News Summarization with BERT-powered encoders

Dang Ngoc Huy
h.dang@mpp.hertie-school.org

Joshua Aje
j.aje@mpp.hertie-school.org

Ole Teutloff
o.teutloff@mpp.hertie-school.org

Abstract

Text summarization is one of the central challenges in the fields of Machine Learning and Natural Language Processing (NLP). Bidirectional Encoder Representations from Transformers (BERT), a new contextual pre-training method for language representations, has been heralded as the state-of-the-art neural network architecture that can outperform any others in over 11 complex NLP tasks at the time of its creation. In this paper, we explore the potential of utilizing BERT as the basis for a document level encoder that can capture and generate a representation for text sentences and meanings, ultimately providing a reliable and accurate automated summarization process of news articles from different international outlets [1].

1. Introduction[2]

Text summarization is the practice of producing a succinct and accurate summary of enormous text that still delivers the central message by concentrating on the essential information and meanings.

In the digital age of today, data is being created, collected and ingested at an unprecedented rate, with increasing variety and scope. Researchers, world leaders, decision makers and normal citizens alike are bombarded with colossal amount of information daily, most of which is often the noise that conceals the vital signal that should have been accessed more easily. Manual summarization, however, is a costly undertaking with a built-in time lag. The development of an automated text summarization system that permits the ease of insight extraction is, therefore, becoming progressively necessary. Not only does it help to reduce the complexity of the document in question, but it also weeds out the unnecessary and redundant information, decreasing the time spent for research and subsequently helping to increase productivity and understanding of useful materials and data.

The recent development and progress in Machine Learning and Deep Learning has provided an assortment of novel techniques for automatic text summarization. However, the broad stroke approaches anchors predominantly in two methods: extractive and abstractive summarization.

Extractive summary

Extractive summarization pulls sentences and words directly from the original document to represent its central meaning. By weighing the similarities, importance, and relevance of different sentences and segments of words, the method is able to produce a shortened representation of the text in question by joining words and sentences of significance together.

Abstractive summary

Abstractive summarization, on the other hand, seeks to generate words and sentences based on the contextual and semantic understanding and meaning of the text in question. Therefore, its goal is not to extract, but to paraphrase, and produce new materials that can accurately reflect the content of the original documents, even with new words and sentences that have not appear previously.

This paper will first provide a review of the current state-of-the-art in the discipline of text summarization using Machine Learning. Subsequently, the proposed methodology section will elucidate on the implementation of the modeling techniques utilized for this paper, with pre-trained encoders based on BERT. Lastly, experiments with the models on the Cornell Newsroom dataset, together with the analysis of the results obtained will be detailed.

[1] The code for this project can be found here: Project Github Repo

[2] The research project serves as a joint project for the courses E1296 and E1326.

2. Related Works

This purpose of this project is to adapt the research conducted by **Yang Liu and Mirella Lapata, Text Summarization with Pretrained Encoders. arXiv 2019**, to apply BERT in the context of obtaining accurate news summary for the Cornell Newsroom dataset.

Building up from the success of Transfer Learning, BERT was a sensational innovation introduced by the Google AI Language team in 2018 in their seminal paper from **J Devlin, MW Chang, K Lee, K Toutanova, BERT: Pretraining of Deep Bidirectional Transformers for Language Understanding, arXiv 2019**. BERT's key technical achievement is the application of the bidirectional training of Transformer, an attention mechanism that learns contextual relations between words in a text. Whereas previous endeavors scrutinized word sequence directionally from left to right, right to left or a combination of both, the Transformer encoder architecture of BERT looks at the entirety of the word sequence simultaneously and thus was able to learn the context and relationships of a word based on all of its surrounding elements, not just exclusively from the left or right direction. This mechanism allows BERT to grasp a deeper understanding of the context of the document and text in question as well as the connective tissues of its genetic makeup of words, sub-words and sentences. With hyperparameter fine-tuning, BERT is able to to perform a variety of NLP tasks using its pre-trained neural network to create word embeddings as features for modeling.

Following the pragmatic approach of BERT, Yang Liu et al. has constructed a pre-trained encoder using fine-tuned BERT which has the capability to capture a succinct representation of the text that needs summarizing. Their models employ approaches of extractive, abstractive and a mixture of both to help create better summaries. For the extractive model, in order to represent different sentences in the document, external tokens are enclosed at the beginning of each sentence while a specialized symbol is adopted to collect features of the preceding sentence. The pre-trained encoder is stacked with multiple Transformer layers, each of which representing distinct document-level features such as adjacent sentences or multi-sentence discourse. The abstractive model follows an encoder-decoder architecture, using the pre-trained encoder with randomly-initialized Transformer decoder. The final model combines both extractive and abstractive approaches, utilizing two stage progression with the encoder being fine-tuned twice through extractive and abstractive summarization tasks respectively.

To evaluate for their models, three different news dataset from CNN/DailyMail news highlights, the New York Times Annotated Corpus, and XSum were employed since each represents a different style of text summary from long format to brief one sentence summaries. For each dataset, Yang Liu et al. was able to demonstrate that the models can accomplish state-of-the-art results for both extractive and abstractive purposes.

