darkner, and Desmond Elliott. 2022. Revisiting transformer-based models for long document classification. *arXiv preprint arXiv:2204.06683*.

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

Ming Ding, Chang Zhou, Hongxia Yang, and Jie Tang. 2020. Cogltx: Applying bert to long texts. *Advances in Neural Information Processing Systems*, 33:12792–12804.

Jun He, Liqun Wang, Liu Liu, Jiao Feng, and Hao Wu. 2019. Long document classification from local word glimpses via recurrent attention learning. *IEEE Access*, 7:40707–40718.

Yongli Hu, Puman Chen, Tengfei Liu, Junbin Gao, Yanfeng Sun, and Baocai Yin. 2021. Hierarchical attention transformer networks for long document classification. In *2021 International Joint Conference on Neural Networks (IJCNN)*, pages 1–7. IEEE.

Weichun Huang, Ziqiang Tao, Xiaohui Huang, Liyan Xiong, and Jia Yu. 2021. Hierarchical self-attention hybrid sparse networks for document classification. *Mathematical Problems in Engineering*, 2021.

Snehal Ishwar Khandve, Vedangi Kishor Wagh, Apurva Dinesh Wani, Isha Mandar Joshi, and Raviraj Bhuminand Joshi. 2022. Hierarchical neural network approaches for long document classification. In *2022 14th International Conference on Machine Learning and Computing (ICMLC)*, pages 115–119.

Liu Liu, Kaile Liu, Zhenghai Cong, Jiali Zhao, Yefei Ji, and Jun He. 2018. Long length document classification by local convolutional feature aggregation. *Algorithms*, 11(8):109.

Raghavendra Pappagari, Piotr Zelasko, Jesús Villalba, Yishay Carmiel, and Najim Dehak. 2019. Hierarchical transformers for long document classification. In *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 838–844. IEEE.

Hyunji Hayley Park, Yogarshi Vyas, and Kashif Shah. 2022. Efficient classification of long documents using transformers. *arXiv preprint arXiv:2203.11258*.

Ruyu Qin, Min Huang, Jiawei Liu, and Qinghai Miao. 2022. Hybrid attention-based transformer for long-range document classification. In *2022 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE.

Yuqi Si and Kirk Roberts. 2021. Hierarchical transformer networks for longitudinal clinical document classification. *arXiv preprint arXiv:2104.08444*.

Chi Sun, Xipeng Qiu, Yige Xu, and Xuanjing Huang. 2019. How to fine-tune bert for text classification? In *China national conference on Chinese computational linguistics*, pages 194–206. Springer.

Yi Tay, Mostafa Dehghani, Samira Abnar, Yikang Shen, Dara Bahri, Philip Pham, Jinfeng Rao, Liu Yang, Sebastian Ruder, and Donald Metzler. 2020. Long range arena: A benchmark for efficient transformers. *arXiv preprint arXiv:2011.04006*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Vedangi Wagh, Snehal Khandve, Isha Joshi, Apurva Wani, Geetanjali Kale, and Raviraj Joshi. 2021. Comparative study of long document classification. In *TENCON 2021-2021 IEEE Region 10 Conference (TENCON)*, pages 732–737. IEEE.

Lulu Wan, George Papageorgiou, Michael Seddon, and Mirko Bernardoni. 2019. Long-length legal document classification. *arXiv preprint arXiv:1912.06905*.

Zhengjue Wang, Chaojie Wang, Hao Zhang, Zhibin Duan, Mingyuan Zhou, and Bo Chen. 2020. Learning dynamic hierarchical topic graph with graph convolutional network for document classification. In *International Conference on Artificial Intelligence and Statistics*, pages 3959–3969. PMLR.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. *Advances in neural information processing systems*, 32.