2024 IEEE International Conference on Interdisciplinary Approaches in Technology and Management for Social Innovation (IATMSI) | 979-8-3503-6052-3/24/$31.00 ©2024 IEEE | DOI: 10.1109/IATMSI60426.2024.10502898

BeTS: Abstractive Text Summarization with Transfer Learning

Meher Bhardwaj
Dept. of Computer Science and Engineering
IIIT Manipur
Imphal, India
bhardwajmeher01@gmail.com

Hrishikesh Ethari
Dept. of Computer Science and Engineering
IIIT Manipur
Imphal, India
hrishikeshethari@gmail.com

Annepu Sai Charan
Dept. of Computer Science and Engineering
IIIT Manipur
Imphal, India
saicharan0662@gmail.com

Dennis Singh Moirangthem
Dept. of Computer Science and Engineering
IIIT Manipur
Imphal, India
mdennissingh@gmail.com

***Abstract*—We introduce BeTS, which stands for Bert Transformer Summarizer, an abstractive summarization model that is built using transfer learning. Our model consists of a pre-trained BERT encoder and a generative transformer decoder to produce abstract summaries. The model is trained by a simple two step process, where the decoder is first trained and then the entire model is fine-tuned end-to-end. We train the model with CNN/Daily Mail summarization dataset and use the trained weights for transfer learning on the AMI dialog summarization dataset. We successfully demonstrate that our model outperforms the baseline models by at least 7.65 ROUGE points on the AMI data. We also demonstrate that our model shows good generalization performance by testing on different unseen input data and compare the generated results. We are able to generate significantly better summaries on inputs such as dialog compared to the current state-of-the-art model.**

***Index Terms*—Text Summarization, Natural Language Processing, Large Language Models, Transfer Learning**

TABLE I
GENERATED SUMMARIES USING UNSEEN DIALOG DATA. THE MULTI-TURN DIALOG IS MADE INTO A SINGLE SEQUENCE OF UTTERANCES AND FED INTO THE MODELS WITHOUT THE SPEAKER INFORMATION.

Input
A: Hi how are you? Where are you now? B: I am in Hong Kong now. I am doing fine. A: How is the weather there? B: It is raining here.
Summary
BertSumExtAbs: [12]
i am in hong kong now . how is the weather there ? i was in the hong kong city of hong kong .
BertSumAbs: [12]
i am in hong kong now . how is the weather there ? hong kong is a hot spot in the north of the country .
Our Model (BeTS):
hong kong resident says weather in hong kong is raining .

I. INTRODUCTION

Automatic text summarization is a natural language processing (NLP) task of generating concise and precise version of a text document while retaining the essential information. Machine learning algorithms have been applied to this task to generate summaries from news, reviews, dialogs, or even scientific articles [1]–[10] . This task has been typically classified into two categories, *extractive* summarization and *abstractive* summarization. Extractive summarization focuses on selecting important texts (usually sentences) from the original document and then concatenating them together in a summary form, whereas abstractive summarization involves generating novel sentences to form a summary from information extracted from the text. Abstractive summarization involves both; a knowledge modeling process and a language generation process, making it significantly more complex than extractive summarization. The improvement on automatic ROUGE [11] metrics on the abstractive summarization task has reached a bottleneck due to its complexity. Recently, pre-trained language models have become popular for NLP tasks. Bidirectional Encoder Representations from Transformers (BERT) [13] is one of the most popular pre-trained language models, which have been widely used NLP tasks. [14] has modified the BERT model to make it usable for extractive summarization and [12] included a decoder to produce abstractive summaries with the help of a multi-optimizer training mechanism. However, the performance of this model is highly dependent on the dataset as in case of news text summarization where the gold summaries are highly extractive in nature.

In this paper, we put forward a simple abstractive summarization model named BeTS (Bert Transformer Summarizer), which consists of BERT as an encoder and a Transformer as a decoder. Unlike the existing models that use a modified version of BERT for summarization, our model adopts vanilla BERT without any modifications to the original architecture. We argue that even in its original form, with its pre-training on a huge dataset and its ability to extract rich and complex features, BERT can further boost the performance of abstractive summarization.