For our research, we want to apply the methodology employed by Yang Liu et al. to investigate the extent to which the BERT model is applicable for a different, more diverse dataset from Cornell Newsroom, which is comprised of over 1.3 million articles and summaries from a variety of news outlets, focusing on different issues and styles of writing. The state-of-the-art result for this particular dataset is currently achieved by **Tian Shi, Ping Wang, Chandan K. Reddy, LeafNATS: An Open-Source Toolkit and Live Demo System for Neural Abstractive Text Summarization, arXiv 2019**. Their paper outlines the design and implementation of the LeafNATS toolkit, a learning framework based on Neural abstractive text summarization with sequence-to-sequence model (NATS). In the NATS paradigm, the model is trained on a large amount of data to learn the connection between the feature text and the target summary with an attention-based encoder-decoder mechanism, using sequence-to-sequence framework modeled through a Recurrent Neural Network. After the models are trained and the weights and parameters are recorded, the decoder can then outputs a summary as a sequence of words, utilizing the Beam Search algorithm. Different from Greedy Search which only chooses one best output results for each individual time step, the Beam Search algorithm opts for multiple choices for an output sequence at each time step based on conditional probability, given the input sequence, and each subsequent choice of words is paired with the previous choices, ultimately producing multiple output sentences with varying degrees of probabilities.

Each of the methods above has their own strengths and weaknesses and thus, it would be beneficial to learn how they may be compared to one another on performance for the same data. Our research, therefore, seeks to answer this question by implementing the BERT pre-trained encoder for the Cornell Newsroom dataset to understand how it fares in juxtaposition with the NATS approach.

3. Motivation

Automatic text summarization can be used in a great variety of fields (such as for example journalism, blogging or research) and accurate summarization models come with significant benefits in terms of efficiency and cost-effectiveness. Moreover, automatic text summarization

serves as a building block for more complex AI applications. Therefore, it represents a promising and exciting frontier of NLP research.

Furthermore, automatic text summarization with BERT offers several unique learning opportunities for us. Firstly, we learn how to implement a complex state-of-the-art deep learning architecture for NLP. The data as well as a performance leaderboard are freely available on Summari.es allowing us to compare our results with the top implementations in the field. Secondly, we learn how to implement a complex project including automated data retrieval, analysis and presentation (dashboard). This allows us to combine Python programming tasks (data collection and dashboard) with an advanced NLP application.

4. Evaluation

Automatic text summarization models are commonly evaluated using Recall-Oriented Understudy for Gisting Evaluation (ROUGE scores) [1]. ROUGE scores capture how an automatically generated text summary compares to an "ideal" summary (in most cases written by humans). ROUGH scores are calculated by counting the overlapping units (n-grams, word pairs or word sequences) between the automatic summaries and the ones created by humans. Several types of ROUGE scores exist. R-1, R-2 and R-L are commonly used metrics [3, 2]. R-1 and R-2 stand for the overlap of unigrams and bigrams. R-L represents the longest common subsequence that overlaps [1]. Used by most researchers in the field and the leaderboard of Summari.es - the website that contains the Cornell Newsroom dataset and all information on efforts and researches to provide accurate automatic summarization for this data, we will focus on these three ROUGE scores to evaluate the performance of our model.

A successful project would, on the one hand, lead to a sufficiently accurate summarization model using BERT-powered encoders that achieves summary performance similar or above the performance achieved by the state-of-the-art approaches featured on the leaderboard of Summari.es. On the other hand, we intent to build a dashboard and API infrastructure that allows us to retrieve, summarize and visualize newspaper articles from several international outlets.

5. Resources

The dataset we intent to use for this project is the "Cornell Newsroom" data which is available as an open-source dataset for use of research on Summari.es. The dataset consists of 1.3 million articles and summaries from 38 major publications written between 1998 and 2017, ranging in different styles, subject matters and targeted audience. The dataset has not been used for evaluation of the pre-trained BERT encoders model and thus would be a suitable candidate as a comparison for how our model can perform better or worse than the current state-of-the-art for this data.

For the analysis we plan on using Google Colab, in tandem with Kaggle GPU kernels and possibly AWS. To implement the NLP model we will rely on PyTorch.

6. Contributions

All efforts in data preprocessing, experimentations, analysis, dashboard creation and progress report writing will be divided equally among the group members. For the purpose of this report, responsibilities can be shared as followed:

Joshua Aje will lead the pre-processing, and wrangling of the dataset to shape it to a form manageable by the model, and test-run the baseline model.

Ole Teutlof will be in charge of experimenting with the model and work on the analysis and evaluation.

Dang Ngoc Huy will be responsible for working on the dashboard to pull news article with official API keys from established news organization, and visualize the summarized news items.

References

[1] C.-Y. Lin. Rouge: A packagefor automatic evaluation of summaries. In *ProceedingsofWorkshop on Text Summarization Branches Out, Post2Conference Workshop of ACL*, 2004.

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

[3] T. Shi, P. Wang, and C. K. Reddy. LeafNATS: An open-source toolkit and live demo system for neural abstractive text summarization. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, pages 66–71, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.