In our work, we focus on transfer learning using our pre-

979-8-3503-6052-3/24/$31.00 ©2024 IEEE

trained model to abstractively summarize multiple kinds of input texts. We first train our model using the CNN/Daily Mail summarization corpus and then perform transfer learning on the AMI dialog summarization corpus. We also demonstrate that our model is significantly better in generalizing on unseen data and generates better abstractive summaries as shown in Table I.

II. Related Works

A. Transformer

Transformer [15] is the first fully attention-based model to be used in sequence-to-sequence transformation tasks. Previous works relied on the sequence-modeling power of Recurrent Neural Networks [16], [17] for input and output representations. However, Transformer surpasses the existing models with the introduction of a 'self-attention' mechanism for input and output representations and positional encoding for sequence ordering. This enabled Transformer to be exclusively reliant on attention. A scaled dot-product attention mechanism is used in order to map a query Q and available information, in the form of key-value pairs K-V, to an output. This mapping is shown in Eq. (1):

$$Att(Q,K,V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V \qquad (1)$$

where d_k is the dimension of K.It should be noted that in self-attention mechanism, the Q, K, V information is originated from the same input.

B. BERT for Summarization

The advent of Transformer paves the way for BERT [13] , which is a powerful language representation model that is able to achieve state of the art performance on various natural language processing tasks [13] . BERT attains impressive results with the development of a two phase pre-training task: prediction and masking of a limited percentage of the input tokens, and prediction of how likely it is that sentence A is followed by sentence B. The combination of these two strategies increases context awareness tremendously, making it a high performing language model. BERT can also be used for the extractive summarization task, as shown by [14] . To use BERT for extractive summarization [14], modified the input sequence and embeddings of BERT to make it possible for extracting summaries. However, since BERT does not have a decoder for generation, it cannot be used for abstractive summarization as such. Therefore, [12] extended their model for extractive summarization to produce abstracts using a decoder. Their model requires multiple training process using the extractive labels as well as the abstractive targets. This model shows the state-of-the-art performance in several news summarization corpora. However, the performance of this model highly depends upon the amount of extractive summaries present in the actual gold summaries. This is true for news summarization where the gold summaries are mostly extractive in nature. On the other hand, in tasks such as dialog summarization, since extractive summaries are not a good representation of the text and thus have no significant meaning, models depending on extractive signals may not perform well.

III. Proposed Model

BERT has been used for multiple NLP tasks such as feature extraction, classification, etc. It being just a pre-trained encoder, BERT by itself does not have any text generation capability. Therefore, BERT in its original form cannot be used for generation tasks such as abstractive summarization. However, it can perform extractive summarization using its classifier as shown in [14].

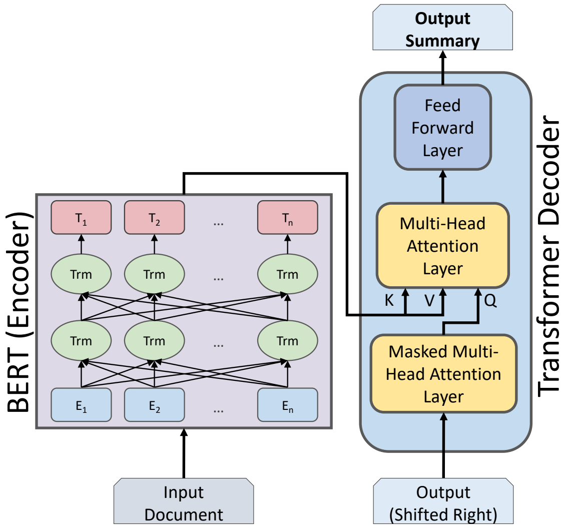

Fig. 1. Proposed BeTS model as an encoder-decoder framework for abstractive summarization.

In order to utilize the pre-trained BERT for abstractive summarization, we need to develop a new model utilizing BERT. In this work, we propose a very simple abstractive summarization model in the standard encoder-decoder framework with a vanilla BERT (encoder) and a Transformer decoder for the abstractive summarization task as shown in Figure 1. The decoder is a 12-layered Transformer with 768 units initialized randomly. This allows us to generate summaries using the decoder from the rich features of BERT.

In our model, we expected a notable disproportion in the training process, since BERT is pre-trained whereas the decoder is initialized arbitrarily and has to be trained from scratch. In order to tackle this problem, we execute a simple two-step training process to train our model. We initially train only the decoder network using the BERT encoder features till we get a stable performance. Then, we proceed by training the entire encoder-decoder network on the final negative log-likelihood (NLL) loss. For this, we use the Adam optimizer with a learning rate decay schedule and a warmup-step of $60,000$. As expected, because of the decay schedule, the learning rate will be small during the second stage of the training progress where we fine-tune the BERT weights. This is desirable as the encoder and decoder need to be jointly trained with more accurate gradients.

We use the aforementioned two-step training process for the training on the CNN/Daily Mail summarization corpus only. Once our model has been trained, we can use it for transfer learning on other datasets in a standard single-step training process.

IV. Experiments and Results

In this section, we describe the datasets, our experiments, and the results of our proposed model in detail.

A. Dataset

We conduct experiments on two summarization datasets.

a) CNN/Daily Mail: Proposed by [18], the dataset consists of 90k news articles from CNN and 197k from Daily Mail along with their respective summaries. The joint CNN/Daily Mail dataset is split into 11,487 test, 13,368 validation and 286,817 training pairs. We utilize this dataset during the two-step training stage of our model.

b) AMI DialSum: Dialog summarization dataset used by [19] and based on the AMI Meeting Corpus [20]. This dataset consists of 7,824 meeting transcripts and summaries split into 400 test/validation and 7,024 train samples. We perform transfer learning on this dataset using our summarization model, which has been pre-trained on CNN/Daily Mail.

B. Experiment Settings

We use the AdamOptimizer [21] with a "weight decay rate" of 0.01, β_1 of 0.9, β_2 of 0.999, and ϵ of $1e$-6. We also use gradient clipping with "clip by global norm" of 1.0. Table II shows the remaining details of the model and training parameters. Our model is trained on 2 Nvidia Quadro V100 GPUs.

TABLE II
Model and training parameters of BeTS.

Parameter	Value
BERT Model	BERT-Base Uncased
Decoder Heads	16
Decoder blocks	12
Decoder Dim	768
Batch Size	24
Optimizer	Adam
Warmup-steps	60,000

C. Results on Summarization

a) CNN/Daily Mail: We compare the performance of our model to the current baselines and state-of-the-art models. Table III shows the ROUGE evaluation results in two categories. The models using just abstractive summarization training in an end-to-end manner and the models using both extractive label signals as well as the abstractive loss for training. Our model performs comparable to the existing models. The BertSumExtAbs [12] with extractive and abstractive training achieves the good performance. This is expected as the CNN/Daily Mail summaries are highly extractive in many cases and the BertSumExtAbs model can show better performance on such data. On the other hand, this model will not be able to show such significant performance gains on data that do not have extractive summaries, such as dialog data. Our model is also performing at par if not better, compared to other abstractive summarization methods. While BertSumAbs [12] also uses BERT, our model is able to outperform it with the help of transfer learning.

TABLE III
ROUGE F1-scores for Summarization on the CNN/DailyMail Corpus.

Model	R-1	R-2	R-L
Extractive+Abstractive			
RSSR [22]	40.88	17.80	38.54
BertSumExtAbs [12]	42.13	**19.60**	39.18
Abstractive			
PointerGen [2]	36.44	15.66	33.42
PointerGen+Coverage [2]	39.53	17.28	36.38
DRM [3]	39.87	15.82	36.90
BottomUp [23]	41.22	18.68	38.34
DCA [6]	41.69	19.47	37.92
TransformerAbs [12]	40.21	17.76	37.09
BertSumAbs [12]	41.72	19.39	38.76
BeTS (Ours)	**42.91**	19.58	**39.27**

TABLE IV
ROUGE F1-scores for Summarization on the AMI dialog summarization Corpus.

Model	R-1	R-2	R-3	R-L
AttSeq2Seq [24]	34.74	25.15	21.35	34.70
PointerGen [2]	31.21	26.35	25.22	31.21
BertSumExtAbs [12]	20.95	14.89	9.77	20.61
BertSumAbs [12]	16.11	12.35	8.11	15.88
BeTS (Ours)	**42.39**	**35.36**	**31.99**	**42.34**

b) AMI DialSum: Table IV displays the ROUGE scores of BeTS and baseline models regarding abstractive summarization on AMI DialSum. The ROUGE scores for BeTS, BertSumAbs, and BertSumExtAbs are obtained by fine-tuning the pre-trained CNN/Daily Mail models. The scores for both AttSeq2Seq [24] and PointerGen [2] baseline models are obtained from work done by [19] . Our model outperforms all baseline models by at least 7.65 ROUGE points. Note that [19] achieves higher ROUGE scores by using additional dialog act information. Although we acknowledge that high scores are important, our aim is not solely to increase those scores, but to improve the summary generation capability of our model and to achieve better generalization performance with transfer learning. Dialogs of different length are summarized and the generated results are shown in Table VI (short dialog) and Table VII (long dialog). Summaries obtained from BertSumExtAbs and BertSumAbs [12] are highly extractive, whereas ours is able to summarize the information given in a more abstract way. Note, however, that due to the lack of speaker information, name addressing may not be accurate.

As demonstrated in the results, even though BertSumExtAbs scores higher ROUGE scores on the CNN/Daily Mail dataset, it is not able to maintain that performance through transfer learning when fine-tuned with the AMI dataset, with our model performing twice as much. This is because the BertSumExtAbs model is too dependent on information derived from extractive

TABLE V

GENERATED SUMMARIES USING UNSEEN INPUT DATA (AN INTRODUCTION OF A SCIENTIFIC ARTICLE).

Input: Introduction of the paper "Character-aware neural language models" by Kim et. al., 2016
Language modeling is a fundamental task in artificial intelligence and natural language processing (NLP),... Unlike previous works that utilize subword information via morphemes, our model does not require morphological tagging as a pre-processing step.... Given that most of the parameters in NLMs are from the word embeddings, the proposed model has significantly fewer parameters than previous NLMs, making it attractive for applications where model size may be an issue....
Summary
BertSumExtAbs:
language models are simple to train , but probabilities of rare n-grams can be poorly estimated due to data sparsity . the parameters are learned as part of the word embeddings obtained through neural network . unlike previous works that utilize subword information via morphemes , our model does not require managerial tagging as a pre-processing step . but unlike the recent line of work that combines input word embeddings with features from a character-level model.
BertSumAbs:
language model is formalized as a probability distribution over a sequence of strings (words). traditional methods usually involve making an n-th order markov assumption and estimating n-gram probabilities via counting . the count-based models are simple to train , but probabilities of rare n-grams can be poorly estimated due to data sparsity.
Our Model (BeTS):
embeddings of rare words can be poorly estimated , leading to high perplexities for rare words (and words surrounding them) . this is especially problematic in morphologically rich languages with long-tailed frequency distributions or domains with dynamic vocabularies . the proposed model has significantly fewer parameters than previous nlms . unlike previous works that utilize subword information via morphemes , our model does not require morphological tagging as a pre-processing step . unlike the recent line of work which combines input word embeddings with features from a character-level model , our model does not utilize word embeddings at all in the input layer.

TABLE VI

GENERATED SUMMARIES USING UNSEEN DIALOG DATA (SHORT DIALOG).

Input
A: Good evening . Welcome to Cherry's . Do you have a reservation ? B: No , we don't . A: How many of you , please ? B: Six , including two kids . A: I'm afraid all the big tables are taken .
Summary
BertSumExtAbs [12]:
do you have a reservation ? no , we do n't . how many of you , please ?
BertSumAbs [12]:
do you have a reservation ? no , we do n't . how many of you , please ?
Our Model (BeTS):
cherry ' s has a reservation for six members , including two kids .

TABLE VII

GENERATED SUMMARIES USING UNSEEN DIALOG DATA (LONG DIALOG).

Input
A: I am a student at Cambridge University . I read your ad and would like to know something more about your room please . B: It's a big bedroom with a drawing room facing a beautiful patio . A: Is there a bathroom ? B: No, but there is one downstairs, which my daughter used some years ago . A: It sounds good . Could I go and see it myself ? B: Certainly , you're welcome anytime . A: See you later !
Summary
BertSumExtAbs [12]:
i 'm a student at cambridge university . i read your ad and would like to know something more about your room please .
BertSumAbs [12]:
the cambridge university student is a student at cambridge university . the room is a big bedroom with a drawing room facing a beautiful patio .
Our Model (BeTS):
cambridge university student ' s daughter used the room years ago and would like to see it .

summaries, making it very difficult to adapt to other datasets different from the one it has been pre-trained on. This is especially true for the dialog summarization dataset where extractive labels are available. In order to further show the significance of our model, we also compare with the pure abstractive BertSumAbs model that do not use extractive label information for training. Even though BertSumAbs is a purely abstract approach, our proposed model significantly outperforms it.

c) Generalization Performance: In order to demonstrate the generalization performance, we compare the summaries generated with BeTS against the ones generated with the state-of-the-art BertSumExtAbs model and the pure abstractive BertSumAbs model, with all three models being pre-trained on the CNN/Daily Mail dataset. Table I shows the generated summaries on a dialog input data, and Tables V, VIII and IXshow the results of inputs from scientific articles. We also summarize abstracts from scientific papers. Different lengths of abstract are shown in Table X (short) and Table XI (long). The summaries generated by BertSumExtAbs [12] are, as expected, highly extractive in both short and long inputs. Both BertSumAbs and BertSumExtAbs generated highly extractive summarization. Additionally, a comparison between the generated summaries indicate that BertSumExtAbs and BertSumAbs both have difficulty in summarizing short inputs, performing better on longer text. Our BeTS model shows reasonably good performance on both long and short inputs.

As demonstrated, the generated summaries of our model are significantly better than the competing models. The performance difference is significant on the dialog data shown in Table I, Table VI and Table VII as the BertSumExtAbs model depends on the extractive features of the encoder. However, in dialogs and conversations, extraction is usually not desirable as they make no sense in the summaries. Furthermore, the purely abstract BertSumAbs model also did not perform well in comparison to our model. This is illustrated in Table I, where the generated summary is still very extractive, similar to BertSumExtAbs, and it doesn't encompass relevant information mentioned in the conversation. On the contrary, the generated summary opposes what is said in the dialog, whereas

TABLE VIII
GENERATED SUMMARIES USING UNSEEN INPUT DATA (AN ABSTRACT OF A SCIENTIFIC ARTICLE).

Input: Introduction of the paper “Seeding the singularity for A.I” by Kraikivsk, 2019
The singularity refers to an idea that once a machine having an artificial intelligence surpassing the human intelligence capacity is created, it will trigger explosive technological and intelligence growth. I propose to test the hypothesis that machine intelligence capacity can grow autonomously starting with an intelligence comparable to that of bacteria - microbial intelligence. …The algorithms must demonstrate a rapid growth in skills of dataprocessing and analysis and gain qualitatively different functionalities, at least until the current computing technology supports their scalable development. The existing algorithms …will be reviewed in this work. Future computational tests could support or oppose the hypothesis that artificial intelligence can potentially grow to the level of super-intelligence which overcomes the limitations in hardware by producing necessary processing resources or by changing the physical realization of computation from using chip circuits to using quantum computing principles.
Summary
BertSumExtAbs:
i propose to test the theory that machine intelligence capacity can grow autonomously starting with an intelligence comparable to that of bacteria - microbial intelligence . the goal will be to demonstrate that rapid growth in intelligence capacity could be realized at all in artificial computing systems . future artificial intelligence tests could support or oppose the hypothesis that artificial intelligence can potentially grow to the level of superintelligence .
BertSumAbs:
i propose to test the theory that machine intelligence capacity can grow autonomously starting with an intelligence comparable to that of bacteria . the goal will be to demonstrate that rapid growth in intelligence capacity could be realized at all in artificial computing systems . the algorithms must demonstrate a rapid growth of dataprocessing and analysis until the current computing technology supports their scalable development .
Our Model (BeTS):
the title refers to an idea that a machine with an artificial intelligence capacity is created , will trigger explosive tech and intelligence growth . an artificial intelligence can grow autonomously . i propose the three properties that may allow an artificial intelligence to exhibit a steady growth in its intelligence capacity .

our model is able to correctly and concisely convey all the necessary information.

TABLE IX
GENERATED SUMMARIES USING UNSEEN INPUT DATA (AN INTRODUCTION OF A SCIENTIFIC ARTICLE).

Input: Introduction of the paper by Jiang and de Marneffe, 2019
Prediction of speaker commitment is the task of determining to what extent the speaker is committed to an event …There has been work on factors leading to speaker commitment in theoretical linguistics and computational linguistics, …the CommitmentBank, a dataset of naturally occurring sentences annotated with …It allows us to evaluate whether current speaker commitment models achieve robust language understanding, by analyzing their performance on specific challenging linguistic constructions.
Summary
BertSumExtAbs:
prediction of speaker commitment is the task of determining what extent the speaker is committed to an event in a sentence as actual , non-actual , or uncertain . the commitmentbank is a dataset of naturally occurring sentences annotated with speaker commitment towards the content of complements of clause-embedding verb under canceling-entailment environments .
BertSumAbs:
prediction of speaker commitment is the task of determining to what extent the speaker is committed to an event as actual , non-actual , or uncertain . this matters for downstream nlp applications , such as information extraction or question answering . the commitmentbank , restricted to specific linguistic constructions , is a good test case .
Our Model:
prediction of speaker commitment is the task of determining to what extent the speaker is committed to an event in a sentence . the commitmentbank is a dataset of naturally occurring sentences annotated with speaker commitment . it includes study of speech commitment . we use it to evaluate whether current speaker commitment models achieve robust language understanding .

TABLE X
GENERATED SUMMARIES USING UNSEEN INPUT DATA (A SHORT ABSTRACT OF A SCIENTIFIC ARTICLE).

Input: Abstract of the (Marwala, 2015) paper on Artificial Intelligence
Artificial intelligence has impacted many aspects of human life. This paper studies the impact of artificial intelligence on economic theory. In particular we study the impact of artificial intelligence on the theory of bounded rationality, efficient market hypothesis and prospect theory.
Summary
BertSumExtAbs:
this paper studies the impact of artificial intelligence on economic theory .
BertSumAbs:
artificial intelligence has impacted many aspects of human life . this paper studies the impact of artificial intelligence on economic theory .
Our Model (BeTS):
the study looks at the impact of artificial intelligence on the theory of bounded rationality.

TABLE XI
GENERATED SUMMARIES USING UNSEEN INPUT DATA (A LONG ABSTRACT OF A SCIENTIFIC ARTICLE).

Input: Abstract of the (Şeker, 2019) paper on Artificial Intelligence
Nowadays, considering the speed of the processes and the amount of data used in cyber defense, it cannot be expected to have an effective defense by using only human power without the help of automation systems. However, for the effective defense against dynamically evolving attacks on networks, it is difficult to develop software with conventional fixed algorithms. This can be achieved by using artificial intelligence methods that provide flexibility and learning capability. The likelihood of developing cyber defense capabilities through increased intelligence of defense systems is quite high. Given the problems associated with cyber defense in real life, it is clear that many cyber defense problems can be successfully solved only when artificial intelligence methods are used. In this article, the current artificial intelligence practices and techniques are reviewed and the use and importance of artificial intelligence in cyber defense systems is mentioned. The aim of this article is to be able to explain the use of these methods in the field of cyber defense with current examples by considering and analyzing the artificial intelligence technologies and methodologies that are currently being developed and integrating them with the role and adaptation of the technology and methodology in the defense of cyberspace.
Summary
BertSumExtAbs:
the likelihood of developing cyber defense capabilities is quite high . many cyber defense problems can be solved only when artificial intelligence is used . in this article , the current artificial intelligence practices and techniques are reviewed and the importance of artificial intelligence in cyber defense systems is mentioned
BertSumAbs:
in this article , the current artificial intelligence practices and techniques are reviewed and the importance of artificial intelligence in cyber defense systems is mentioned . the aim of this article is to be able to explain the use of these methods in the field of cyber defense with current examples .
Our Model (BeTS):
nowadays , the chance of developing cyber defense capabilities is quite high . it can be achieved by using artificial intelligence methods that provide flexibility and learning capability . this article will help explain the use of these technologies in cyber defense .

V. DISCUSSION

Even though BeTS, BertSumAbs, and BertSumExtAbs use BERT, BeTS is significantly different both in terms of network architecture as well as the training process. Both BertSumAbs and BertSumExtAbs have additional encoder layers over the vanilla BERT encoder layers but with a smaller decoder. Both models use two different optimizers for training, which requires careful setting of two learning rates. BertSumAbs is trained only on pure abstractive targets whereas BertSumExtAbs uses both extractive labels as well as abstractive targets for training. On the other hand, our BeTS model uses vanilla BERT as the encoder and the decoder is of a similar size to the encoder. Our model is trained using a single optimizer in a simple two step optimization process as described in Section III. It is trained only on abstractive targets. Our model is much simpler in terms of both the network architecture as well as the training process and the hyper-parameter settings. These differences between our proposed model and the competing models explain the significant differences in the generated outputs as well as the generalization capability of the models. Recently, BART [25] type of models that train the entire encoder-decoder framework on abstractive summary generation can perform very well. However, such models need a lot of data and computing resources, whereas our model can be used efficiently for low resource languages with transfer learning; since Large Language Models (LLMs) like BERT are available for multiple languages including Indian languages, such as IndicBERT [26], MuRIL [27].

VI. Conclusion

In this paper, we have observed and aimed to improve the task of abstractive summarization using transfer learning. We proposed an encoder-decoder framework named BeTS that utilizes BERT as it is, without any modifications. We show that the performance of our model is on a par with current state-of-the-art model, however our proposed model is simpler and easier to train. We also achieve better performance on a dialog summarization dataset using transfer learning with our pre-trained model. Moreover, we have demonstrated how our model can generalize to generate better summaries on unseen input texts such as dialogs etc. compared to the current state-of-the-art model.

In the future, we will work to incorporate context understanding for dialog summarization and summarization tasks for low resource Indian languages with the help of multi-lingual LLMs. We also plan to adapt to the newer pre-trained models, such as Pegasus [25], [28], to enhance the performance of our summarization model.

References

[1] A. M. Rush, S. Chopra, and J. Weston, “A neural attention model for abstractive sentence summarization,” *arXiv preprint arXiv:1509.00685*, 2015.

[2] A. See, P. J. Liu, and C. D. Manning, “Get to the point: Summarization with pointer-generator networks,” *arXiv preprint arXiv:1704.04368*, 2017.

[3] R. Paulus, C. Xiong, and R. Socher, “A deep reinforced model for abstractive summarization,” *arXiv preprint arXiv:1705.04304*, 2017.

[4] A. Cohan, F. Dernoncourt, D. S. Kim, T. Bui, S. Kim, W. Chang, and N. Goharian, “A discourse-aware attention model for abstractive summarization of long documents,” *arXiv preprint arXiv:1804.05685*, 2018.

[5] D. S. Moirangthem and M. Lee, “Hierarchical and lateral multiple timescales gated recurrent units with pre-trained encoder for long text classification,” *Expert Systems with Applications*, vol. 165, p. 113898, 2021.

[6] A. Celikyilmaz, A. Bosselut, X. He, and Y. Choi, “Deep communicating agents for abstractive summarization,” *arXiv preprint arXiv:1803.10357*, 2018.

[7] M. Kim, M. D. Singh, and M. Lee, “Towards abstraction from extraction: Multiple timescale gated recurrent unit for summarization,” *ACL 2016*, p. 70, 2016.

[8] S. Erera, M. Shmueli-Scheuer, G. Feigenblat, O. P. Nakash, O. Boni, H. Roitman, D. Cohen, B. Weiner, Y. Mass, O. Rivlin *et al.*, “A summarization system for scientific documents,” *EMNLP-IJCNLP 2019*, p. 211, 2019.

[9] G. Cunha Sergio, D. S. Moirangthem, and M. Lee, “Attentively embracing noise for robust latent representation in BERT,” in *Proceedings of the 28th International Conference on Computational Linguistics*. Barcelona, Spain (Online): International Committee on Computational Linguistics, Dec. 2020, pp. 3479–3491.

[10] D. S. Moirangthem and M. Lee, “Abstractive summarization of long texts by representing multiple compositionalities with temporal hierarchical pointer generator network,” *Neural Networks*, vol. 124, pp. 1–11, 2020.

[11] C.-Y. Lin, “Rouge: A package for automatic evaluation of summaries,” in *Text summarization branches out*, 2004, pp. 74–81.

[12] Y. Liu and M. Lapata, “Text summarization with pretrained encoders,” *arXiv preprint arXiv:1908.08345*, 2019.

[13] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “Bert: Pre-training of deep bidirectional transformers for language understanding,” *arXiv preprint arXiv:1810.04805*, 2018.

[14] Y. Liu, “Fine-tune bert for extractive summarization,” *arXiv preprint arXiv:1903.10318*, 2019.

[15] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, “Attention is all you need,” in *Advances in neural information processing systems*, 2017, pp. 5998–6008.

[16] S. Chopra, M. Auli, and A. M. Rush, “Abstractive sentence summarization with attentive recurrent neural networks,” in *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2016, pp. 93–98.

[17] R. Nallapati, B. Zhou, C. Gulcehre, B. Xiang *et al.*, “Abstractive text summarization using sequence-to-sequence rnns and beyond,” *arXiv preprint arXiv:1602.06023*, 2016.

[18] K. M. Hermann, T. Kocisky, E. Grefenstette, L. Espeholt, W. Kay, M. Suleyman, and P. Blunsom, “Teaching machines to read and comprehend,” in *Advances in neural information processing systems*, 2015, pp. 1693–1701.

[19] C.-W. Goo and Y.-N. Chen, “Abstractive dialogue summarization with sentence-gated modeling optimized by dialogue acts,” in *2018 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 2018, pp. 735–742.

[20] I. McCowan, J. Carletta, W. Kraaij, S. Ashby, S. Bourban, M. Flynn, M. Guillemot, T. Hain, J. Kadlec, V. Karaiskos *et al.*, “The ami meeting corpus,” in *Proceedings of the 5th International Conference on Methods and Techniques in Behavioral Research*, vol. 88, 2005, p. 100.

[21] D. P. Kingma and J. Ba, “Adam: A method for stochastic optimization,” *arXiv preprint arXiv:1412.6980*, 2014.

[22] Y.-C. Chen and M. Bansal, “Fast abstractive summarization with reinforce-selected sentence rewriting,” in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2018, pp. 675–686.

[23] S. Gehrmann, Y. Deng, and A. M. Rush, “Bottom-up abstractive summarization,” *arXiv preprint arXiv:1808.10792*, 2018.

[24] R. Nallapati, B. Zhou, C. N. dos santos, C. Gulcehre, and B. Xiang, “Abstractive text summarization using sequence-to-sequence rnns and beyond,” 2016.

[25] M. Lewis, Y. Liu, N. Goyal, M. Ghazvininejad, A. Mohamed, O. Levy, V. Stoyanov, and L. Zettlemoyer, “Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension,” 2019.

[26] S. Doddapaneni, R. Aralikatte, G. Ramesh, S. Goyal, M. M. Khapra, A. Kunchukuttan, and P. Kumar, “Towards leaving no indic language behind: Building monolingual corpora, benchmark and models for indic languages,” 2023.

[27] S. Khanuja, D. Bansal, S. Mehtani, S. Khosla, A. Dey, B. Gopalan, D. K. Margam, P. Aggarwal, R. T. Nagipogu, S. Dave, S. Gupta, S. C. B. Gali, V. Subramanian, and P. Talukdar, “Muril: Multilingual representations for indian languages,” 2021.

[28] J. Zhang, Y. Zhao, M. Saleh, and P. J. Liu, “Pegasus: Pre-training with extracted gap-sentences for abstractive summarization,” *arXiv preprint arXiv:1912.08777*, 2019.